Deciphering cellular and molecular mechanisms driving malignant transformation in Neurofibromatosis Type 1 using a reproducible computational workflow for single-cell transcriptomics

Audrey Onfroy

▶ **To cite this version:**

Audrey Onfroy. Deciphering cellular and molecular mechanisms driving malignant transformation in Neurofibromatosis Type 1 using a reproducible computational workflow for single-cell transcriptomics. Bioinformatics [q-bio.QM]. Paris 12, 2024. English. ⟨NNT : 2024PA120090⟩. ⟨tel-05293125⟩

HAL Id: tel-05293125

https://hal.science/tel-05293125v1

Submitted on 1 Oct 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY-SA 4.0 - Attribution - ShareAlike - International License

Paris-Est Créteil University

Doctoral School Life and Health Sciences

Mondor Institute for Biomedical Research

Deciphering cellular and molecular mechanisms driving malignant transformation in Neurofibromatosis Type 1 using a reproducible computational workflow for single-cell transcriptomics

By Audrey Onfroy

PhD thesis in Cellular and molecular biology

Presented and publicly supported on September 27, 2024

In front of a jury composed of:

Dr. **Loredana Martignetti**	Curie Institute	Reporter
Dr. **Eduard Serra**	Institut Germans Trias i Pujol	Reporter
Pr. **Sarah Cohen-Boulakia**	Paris-Saclay University	Examiner
Dr. **Raphaël Margueron**	Curie Institute	Examiner & President
Pr. **Denis Thieffry**	École Normale Supérieure	Thesis supervisor
Pr. **Piotr Topilko**	Paris-Est Créteil University	Thesis supervisor

UPEC
UNIVERSITÉ PARIS-EST CRÉTEIL VAL DE MARNE
Connaissance - Action

PARIS EST SUP — ÉCOLE DOCTORALE
Sciences de la Vie et de la Santé

Université Paris-Est Créteil

École Doctorale Sciences de la Vie et de la Santé

Institut Mondor de Recherche Biomédicale

Identification des mécanismes cellulaires et moléculaires gouvernant la transformation maligne dans la Neurofibromatose de Type 1, en assurant la reproductibilité des analyses transcriptomiques

---

Par Audrey Onfroy

Thèse de doctorat en Biologie moléculaire et cellulaire

Présentée et soutenue publiquement le 27 septembre 2024

Devant un jury composé de :

Dr. **Loredana Martignetti**	Institut Curie	Rapportrice
Dr. **Eduard Serra**	Institut Germans Trias i Pujol	Rapporteur
Pr. **Sarah Cohen-Boulakia**	Université Paris-Saclay	Examinatrice
Dr. **Raphaël Margueron**	Institut Curie	Examinateur & Président
Pr. **Denis Thieffry**	École Normale Supérieure	Directeur de thèse
Pr. **Piotr Topilko**	Université Paris-Est Créteil	Directeur de thèse

Inserm
La science pour la santé
From science to health

IMRB INSTITUT MONDOR DE RECHERCHE BIOMÉDICALE

IBENS
Institut de Biologie de l'Ecole Normale Supérieure

Identification des mécanismes cellulaires et moléculaires gouvernant la transformation maligne dans la Neurofibromatose de Type 1, en assurant la reproductibilité des analyses transcriptomiques
Audrey Onfroy © 27 septembre 2024
Thèse de doctorat en biologie moléculaire et cellulaire
Rapporteurs : Loredana Martignetti et Eduard Serra
Examinateurs : Sarah Cohen-Boulakia et Raphaël Margueron
Président : Raphaël Margueron
Directeurs de thèse : Denis Thieffry et Piotr Topilko
Université Paris-Est Créteil
École doctorale Sciences de la Vie et de la Santé

Équipe : Oncogenèse des lymphomes et des tumeurs de la Neurofibromatose de Type 1
Institut Mondor de Recherche Biomédicale, UMR 955 Inserm - Université Paris-Est Créteil
8 rue du Général Sarrail
94010 Créteil cedex

Sauf indication contraire, ce travail est sous licence
Except where otherwise noted, this work is licensed under
https://creativecommons.org/licenses/by-nc-sa/4.0/

Acknowledgments

I would like to express my gratitude to the members of the thesis committee, **Dr Loredana Martignetti** and **Dr Eduard Serra**, for accepting to read and comment on my manuscript ; **Professor Sarah Cohen-Boulakia** and **Dr Raphaël Margueron**, for evaluating my work ; and, during my thesis follow-up committee, Dr Raphaël Margueron and **Dr Florence Cavalli**, for their valuable feedback and insights that have helped me refine my research and improve the quality of this thesis.

Je tiens à exprimer une sincère reconnaissance envers **Sarah Cohen-Boulakia**, qui a contribué bien plus qu'en tant qu'examinatrice. Merci de m'avoir acceptée en tant qu'étudiante au sein du Master de Biologie Computationnelle d'Orsay, malgré une candidature très tardive. Les enseignements du Master ont été une découverte et une révélation pour moi. Ils m'ont permis de définir le début de ma carrière professionnelle, en tant que biologiste computationnelle. Merci également de m'avoir mise en contact avec Piotr pour postuler à son offre de thèse de doctorat, qui correspond exactement à mes centres d'intérêt professionnels. Désormais au terme des trois ans de thèse, merci d'avoir accepté le rôle d'examinatrice. Je n'aurais pas pu achever ce parcours sans cette sollicitation pour évaluer mon travail.

Je souhaite remercier les **Professeurs Philippe Gaulard** et **Nicolas Ortonne** de m'avoir accueillie au sein de leur équipe et le **Professeur Pierre Wolkenstein** pour son soutien. Je voudrais ensuite exprimer une profonde gratitude envers mes directeurs de thèse, les **Professeurs Piotr Topilko et Denis Thieffry**. Je vous remercie pour votre soutien et votre disponibilité, de m'avoir permis de présenter mon travail à de nombreuses reprises et de développer une expérience d'enseignement dans le cadre de l'École de Bioinformatique de Roscoff. **Piotr**, merci de m'avoir acceptée au sein de ton laboratoire, et donné l'unique opportunité de conduire cette thèse dans un projet de recherche d'équipe qui correspond exactement aux thématiques qui m'intéressent. Merci pour ta patience, ta bienveillance et ton enthousiasme, qui m'ont accompagnée durant ces trois ans. **Denis**, merci pour nos enrichissantes discussions, qui vont bien au-delà de la thèse, tes nombreux conseils, ta disponibilité internationale, et de m'avoir fait découvrir les bibimbaps.

Je remercie ensuite très chaleureusement le président de l'**Association Neurofibromatoses et Recklinghausen**, Monsieur Jean-Michel Dubois, de m'avoir permis

de participer aux Journées Partage et de m'octroyer un financement afin que je puisse poursuivre mes travaux au sein du laboratoire encore quelques temps.

Merci ensuite à tous les membres de l'**équipe 9**. Tout d'abord, je tiens à remercier l'équipe MPNST. **Kasia**, thank you for transmitting all your knowledge about MPNSTs and trying to make sense of the data. Your requirements on the clarity of the computational analyses and the quality of figures were challenging to establish, but finally I keep them as an objective while conducting research. **Fanny**, merci d'avoir pris le temps de m'expliquer la production des données de A à Z. C'est une chance de pouvoir comprendre (et voir !) l'intégralité du processus à l'origine des données. Merci pour la qualité exceptionnelle et toujours grandissante des données. Le prochain échantillon 23 comptera ses 20 000 parfaites cellules ! **Laure**, merci d'apprivoiser les cellules malignes et persister à toujours trouver les meilleures conditions pour qu'elles s'expriment. Sans cellules, pas de données ! Merci aussi pour nos échanges, très instructifs sur la recherche académique, ou plus digressifs sur la Normandie : y'a une éclaircie là ? Enfin, **Tatiana**, qui tente désormais de percer les mystères des cellules pré-malignes, merci pour tes conseils. Je n'aurais pas pu conduire ma thèse sans vous: la bio-informatique sans biologie, ça reste juste de l'informatique.

Ensuite, je tiens à remercier **Myriam** et Pernelle. Il faut toujours des doctorants d'année supérieure pour transmettre aux nouveaux le protocole du déroulement de la thèse ! Je remercie particulièrement **Pernelle**, pour avoir été une co-doctorante exceptionnelle, pour notre soutien mutuel dans la période de rédaction, et pour tous les moments drôles (*a bo'oh'o'wa'er* !).

L'année dernière, j'ai eu la chance d'accompagner **Emeline**, lors de son stage de Master. Merci d'avoir été une brillante stagiaire, formidablement efficace et pleines de ressources ! Je profite de ce point pour remercier aussi les membres de la DSI de l'UPEC, en particulier **Yahya**, **Vincent** et **Jimmy**, pour leur enthousiasme sur ce projet.

Enfin, merci à tous les autres membres de l'équipe, pour nos moments de partage dans la salle de pause ou pendant les excursions en plein air, les "Vous avez reçu des BL ?" et les "Vous savez où ça se range ?", et pour mon super anniversaire. Merci aux gestionnaires, **Claudia Grangetaud** puis **Mélissa Glasson**, pour votre appui indispensable au fonctionnement de l'équipe.

Pas si loin du labo, je tiens à remercier particulièrement la **Professeure Sophie Hüe**. Juste avant, je remercie à nouveau Piotr de m'avoir permis de découvrir une autre thématique de recherche et d'y consacrer du temps. Sophie, merci pour ta patience à transmettre tes connaissances sur la ~~HD~~ **HS**, pour toutes tes idées et pour ta curiosité à comprendre comment explorer les données. Nous avons formé un binôme très

complémentaire, entre immunologie, fonctionnement du follicule pileux, et analyses transcriptomiques. Cette étroite collaboration a permis d'avancer efficacement sur une thématique biologique que je connais mal. Merci ensuite pour les diverses opportunités et nos discussions sur la thèse. Enfin, merci pour ta disponibilité instantanée et ton enthousiaste devant ces données fantastiques. Pendant ces trois ans de thèse, j'ai eu la chance de travailler avec beaucoup de personnes disponibles et enthousiastes, et cela est très contagieux ! et merci pour les mochis.

Je m'éloigne maintenant de quelques kilomètres de l'hôpital, pour aller à la plateforme de génomique de l'IBENS. Merci à tous pour m'avoir accueillie chaleureusement tous ces vendredis, notamment **Laurent**, **Sophie**, **Ali** et **Salomé**. Plus particulièrement, je tiens à exprimer une profonde reconnaissance et beaucoup de gratitude envers **Morgane Thomas-Chollier**, qui a compté comme une troisième encadrante dans le déroulement de ma thèse. Merci pour ton temps, ta gentillesse et ta générosité. Merci de m'avoir fait me poser les bonnes questions, qui ont été très structurantes. Enfin, merci pour les fichiers de configuration LaTeX ! :)

Direction désormais la Bretagne, à **Roscoff**, pour remercier tous les membres de la formation EB(A)II de m'avoir intégrée à l'équipe "single-cell", et, récemment à l'équipe "bulk". C'est vraiment chouette de pouvoir échanger avec passion à propos de nos packages préférés (ggplot et "dépliir"), et transmettre cette passion à des élèves toujours plus enthousiastes ! Un merci particulier à **Bastien** et **Marine** pour nos échanges sur les analyses scRNA-Seq.

Et comme le code, ça ne fonctionne pas toujours du premier coup, je remercie tous les anonymes de la **communauté** (bio-)informatique, sur divers forums Stack en ligne ou dans les issues Github, pour l'entraide sur tous les aspects de la programmation et le partage de toujours plus d'outils en libre accès.

Je remercie également mes amis : tous les **camarades du lycée**, mais aussi **Myriame**, **Marie** et **Isabelle**. Plus récemment, merci à la **reprise du mardi soir** pour ces moments d'évasion équestre et de partage.

Je voudrais aussi remercier toute ma famille. En particulier, merci à **mes parents** de m'avoir encouragée et soutenue dans la poursuite de longues études. Je transmets maintenant le ~~flambeau~~ clavier à mon **frère Alexis**, qui commence sa thèse 2 jours après que la mienne s'achève. Puisse cette période t'apporter beaucoup de réussite !

Enfin, merci à **Angie** (le chien) et **Odyssée** (le chat), qui ont suivi de très *très* très près la rédaction de cette thèse. Pour clôturer ce générique, merci à **Vincent** pour son soutien quotidien et sa patience olympienne.

Abstract

DECIPHERING CELLULAR AND MOLECULAR MECHANISMS DRIVING MALIGNANT TRANSFORMATION IN NEUROFIBROMATOSIS TYPE 1 USING A REPRODUCIBLE COMPUTATIONAL WORKFLOW FOR SINGLE-CELL TRANSCRIPTOMICS

Neurofibromatosis Type 1 (NF1) is a syndrome caused by *NF1* haploinsufficiency and affecting 1 in 3,000 births worldwide. NF1 patients are predisposed for peripheral nerve sheath tumors (PNSTs) development, arising from the Schwann cell lineage that carry a genetic loss of function in the second *NF1* allele. Among PNSTs, plexiform neurofibromas (pNFs) are slow-growing benign tumors that can become disfiguring and painful. These pNFs can transform into malignant PNSTs (MPNSTs) via a transient dysplastic stage (dyNF). Limited therapeutic options persist due to a poor understanding of the mechanisms driving their initiation.

Our laboratory has developed a mouse model that faithfully recapitulates the pathogenesis of diverse types neurofibromas, including the malignant transformation. In this model, NF1 protein inactivation is confined to boundary cap cells, a specific type of peripheral glial cell likely to be the origin of these tumors. This model offers a unique opportunity to study tumor initiation and progression. To explore the cellular heterogeneity and biological activity of these tumors, we employed the single-cell RNA sequencing (scRNA-Seq) technology. The biological aim of my thesis was to analyze these data to decode the cellular and molecular mechanisms governing tumor initiation and malignant progression. The computational aim was to ensure the reproducibility of the analyses, at the level of the scientific publications.

The scRNA-Seq data processing consists of two main steps, joined by the genes-by-cells count matrix. The first step is reproducible and re-usable in the laboratory. From the count matrices, a computational framework was developed for modular, reproducible, and readable analyses.

We generated a unique scRNA-Seq atlas of peripheral nerves, revealing the early mechanisms driving pNF initiation. Nerves preceding pNFs are thicker (hyperplastic) and characterized by pro-fibrotic activity in tumor Schwann cells, a pro-inflammatory profile in endoneurial fibroblasts, and the expansion of a specific subpopulation of

endoneurial macrophages. This phenotype is triggered by *Nf1* haploinsufficiency and chronic skin trauma.

In our mouse model, MPNSTs developed only in the context of chronic skin trauma, highlighting the role of inflammation in tumor progression. We confirmed the dyNF origin of MPNSTs through allografting experiments. The malignant transformation is characterized by the loss of glial markers in tumor Schwann cells and the acquisition of mesenchymal features. Further research demonstrated that our mouse model replicates the genetic loss of *Cdkn2a*, the initial genetic event in tumor progression in NF1 patients. We identified the transcription factor SOX9 as a biomarker for MPNSTs. Sustained SOX9 expression is associated with Ras signaling pathway activity, positioning it as a potential therapeutic target in MPNSTs.

In conclusion, our work presents the first model that recapitulates the stepwise progression of PNSTs, triggered by chronic skin trauma. We provided a comprehensive transcriptomic atlas of the early stages preceding pNF formation and the malignant transformation, on two genetic backgrounds. While *Nf1* haploinsufficiency induced a more severe phenotype in hyperplastic nerves, it does not affect the composition of MPNSTs. Our findings offer new insights into the pathogenesis of PNSTs and guide the development of novel therapeutic strategies.

Résumé

IDENTIFICATION DES MÉCANISMES CELLULAIRES ET MOLÉCULAIRES GOUVERNANT LA TRANSFORMATION MALIGNE DANS LA NEUROFIBROMATOSE DE TYPE 1, EN ASSURANT LA REPRODUCTIBILITÉ DES ANALYSES TRANSCRIPTOMIQUES

La Neurofibromatose de Type 1 (NF1) est un syndrome causé par une haploinsuffisance du gène *NF1* et affecte une naissance sur 3 000 dans le monde. Les patients sont prédisposés au développement de tumeurs des gaines nerveuses périphériques (PNSTs), issues de la lignée des cellules de Schwann ayant subi une perte de fonction de la seconde copie de *NF1*. Parmi les PNSTs, les neurofibromes plexiformes (pNFs) sont des tumeurs bénignes à croissance lente pouvant devenir défigurantes et douloureuses. Ces pNFs peuvent se transformer en PNSTs malins (MPNSTs) via un stade transitoire (neurofibrome dysplasique, dyNF). Les options thérapeutiques sont limitées en raison d'une faible caractérisation des mécanismes conduisant à cette progression tumorale.

Notre laboratoire a développé un modèle murin qui reproduit fidèlement la pathogénie de divers types de neurofibromes, y compris la transformation maligne. Dans ce modèle, l'inactivation de la protéine NF1 est restreinte aux cellules des capsules frontières, un type spécifique de cellule gliale périphérique probablement à l'origine de ces tumeurs. Ce modèle offre une opportunité unique pour étudier l'initiation et la progression tumorale. Nous avons utilisé la technologie de séquençage d'ARN sur cellules uniques (scRNA-Seq) pour étudier l'hétérogénéité cellulaire de ces tumeurs. L'objectif biologique de ma thèse était d'analyser ces données afin de décoder les mécanismes cellulaires et moléculaires régissant l'initiation et la progression maligne des tumeurs. L'objectif computationnel était d'assurer la reproductibilité des analyses, au niveau des publications scientifiques.

Le traitement des données scRNA-Seq se compose de deux étapes principales, s'accentuant autour d'une matrice de comptage gènes-par-cellules. La première étape est reproductible et réutilisable au sein du laboratoire. À partir des matrices de comptage, un cadre informatique a été développé pour des analyses modulaires, reproductibles et lisibles.

Nous avons généré un atlas transcriptomique unique des nerfs périphériques, révélant les mécanismes précoces conduisant à l'initiation des pNFs. Les nerfs précédant les pNFs sont plus épais (hyperplasiques) et caractérisés par une activité pro-fibrotique dans les cellules de Schwann tumorales, un profil pro-inflammatoire dans les fibroblastes endoneuriaux, et l'expansion d'une sous-population spécifique de macrophages endoneuriaux. Ce phénotype est accentué par l'haploinsuffisance de *Nf1* et les traumas cutanés chroniques. Dans notre modèle murin, les MPNSTs se développent uniquement dans ce contexte de traumas, soulignant le rôle de l'inflammation dans la progression tumorale.

Nous avons confirmé l'origine dyNF des MPNSTs par des expériences d'allogreffes. La transformation maligne se caractérise par la perte de marqueurs gliaux dans les cellules de Schwann tumorales et l'acquisition de caractéristiques mésenchymateuses. Des recherches supplémentaires ont démontré que notre modèle murin reproduit la perte génétique de *Cdkn2a*, l'événement génétique initial dans la progression tumorale chez les patients. Nous avons identifié le facteur de transcription SOX9 comme biomarqueur des MPNSTs. L'expression soutenue de SOX9 est associée à l'activité de la voie de signalisation Ras, la positionnant comme une cible thérapeutique potentielle dans les MPNSTs.

En conclusion, notre travail présente le premier modèle qui récapitule la progression progressive des PNSTs, déclenchée par des traumas cutanés chroniques. Nous avons généré un atlas transcriptomique unique des premiers stades précédant la formation des pNFs et de la transformation maligne, sur deux fonds génétiques. Tandis que l'haploinsuffisance de *Nf1* induit un phénotype plus sévère dans les nerfs hyperplasiques, elle n'affecte pas la composition des MPNSTs. Nos résultats offrent de nouvelles perspectives sur la pathogénie des PNSTs et guident le développement de nouvelles stratégies thérapeutiques.

Contents

Acknowledgments i

Abstract iv

Résumé vi

Contents vii

List of figures xiii

List of tables xvi

List of abbreviations xviii

I Introduction 1

1 From peripheral nervous system to Schwann cells 2

1.1 Overview of the peripheral nervous system 2

1.2 Structure and composition of peripheral nerves 3

1.2.1 Axons support information transmission 3

1.2.2 Schwann cells and satellite cells are the glial cells 4

1.2.3 Peripheral nerves contain three fibroblast subtypes 5

1.2.4 Macrophages are the main immune type 6

1.2.5 Vasculature accompanies peripheral innervation 7

1.3 Schwann cells originate from the neural crest 8

1.3.1 Origin and development of the neural crest 8

1.3.2 Boundary cap cells form a transient neural crest-derived population 11

1.3.3 Schwann cells precursors are multipotent glial cells 13

1.3.4 Immature SCs initiate the final structure of peripheral nerves 14

1.4 The multiple facets of Schwann cells in the adult nerves 15

1.4.1 Insight into Schwann cells in the physiological state 15

1.4.2 Mature SCs switch between steady-state and repair phenotype 18
1.5 Overview of the chapter 22
2 NF1, a monogenic but complex syndrome 23
2.1 Generalities about the disease 23
2.1.1 Symptoms, treatment and diagnosis 23
2.1.2 Molecular implications behind NF1 disease 24
2.1.3 Diversity of benign neurofibromas 28
2.2 Plexiform neurofibromas are benign nerve sheath tumors 29
2.2.1 Diagnosis and treatments 29
2.2.2 Cellular origin 30
2.2.3 Cellular and molecular features of pNFs 31
2.2.4 No model of the spontaneous malignant transformation of pNFs 34
2.3 Dysplastic neurofibromas are pre-malignant lesions 35
2.3.1 Terminology 35
2.3.2 Cellular composition 36
2.3.3 Loss of *CDKN2A* 36
2.3.4 Lack of model recapitulating the pNF-to-dyNF transition 36
2.4 MPNST are aggressive soft tissue sarcomas 37
2.4.1 Management of NF1 patients with MPNSTs 37
2.4.2 Half of MPNSTs develop in sporadic context 38
2.4.3 MPNSTs display epigenetic and genetic anomalies 39
2.4.4 Cellular composition and molecular identity of MPNSTs 43
2.4.5 Cell lines and genetically engineered animal models for MPNSTs 46
2.5 A *Nf1*-KO mouse model recapitulates several features of NF1 47
2.6 Overview of the chapter 48
3 Sequencing breakthroughs revolutionize biology 49
3.1 Overview of genomic approaches 49
3.2 scRNA-Seq enables single-cell level analysis of tissue 52
3.2.1 Preparation of the single cells suspension 52
3.2.2 Tracing the cellular origin of transcripts 54
3.2.3 Library preparation and sequencing with Illumina 57
3.3 Computational processing of scRNA-Seq 58
3.3.1 From the reads to a count matrix 59
3.3.2 The count matrix as a wealth of information 62
3.4 Outlook and feedback 68

4 Reproducibility challenges in computational biology **69**
4.1 Why making reproducible research ? 69
4.2 Reproducibility crisis 70
4.2.1 Global reasons for the reproducibility crisis 70
4.2.2 The case of computational biology 72
4.3 Promote reproducible research 73
4.3.1 Global solutions to overcome the crisis 73
4.3.2 Specific solutions for computational biology 75
4.4 Outlook and feedback 79

II Aim of the thesis **81**

III Materials and Methods **85**

1 Biological sample presentation **86**
1.1 Mice data produced in the lab 86
1.1.1 Presentation of the mouse models 86
1.1.2 Experimental designs 86
1.1.3 Available data overview 88
1.2 Patients data 90
1.3 Data from public repositories 91

2 Reproducibility of the computational analyses **92**
2.1 Traceability 92
2.2 R package 94
2.3 Computational environment control 95
2.4 Reproducibility evaluation 96

3 Construction of the datasets **97**
3.1 Individual processing 97
3.1.1 Obtaining the individual count matrices 97
3.1.2 Gene names and Ensembl ID 98
3.1.3 Count matrix to annotated and visualisable Seurat object 98
3.2 Atlas and cell population datasets 99

4 Analysis of the datasets **102**
4.1 Gene expression analyses 102
4.1.1 Differential expression analysis 102
4.1.2 Enrichment analysis 102
4.2 Trajectory inference and RNA velocity 104

4.2.1 Trajectory inference from a topology 104
4.2.2 RNA velocity 104
4.2.3 Expression pattern along pseudotime 105
4.3 Inference of CNV from scRNA-Seq data 106
5 Tools to write this thesis 108

IV Results 111
1 Computational results 114
1.1 Automation of the custom reference construction 114
1.2 Improvement of the pre-existing pipeline for reads processing 115
1.2.1 Motivations and global features 115
1.2.2 Architecture of the pipeline 116
1.3 The aquarius package: a wrapper for simplified analyses 117
1.3.1 Example of functions 118
1.3.2 A marker-based, single-cell level, cell type annotation 119
1.4 Establishment of a framework for reproducible analyses 121
1.4.1 Use of R Markdown notebooks to ensure traceability 121
1.4.2 Notebooks tree for flexible analyses 123
1.4.3 Containerized compilation ensures reproducible analyses 125
1.4.4 Index page enables easy navigation through the compiled notebooks 125
1.5 Summary 127
2 Article 1 128
2.1 Overview of the study 128
2.2 Exploring the mechanisms governing the initiation of plexiform neurofibromas using Prss56Cre, *Nf1*-KO mouse model at a single-cell resolution ... 129
2.3 Current and future directions 151
2.3.1 Code availability 151
2.3.2 Combination of skin trauma and *Nf1* haploinsufficiency in nerves 152
3 Article 2 154
3.1 Overview of the study 154
3.2 A mouse model uncovers mechanisms of malignant transformation in Neurofibromatosis Type 1 155
3.3 Complementary results 185
3.3.1 Histological assessment of tumor type and single-cell transcriptomics 185
3.3.2 Nodular pNFs are likely to be the origin of dyNF 186
3.3.3 Discrepancies between RNA velocity and trajectory inference 188

3.3.4 CNV informs about the intra-tumoral clonality 189
4 Article 3 ..... **192**
4.1 Overview of the study 192
4.2 The molecular signature of hair follicles classifies *Hidradenitis Suppurativa* patients into inflammatory, keratinizing, and mixed endotypes 193
4.3 Outlook 217

V Discussion **219**
1 Summary of the contributions **220**
1.1 Computational contributions 220
1.2 Biological contributions 221
2 Discussion on the computational part **222**
2.1 Hindsight regarding the analysis of scRNA-Seq data 222
2.1.1 Biological relevance 222
2.1.2 Variability of the discrete cell clustering 222
2.1.3 Trajectory inference as a continuous clustering 223
2.1.4 Limits of the gene sets enrichment analyses 223
2.2 Perspectives about the methods 224
2.2.1 Benefits and future directions of the cell type annotation method 224
2.2.2 Methodological improvement 225
2.2.3 The ideal tool is not always the one being used 225
2.3 Missing documentation for re-usability 226
2.3.1 Limitations of the aquarius package 226
2.3.2 Re-usability of the analysis framework 227
2.4 Making research reproducible 227
2.4.1 Chronology of the five pillars 227
2.4.2 The reproducibility as a methodological step in computational biology 229
3 Discussion on the biological part **230**
3.1 The *Nf1*-KO mouse model 230
3.1.1 The *Nf1*-KO mouse model recapitulates several features of NF1 disease 230
3.1.2 Limitation of PNSTs studies to male mice 230
3.2 Fate of boundary cap cells in the peripheral nervous system 232
3.2.1 Distinct fate between BC-derived and NC-derived cells 232
3.2.2 Predisposition of BC-derived cells to form neurofibromas 232
3.3 Triggers of tumor development and malignant progression 233

3.3.1 *Nf1* haploinsufficiency in the micro-environment 233
3.3.2 Skin trauma and inflammation 234
3.4 From hN to pNF: the hypothesis of unrepaired wound 235
3.4.1 Repair phenotype in hyperplastic nerves and pNFs 235
3.4.2 From nerve-associated fibroblasts to cancer-associated fibroblasts .. 235
3.4.3 Repair phenotype of tumor SCs 236
3.4.4 Late events of pNF initiation 237
3.5 From pNF to dyNF: a trigger in the nerve micro-environment 237
3.6 From dyNF to MPNST: the glial-to-mesenchymal transition 238
3.6.1 Chronology of genetic events and GMT 239
3.6.2 The hypothesis of cancer stem cells 240
3.6.3 A population-level transition or a cellular switch ? 240
3.6.4 Relevance in the sporadic context 241
3.7 Therapeutic hopes 242
3.7.1 Target the fibrosis via periostin 242
3.7.2 A therapeutic window through the immune infiltrate of dyNF 243
3.7.3 Ras signaling and MPNST progression 243

VI Appendix **247**

1 List of genes and proteins **248**

2 R packages in the R 3.6.3 Singularity container **255**

3 Composition of Conda environments **261**

4 Supplementary materials **262**
4.1 The Seurat object format 262
4.2 Additional assays 263
4.3 Differential expression analyses 264

5 Supplementary results **266**

6 Participation to scientific events **268**
Poster 1: JOBIM 2022 and MaDICS Symposium 2022 269
Poster 2: Doctoral School Day 2022 270
Poster 3: 2023 Global NF Conference 271
Poster 4: 2024 Global NF Conference 272

References **275**

List of Figures

1 The peripheral nervous system is divided into subsystems 2
2 Peripheral nerve structure 3
3 Structure of a neuron 4
4 Classification of fibroblasts found in peripheral nerves, and main markers . 5
5 Classification of immune cells found in peripheral nerves, and main markers 6
6 Classification of vascular cells found in peripheral nerves, and main markers 7
7 From neural plate induction to neural crest delamination 8
8 Molecular cascade controlling trunk NCCs delamination and migration 10
9 Neural crest migration and trunk derivatives 11
10 Boundary cap cells is a vestige of NCCs giving rise to Schwann cells 11
11 Construction of *Prss56*$^{Cre,+}$; Rosa26$^{Tom,+}$ mouse strain 12
12 Schwann cells precursors are multipotent glial cells 14
13 Structure of myelinating and non-myelinating Schwann cells 16
14 Schwann cells functions from nerve injury to recovery 19
15 Signaling pathways involved in the identity of repair Schwann cells 20
16 Implication of CDKN2A and CCND1 in cell cycle 21
17 Schwann cells are glial cells of the peripheral nervous system 22
18 NF1 disease covers a large range of symptoms 24
19 *NF1* gene and protein structure 25
20 RAS molecular switch and signaling 25
21 Neurofibromin signaling pathways and relation with RASopathies 26
22 Various types of neurofibromas 28
23 Cellular and molecular features of plexiform neurofibromas 31
24 Risk factors and patients care 37
25 Etiology of MPNSTs 38
26 Molecular events underlying MPNST progression 40
27 Recruitment and structure of the Polycomb Repressive Complex 2 42
28 PRC2 and histone modifications 42
29 Biomarkers and signaling pathways defining MPNSTs 43
30 *Prss56*^{Cre} *Nf1*-KO mouse model recapitulates several features of NF1 47

31 Human genome composition 49
32 DNA to protein processing and genomic approaches of interest 50
33 Advantages and limitations of scRNA-Seq over bulk RNA-Seq 51
34 Single cells suspension preparation 52
35 Tracing the cellular origin of transcripts with barcoded droplets 54
36 RNA transcription and processing 56
37 Illumina-specific library preparation and sequencing 57
38 Overview of scRNA-Seq data processing 58
39 Main processing steps for reads to count matrix 59
40 Build the count matrix from reads 61
41 Standard pipeline to prepare the count matrix for analysis 63
42 Multiple directions make sense of scRNA-Seq data 65
43 Representations are a glimpse of high-dimensional data 67
44 Factors contributing to irreproducible research 71
45 Reproducibility issues in computational biology 72
46 Solutions promoting reproducible research 74
47 Reproducible computational biology relies on five pillars 75

48 Management of spontaneous tumors in *Nf1*-KO male mice 87
49 Management of biobanked cells from *Nf1*-KO male mice 87
50 Metadata of scRNA-Seq from mice, generated in the laboratory 89
51 Information regarding all patients data presented in this thesis 90
52 Framework for reproducible computational analyses 92
53 From R Markdown notebook to their HTML compiled version 93
54 Selection of the number of dimensions to generate 2D representation 100
55 Influence of the number of diffusion components on the UMAP appearance 101
56 Gene expression patterns along pseudotime 105
57 Metrics used to order features on the heatmap 106

58 Nextflow workflow structure to build the custom reference 115
59 Nextflow workflow structure to obtain the count matrix 117
60 Cell type annotation method in aquarius::cell_annot_custom function 120
61 Best practices in the development of R Markdown notebooks 122
62 Directory tree 123
63 Notebook tree 124
64 Structure of the index.html page 126
65 Three-step pipeline to generate the index.html page 127

66 Graphical abstract of the Article 1 129
67 Screenshot of the repository page associated with the Article 1 151
68 Screenshot of a page from the website associated with the Article 1..... 152
69 Graphical abstract of the Article 2 155
70 *Sox10*+ and *Sox9*+ tumor cells balance is associated with tumor type..... 185
71 Nodular pNFs contain cancer-associated fibroblasts 186
72 Nodular pNFs tumor cells share identity with dyNF tumor cells 187
73 Model of pNF progression 188
74 Comparison between trajectory inference and RNA velocity 189
75 Distinct tumor subpopulations are associated with specific CNV in 2021_33 190
76 Graphical abstract of the Article 3 193

77 Fixing the analyses can lead to non robust results 229
78 *Nf1*-KO mouse model recapitulates the development of PNSTs 231

79 Overview of the Seurat object structure, in Seurat version 3 262
80 Construction of additional assays and their usage..... 263
81 Whole Exome Sequencing genomic profile of 2021_33..... 266
82 Distinct tumor subpopulations are associated with specific CNV in 2022_26 267

List of Tables

1 Nomenclature and associated histological features of NF1-related PNSTs. 35
2 Comparison of Conda with Docker and Singularity 77

3 Mouse models of interest in the laboratory 86
4 Overview of the reanalyzed published scRNA-Seq data 91
5 Issues with existing solutions to build HTML pages from (R) Markdown files 94
6 Content of the Conda environments 95
7 Content of the individual Seurat object 98
8 Two-dimension projections to visualize cells 101
9 Tools to write this thesis 108

10 Features underlying the improvement of the reads-to-count matrix pipeline 116
11 Content of the aquarius package, version 0.1.5 117
12 Example of wrappers in the aquarius package 118
13 Example of aquarius functions for visualization 118
14 Validation datasets for the set of markers used in cellular annotation 121

15 Conda environment to use velocyto 261
16 Conda environment to use scvelo 261
17 Output of differential expression analysis, using Seurat::FindMarkers..... 265

List of abbreviations

BC(C)	boundary cap (cell)
CAF	cancer-associated fibroblast
CB	cell barcode
cNF	cutaneous neurofibroma
CNV	copy number variation
CPU	central processing unit
DE	differential(ly) express(ion)(essed)
DEA	differential expression analysis
DNA	deoxyribonucleic acid
DRG	dorsal root ganglion
dyNF	dysplastic neurofibroma
ECM	extra-cellular matrix
EMT	epithelial-to-mesenchymal transition
endoFb	endoneurial fibroblast
epiFb	epineurial fibroblast
FACS	fluorescence-activated cell sorting
FAIR	findable, accessible, interoperable and reusable
GAP	GTPase activating protein
GB	gigabyte
GDP	guanosine diphosphate
GMT	glial-to-mesenchymal transition
GSEA	gene set enrichment analysis
GTP	guanosine triphosphate
HF	hair follicle
HVG	highly variable genes
iSC	immature Schwann cell
mDC	monocyte-derived dendritic cell
MPNST	malignant peripheral nerve sheath tumor
mSC	myelinating Schwann cell

NAF	nerve-associated fibroblast
NC(C)	neural crest (cell)
NF	neurofibroma
NF1	neurofibromatosis type 1
nmSC	non-myelinating Schwann cell
NS	nervous system
OS	operating system
PCA	principal component analysis
periFb	perineurial fibroblast
pNF	plexiform neurofibroma
PNS	peripheral nervous system
PNST	peripheral nerve sheath tumor
PRC2	polycomb repressive complex 2
QC	quality control
RNA	ribonucleic acid
rSC	repair Schwann cell
RTK	receptor tyrosine kinase
SC	Schwann cell
SCP	Schwann cell precursor
scRNA-Seq	single cell RNA sequencing
snRNA-Seq	single nuclei RNA sequencing
SNV	single nucleotide variant/variation
SNP	single nucleotide polymorphism
TF	transcription factor
(T)ME	(tumor) micro-environment
tSNE	t-distributed stochastic neighbor embedding
UMAP	uniform manifold approximation and projection
UMI	unique molecular identifier
WES	whole exome sequencing

I Introduction

Malignant peripheral nerve sheath tumors are aggressive tumors that occur particularly in patients with Neurofibromatosis Type 1 (NF1). My work belongs to a team project aiming to study the cellular and molecular mechanisms that control tumorigenesis and malignant progression.

First, Section 1 presents a biological overview essential for understanding the biology behind tumors. Subsequently, Section 2 presents the NF1 disease. NF1 tumors exhibit heterogeneity in their composition and evolve over time. To gain more precise understanding of the heterogeneity of tumor samples, the laboratory took advantage of single-cell RNA sequencing technology. Therefore, Section 3 provides an overview of the most recent analytical techniques and the scope of interpretations enable by such data. Since analyses are conducted through many successive steps, Section 4 addresses the challenges of reproducibility in computational biology.

Except Figures 2, 10 and 45, all figures in the introduction are originally created using BioRender.com, a software to create scientific illustrations. All genes and proteins are explicitely named in the Appendix, Section 1.

1 From peripheral nervous system to Schwann cells

1.1 Overview of the peripheral nervous system

The nervous system (NS) orchestrates the coordination of movements and sensory information. In vertebrates, the NS is organized into the central (CNS) and the peripheral nervous systems (PNS) (Figure 1).

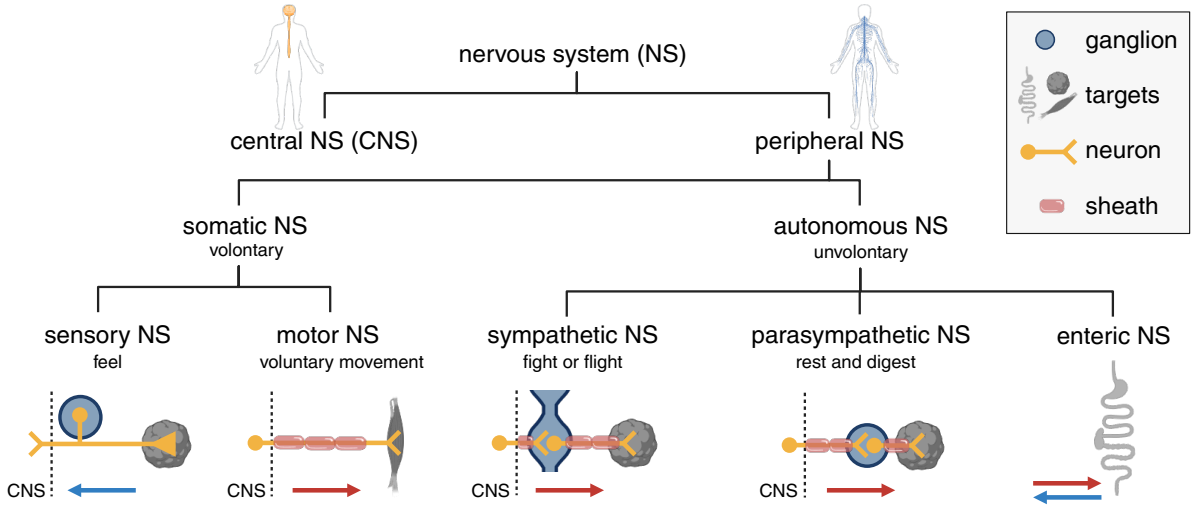


Figure 1 – The peripheral nervous system is divided into subsystems. The PNS is divided into two systems: the somatic and the autonomous. On the bottom, the most typical neural structure for each subsystem is demonstrated. The arrows portray the electric signal's direction. Inspired from Calabrò *et al.* (2023).

The CNS, consisting of the brain and the spinal cord, processes and integrates information from various parts of the body, to initiate appropriate responses. The PNS, composed of ganglions and peripheral nerves, is responsible for transmitting information between the CNS and the rest of the body. Each function is associated to a specific subsystem (Figure 1).

1.2 Structure and composition of peripheral nerves

In the NS, the functional unit is the neuron (Section 1.2.1). Their axons are protected by a series of nested protective layers (Figure 2). First, they are wrapped by Schwann cells, which, according to the axon diameter, produce myelin or not (Section 1.4). The axon-Schwann cells pair forms a nerve fiber. Nerve fibers that converge in the same region of the body are assembled to form nerve fascicle (Stewart, 2003), which is maintained by the perineurium. The perineurium defines an inner space called the endoneurium. Finally, several fascicles are protected by the epineurium, delineating the nerve. A vascular plexus is organized as a net around the nerve, and go through each protective layer to supply the nerve in nutrients.

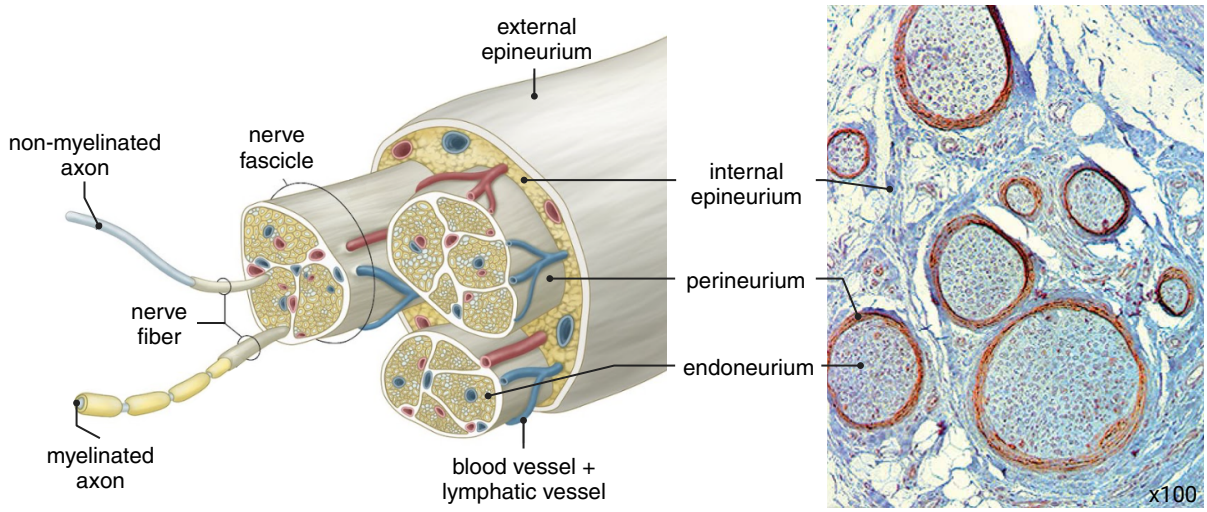


Figure 2 – **Peripheral nerve structure**. **A.** The main structures of a peripheral nerve are annotated. Adapted from *Nerve tissue, synapses, and neurotransmitters* (2022). **B.** Transversal nerve section showing nerve fascicles. The perineurium is revealed by the epithelial membrane antigen. Adapted from Piña-Oviedo and Ortiz-Hidalgo (2008).

1.2.1 Axons support information transmission

A neuron is an excitable cell capable of initiating and conducting nerve impulses via an extension from its cell body, known as the axon (Figure 3). The axon is a specialized tubular extension of the cell body whose architecture is maintained by elements of the

cytoskeleton. Besides its structural function, the cytoskeleton supports the transport of molecules and organelles between the cell body and the distal region.

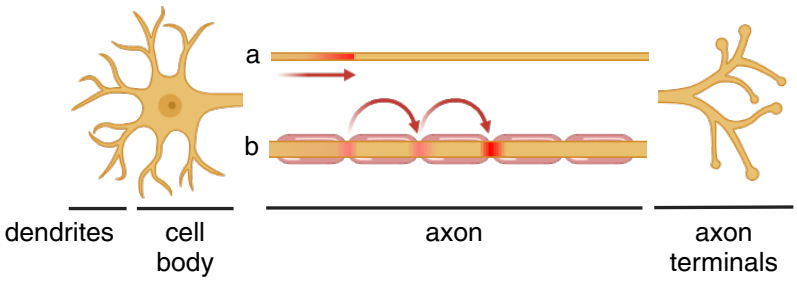


Figure 3 – **Structure of a neuron**. Red arrows depict the direction of signal transduction. In small-diameter axons (**a**), the signal is progressively transduced, while large-diameter axons (**b**), an isolating sheath leads to a saltatory conduction.

The dendrites are short cytoplasmic extensions that surround the cell body of certain neurons. Unlike the dendrites, the diameter of the axons does not vary for a given neuron and ranges from 0.2 to 20 µm. There are two main types of neurons: motor neurons and sensory neurons. Sensory neurons are T-shaped, called bipolar, have small-diameter axon, below 1.5 µm, and the axons of several neurons are gathered together by Schwann cells (next Sections). On the opposite, motor neurons are multipolar, due to their many dendrites, and have a large-diameter axon, surrounded by Schwann cells that make several loops around axons and produce myelin. The myelin sheath allows for saltatory electrical signal transduction, faster compared to a progressive transduction (Figure 3b). Axons transmit information in the form of action potentials, which are discrete electrical impulses that travel along the membrane (Figure 3, arrows). In the terminal region, the axon divides into many branches called axon terminals (Figure 3). Each axon terminal forms a synapse with its target. This allows a single neuron to send signals to multiple targets. The action potential is converted into a chemical signal, called neurotransmitters, which circulates in the extracellular space between the source and the target cells, known as a synapse. Neuronal targets receive information through their dendrites, and possibly from multiple sources. If the collected information is sufficient, it can trigger a new action potential that is propagated along an eventual post-synaptic axon.

1.2.2 Schwann cells and satellite cells are the glial cells

Neurons are in direct contact only with glial cells, which surround them as a protective layer. All glial cells originate from the neural crest (Section 1.3). In the adult PNS, two types of glial cells are found: **Schwann cells** (Section 1.4), and **satellite glial cells**.

Satellite glial cells (SGCs) are found exclusively in the peripheral ganglions. They form a tight sheath covering neuronal cell bodies, usually separating each one (Hanani & Spray, 2020). The sheath is maintained by adhesive, tight and gap junctions between SGCs. In particular, gap junctions allow SGCs to be synchronized for a given cell body, to communicate and exchange molecules.

1.2.3 Peripheral nerves contain three fibroblast subtypes

A fibroblast is notably responsible for the extra-cellular matrix (ECM) modeling, and is found in connective tissue. Its shape is flat, elongated and not polarized. In nerve, three fibroblasts subpopulations are characterized by their location, functions, and expression of specific markers (Figure 4). They are organized in three nested layers, from the inner to the outer: endoneurium (purple), perineurium (green) and epineurium (blue).

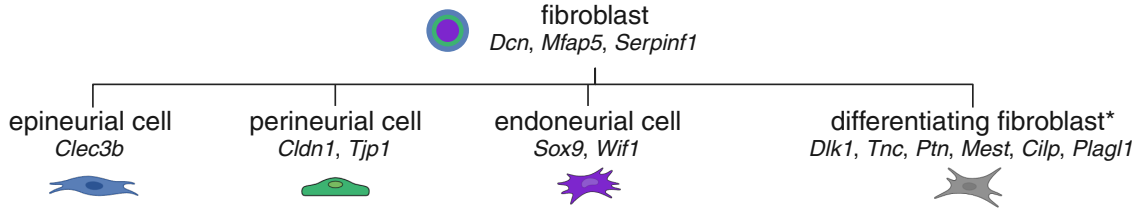


Figure 4 – Classification of fibroblasts found in peripheral nerves, and main markers. Markers come from scRNA-Seq studies made on sciatic nerves (Carr *et al.*, 2018; Toma *et al.*, 2020; Zhao *et al.*, 2022). Star (*): Differentiating fibroblasts are a transient population present after nerve injury.

Endoneurial fibroblasts (endoFb) fill the endoneurium and are in direct contact with Schwann cells. They are responsible for collagen synthesis, a component of the ECM. They may be implied in immune surveillance (Richard *et al.*, 2012).

Perineurial fibroblasts (periFb) delineate nerve fascicles. They define the border between endoneurial (inside) and epineurial (outside) environments. As a physical protective layer, they express tight junctions proteins, such as TJP1 and CLDN1 (Kucenas, 2015). PeriFb protect axons from ionic flux, toxins and infection. They belong to the blood-nerve barrier by protecting endoneurium from damaging substances from the bloodstream (Reinhold & Rittner, 2020). Thus, in diseases that result in perineurium decompaction, patients are exposed to serious viral or bacterial infections that lead to inflammatory diseases. Finally, periFb have been shown to be the first cellular element to bridge the injury gap after sciatic nerve section (Kucenas, 2015).

Epineurial fibroblasts (epiFb) constitute the outermost layer of the nerve and are poorly characterized. In the context of nerve injury, due to new proximity between epiFb and Schwann cells, secreted factors from epiFb affect the repair ability of Schwann cells (Neerven *et al.*, 2013).

All three types of fibroblasts have been shown to participate to nerve repair. Particular, an injury-specific population has been identified in Parrinello *et al.* (2010). Specific markers for these **differentiating fibroblasts** have been identified at a transcriptomic level (Carr *et al.* (2018), Figure 4).

1.2.4 Macrophages are the main immune type

Two types of immune cells are found in the nerves. **Resident** immune cells ensure surveillance, while **circulating** immune cells are present in the vessels. They can come out of vessels, when recruited in the nerve or in pathological conditions. Although macrophages are the main immune population found in nerves, higher heterogeneity has been identified through single-cell RNA Sequencing (scRNA-Seq) analyses (Figure 5, Gerber *et al.* (2021) and Chen *et al.* (2021)).

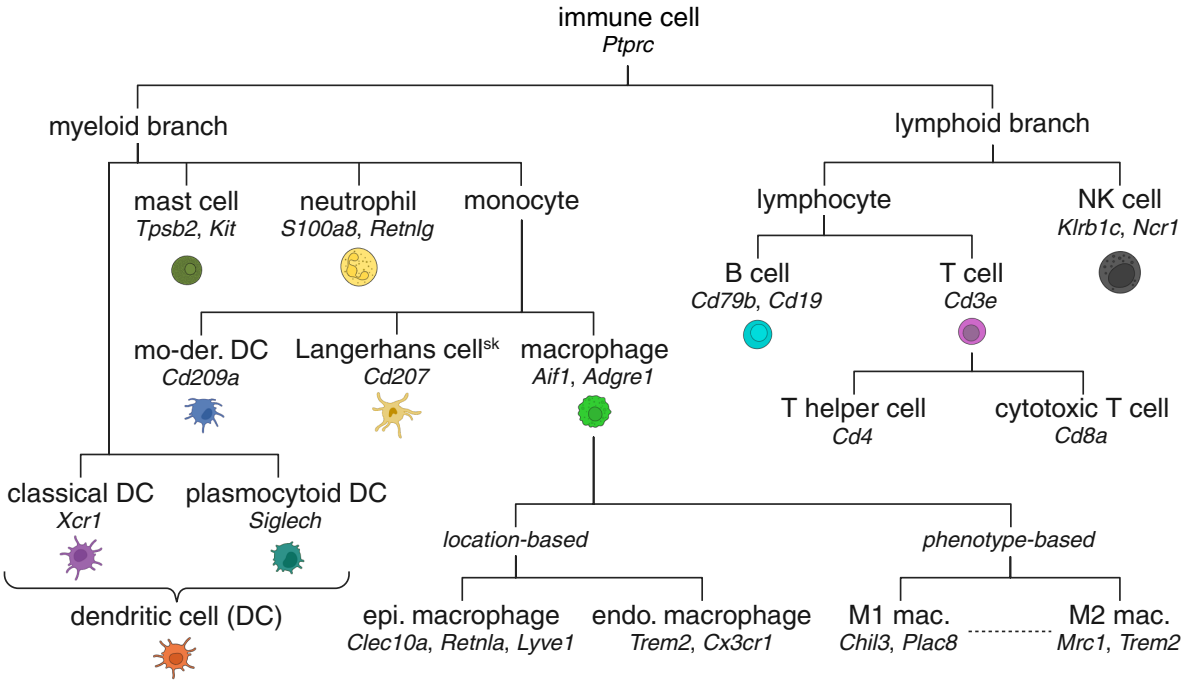


Figure 5 – Classification of immune cells found in peripheral nerves, and main markers. Markers come from scRNA-Seq studies made on sciatic nerves. Dotted line between macrophages highlights a continuum between classical macrophage subtypes (Katkar & Ghosh, 2023). Dendritic cells (DCs) originate from monocytes (mo-der.: monocyte-derived) or not. DC subtypes and markers are described in Eisenbarth (2019). epi.: epineurial, endo.: endoneurial, [sk]: Langerhans cell is a specific population located in the skin.

In the endoneurium, resident immune cells are mainly macrophages (Stierli *et al.*, 2018). Classically, macrophages are divided into anti-inflammatory "M1" and pro-inflammatory "M2" subtypes. However, this binarization is under active discussion since macrophages may more represent a continuum between several states (Katkar & Ghosh, 2023). Beyond a phenotype-based classification, macrophages have also been characterized from their location and have distinct response to injury (Amann & Prinz, 2020; Ydens *et al.*, 2020). Indeed, outer **epineurial** macrophages do not react to sciatic nerve crush injury while inner **endoneurial** macrophages, close to Schwann cells and axons, express genes involved in circulating monocytes recruitment, that, once in the nerve, rapidly differentiate into macrophages (Ydens *et al.*, 2020). Finally, using scRNA-Seq technologies, monocytes and macrophages are challenging to distinguish and are often grouped as a "monocyte-macrophage" population.

1.2.5 Vasculature accompanies peripheral innervation

Nerves are vascularized by vessels that form a plexus around nerves (Figure 2). In a stepwise manner, they ramify and traverse the layer of fibroblasts to provide oxygen, glucose and nutrients up to the innermost part (Yeoh *et al.*, 2022). **Vascular cells**, forming lymphatic or blood vessels, classically comprise **endothelial cells** and **mural cells** (Figure 6). Both populations depict high heterogeneity based on their location: epineurial or endoneurial ; the diameter of the vessel they are surrounding: from arteries to arterioles, veins to venules, and capillaries ; and the fluid type conveyed: lymph, oxygen-rich blood and oxygen-deprived blood.

During nerve injury, blood flow is slowed down (nerve compression) or disrupted (nerve section), leading to oxygen privation, known as hypoxia (Yeoh *et al.*, 2022). The hypoxia participates to the recruitment of circulating macrophages at the injury site. They are required for other nerve cell types to regenerate the injury site.

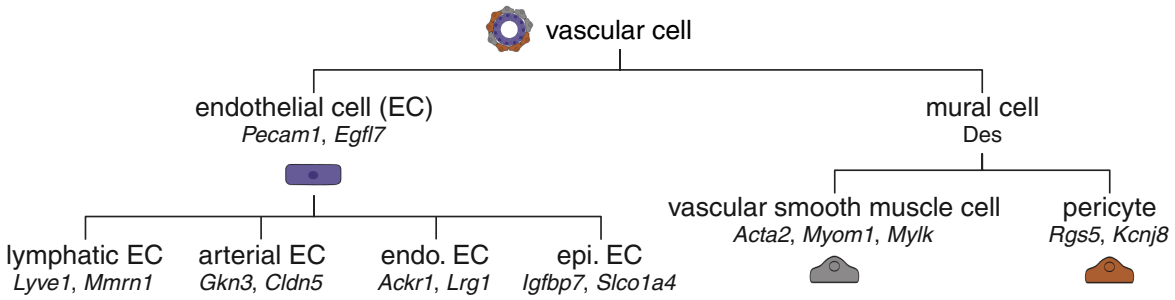


Figure 6 – **Classification of vascular cells found in peripheral nerves, and main markers**. Markers come from Kalucka *et al.* (2020) and Chen *et al.* (2021) and Chandrasekaran *et al.* (2022). epi.: epineurial, endo.: endoneurial.

1.3 Schwann cells originate from the neural crest

This thesis concentrates on tumors arising in peripheral nerves, and tumor cells originate from Schwann cells (SCs). Therefore, this section concentrates on the developmental biology of SCs, and next section outlines their various roles in the adult PNS (Section 1.4). In next paragraphs, I refer to the stages of human embryonic development in weeks using "W", while for mice, I use "E" to represent embryonic day.

1.3.1 Origin and development of the neural crest

Schwann cells (SCs) come from the **neural crest**, a specific region of the neural tube. The neural tube originates from the folding of the neural plate, a specific region of the ectoderm (Figure 7A). At W3 or E6, ectoderm undergoes medio-lateral specification in neural ectoderm and non-neural ectoderm, separated by the neural border. The neural ectoderm further thicken and becomes the neural plate (Figure 7B). Then, the neural plate curves inward, and both sides of neural border fuse, forming the neural tube (Figure 7C). After the neural plate folding, the neural plate borders form the neural crest (Figure 7D).

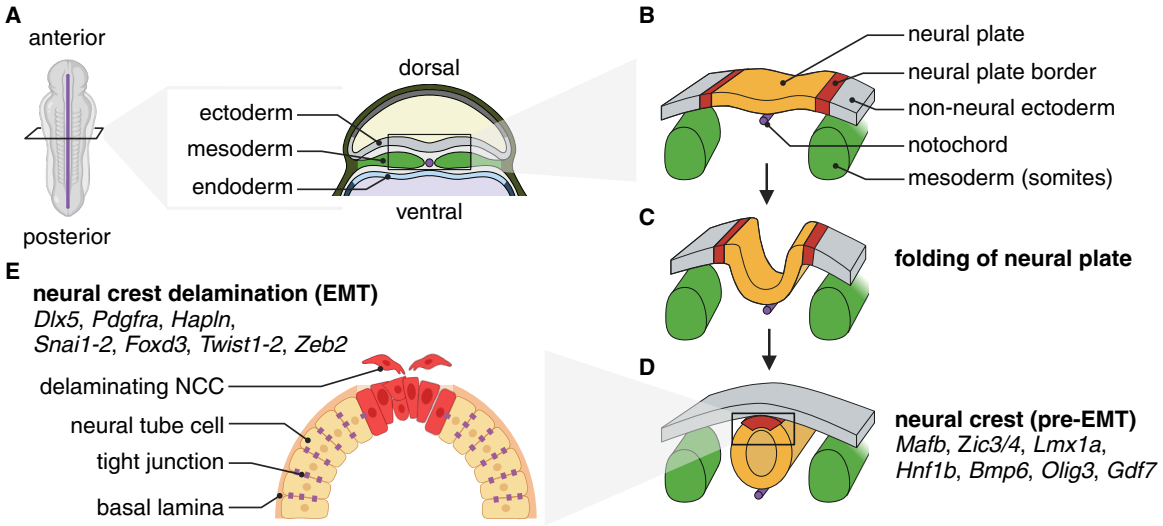


Figure 7 – From neural plate induction to neural crest delamination. **A**. Antero-posterior view of an embryo with delineated notochord in purple. Transversal section of the trunk is represented on the right. The rectangle delimits the dorsal region represented in cross sections (B-D). Adapted from Carlson (2014). **B**. The notochord (purple) induces the definition of the neural plate borders (red). The beginning of the neurulation is characterized by the neural plate thickening. **C**. Folding of the neural plate. **D.** Both sides of neural border fuse, forming the neural tube, marking the end of neurulation. The rectangle delimits the neural tube region represented in cross section E. **E**. Neural crest cells (NCCs, in red) undergo an epithelial-to-mesenchymal transition (EMT) by down-regulating tight junctions and up-regulated mesenchymal genes, and delaminates from the neural tube. NCCs marker genes for each developmental stage come from Soldatov *et al.* (2019). Figures adapted from Simões-Costa and Bronner (2015), Mayor and Theveneau (2013), Furlan and Adameyko (2018), and Erickson *et al.* (2023).

Cells from the neural tube are epithelial cells: they are polarized. The basal pole produces and is in contact with a basal lamina, a specific ECM, which separates them from the underlying connective tissue. The apical pole corresponds to the inner surface of the neural tube. Neurepithelial cells are strongly maintained by tight junctions, which make them a barrier between the outer and the inner part they define (Figure 7E).

At W4 or E7, neurepithelial cells undergo an **epithelial-to-mesenchymal transition** (EMT) and become neural crest cells (NCCs). A key feature of EMT is a cadherin switch, i.e. the loss of E-cadherin (*Cdh1*) in favor to type II cadherins. EMT corresponds to a global change in adhesion properties, leading cells to degrade their basal lamina. This event in called **delamination** (Mayor & Theveneau, 2013).

Box 1: The SoxE group, driver of cell fate

During delamination and migration, it is noteworthy that NCCs simultaneously express two transcription factors (TFs) of **SoxE** gene family: *Sox9* and *Sox10*, both of interest during my thesis. In humans and mice, there are 20 Sox genes forming a family of TFs. They are defined by a high sequence homology to the DNA binding domain of SRY protein, called high mobility group (HMG). Due to their sequence similarity, Sox genes are categorized into 9 groups. The SoxE group contains *Sox8*, *Sox9* and *Sox10* genes (Julian *et al.*, 2017).

Sox10 is a key regulator in the differentiation of the peripheral glial cells (Britsch *et al.*, 2001), and persists in mature glial cells. It also participates to the development of enteric nervous system and melanocytes (Qi *et al.*, 2022). Despite all SoxE genes have a dimerization domain, SOX10 binds to DNA both either as a monomer or a dimer. According to its binding partner, SOX10 functions vary. Together with MITF, it controls melanocyte development, while with PAX3, it participates to neural crest development. In SCs, by interacting with KROX20, it is responsible for myelination (Qi *et al.*, 2022).

Sox9 is implicated in the development of embryonic layers, in adult and in many pathologies (Jo *et al.*, 2014). SOX9 requires a binding partner, which define its activity as an activator or repressor of the transcription. Notably, it can work as a homodimer and act as a self-activator. Signaling pathways up-regulating *Sox9* expression include Wnt, Bmp, Shh, Fgf and TGF-β, while Notch inhibits *Sox9* expression. SOX9 participates to many cell types development, including chondrogenesis. Activation of *Sox9* can leads to EMT, in NCCs as we discussed above, but also in cancer. Finally, SOX9 is implicated in excessive ECM deposition, know as fibrosis, particularly through collagens synthesis (Jo *et al.*, 2014).

The Wnt pathway, implying *Wnt1* and *Bmp4*, initiates the EMT (Figure 8, Theveneau and Mayor (2012)). While *Wnt1* favors cell cycle progression, *Bmp4* is responsible for basal lamina degradation, through the activation of metalloproteases that reshape the ECM. Particularly, *Bmp4* activates several transcription factor (TFs), including *Snai2* and *Sox9* (Box 1), two mesenchymal cells markers. *Sox9* inhibits *Wnt1* and *Bmp4*, in a negative retro-control loop manner, and activates two other TFs: *Foxd3* and *Sox10* (Box 1). All three TFs are directly responsible for the cadherin switch and remodeling of the ECM. Due to change in their interaction with ECM and modification of their cytoskeleton, delaminating NCCs have **migratory** capacities, starting from W4 or E9.

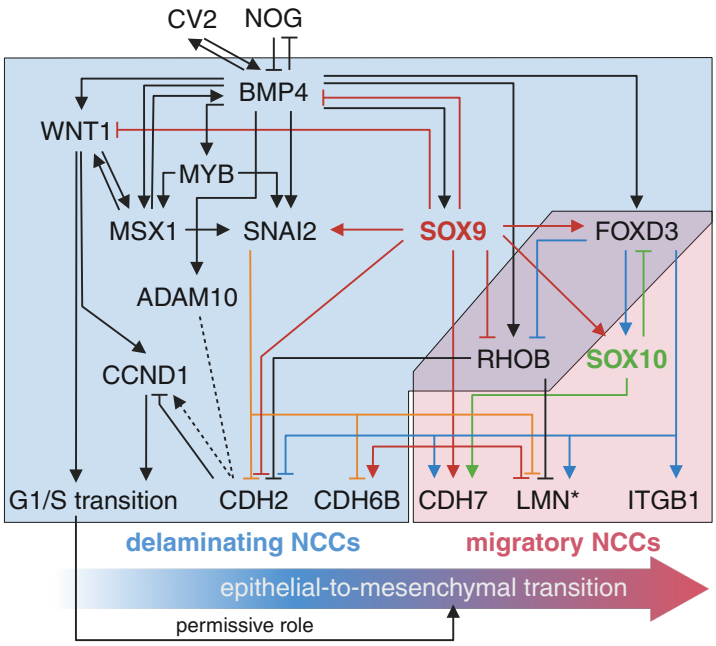


Figure 8 – Molecular cascade controlling trunk NCCs delamination and migration. Neurepithelial cells undergo epithelial-to-mesenchymal transition through activation of Wnt pathway. Neural crest cells degrade their basal lamina and stop expressing tight junction proteins, thus stop being epithelial cells. This detachment facilitates their proliferation. By remodeling their surrounding ECM, cells start to migrate and become mesenchymal cells. Adapted from Theveneau and Mayor (2012).

Based on contact inhibition and mutual cell attraction mechanisms, NCCs migrate collectively along streams (Figure 9A, Mayor and Theveneau (2013)). Indeed, negative signals repel NCCs from stream borders while positive signals attract them along a migratory corridor and to each target region. According to their local environment in the streams, NCCs give rise to a plethora of **derivatives**, including SCs and fibroblasts (Figure 9B, Soldatov *et al.* (2019)). Particularly, NCCs give rise to boundary cap cells at E9 (Section 1.3.2) and SC precursors (Section 1.3.3), and both can differentiate into mature SCs (Section 1.4).

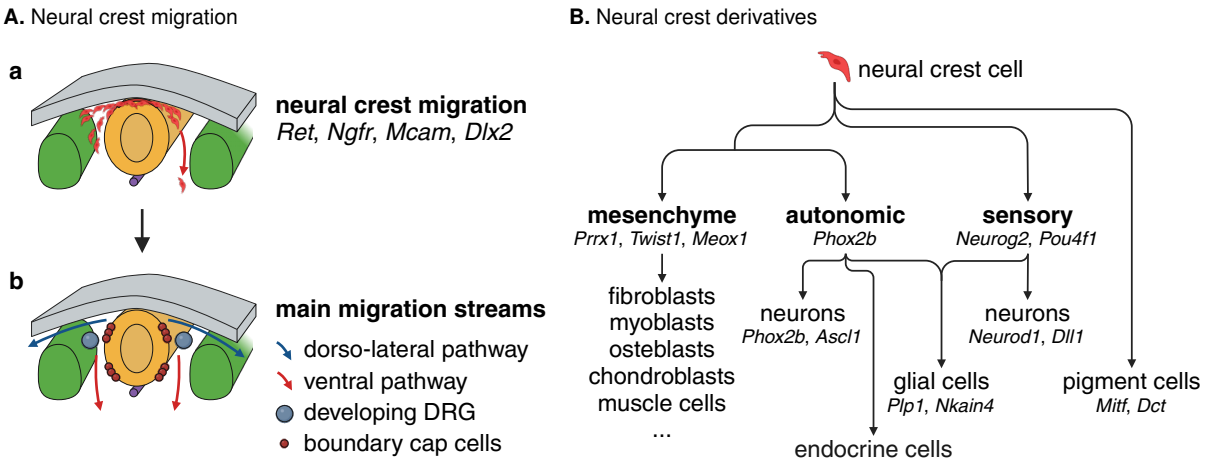


Figure 9 – **Neural crest migration and trunk derivatives**. **A**. Main migration stream of the neural crest cells are ventral (**a**), then dorso-lateral (**b**). A subset of NCCs stops close to the neural tube to form the future dorsal root ganglion (DRG) and boundary cap cells. **B**. NCCs lead to a variety of derivatives. All markers come from Soldatov *et al.* (2019). Figures adapted from Simões-Costa and Bronner (2015), Mayor and Theveneau (2013), Furlan and Adameyko (2018), and Erickson *et al.* (2023).

1.3.2 Boundary cap cells form a transient neural crest-derived population

Boundary cap cells (BCCs) are organized as caps that prevent motor neuron cell bodies from escaping the neural tube, while their axons navigate toward their targets. From E12, they start to migrate along the peripheral nerves, and for some of them, up to nerve endings in the skin (Figure 10A-B). No BCCs remain in the adult organism (Gresset *et al.*, 2015).

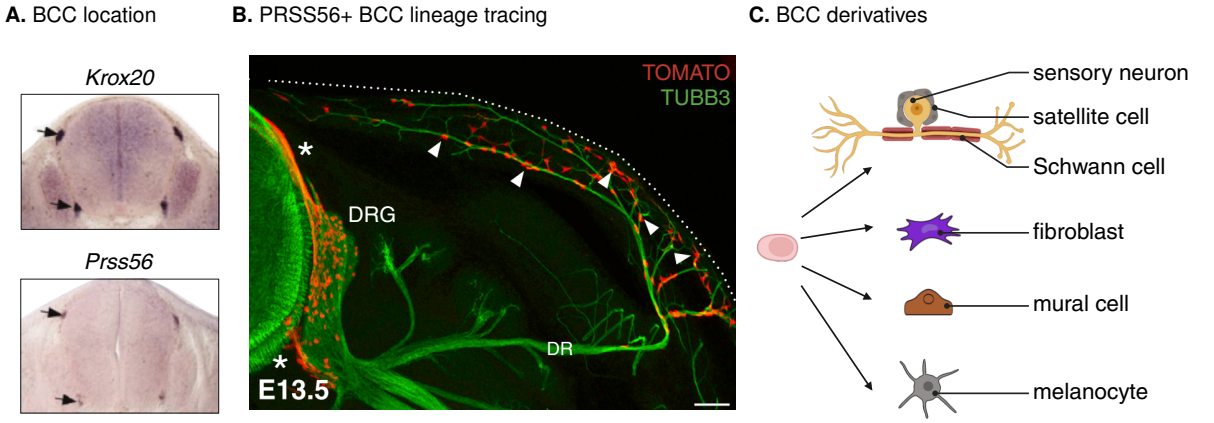


Figure 10 – **Boundary cap cells is a vestige of NCCs giving rise to Schwann cells**. **A**. BCCs are stained for *KROX20* or *PRSS56* on transversal section of a mouse embryo at E12.5. Adapted from Coulpier *et al.* (2009). **B**. Transversal section of *Prss56*$^{Cre,+}$;*Rosa26*$^{Tom,+}$ mouse embryos, where TOMATO labels all cells expressing or having expressed *Prss56* and β3-tubulin (TUBB3) mark axonal cytoskeleton microtubules, shows that BCCs migrate from the PNS/CNS interface, markers by *, along the peripheral nerves up to nerve endings in the skin. Dotted lines delineate the embryo. DRG: dorsal root ganglion. Adapted from Gresset *et al.* (2015). **C**. Fate mapping of *Prss56*-expressing BCCs identified their derivatives (Radomska & Topilko, 2017).

BCCs markers include KROX20 and PRSS56 and correspond to two partially overlapping BCCs subsets giving rise to distinct derivatives. Unlike KROX20, which is expressed in other many cell types (Liao *et al.*, 2019), PRSS56 is specific to BCCs, as it is transiently expressed, and never reactivated *de novo* (Coulpier *et al.*, 2009). The role of PRSS56 is unknown and its inactivation does not alter BCCs functions and migration (Gresset *et al.*, 2015). Due to its BCC-restricted expression, PRSS56 is a good candidate to study their fate. Fate mapping of PRSS56-positive BCCs, through the Cre recombinase system (Box 2), revealed their multipotent nature (Figure 10C).

Box 2: Cre-Lox recombination targets lineage-specific DNA modifications

The Cre-Lox recombination is a technology used to target a specific cell lineage for DNA modifications, such as deletion. LoxP sequences are palindromic DNA sequences recognized by the CRE recombinase, which cuts and recombines two loxP sites (Figure 11).

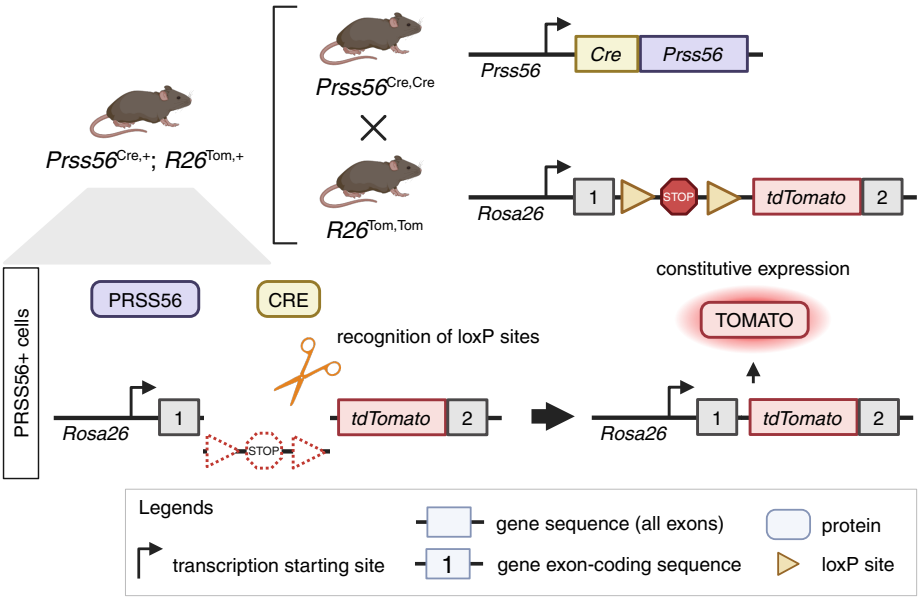


Figure 11 – Construction of *Prss56*$^{Cre,+}$; Rosa26$^{Tom,+}$ mouse strain. *Prss56*$^{Cre,+}$; Rosa26$^{Tom,+}$ mouse strain is obtained by crossing *Prss56*^{Cre,Cre} mice, where *Prss56* promoter is followed by the Cre recombinase coding sequence, with *Rosa26*^{Tom,Tom} mice, where *Rosa26* locus, ubiquitously expressed, contains *Tomato* gene preceded by a stop codon coding sequence flanked by loxP sequences. Cre recombinase particularly recognizes loxP DNA sites, cut them and repair DNA by homologous recombination. CRE protein is only expressed in PRSS56+ cells. Thereby, all cells expressing or having expressed *Prss56* have lost the loxP-Stop-loxP sequence, leading to the constitutive and definitive expression of *Tomato*. This construction makes all PRSS56+ cells and their derivatives TOMATO-positive, even if *Prss56* is no more expressed.

1.3.3 Schwann cells precursors are multipotent glial cells

Schwann cell precursors (SCPs) is a population of multipotent progenitors, emerging in the embryo at W4 or E12. They directly originate from NCCs, or from the intermediate stage formed by BCCs (Solovieva & Bronner, 2021). Since transition from NCCs to SCPs is slow and continuous, Furlan and Adameyko (2018) suggested to name SCPs from the time they start to be associated with peripheral axons.

The main function of SCPs is to support the development of the PNS. From E12, neurons stop to proliferate. No more neurons are found in adult organism compared to embryo at E12. Their cell bodies are at their definitive location in the organism. The PNS formation begins when axons start to migrate up to their targets. Despite SCPs are not required for axons migration, they are essential for guiding them up to their targets and neurons that fail to reach their targets will die (Jessen *et al.*, 2015). Thereby, axons sustain SCPs survival, notably based on the secretion of β-neuregulin 1 (NRG1), received by SCPs through the ERBB2-ERBB3 proteins family. NRG1 signaling is present in the whole trajectory of SCs.

Additionally to sustain their survival, NRG1 signaling promotes the proliferation of SCPs. The highly proliferative nature of SCPs make them able to follow axon growth and cover them from the cell body up to their targets. Once targets are reached, SCPs persist to proliferate, leading them to surround fewer and fewer axons.

Another role of SCPs in PNS development relies in the vascularization. Indeed, axonal growth requires nutrients, which is provided by blood vessels. Thereby, SCPs secrete VEGF and CXCL12 to attract vessels and promote the neurovascular alignment (Furlan & Adameyko, 2018).

Unlike their derivatives, SCPs specifically express cadherin 19 (*Cdh19*), AP2A, encoded by *Tfap2a*, and integrin α4 (*Itga4*) (Figure 12A). However, only CDH19 is specific to SCPs because other are also expressed in NCCs. Cadherin 19 acts as a adhesion molecule to promote cell-cell interaction and adhesion.

Despite their name suggest it, SCPs do not only give rise to SCs, but to many derivatives (Solovieva & Bronner, 2021). By performing scRNA-Seq at various embryonic and post-natal stages, Kastriti *et al.* (2022) demonstrated that SCPs form a hub of progenitor cells and highlighted specific markers along their differentiation (Figure 12B). Thus, *Sox10* expression persists is SCPs giving rise to glial cells and melanocytes. In melanocytes, *Sox9* is re-activated and co-expressed with *Sox10* (Harris *et al.*, 2010).

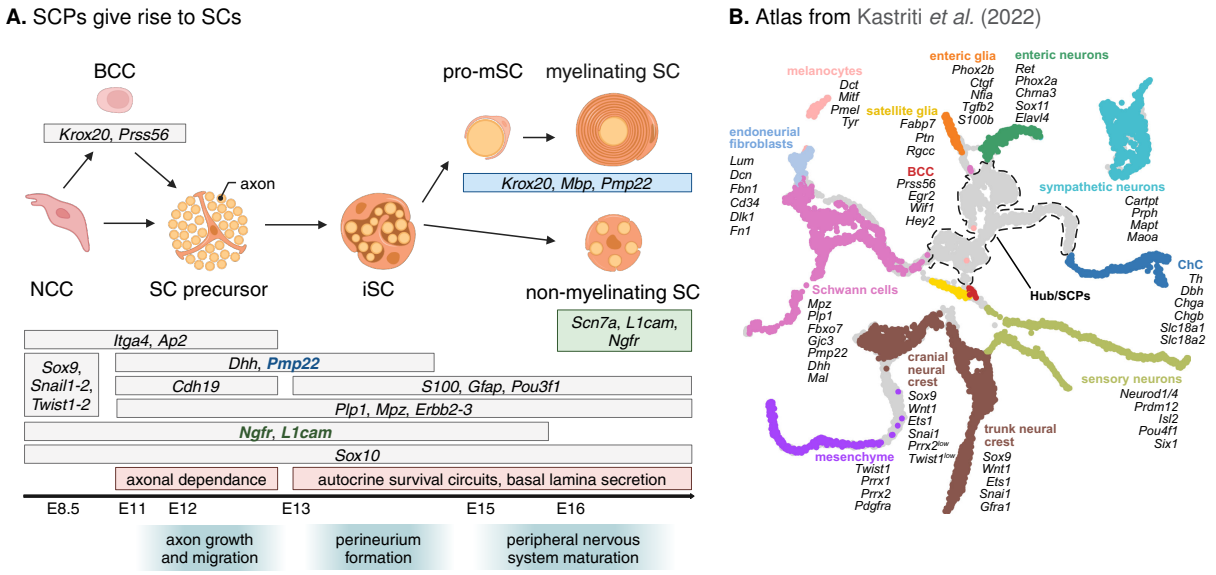


Figure 12 – Schwann cells precursors are multipotent glial cells. **A**. All SCs come from neural crest cells (NCC). Specific markers for each stage are annotated in gray. Blue markers are specific to myelinating SCs while green markers are specific to non-myelinating SCs. Mouse embryonic developmental days and marker genes are indicated, and come from Ackerman and Monk (2016), Jessen *et al.* (2015), and Jessen and Mirsky (2019a). Bottom boxes indicate PNS development stages. **B**. SCPs from a hub of progenitor cells (gray), giving rise to various cell types (colors) (Kastriti *et al.*, 2022).

1.3.4 Immature SCs initiate the final structure of peripheral nerves

From E13, a subset of SCPs converts to **immature Schwann cells** (iSCs) (Jessen & Mirsky, 2019a). Unlike SCPs, all SCs derivatives can survive without axonal contact. Furthermore, iSC change their expression profile to secrete a basal lamina, a feature maintained in mature SCs (Section 1.4.1.1). In relation with their changing activity, SCPs and iSCs differ by their expression profile: iSCs express *Gfap* and *S100b*, respectively encoding a cytoskeleton protein and responsible for calcium metabolism, and down-regulate some SCP markers (Figure 12A).

While the axonal-dependent survival of SCPs during development is a mechanism to maintain the contact between SCPs and axons, the ability of SCs to survive in the absence of axonal contact is essential for nerve repair in adult nerves. Indeed, both immature and mature SCs have long term survival ability due to autocrine stimulations, particularly through laminins, a component of the basal lamina, neurotrophin 3 (*Nt3*), insulin-like growth factors and *Pdgfb* (Jessen *et al.*, 2015).

The differentiation of SCPs into iSCs coincides with the **development of the PNS**. Indeed, iSCs participate in the recruitment of the remaining cellular components of the future peripheral nerve. Through DHH signaling, iSCs recruit fibroblasts forming the

perineurium. Secreted DHH interacts with PTCH receptor of perineurial cells precursors, which activates a second membrane protein, SMO, to promote their differentiation into cells delineating the perineurium (Sundaram *et al.*, 2019). Moreover, as SCPs, iSCs promote vascularization through VEGF (Monk *et al.*, 2015). Once nerves are protected by fibroblast layers and fed with oxygen and nutrients, they are functional.

A second mechanism occurring in iSCs is the **radial sorting**. Indeed, the maturation of iSCs relies on the diameter of axons they are ensheathing, and two subtypes of SCs emerge from iSCs (Figure 12A). Initially, iSCs enwrap several axons. By modulating SCs survival, proliferation and death, large-diameter axons restrict SCs to be specific to a single axon (Jessen & Mirsky, 2019a; Monk *et al.*, 2015). Such SC are called **pro-myelinating SC** (Figure 12A). They extend their membrane to form several spires around axons, and start to produce myelin. They finally differentiate to **myelinating SC**. On the opposite, SCs around small diameter axons remain **non-myelinating SCs**. The myelination of axons take several weeks, thus, iSCs are still present in the early postnatal period (Mirsky *et al.*, 2002).

1.4 The multiple facets of Schwann cells in the adult nerves

1.4.1 Insight into Schwann cells in the physiological state

SCs are the only cells in direct contact with axons (Figure 13). Their role is to provide a metabolic support to the axons. They maintain a healthy environment by recycling neurotransmitters released by the axons and other waste products, regulating the ion balance and supplying nutrients to axons (Kidd *et al.*, 2013).

Both **myelinating** (**mSCs**) and **non-myelinating** (**nmSCs**) SCs express membrane proteins called integrins, encoded by *Itg** genes (Figure 13C) and a basal lamina (Section 1.4.1.1). Myelinating SCs differ from nmSCs by their production of myelin (Section 1.4.1.2). Finally, nmSCs are characterized by the specific expression of SCN7A, LKB1 and TFAM (Figure 13C, Harty and Monk (2017)).

1.4.1.1 Basal lamina is a ECM layer protecting SCs

SCs are surrounded by a specific extra-cellular matrix they produce, called the **basal lamina** (Figure 13C). Close to SC membrane, the basal lamina *lucida* is laminin-rich (encoded by *Lmn** genes), whereas the basal lamina *densa* is collagen-rich (encoded by *Col** genes) (Court *et al.*, 2006). The basal lamina provides a structural support to SCs, participate to the myelination process, support saltatory signal transduction, filter

elements from the surrounding connective tissues, and protects axons from compression and extension forces (Court *et al.*, 2006; Kidd *et al.*, 2013). The composition of the basal lamina evolves during the maturation of SCs (Court *et al.*, 2006).

1.4.1.2 Myelin sheath is a particularity of mSCs

Large-diameter axons are wrapped in a spiral manner by a multilayered sheath of compacted plasma membranes called **myelin**, produced by mSC (Figure 13A, left). Initially, mSCs surround multiple times an axon in an uncompacted way. They further over-express myelin-related genes, *Mpz*, *Mbp* and *Pmp22* (Figure 13C). During a compaction period, MPZ proteins on each membrane layer form multimeric interactions, like the teeth of a zip, bringing the two membranes closer together (Figure 13B, "M"

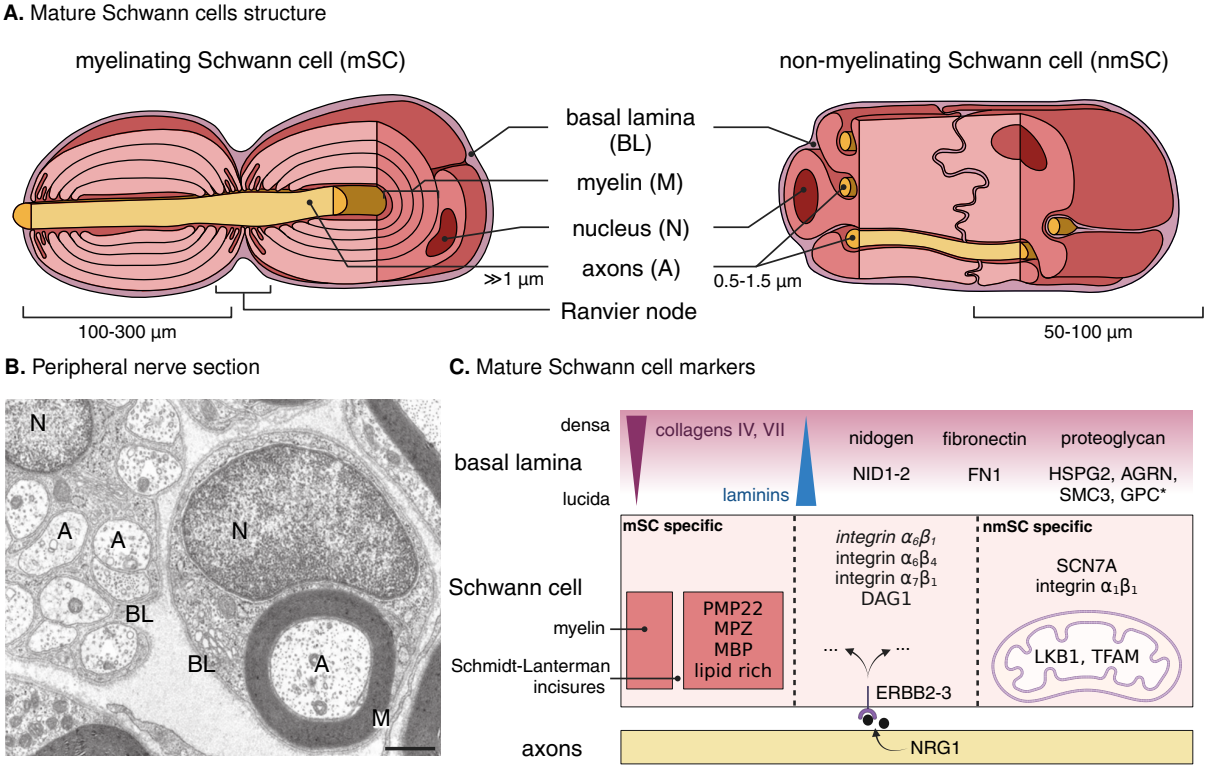


Figure 13 – **Structure of myelinating and non-myelinating Schwann cells**. **A**. There are two types of mature SCs, characterized by their width (Kidd *et al.*, 2013), the presence of myelin sheath, and the number and diameter of the axons they surround (Griffin & Thompson, 2008). Inspired from "Anatomy of Peripheral Nerve" image ID 4069. **B**. Transversal section of a nerve by electron microscopy showing a mSC and a nmSC, adapted from Kidd *et al.* (2013). Scale bar is approximately 1 µm. **C**. The basal lamina is common to both mSCs and nmSCs. Specific markers to mSCs are depicted on the left while nmSCs markers are on the right. Common markers are indicated in the middle. Italicized integrin α6β1 is expressed from the immature stage. N: SC nucleus, A: axon, BL: basal lamina, M: myelin sheath.

annotation). Although their cytoplasm and nucleus are pulled at the periphery, mSCs still communicate with axon through discontinuities in the myelin sheath, known as Schmitt-Lanterman incisures (Figure 13C, Salzer (2015)).

As lipid-rich, the myelin sheath acts as an **electrical insulator**, preventing the transmission of action potentials along the axonal membrane. Each mSC is separated from the others by a small space, called the Ranvier node, where axon is thinly protected (Figure 13A, left). As a result, the signal jumps between successive Ranvier nodes, which is known as saltatory conduction (Figure 3). Due to the enlargement of the global fiber diameter caused by the myelin sheath and saltatory conduction, signal transmission speed is increased as compared to non-myelinated axons. Fast transmission is crucial for urgent information, such as related to flight and requiring rapid movement. Moreover, the myelination prevents any axonal sprouting, i.e. longitudinal growth, thereby stabilizing connections (Griffin & Thompson, 2008).

1.4.1.3 Unmyelinated fibers are a particularity of the peripheral nervous system

The myelin sheath is not found around all axons, and there is approximately 4 times more unmyelinated than myelinating axons in cutaneous nerves (Griffin & Thompson, 2008). Unmyelinated fibers are a particularity of the PNS, since all axons are myelinating in the CNS. The absence of myelin confers **plasticity** to axons because they can sprout, particularly at nerve endings, and grow, notably after injury. Finally, since several axons are packed together by nmSC, physical space is saved (Figure 13B).

However, all nmSC have the possibility to myelinate or not, if they received the appropriate signals (Griffin & Thompson, 2008). Thus, the myelination process is reversible. Indeed, SCs myelination is due to a balance between myelin activating and repressing signals (Jessen & Mirsky, 2008). Signals promoting myelin formation include NRG1 released by axons, while signals repressing myelination includes down-regulation of cell adhesion molecules L1CAM and NCAM (Taveggia *et al.*, 2010; Glenn & Talbot, 2013; Kidd *et al.*, 2013).

Mature SCs are maintained by extrinsic signals, provided by direct contact with axons and the basal lamina, and intrinsic signals. Changes in these signals, such as nerve injury, result in mature SC switching to a repair phenotype, termed repair SCs (rSC).

1.4.2 Mature SCs switch between steady-state and repair phenotype

Given the involvement of the PNS in the functioning of the body, any impairments within this system can lead to severe consequences. Fortunately, unlike the CNS, the PNS possesses unprecedented regenerative abilities. In case of a nerve fiber injury, the PNS can regenerate and restore its function, thanks to the presence of SCs, which facilitate repair and regrowth. The repair process of peripheral nerve can be segmented in five steps (Figure 14):

1. **Acquisition of repair phenotype** - Nerve injury, for instance crushing or cutting, leads to alteration in SC-axon contact, and SCs detachment from axons. This process is particularly driven by the transcription factor (TF) c-Jun, encoded by *Jun* gene (Arthur-Farraj *et al.* (2012) and next section), and leads mature SCs to switch toward a repair phenotype (rSC). Particularly, mature SCs demyelinate, repress the myelination process and detach from axons. Both rSCs and resident macrophages express pro-inflammatory cytokines to recruit circulating macrophages.
2. **Wallerian degeneration** corresponds to the digestion of axonal debris and myelin sheath. Such myelinophagy has been described in both macrophages and rSCs (Jessen & Mirsky, 2019b).
3. **Growth stimulation** - Once the injured zone is cleaned from its debris, the second phase of reparation starts. rSCs secrete growth factors, and macrophages switch from pro-inflammatory to anti-inflammatory phenotype (Wu *et al.*, 2024).
4. **Axon regrowth** - While macrophages secrete anti-inflammatory cytokines, decreasing immune cells infiltration, rSCs organized as bands of Büngner. The latter corresponds to longitudinally aligned rSCs, forming regeneration tracks and guiding axonal regrowth (Figure 14). Such a structure is controlled by the TF c-Jun, since its inactivation impair their formation (Arthur-Farraj *et al.*, 2012).
5. **Differentiation** - Once the proximal part of the injured axon reaches the distal part, rSCs re-differentiate to the mature phenotype. They down-regulate repair-related genes, and, based on the axonal diameter they surround, become either mSCs or nmSCs. Finally, macrophages leave the injury site (Figure 14).

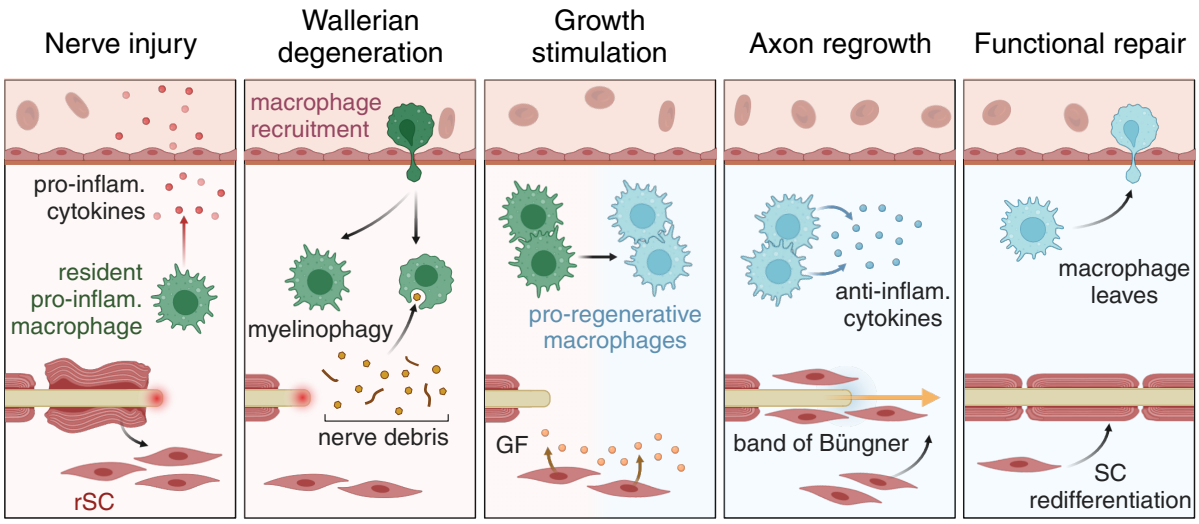


Figure 14 – Schwann cells functions from nerve injury to recovery. This figure focuses on the role of SCs during nerve repair. The implication of vessels and fibroblasts has already been succinctly underlined in previous sections. After nerve injury, SCs demyelinate, detach from axon and become repair SCs (rSC). Together with rSCs, resident pro-inflammatory macrophages secrete cytokines to recruit circulating macrophages. Both rSCs and macrophages participate to the removal of nerve debris, particularly through myelinophagy. After injury site cleaning, rSCs secrete growth factors, stimulating macrophage switching to a pro-regenerative phenotype and axonal survival. Then, rSCs organize into Büngner bands to guide axonal growth. After axon reaches the distal part of the injury site, rSCs differentiate back into mature SCs. Adapted from "Inflammation and Regeneration of Peripheral Nerves", by BioRender.com (2023). Retrieved from `https://app.biorender.com/biorender-templates`.

Sometimes, the nerve do not recover after injury. Firstly, misrouting of axons leads to re-innervation errors. Secondly, excessive ECM deposit, known as fibrosis, can hinder nerve recovery. The fibrosis leads the affected area to become stiff, thick and axon can not growth. In chronic denervation, the distal part progressively loose its ability to repair, due to a loss of the repair phenotype in SCs (Jessen & Mirsky, 2019b).

1.4.2.1 The central role of c-Jun in repair Schwann cells

The identity of rSCs is characterized by multiple markers (Figure 15). The over-activation of Ras pathways particularly leads to the inhibition of myelination and activation of myelinophagy. Through RAF-MEK-ERK pathway, known as MAP kinases (MAPK) pathway, cells participate to macrophages recruitment. Finally, the central c-Jun is activated by both PI3K-AKT-mTOR and RAC1-MKK7-JNK pathways. The transcription factor (TFs) c-Jun activates the expression of other TFs and growth factors, permitting axonal growth and survival. Epigenetic changes, including the loss of trimethylation on lysine 27 on histone 3, known as H3K27me3 mark, leads to the tran-

scription of genes implied in axonal growth. Loss of H3K27me3 inhibits the expression of gene encoding cyclin-dependent kinase inhibitor 2A (CDKN2A), leading to cell cycle progression (Box 3). Proliferation is also triggered by cyclin D1 (CCND1), a protein that could be activated downstream Ras pathways (Box 3).

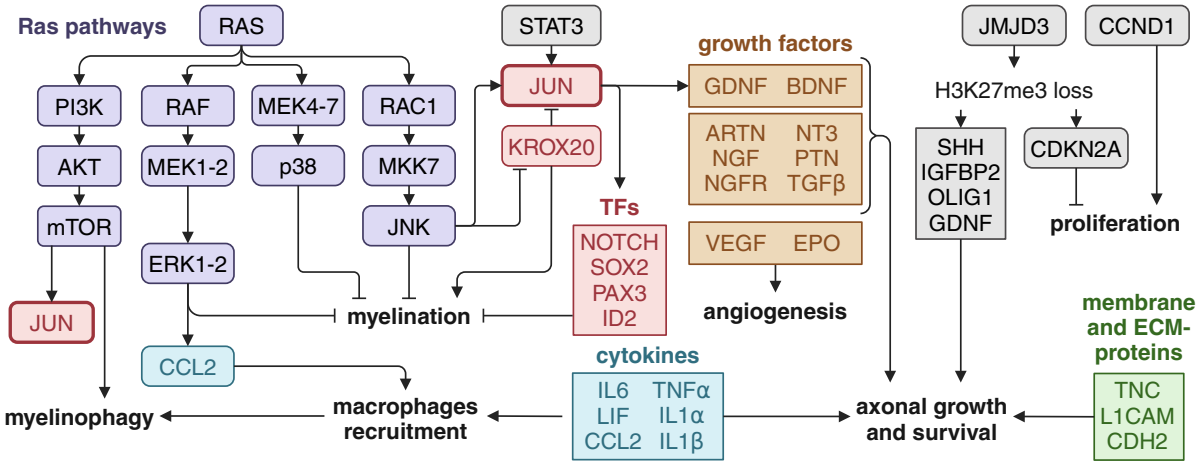


Figure 15 – Signaling pathways involved in the identity of repair Schwann cells. In rSCs, several pathways downstream RAS are activated to down-regulate the myelination, promote the myelinophagy, recruit macrophages and activate c-Jun. In long term maintenance of repair phenotype, STAT3 maintains c-Jun expression (Benito *et al.*, 2017). c-Jun activates TFs responsible for demyelination, and growth factors promoting axonal survival and growth. Beside TFs and growth factors, rSCs produce various cytokines to recruit macrophages, and extra-cellular or membrane proteins interacting with axons. Repair phenotype also relies on epigenetic modification. The loss of H3K27me3 mark induced by JMJD3 participates to the activation of *Cdkn2a* to limit rSC over-proliferation (Gomez-Sanchez *et al.*, 2013). Note that all genes are not simultaneously expressed but sequentially. This diagram is mainly based on the review from Jessen and Arthur-Farraj (2019). Please note there are multiple interconnections between the represented modules, and the signaling pathways correspond more to a signaling network.

1.4.2.2 Specificity of rSCs compared to iSCs

Unlike mature SCs, rSCs have a reduced basal lamina and a similar elongated morphology as iSCs. That is why, rSCs can be mistaken for iSCs. However, additional to dedifferentiated features, rSCs produce growth and neurotrophic factors to promote survival and growth of axons (Figure 15). Indeed, Jessen and Mirsky (2016) defined the transition of mature SCs into a repair phenotype as a **transdifferentiation**, combining **dedifferentation** and **activation**. They highlighted key differences between iSCs and rSCs. Firstly, rSCs are characterized by unique expression profile controlled by c-Jun (Figure 15). Secondly, rSCs are implicated in a local innate immune response, and subsequently express cytokines and attracting and activating macrophages. Cytokines can be interleukins (*Il**), chemokines (*Ccl**, *Cxcl** or , *Cx*cl**) and transforming growth

factors (*Tgf**). Thirdly, rSCs participate to the clearance of the myelin, either through macrophage recruitment or by myelinophagy. Finally, rSCs drastically remodel their cytoskeleton to adopt an elongated shape, particularly through actin polymerization. It allows the formation of bands of Büngner, which permit axon growth guidance.

Box 3: CDKN2A and CCND1 controls G1 to S cell cycle phase progression

The activation of MAPK pathway leads phosho-ERK to stimulate the transcription of *Ccnd1*, whose translation is then promoted by the PI3K-AKT-mTOR pathway (Figure 16, left). CCND1 dimerizes with the cyclin-dependent kinases, CDK4 or CDK6 proteins. By phosphorylating RB, E2F is free and cell cycle progresses.

Cdkn2a is a 3-exons gene ubiquitously expressed. Based on alternative splicing, two main isoforms are described: a member of INK4 family, called $p16^{INK4a}$, and an alternative reading frame (ARF) protein, called $p14^{ARF}$ in human and $p19^{ARF}$ in mice (Figure 16, right). INK4 proteins form a family of inhibitors of cyclin-dependent kinases (CDK4 and CDK6). Thus, expression of $p16^{INK4a}$ leads to cell cycle blockage in G1 phase. $p19^{ARF}$ also act as cell cycle inhibitor, via the inhibition of MDM2 protein, which inhibits P53. Protein 53 (P53), encoded by the tumor suppressor gene TRP53, allows DNA damage recognition. Depending on the reparability of the DNA, P53 alternatively promotes DNA repair or apoptosis.

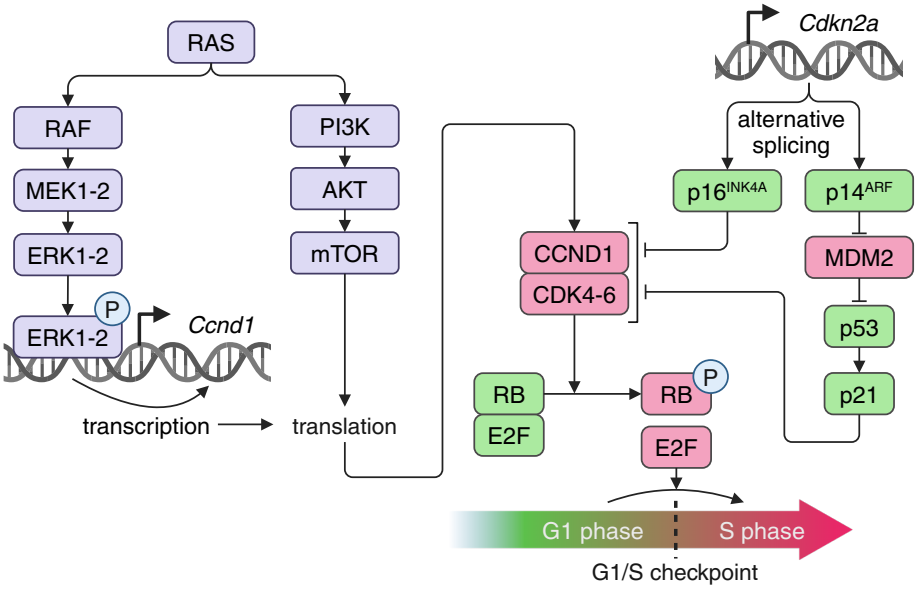


Figure 16 – Implication of CDKN2A and CCND1 in cell cycle. Ras pathways are depicted in purple. Proteins in favor of the G1-S phase progression are in red while proteins leading to cell cycle blockage in G1 phase are in green. P: phosphate. Inspired from Schwaederlé *et al.* (2014) for CCND1 part and Vandenberk *et al.* (2011) for CDKN2A part.

1.5 Overview of the chapter

SCs are highly plastic cells of the peripheral nervous system (PNS) and permit its regeneration after injury (Figure 17). Any deregulation of the pathways balancing SCs plasticity may contribute to pathologies. In Neurofibromatosis Type 1 (NF1), a subset of SCs carries bi-allelic inactivation of *Nf1* gene, encoding a negative regulator of RAS activity. Resulting Ras pathways over-activation induces tumor development.

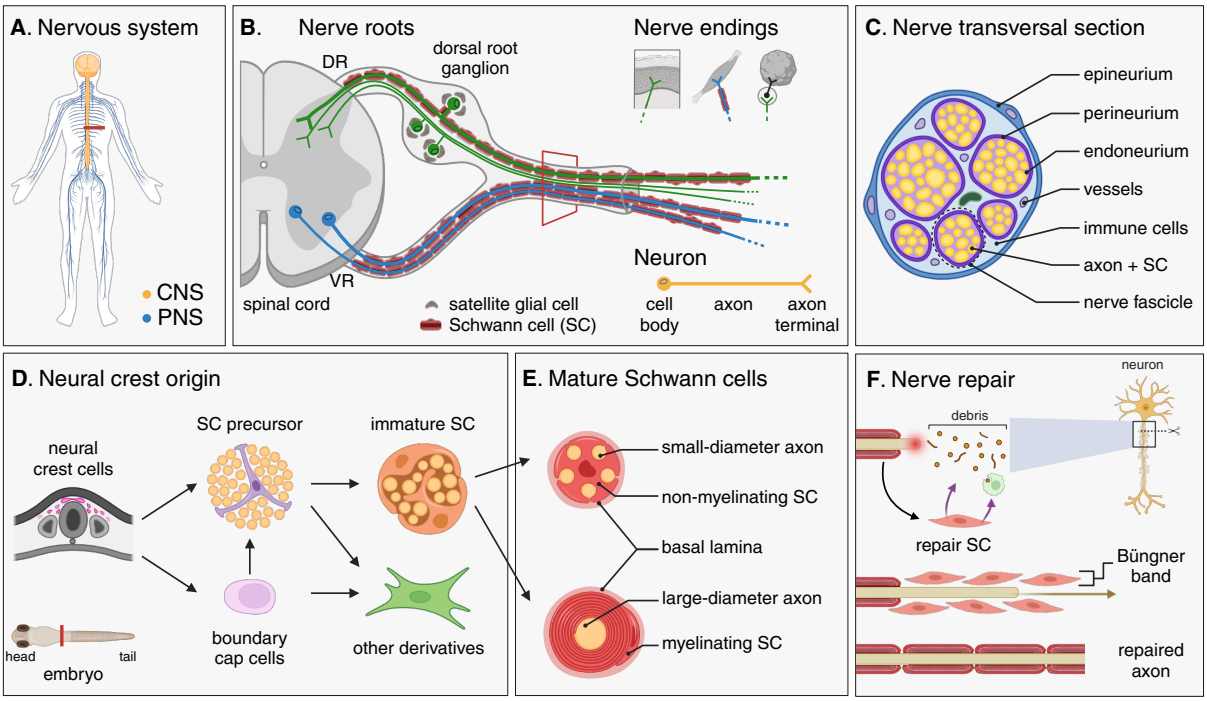


Figure 17 – Schwann cells are glial cells of the peripheral nervous system. **A**. The nervous system is divided between the CNS and the PNS (Section 1.1). **B**. Transversal section from red line in A. The PNS establishes a connection between the CNS, represented by the spinal cord, and various parts of the body, including the skin, muscles and glands. PNS encompasses ganglions and nerves (Section 1.2). The functional unit of the nervous system is neurons, depicted as lines. Their cell bodies are located in either the CNS, for motor neurons (blue) ; or ganglions, for sensory neurons (green). Two types of glial cells protect neurons in the PNS: satellite glial cells surround cell bodies, while Schwann cells (SCs) envelop axons. **C**. Transversal section from red section in B. **D**. SCs originate from the neural crest (Section 1.3.1). Neural crest cells give rise to Schwann cell precursors (Section 1.3.3) and boundary cap cells (Section 1.3.2). Among their multiple derivatives, Schwann cell precursors generate immature SCs, that start to have mature SC characteristics. **E**. Mature SCs comprise non-myelinating or myelinating subtypes, according to the axon diameter they wrap (Section 1.4.1). **F**. PNS as a unprecedented regenerative capacity, due to SCs transdifferentiation toward a repair phenotype after nerve injury. After debris removal, repair SCs organize in Büngner band that enable axon regrowth. Finally, they re-differentiate to a mature state, based on the axon diameter (Section 1.4.2).

2 NF1, a monogenic but complex syndrome

Neurofibromatosis Type 1 (NF1) is also called von Recklinghausen disease, named after the pathologist who characterized the disease in 1882. NF1 is an autosomal dominant monogenic syndrome, characterized by alterations in the neurofibromin 1 gene (Section 2.1). The disease affects equally all populations, genders and ethnicities[1], which a global incidence of 1/3,500 births (Evans *et al.*, 2010). NF1 patients suffer from a panel of symptoms, among which peripheral nerve sheath tumors (PNSTs). Particularly, **plexiform neurofibromas** are benign tumors, but they are difficult to treat and sometimes cannot be removed surgically (Section 2.2). They can transform into pre-malignant tumors called **dysplastic neurofibromas** (Section 2.3) and then into **malignant PNSTs** (MPNSTs), which are aggressive tumors (Section 2.4).

2.1 Generalities about the disease

2.1.1 Symptoms, treatment and diagnosis

NF1 patients are characterized by a **range of symptoms**, with each patient having a unique profile (Figure 18). Symptoms concern different organs and tissues, do not develop at the same time and in the same quantities. For instance, plexiform neurofibromas (pNFs) develop in the peripheral nervous system, in 30 to 50% of NF1 patients, either prenatally or during early childhood (Ferner *et al.*, 2006). On the opposite cutaneous neurofibromas (cNFs) develop during puberty, in almost all patients, and their number can reach hundreds. Finally, symptoms differ by their gravity. Indeed, while cNFs remain benign tumors, pNFs can transform into MPNSTs in young adults, where they are responsible for a shortened life expectancy (Evans *et al.*, 2011).

No **treatment** can reverse NF1, but its symptoms can be addressed individually. More than 50 clinical trials are currently active[1], notably to address neurofibromas. However, while optic gliomas may regress spontaneously (Parsa *et al.*, 2001), MPNSTs are resistant to radiotherapy and chemotherapy. For such symptoms, improvement of early detection is crucial for a better care.

NF1 diagnosis criteria were established in 1988 and revised by Legius *et al.* (2021). They include hereditary features and the number of NF1-specific symptoms (Figure 18, blue marks). When at least 2 of 7 criteria are met, decision is made to sequence *NF1* gene. Alterations in one allele are sufficient to diagnose NF1 disease.

[1]Based on `https://www.ctf.org/`, accessed in June 2024.

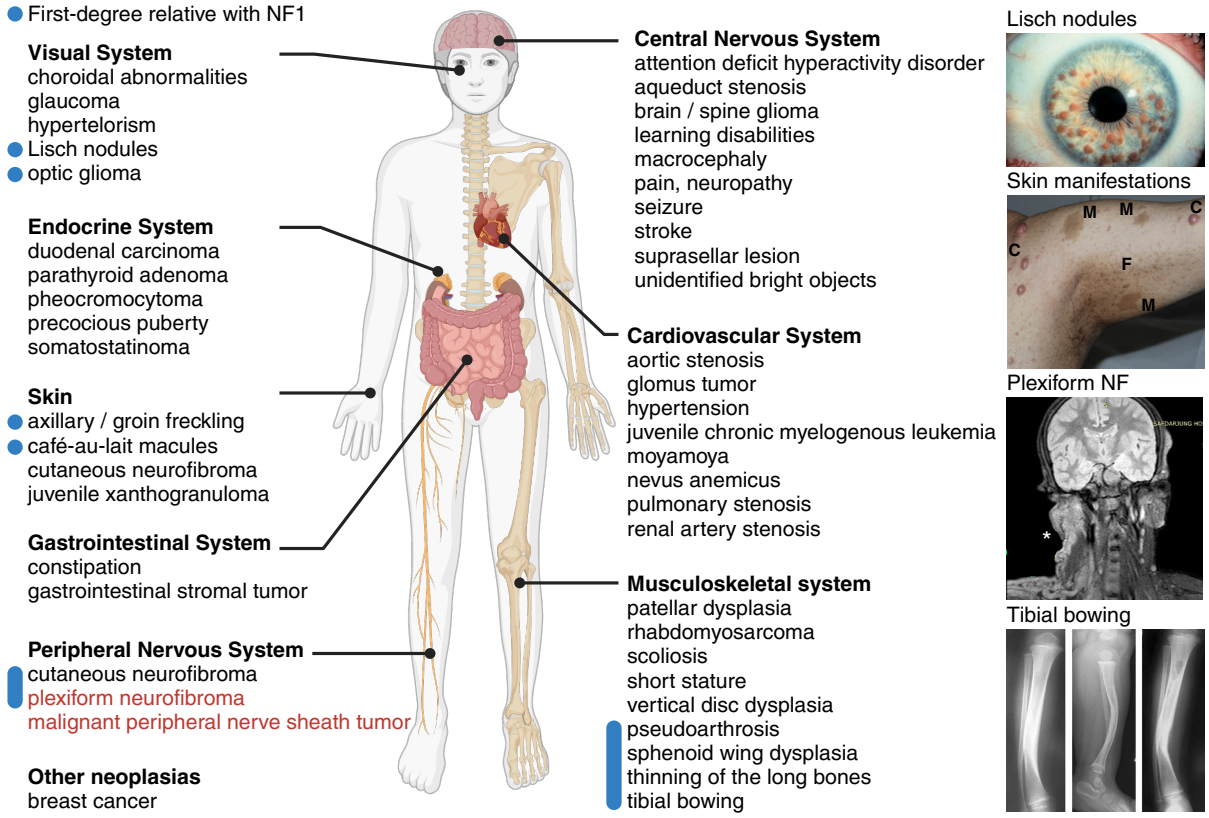


Figure 18 – NF1 disease covers a large range of symptoms. Each NF1 patient depicts a unique combination of possible symptoms. Symptoms written in red are detailed in next sections. Symptoms marked in blue constitute diagnostic criteria. The right part shows representative pictures of 4 diagnosis criteria: Lisch nodules (from Dahl (2021)), skin manifestations, including cutaneous neurofibromas (C), café-au-lait macules (M) et axillary freckles (F) (from Chernoff and Schaffer (2016)), plexiform neurofibroma of the neck (from Grover *et al.* (2021)), marked by the white star, and tibial bowing (from Elefteriou *et al.* (2009)). Inspired from Báez-Flores *et al.* (2023).

2.1.2 Molecular implications behind NF1 disease

In this section are succinctly described the structure and functions of NF1 protein, in physiological condition, and the genetic alterations identified in NF1 patients.

2.1.2.1 NF1 gene and protein structure

NF1 in *Homo sapiens*, or *Nf1* in *Mus musculus*, is located on chromosomes 17 and 11, respectively. The **gene** spans over 350,000 nucleotides and contains 61 exons, including 57 constitutive and 4 subjected to alternative splicing (gray on Figure 19).

The most abundant form of NF1 protein contains the 57 constitutive exons, leading to a **2818-amino-acids protein** in *Homo sapiens*. Although broadly expressed, the highest levels of the protein are found in the nervous system, including non-myelinating

Schwann cells (Bergoug *et al.*, 2020). NF1 is a **multi-domain protein**, with currently 11 domains described, based on their amino acid composition, sequence homology with other proteins or binding affinity. Despite NF1 displays a plethora of interacting partners (Bergen *et al.*, 2020), the best characterized interaction concerns RAS proteins, through the GAP-related domain (GRD) (Figure 19, in red).

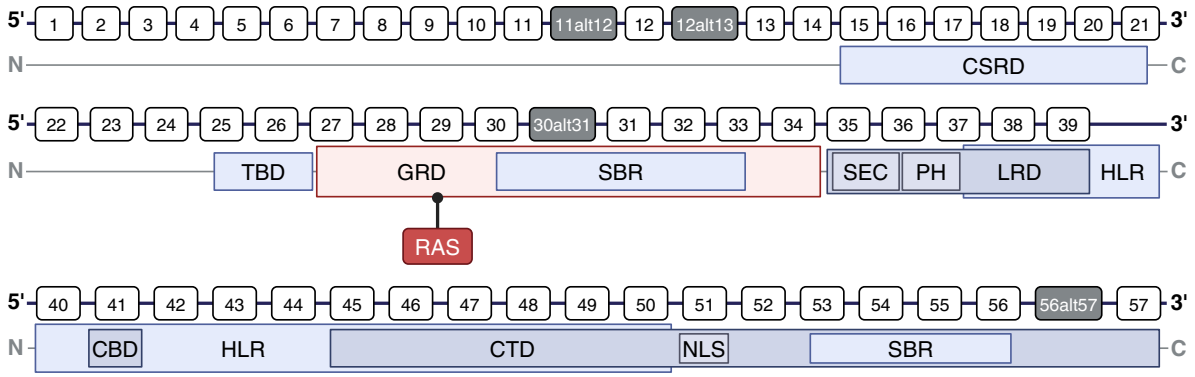


Figure 19 – ***NF1* gene and protein structure**. The upper line corresponds to the gene structure and the lower line to protein domains. Exons are numbered according to the *Homo sapiens* nomenclature of Anastasaki *et al.* (2017). Exons in gray are prone to alternative splicing. Among the interacting partners of NF1, RAS proteins bind to the GRD. CSRD: cysteine-serine-rich domain, TBD: tubulin-binding domain, GRD: GAP-related domain, SBR: syndecan-binding region, SEC: Sec14 homologous domain, PH: pleckstrin homologous domain, LRD: leucine-rich domain, HLR: HEAT-like repeat region, CBD: caveolin-binding domains, CTD: C-terminal domain, NLS: nuclear localization signal.

2.1.2.2 NF1, a GTPase activating protein

Through its GRD domain, NF1 acts as a GTPase activating protein (GAP) of RAS proteins family. RAS constitutes a family of proto-oncogene proteins, as it gathers the most frequently mutated genes in human cancers (Moore *et al.*, 2020).

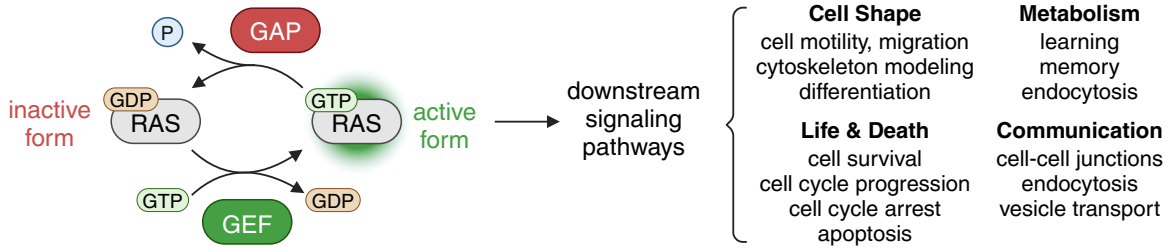


Figure 20 – **RAS molecular switch and signaling**. RAS acts as a small GTPase and functions as a binary switch with inactive and active states. In its active form, RAS is implicated in many signaling pathways. GAP: GTPase-activating protein (eg. NF1), GEF: guanine nucleotide exchange factor.

RAS proteins are small GTPases, meaning they are enzymes catalyzing the hydrolysis of GTP into GDP. They function as a binary switch, being inactive when bound to GDP or active when bound to GTP (Figure 20, left). The molecular switch is facilitated by guanine nucleotide exchange factors (GEFs) and GTPase-activating proteins (GAPs), including NF1. Since, **NF1 inactivates the activity of RAS proto-oncogene proteins**, *NF1* is considered a tumor suppressor gene.

In its active form, RAS triggers the phosphorylation of proteins, leading to their activation. Downstream signaling pathways form a complex network of positive and negative feedback loops and interactions between signaling cascades, leading to a plethora of consequences (Figure 20, right). For instance, the RAS-MAPK pathway, delineated by a red shape on Figure 21, leads to the phosphorylation of ERK, which is implicated in the cell cycle progression and the activation of transcription factors.

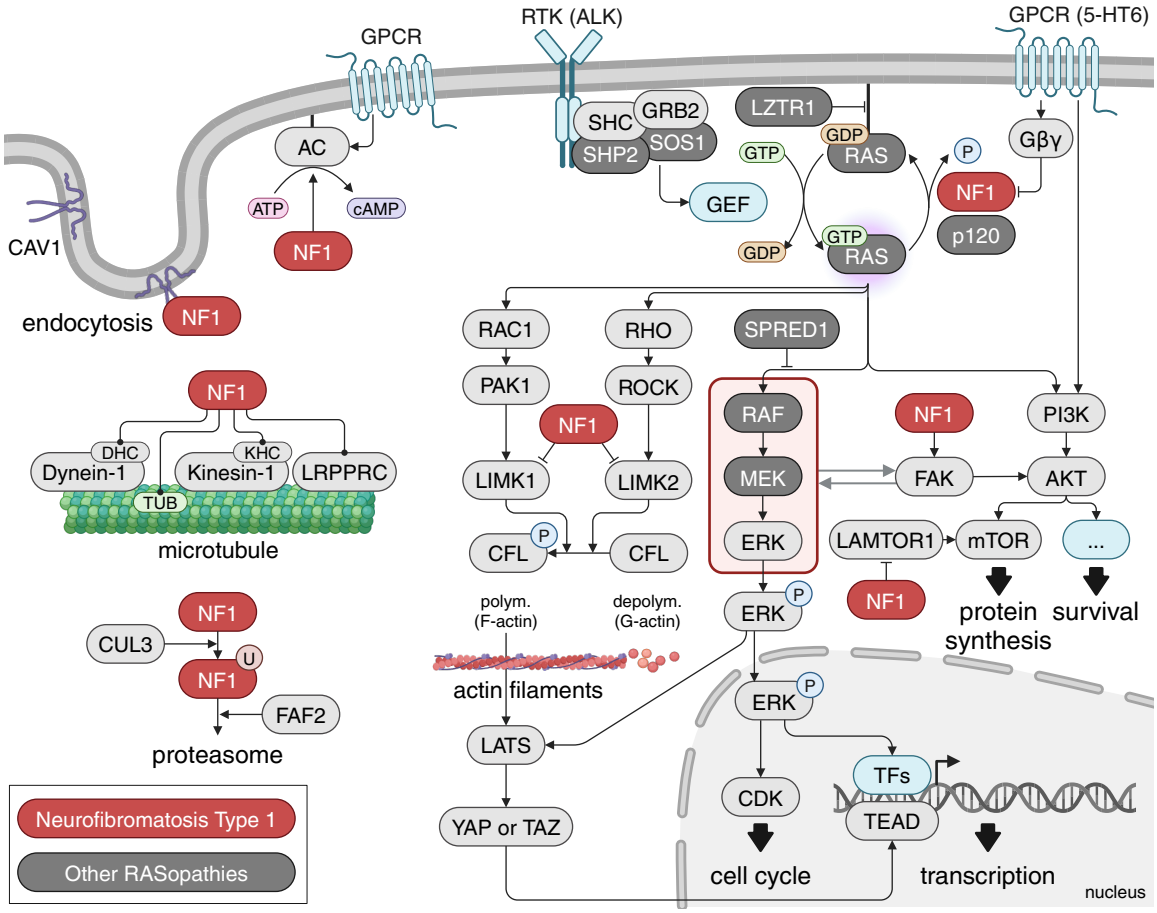


Figure 21 – Neurofibromin signaling pathways and relation with RASopathies. The focus is on pathways related to neurofibromin (in red), from Bergoug *et al.* (2020), Mo *et al.* (2022), Rauen (2022), and Báez-Flores *et al.* (2023). The RAS-MAPK pathway is highlighted by a red shape. Proteins affected in RASopathies are shown as dark gray elements. Blue elements represent protein families. RTK: receptor with tyrosine kinase activity.

Through its many interacting partners, NF1 is implied in **numerous other pathways** (Figure 21, Bergoug *et al.* (2020)). Close to the membrane, NF1 participates to the endocytosis, while, associated to the cytoskeleton, it regulates the vesicules transport within cells. Furthermore, NF1 contributes to the positive regulation of cycling AMP (cAMP) levels. The latter then activates protein kinase A (PKA), which promotes growth arrest and cell differentiation. In NF1-deficient cells, low levels of cAMP can be responsible for tumor formation (Bergen *et al.*, 2020). Increasing cAMP levels is a therapeutic strategy to address cutaneous neurofibromas (Mazuelas *et al.*, 2024).

The activity of NF1 is **regulated** by post-translational modifications (Figure 21, left). For instance, the CUL3 protein catalyzes NF1 ubiquitination, followed by FAF2 which promotes its proteolysis (Bergoug *et al.*, 2020; Báez-Flores *et al.*, 2023). As they contribute to decrease NF1 activity, and NF1 already suffering from loss of functions, CUL3 has been proposed as therapeutic target (Bergoug *et al.*, 2020).

2.1.2.3 NF1 patients genetic

Due to the involvement of *NF1* in the RAS-MAPK pathway, NF1 disease is categorized as a **RASopathy** (Rauen, 2022). RASopathies denote a group of clinically overlapping symptoms that share a common molecular origin and involve the RAS-MAPK pathway. Proteins whose genes are affected by germline mutations in RASopathies are shown in dark gray on Figure 21.

While in other RASopathies, mutations can consist in gain of functions, NF1 is characterized by **loss of functions of NF1 protein**. More than 90% of patients are characterized by **intragenic *NF1* mutations**, which can be either small nucleotide variations (insertion, deletion, mutation) or exon copy number variations. There is no mutation hotspot in the gene, meaning that the whole gene is susceptible to mutations leading to NF1-associated symptoms (Napolitano *et al.*, 2022). The remaining 10% of patients suffer from ***NF1* microdeletion syndrome**, which corresponds to the deletion of the whole *NF1* chromosomal region (Imbard *et al.*, 2015). Such a copy deletion may also involve the *SUZ12* gene, a neighbor of *NF1* (Section 2.4.3.2). The correlations between the genotype and the clinical manifestations of NF1 patients concern few *NF1* variants (Imbard *et al.*, 2015; Koczkowska *et al.*, 2020; Napolitano *et al.*, 2022; Pacot *et al.*, 2024). It implies that symptoms cannot be predicted from the *NF1* genetic profile of patients.

All NF1 patients are characterized by a **germline** heterozygous inactivation of the *NF1* gene. *NF1* haploinsufficiency is transmitted in half of the cases in a hereditary

way or *de novo* due to a mutation occurring during gametogenesis. NF1 disease is a tumor predisposition syndrome because any **somatic** alteration of the second allele of *NF1* can lead to additional symptoms (Spyk *et al.*, 2011). Among them, **neurofibromas** are the most common benign peripheral nerve sheath tumors.

2.1.3 Diversity of benign neurofibromas

Neurofibromas (NFs) develop from nerve sheath tumors, so they can arise throughout the entire body. They are classically divided into two subtypes based on their location and clinical features (Figure 22, Fisher *et al.* (2022)).

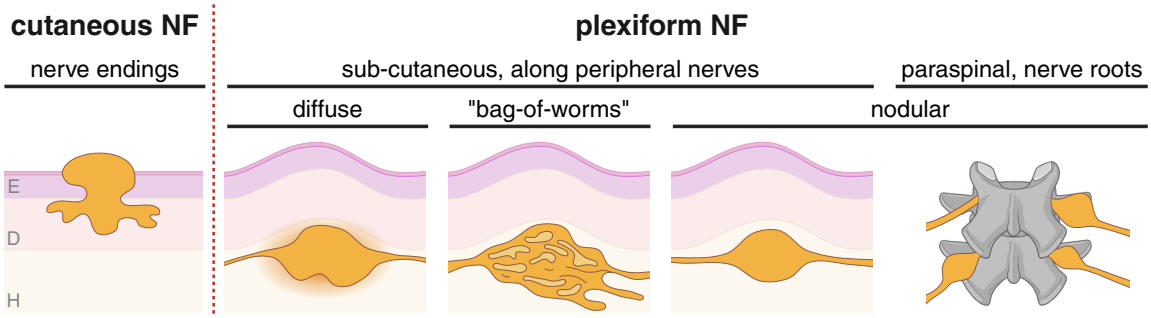


Figure 22 – **Various types of neurofibromas**. NFs are classified based on their location and clinical aspects. E: epidermis, D: dermis, H: hypodermis. Adapted from `https://amanf.org.br/2015/09/qual-a-melhor-tecnica-para-retirar-os-neurofibromas-parte-1/` (accessed March 2024), in coherence with Rasulic *et al.* (2021).

On one hand, dermal or **cutaneous neurofibromas** (cNFs) develop in the dermis at nerve endings and never become malignant. Their growth is transient and they remain small tumors. However, their number can reach hundred and severely impact the quality of life of patients.

On the other hand, **plexiform neurofibromas** (pNFs) "develop within peripheral nerves and their perineurial sheaths. [They] can invade adjacent tissues by disrupting the perineurium, remaining non-metastatic and clinically benign, but causing significant morbidity and increased risk of mortality when symptomatic" (Ratner & Miller, 2015). Indeed, pNFs are characterized by a constant slow growth and can reach kilograms. Based on their morphology, depth and relationship to adjacent tissues, subtypes are further reported (Fisher *et al.*, 2022). **Nodular** pNFs are well-defined tumors without alteration of the perineurium. **Diffuse** pNFs display alteration of the perineurium, leading to the escape of nerve fascicle populations from the nerve. Finally, pNFs with a **bag-of-worms** aspect are characterized by "multiple intertwined hypertrophic nerve fascicles" and "serpentine pattern with multiple nodules". (Rasulic *et al.*, 2021).

As reviewed by Carroll (2012), "despite major differences in their clinical behavior, all neurofibromas subtypes are composed of an identical but complex mixture of cell types" in patients. Indeed, as their name suggests, neurofibromas have the **same cellular composition as nerves**. They are composed of a majority tapered Schwann cells, fibroblasts, vascular elements, an immune infiltrate of mast cells and axons (Carroll, 2012; Meyer, 2020; Messersmith & Krauland, 2019). They further depict abnormal extra-cellular matrix (ECM) deposit, known as **fibrosis**, responsible for the large volume of some pNFs (Jiang *et al.*, 2021).

Only pNFs have the potency (5-15%) for malignant transformation (Carroll, 2012; Meyer, 2020; Miettinen *et al.*, 2017; Messersmith & Krauland, 2019). However, malignant transformation of diffuse pNFs is extremely rare (Schaefer & Fletcher, 2015). It remains to be characterized if the diffuse and nodular subtypes of pNFs form distinct or successive stages in their development.

2.2 Plexiform neurofibromas are benign nerve sheath tumors

In this section, I provide an overview of pNFs (Section 2.2.1), the effort of animal modeling to identify their cellular origin (Section 2.2.2) and their cellular and molecular features (Section 2.2.3).

2.2.1 Diagnosis and treatments

Despite they can develop in adults in sporadic context (Ho *et al.*, 2022), pNFs constitute a diagnosis criterion and are a hallmark of NF1 patients. Between 30% and 50% of patients develop pNFs (Ferner *et al.*, 2006), during fetal development, after birth or during early childhood (Akshintala *et al.*, 2020). As internal tumors, pNFs are reported and detected only when symptomatic. After tumor biopsy, the **diagnostic criteria** of pNFs include their macroscopic pattern of tumor (clinical aspect) and microscopic tissue characterization (histological aspect) (Meyer, 2020).

NF1 patients are followed throughout their lives for the metabolic activity of pNFs, monitored by magnetic resonance imaging (MRI). In case of rapid growth, large size leading to damage of vital structures, and risk of malignant transformation, pNFs may be removed (Miettinen *et al.*, 2017). While a **complete excision** is curative, it can be achieved in only 15% of cases (Fisher *et al.*, 2022). Indeed, due to an inaccessible location, the involvement of vital structures or a diffuse nature that hinders complete removal, the surgery is not always possible. More than 20 trials to cure pNFs were conducted and failed (Fisher *et al.*, 2022). To date, selumetinib is the only drug

approved for pediatric NF1-related unresectable pNFs. It acts as an inhibitor of MEK proteins, which are part of the RAS-MAPK pathway (Figure 21, red shape). In this context of **lack of therapeutic options**, there is a need to gain knowledge in the mechanisms governing pNF initiation.

2.2.2 Cellular origin

In the context of NF1, the loss of *NF1* heterozygosity has only been observed in the Schwann cell (SC) lineage, whereas all other cells remain $NF1^{+/-}$ (Perry *et al.*, 2001; Zhu *et al.*, 2002). *NF1* alterations have also been identified in SCs from sporadic pNFs, (Perry *et al.*, 2001; Beert *et al.*, 2012), suggesting a common cellular and genetic origin of both pNFs.

Although pNFs are thought to arise from $NF1^{-/-}$ **tumor** SCs, the specific time, during development, when the second mutation occurs, can not be studied in patients. Such a study has been extensively conducted in **genetically engineered mouse models** (Jacks *et al.*, 1994; Gitler *et al.*, 2003; Zhu *et al.*, 2002; Zheng *et al.*, 2008; Wu *et al.*, 2008; Le *et al.*, 2011; Mayes *et al.*, 2011; Ribeiro *et al.*, 2013; Chen *et al.*, 2014; Brosseau *et al.*, 2018; Radomska *et al.*, 2019; Chen *et al.*, 2019; Mo *et al.*, 2020) and recently reviewed (Ge *et al.*, 2022). Most models are in the form of:

$$\text{Driver}^{\text{Cre or CreERT}};Nf1^{\text{fl/fl or fl/-}};Rosa26^{\text{reporter}}$$

Indeed, the Cre recombinase system (Box 2) can be used to target the bi-allelic loss of *Nf1* in cells expressing a driver gene. Each model differs in the timing at which *Nf1* is altered, the location and number of cells affected. The driver gene defines the **cellular origin** of potential symptoms. To track all mutant cells by immunostaining, a **fluorescent reporter** can be added to the ubiquitous and strong promoter of *Rosa26*. Models further differ by their **genetic background**. In the $Nf1^{fl/fl}$ models, all cells are initially wild type for *Nf1*, whereas in the $Nf1^{fl/-}$ models, they are $Nf1^{+/-}$, thus better reflecting the condition of NF1 patients.

Targeting a bi-allelic loss of *Nf1* is all **SC precursors** (SCPs), through *Sox10* driver gene, leads to neonatal lethality, precluding the analysis of pNF formation (Mo *et al.*, 2020). However, same authors showed that loss of *Nf1* is all glial (*Sox10*+) cells in adult mice, using CreERT system, leads to pNFs development. This highlights that glial cells are at the origin of pNFs. Similarly, inducing *Nf1* loss in *Plp1* lineage in some SCPs and **immature SCs** leads to pNFs in mice on both genetic backgrounds (Le *et al.*, 2011; Mayes *et al.*, 2011). A mutation of *Nf1* restricted to **myelinating SCs** (mSCs), through the *Mpz* driver gene, is not sufficient to drive pNF formation, in both

Nf1$^{fl/fl}$ and *Nf1*$^{fl/-}$ backgrounds (Ribeiro *et al.*, 2013). This results suggest that mSCs are not likely to be at the origin of pNFs. Additionally, histologic and immunostaining of murine pNFs further highlighted the hyper proliferation of **non-myelinating SCs** (nmSCs), with specific staining for SOX10, S100B and GAP43, markers of the SC lineage, NGFR and GFAP, restricted to the nmSC lineage, and no or low staining for MBP, BFABP, the latter being a marker for glial progenitors (Jessen & Mirsky, 2005). Altogether, these observations suggest that **nmSCs are the cellular origin of pNFs**.

2.2.3 Cellular and molecular features of pNFs

The cellular and molecular mechanisms characterizing pNFs are mainly known from mouse models (Ge *et al.*, 2022). They are are summarized in Figure 23, which serves as support for next sections.

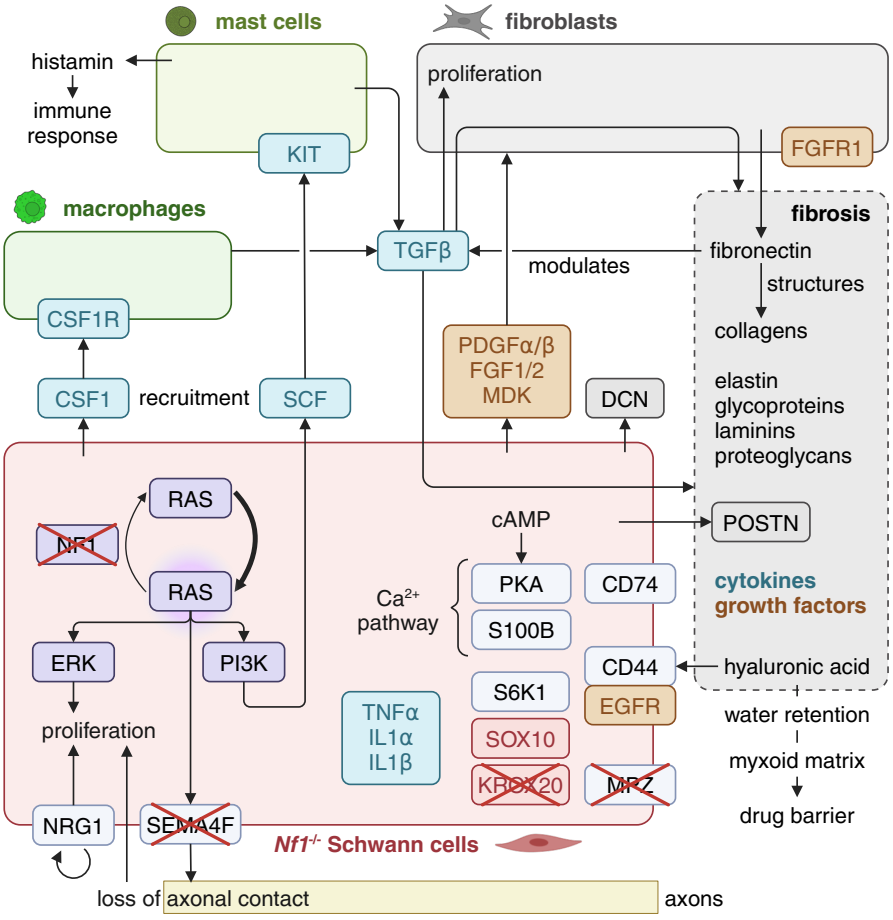


Figure 23 – Cellular and molecular features of plexiform neurofibromas. The figure summarizes elements from Carroll (2012), Chen *et al.* (2019), Ge *et al.* (2022), Kershner *et al.* (2022), and Jiang *et al.* (2023). Each cellular compartment is represented by a large colored shape. The extra-cellular matrix is represented by a dotted shape. Proteins are related to Ras pathways (purple), pro-inflammatory signals (blue), growth factors (brown) or extra-cellular matrix (dark gray). Red crossed molecules correspond to non functional or down-regulated proteins. Arrows indicate positive effects.

Autocrine mechanisms concern ***NF1*$^{-/-}$ tumor SCs** (Section 2.2.3.1). Paracrine mechanisms rely on the *NF1* **genetic background** and the signaling between cells from the **micro-environment** and tumor cells. Such cross-talks can be modulated by the **extra-cellular matrix** (Section 2.2.3.2). A trigger of pNFs development can also come from **external stimuli** (Section 2.2.3.3).

2.2.3.1 Intrinsic mechanisms in tumor Schwann cells

Inactivation of NF1 in SCs leads their **abnormal proliferation**, recruitment of **immune cells** and contribution to **fibrosis**. Indeed, tumor SCs contribute to inflammation through secretion of pro-inflammatory molecules (Figure 23, blue proteins). The abnormal proliferation results from (i) the over-activation of the RAS-MAPK pathway (Figure 23, purple proteins), (ii) the expression of NRG1, leading to autocrine stimulation of proliferation and (iii) the down-regulation of semaphorin 4F (SEMA4F), contributing to loss of axonal contact (Figure 23, bottom). Finally, tumor SCs express *SOX2*, *NOTCH1*, *ID2*, *EGR2*, *EGR3* and *TNC*, also expressed by repair SCs after injury (Gomez-Sanchez *et al.* (2013), Section 1.4.2).

As suggested in the literature, the repression of mSC identity, axonal detachment, proliferation and activation of inflammatory pathways can be compared to a transition to a repair SC phenotype (Parrinello & Lloyd, 2009; Gomez-Sanchez *et al.*, 2013).

2.2.3.2 Contribution of the micro-environment

Some mouse models are developed on both *Nf1*$^{fl/fl}$ and *Nf1*$^{fl/-}$ genetic backgrounds, enabling to study the impact of *Nf1* loss in the micro-environment on pNF development. Indeed, *Nf1* loss in *Krox20* lineage conducted to few enlarged nerves but no pNF, in *Nf1*$^{fl/fl}$ mice. On the opposite, pNFs were observed on *Nf1*$^{fl/-}$ background, suggesting that *Nf1* haploinsufficiency triggers pNF development (Zhu *et al.*, 2002).

An overview of the cellular heterogeneity of pNFs has been conducted by Kershner *et al.* (2022), who performed scRNA-Seq experiments on pNFs from the *Dhh*$^{Cre,+}$;*Nf1*$^{fl/fl}$ mouse model. Authors confirmed that pNFs share a similar composition with peripheral nerves. T cells, dendritic cells and vascular cells are considered are minor populations and are poorly characterized (Kershner *et al.*, 2022). Hereafter, I detail the contribution of each micro-environmental compartment in pNF biology.

Several **fibroblast** subtypes compose pNFs, including perineurial and endoneurial fibroblasts, characteristic of peripheral nerves (Kershner *et al.*, 2022). TGF-β released

by both tumor SCs and immune cells, and the excessive secretion of fibroblast growth factor 2 (FGF2) and platelet-derived growth factor (PDGF) by tumor SCs, stimulates fibroblasts to produce an excessive ECM (Jiang *et al.* (2021), Figure 23). Kershner *et al.* (2022) further highlighted the contribution of tumor SCs to fibrosis, through the secretion of periostin (POSTN).

In pNF, the **fibrosis** consists mainly of collagens and proteins for the basal lamina of SCs (Jiang *et al.* (2023), Figure 23). It may contain hyaluronic acid, with can further active pathways downstream EGFR, including Ras pathways, in tumor SCs. Additionally, hyaluronic acid contributes to water retention and myxoid matrix, the latter referring to an ECM without organized spatial arrangements. It can impede drug diffusion, thus treatment efficacy.

Jiang *et al.* (2021) highlighted the **mast cells** as a hallmark feature of pNF, where they are recruited by mutant SCs through stem cell factor (SCF) (Figure 23). Mast cells are pro-tumorigenic by participating in inflammation and fibrosis. These features are more severe in *Nf1*$^{+/-}$ cells. However, depletion of *SCF* in mutant SCs from *Plp*$^{CreERT,+}$;*Nf1*$^{fl/fl}$ mice does not impair the development of pNF, although it decreases mast cell influx (Liao *et al.*, 2018). This indicates that mast cells are not required or that alternative mechanisms are established to promote tumorigenesis. The limited contribution of mast cells to pNF development was confirmed by Burks *et al.* (2019).

Macrophages are highly enriched in pNF, where they are recruited by tumor cells through the CSF1-CSF1R axis. In established pNFs, macrophages contribute to the secretion of TGFβ, which further triggers the fibrosis (Figure 23). Prada *et al.* (2012a) further showed a switch in the role of macrophages between the initiation and progression of pNFs, using their *Dhh*$^{Cre,+}$;*Nf1*$^{fl/fl}$ mouse model. Indeed, anti-macrophages treatment, prior to pNF formation, leads to increased tumor volume compared to control mice. On the opposite, late treatment, once pNFs are formed, decreased the tumor volume and induced death in tumor SCs. Such results showed that macrophages are anti-tumorigenic before pNF formation, while they become tumor permissive once pNFs are established.

Brosseau *et al.* (2018) investigated the role of **T cells** in their *Plp*$^{CreERT,+}$;*Nf1*$^{fl/fl\ or\ -}$ mouse model where loss of *Nf1* from the immature SCs. They suggested that *Nf1*$^{+/-}$ T cells contribute to immune surveillance and prevent malignant transformation of pNF, compared to *Nf1*$^{+/+}$ T cells.

Finally, Le *et al.* (2009) highlighted the importance of **nerve micro-environment** by transplanting *Nf1*$^{-/-}$ skin-derived precursors from mice either subcutaneously or in the sciatic nerve of wild-type recipient mice. They observed pNF formation in the sciatic

nerve, whereas no tumors formed out of the nerves. These results showed that such mutant cells are not self-sufficient and required the nerve micro-environment to form tumors.

2.2.3.3 External stimuli: injury and inflammation

Given that pNFs contain many immune cells, fibrosis and that tumor SCs express similar genes than in the context of nerve repair, it has been suggested that pNF are **unrepaired wound of peripheral nerve** (Parrinello & Lloyd, 2009). From the same laboratory, Ribeiro *et al.* (2013) studied *Mpz*$^{CreERT,+}$;*Nf1*$^{fl/fl\ or\ fl/-}$ mouse models, in which mSCs are targeted to loose *Nf1*. Such mice do not spontaneously develop pNFs. Authors performed sciatic nerves injury and showed that pNFs specifically develop at wound site, regardless of the genetic background. This shows that injury is necessary for mSCs to initiate pNF.

All mouse models of pNFs, and particularly young NF1 patients, are not subjected to injury. However, targeting *Nf1* loss in nmSCs is sufficient to recruit immune cells, maintain inflammation and promote fibrosis (Fletcher *et al.*, 2020). Furthermore, the initiation of pNF is accelerated on a *Nf1*$^{+/-}$ background (Brosseau *et al.*, 2018; Fletcher *et al.*, 2020). This shows that the alteration of *NF1* in nmSCs and the *NF1* haploinsufficiency have similar consequence than injury.

2.2.4 No model of the spontaneous malignant transformation of pNFs

Mouse models enable to study the cellular origin, and the cellular and molecular features of pNFs. But, mice have a much shorten life compared to Human, which can preclude the study of the malignant progression of pNFs. A low incidence (10%) of spontaneous malignant transformation of pNFs has been reported in the *Plp*$^{CreERT,+}$;*Nf1*$^{fl/fl}$ mouse model, in which loss of *Nf1* is induced early after birth (Brosseau *et al.*, 2018).

The nearly absence of mouse models recapitulating pNF-to-MPNST transformation suggests that the cellular origin of MPNST is not among the targeted *Nf1*$^{-/-}$ SCs, meaning that the pNFs observed in mice will never transformed, or, the cellular origin of MPNST corresponds to a very small fraction of cells within pNFs, thus malignant transformation of pNF is a very rare event, or, mouse pNFs can transform, but an additional trigger is required, including epigenetic or genetic signals from tumor cells or external signals from the micro-environment. This remains an answered issue.

2.3 Dysplastic neurofibromas are pre-malignant lesions

2.3.1 Terminology

Although neurofibromas (NFs) with atypical histologic features have been defined as **atypical NFs** for decades, they were proposed as pre-malignant tumor recently (Beert *et al.*, 2011). Authors genetically compared atypical NFs to MPNSTs and showed shared features. However, authors did not prove that atypical NFs can transform into MPNSTs. For a better management of patients, (Miettinen *et al.*, 2017) proposed a classification of NF1-associated peripheral nerve sheath tumors (PNSTs) based on histologic features (Table 1). They recognized atypical and **cellular NFs** as to distinct pre-malignant entities, and introduced the atypical neurofibromatous neoplasm of uncertain biologic potential (**ANNUBP**) as the last stage before MPNST. This nomenclature is now used to better classify tumors from patients, and make the decision for their surgical excision (Miettinen *et al.*, 2017).

	Diagnosis	Definition
	neurofibroma (NF)	Benign Schwann cell neoplasm with thin, often wavy nuclei, wispy cell processes, and a myxoid to collagenous matrix. Immunohistochemistry includes extensive but not diffuse S100 and SOX10 positivity and a lattice-like CD34+ fibroblastic network
dysplastic NF	**atypical NF**	NF with atypia alone, most commonly manifesting as scattered bizarre nuclei
dysplastic NF	**cellular NF**	NF with hypercellularity, but retained NF architecture and < 1 mf/50 HPFs
dysplastic NF	**ANNUBP**	Schwann cell neoplasm with at least 2 of 4 features: cytologic atypia, loss of neurofibroma architecture, hypercellularity, mitotic index between 1/50 HPFs and 3/10 HPFs
	low-grade MPNST	Features of ANNUBP, but with mitotic index of 3 to 9/10 HPFs and no necrosis
	high-grade MPNST	MPNST with at least 10 mf/10 HPFs, or 3 to 9 mf/10 HPFs combined with necrosis

Table 1 – **Nomenclature and associated histological features of NF1-related PNSTs**. A neoplasm is an abnormal growth of tissue. Myxoid matrix refers to an ECM without organized spatial arrangements, like a mucus, and composed mainly of glycosaminoglycans, which include hyaluronic acid, keratin sulfate, chondroitin sulfate, dermatan sulfate, heparan sulfate, and heparin (Willems *et al.*, 2010). Collagenous matrix refers to a collagen-rich ECM, in which collagen fibrils are well organized, as "shredded carrots" (Miettinen *et al.*, 2017). Atypia is unusual nuclear appearance, including enlargement and variability in chromatin content. Mitotic figure (mf) count, or mitotic index, corresponds to the number of cells in mitosis, which is characterized by the separation of chromosomes, prior to cell physical division. High-power field (HPF) refers to the field of view of histologic sections, usually 0.2 mm^2. Finally, "loss of NF architecture refers to fascicular growth pattern and/or lack of CD34+ fibroblastic network" (Miettinen *et al.*, 2017). Adapted from Miettinen *et al.* (2017), with mention of dysplastic NF from Valeyrie-Allanore *et al.* (2008).

Prior to the definition of ANNUBP, Valeyrie-Allanore *et al.* (2008) introduced the term **dysplastic neurofibromas** (dyNFs) to characterize NFs with hypercellularity or atypical cells, regardless their architecture. In the laboratory, we therefore used the term dyNF to encompass all atypical NF, cellular NF, and ANNUBP.

2.3.2 Cellular composition

All dyNFs have a similar cellular composition compared to the benign NFs they originate from: S100+ and SOX10+ tumor cells, CD34+ fibroblasts forming a net-like pattern characteristic of NF architecture, and immune cells (Miettinen *et al.*, 2017). However, in contrast to benign NFs, dyNFs may show nuclear atypia and/or hypercellularity, the latter possibly due to lymphocyte infiltration (Table 1).

2.3.3 Loss of *CDKN2A*

Beert *et al.* (2011) showed that 15/16 atypical NFs harbored a deletion of the chromosomal region including *CDKN2A/B* genes, therefore highlighting the first genetic event preceding malignant transformation. Carrió *et al.* (2018) further showed that a mono-allelic loss of *CDKN2A/B* marked the pNF to atypical NF transition, while the complete loss initiate ANNUBP. The loss of CDKN2A may be responsible for excessive proliferation, impeding DNA structure check and further DNA damage repair, leading to accumulation of genetic aberrations (Box 3).

2.3.4 Lack of model recapitulating the pNF-to-dyNF transition

Taking advantage of the identification of *CDKN2A* loss as the first genetic event driving pre-malignant lesions, Rhodes *et al.* (2019) modified the pNF mouse model *Postn*$^{Cre,+}$;*Nf1*$^{fl/fl}$. They induce, in the meantime as the loss of *Nf1*, the loss of p19^{Arf}, one of the two isoforms of CDKN2A. Authors showed that their mice develop ANNUBP at the age of 4 months, but rapidly die of malignant peripheral nerve sheath tumors (MPNSTs). One year later, Chaney *et al.* (2020) showed that the loss of both isoforms of *Cdkn2a* in *Nf1* mutant cells from the *Dhh*$^{Cre,+}$;*Nf1*$^{fl/fl}$ mouse model (Wu *et al.*, 2008) develop paraspinal pNFs and subcutaneous atypical NFs that transform into MPNSTs.

These models highlighted that the loss of p19^{Arf} is sufficient to promote ANNUBP formation and malignant transformation. However, the *Dhh* and *Postn* cell lineages do not spontaneously lose the gene. In fact, to date, no model recapitulates the spontaneous transformation of pre-existing benign NFs into MPNSTs, leading to a poor characterization of the early mechanisms driving malignant transformation.

2.4 MPNST are aggressive soft tissue sarcomas

Plexiform neurofibromas have a 5-15% risk for malignant transformation (Miettinen *et al.*, 2017) and malignant peripheral nerve sheath tumors (MPNSTs) are responsible for a shortened life expectancy of patients with NF1 (Evans *et al.*, 2011). This section provided an overview of these aggressive tumors.

MPNSTs are soft tissue sarcomas. A **sarcoma** is "a type of cancer that begins in bone or in the soft tissues of the body, including cartilage, fat, muscle, blood vessels, fibrous tissue, or other connective or supportive tissue"[2]. Soft tissue sarcomas include approximately 70 subtypes of tumors, diagnosed based on histological features defined by the World Health Organization Classification of Tumors (Sbaraglia *et al.*, 2020).

2.4.1 Management of NF1 patients with MPNSTs

Risk factors for the development of MPNST include genetic factors such as missense mutations affecting codons 844 to 848 of the NF1 protein (Koczkowska *et al.*, 2018), encoded by exon 16 (Figure 19) and *NF1* microdeletion encompassing *SUZ12* (De Raedt *et al.*, 2003; Kehrer-Sawatzki *et al.*, 2017). Family history of MPNST (Reilly *et al.*, 2017), previous radiotherapy (Sharif *et al.*, 2006) and symptoms such as the presence of large internal benign neurofibromas (Mautner *et al.*, 2008; Nguyen *et al.*, 2014), subcutaneous neurofibromas (Tucker *et al.*, 2005) or peripheral neuropathies (Drouet *et al.*, 2004) increase the risk of developing a MPNST (Figure 24).

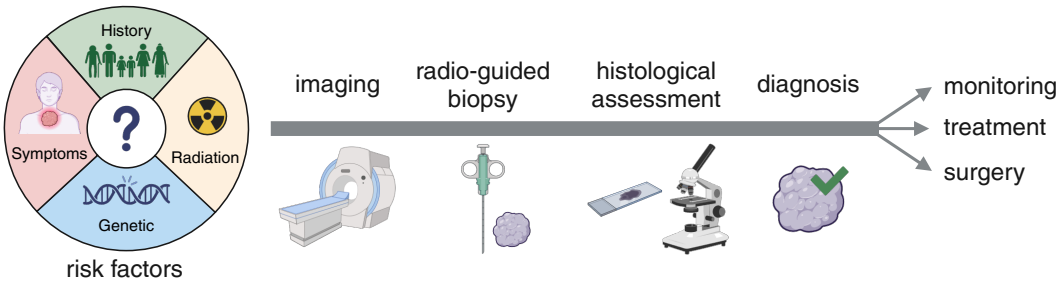


Figure 24 – Risk factors and patients care. The timeline of patients care is schematized. Patients are followed-up when they present risk factors for development a MPNST. After imaging, radio-guided biopsy, histological assessment, the diagnosis is established. A monitoring of tumor growth, treatment and/or surgery can be considered.

Patients with risk factors for developing MPNST are followed closely. When presenting with a symptomatic neurofibroma, i.e. a painful, firm or rapidly growing mass,

[2]From `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/sarcoma`, accessed March 2024

patients undergo imaging (Warbey *et al.*, 2009; Reilly *et al.*, 2017; Fertitta *et al.*, 2024). Based on the tumor size, extent and glucose metabolism, malignant tumor can be assumed (Fisher *et al.*, 2022). Following imaging, patients undergo a radio-guided biopsy to sample the more metabolically active region of the tumor. The biopsy sample is then evaluated by a pathologist to confirm the diagnosis (Figure 24).

MPNST are diagnosed on the basis of **histological features** (Table 1, Miettinen *et al.* (2017) and Somatilaka *et al.* (2022)). These features include uniform spindle cells with hyperchromatic nuclei, perivascular hypercellularity, presence of fascicles. High mitotic indices are the main difference with pre-malignant tumors. In addition, unlike dyNF, in most cases cells within MPNSTs have lost the Schwann cell differentiation markers. (Miettinen *et al.*, 2017). Finally, half of MPNSTs can be distinguished from other soft tissue sarcomas based on the loss of trimethylation of histone 3 at lysine 27 (H3K27me3) epigenetic mark (Pekmezci *et al.*, 2017; Schaefer & Fletcher, 2018).

The complete **surgical excision** of MPNST, with tumor-free margins to prevent local recurrence, is the main curative therapeutic option (Dunn *et al.*, 2013). **Radiotherapy** is performed in 60% of patients with cancer, and be used for a local control of high-grade MPNSTs, in case of large lesions, or to reduce tumor size prior to resection surgery (Gladdy *et al.*, 2010; Reilly *et al.*, 2017). **Chemotherapy** may be considered for metastatic or unresectable MPNSTs due to diffuse nature, involvement of vital structure, or risk of morbidity (Reilly *et al.*, 2017). Current chemical treatments are not specific to MPNSTs, but used in various types of cancer. However, due to high level of toxicity, systemic impact and risk or relapse, chemotherapeutic options have been described as a challenge in MPNSTs (Zehou *et al.*, 2013; Hassan *et al.*, 2021).

Currently, many **pre-clinical** and **clinical trials** are in progress for the treatment of MPNSTs (Martin *et al.*, 2019; Prudner *et al.*, 2020; Yao *et al.*, 2023a). They target specific pathways activated in tumor cells (Section 2.4.4). The better characterization of MPNSTs and the understanding of their initiation and progression are the results of their extensive modeling (Section 2.4.5).

2.4.2 Half of MPNSTs develop in sporadic context

Half of MPNSTs occur in NF1 patients, 45% are sporadic, and remaining 5% are radiotherapy-induced (Somatilaka *et al.* (2022), Figure 25). Although they arise in different contexts, MPNSTs are considered as a single entity for patient care (Valentin *et al.*, 2016).

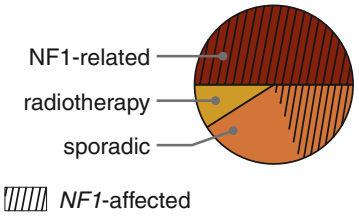


Figure 25 – Etiology of MPNSTs.

Majority of MPNST patients are **diagnosed** in their thirties and there is no gender bias (Mohamad *et al.*, 2021). However, NF1-related MPNSTs are diagnosed earlier compared to sporadic MPNSTs. This may be related to the surveillance of NF1 patients compared to non-NF1 individuals and the fact that MPNSTs often arise from pre-existing neurofibromas in NF1 patients (Valentin *et al.*, 2016). On the opposite, the origin of sporadic MPNSTs remains unknown.

Patients with radiotherapy-induced MPNSTs depict a much lower overall survival (almost 0% at 5 years) compared to other MPNSTs (Gladdy *et al.*, 2010; Yamanaka & Hayano, 2017). Similarly, NF1 patients with MPNSTs have a lower 5-year survival rate (50%) compared to sporadic cases (65%) (Valentin *et al.*, 2016; Hagel *et al.*, 2006). This may be related to three factors: (i) at diagnosis, NF1-associated MPNSTs are more advanced compared to sporadic cases, (ii) they are more often located in deep tissues, whereas sporadic MPNSTs are more superficial or (iii) NF1 and sporadic patients have distinct sensitivity to chemotherapy (Valentin *et al.*, 2016).

2.4.3 MPNSTs display epigenetic and genetic anomalies

Standard therapies for soft tissues sarcomas fail to treat MPNSTs. Epigenetic and genetic profiling may help to improve the diagnosis of benign and malignant tumors, stratify patients into homogeneous subgroups, gain insight into the molecular mechanisms defining MPNST subtypes, and develop a new treatment strategies.

2.4.3.1 DNA methylation, H3K27me3 mark and PRC2

The loss of the **trimethylation of histone H3 at lysine 27** (H3K27me3) is a diagnostic criterion for MPNST. A comparison between H3K27me3-positive and H3K27me3-negative MPNSTs indicated that this loss is associated with inferior survival and thus is a mark of tumor aggressiveness (Cleven *et al.*, 2016).

This epigenetic mark is promoted by the **Polycomb Repressive Complex 2** (PRC2) and is associated with the repression of transcription (Box 4). Notably, *SUZ12* and *EED*, encoding members of the PRC2, are frequently mutated in MPNSTs. This results in a complete loss of PRC2 function, subsequent loss of the H3K27me3 mark, and thus a transcriptional instability (Lee *et al.*, 2014; Zhang *et al.*, 2020a; Suppiah *et al.*, 2023; Cortes-Ciriano *et al.*, 2023).

DNA methylation analyses showed that pNFs and dyNFs share a similar methylation profile while MPNSTs are globally demethylated (Röhrich *et al.* (2016), Nix *et al.* (2018), and Suppiah *et al.* (2023), Figure 26). Among MPNSTs, two groups are further distinguished, in equivalent proportions across patients (Figure 26). One group is characterized by the sustained expression of the H3K27me3 mark, low methylation of CpG islands, few chromosomal rearrangements and the activity of the Wnt/β-catenin/*CCND1* pathways. Conversely, the second group depicts a loss of H3K27me3 mark, over-methylation of CpG islands, more rearrangements and activity of the Shh signaling pathway (Röhrich *et al.*, 2016; Suppiah *et al.*, 2023).

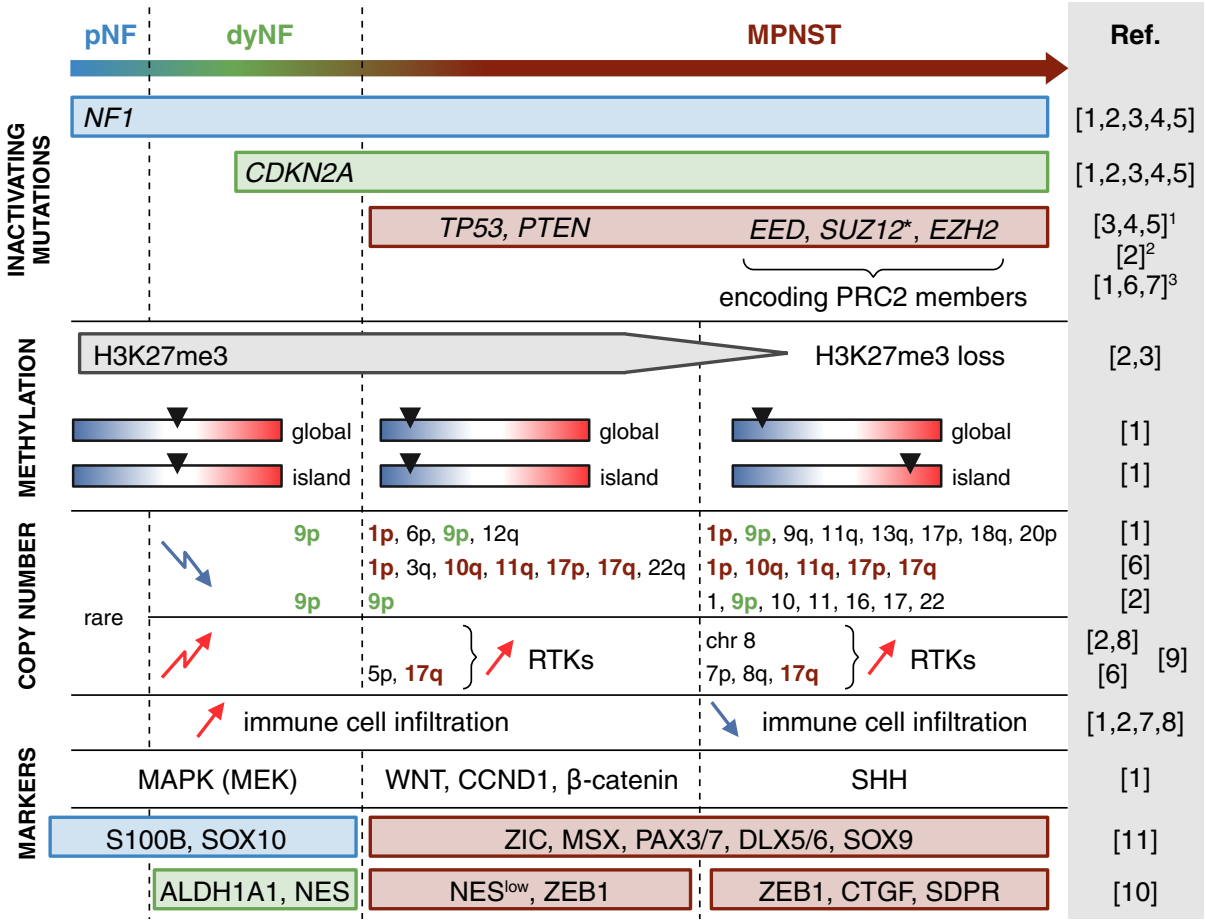


Figure 26 – Molecular events underlying MPNST progression. The illustration summarized findings from Suppiah *et al.* (2023) [1], Cortes-Ciriano *et al.* (2023) [2], Mohamad *et al.* (2021) [3], Zhang *et al.* (2020a) [4], Prudner *et al.* (2020) [5], Röhrich *et al.* (2016) [6], Høland *et al.* (2023) [7], Dehner *et al.* (2021) [8], Somatilaka *et al.* (2022) [9], Wu *et al.* (2022a) [10] and Kochat *et al.* (2021). MPNSTs are described as one group in Mohamad *et al.* (2021), Zhang *et al.* (2020a), and Prudner *et al.* (2020) (annotation [1]), as two distinct groups in Cortes-Ciriano *et al.* (2023) (annotation [2]) and as successive groups in Suppiah *et al.* (2023) and Röhrich *et al.* (2016) (annotation [3]). Star (*): inactivating mutation of *SUZ12* can be found from pNF stage. RTKs: receptors with tyrosine kinase activity.

The loss of H3K27me3 mark is further associated with the down-regulation of immune-related signaling pathways, such as interferon and antigen presentation (Wojcik *et al.*, 2019). The restoration of PRC2 activity in cell lines leads to up-regulation of genes related to interferon signaling and antigen presentation, compared to PRC2-deficient cell lines (Zhang *et al.*, 2022). Therefore, loss of PRC2 is associated with immune evasion and disease progression (Zhang *et al.* (2020a), Figure 26).

2.4.3.2 Point mutations and chromosomic rearrangements

The mutations reported in NF1-related or sporadic MPNSTs concern mainly **loss of functions** and involve first *CDKN2A* and then *TP53*, *EED* and/or *SUZ12* (Figure 26, Cortes-Ciriano *et al.* (2023) and Magallón-Lorenz *et al.* (2023)).

About half of sporadic MPNST have mutations in ***NF1*** (Bottillo *et al.* (2008), Lee *et al.* (2014), and Cortes-Ciriano *et al.* (2023), Figure 25). The loss of function of ***CDKN2A*** is a feature starting in pre-malignant stages and concern almost all MPNSTs, through a bi-allelic loss of copies or a translocation (Magallón-Lorenz *et al.*, 2021; Cortes-Ciriano *et al.*, 2023). ***TP53***, encoding the P53 protein, is a tumor suppressor gene that is altered in a variety of cancers. It is responsible for cell cycle arrest for DNA damage and repair checkpoints (Box 3). Alterations in P53 concern about 60% of MPNSTs (Cortes-Ciriano *et al.*, 2023). Its loss of function can lead to over-proliferation, without genome integrity checkpoints, and thus accumulation of genetic aberrations. Finally, PRC2 activity is negatively altered in several cancers, however with tumor-type specificities (Wassef & Margueron, 2017). Loss-of-function mutations or loss of copies of ***SUZ12*** are found in 75% of MPNSTs and ***EED*** in 50% (Cortes-Ciriano *et al.*, 2023).

MPNST cells further depict **chromosomal rearrangements** (Figure 26). They vary during tumor progression and can be used to better stratify tumors (Szymanski *et al.*, 2021; Sundby *et al.*, 2024). Gains particularly affect the loci of receptors with tyrosine kinase activity (RTKs) (Somatilaka *et al.* (2022), see next section).

In the development of MPNST **animal models**, validation criteria include histologic features and recapitulation of molecular events described in patient MPNST. Regarding genetic events, alterations in specific genes can be verified due to the homology of genomic sequences between humans and mice. However, the physical location of genes on the chromosome, referred to as synteny, differs between mouse and human. In fact, chromosomal rearrangements that occur in mouse models are rarely evaluated.

Box 4: H3K27me3 mark and the Polycomb Repressive Complex 2 (PRC2)

PRC2 is a highly conserved complex consisting of 4 proteins (Figure 27). In both human and mouse genomes, the gene encoding SUZ12 is located close to *NF1*. In NF1 disease, microdeletion of *NF1* may include *SUZ12*, which further increases the risk of developing MPNST (Kehrer-Sawatzki *et al.*, 2017). Between EED and SUZ12 can be found either EZH1 or EZH2. While EZH1 is present in all cells, EZH2 is only found in dividing cells (Margueron & Reinberg, 2011).

PRC2 has binding partners, the AEBP2, JARID2 and PCL1-2-3 (PCL) proteins, which modulate its activity (Holoch & Margueron, 2017). Binding of non-coding RNA to EZH1/2 can also trigger PRC2 recruitment to the DNA (Margueron & Reinberg, 2011).

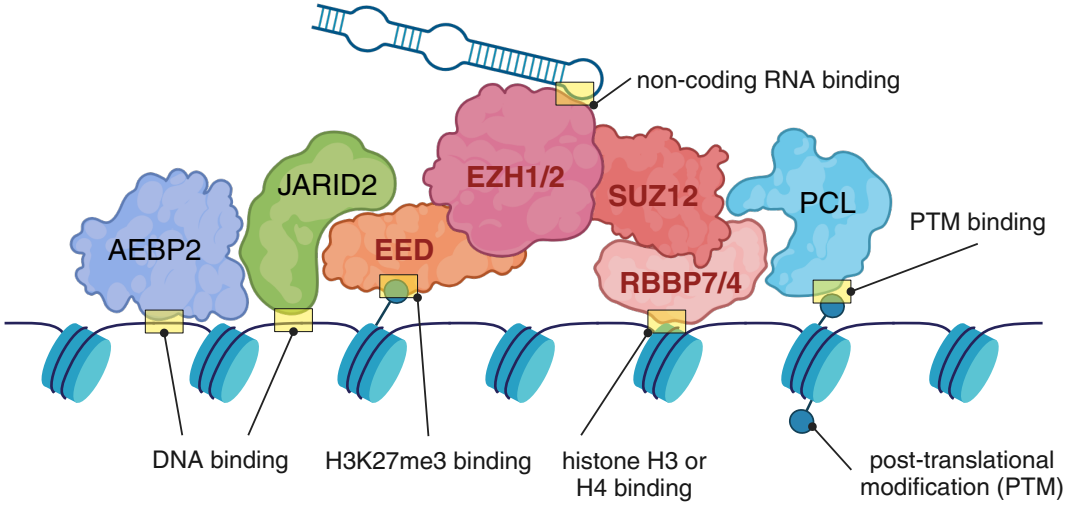


Figure 27 – Recruitment and structure of the Polycomb Repressive Complex 2. PRC2 is shown as red proteins while other proteins recruit PRC2 near the DNA. Binding regions are indicated as yellow boxes. Adapted from Margueron and Reinberg (2011).

PRC2 acts as a methyltransferase, through EED. It promotes successive methylation of histone H3 at lysine 27, known as H3K27me2 and H3K27me3 marks (Figure 28). Increased levels of H3K27me3 correlate with gene silencing (Cao *et al.*, 2002). The binding of EED to H3K27me3 is a positive feedback loop, as it enhances the PRC2 methyltransferase activity.

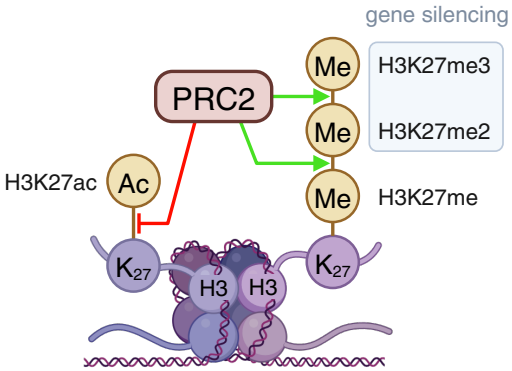


Figure 28 – PRC2 and histone modifications. PRC2 stimulates the methylation of H3K27, with is associated with gene silencing.

2.4.4 Cellular composition and molecular identity of MPNSTs

The Figure 29 serves as support for the next sections, relative to signaling pathways in MPNST tumor cells and the composition of the tumor micro-environment (TME).

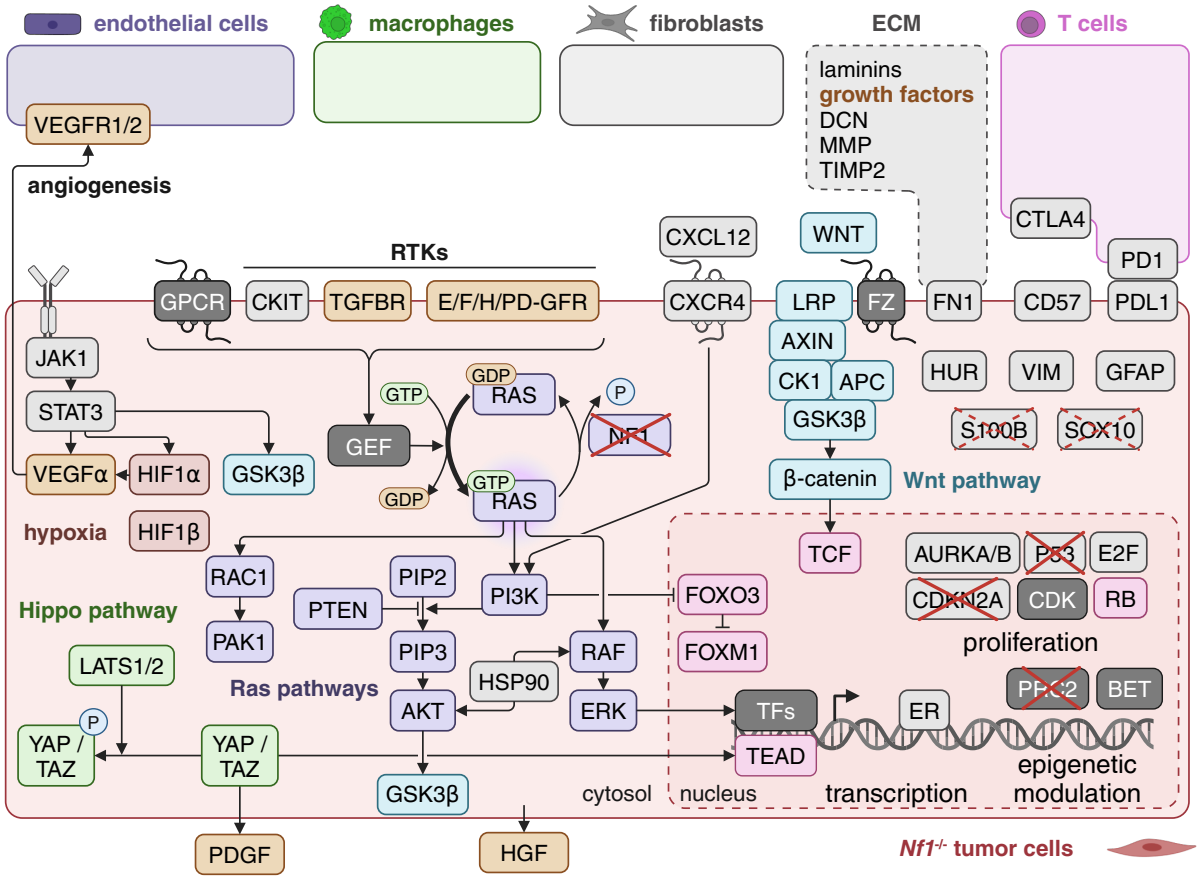


Figure 29 – **Biomarkers and signaling pathways defining MPNSTs**. Proteins are colored according to signaling pathways. Pink elements correspond to transcription factors (TFs). Dark gray elements correspond to protein families or complexes. Red crossed elements may have undergone loss of functions in MPNSTs. Dashed red crossed elements (SOX10 and S100B) are no longer expressed in MPNSTs. The picture summarizes several publications and reviews: tumor-micro environment, extracellular matrix (ECM), VIM and GFAP, (Martin *et al.*, 2020), angiogenesis (Walker & Upadhyaya, 2018; Somatilaka *et al.*, 2022; Bonsang *et al.*, 2022), Hippo pathway (Walker & Upadhyaya, 2018; Martin *et al.*, 2019; Somatilaka *et al.*, 2022), Wnt pathway (Mohamad *et al.*, 2021; Somatilaka *et al.*, 2022), JAK-STAT signaling (Walker & Upadhyaya, 2018), relationships between PI3K and FOX transcription factors (Martin *et al.*, 2020), HSP90 acting as a chaperone for AKT and RAF (Prudner *et al.*, 2020), the CXCL12-CXCR4 axis (Walker & Upadhyaya, 2018; Martin *et al.*, 2020), the presence of estrogen receptor (ER) (Martin *et al.*, 2019), the multiple roles of HUR (Palomo-Irigoyen *et al.*, 2020; Mohamad *et al.*, 2021), and proliferation (Martin *et al.* (2020) and Somatilaka *et al.* (2022) and Box 3). Ras signaling is addressed in most reviews. BET: bromodomain and extra-terminal motif proteins; CDK: cyclin-dependent kinases; GEF: guanine exchange factors; GPCR: G-protein coupled receptors; FZ: frizzled transmembrane proteins; TFs: transcription factors; PRC2: polycomb repressive complex 2.

2.4.4.1 Pathways in MPNST tumor cells

Chromosomal amplifications in MPNSTs lead to over-expression of gene expression, particularly in receptors with tyrosine kinase activity (RTKs): *KIT*, *PDGFR*α, *EGFR*, *HGFR* (also known as *Met*), and its ligand *HGF* (Figure 26, Somatilaka *et al.* (2022)). Since RTKs are upstream of RAS, their amplification may contribute to the over-activation of several **Ras pathways** (Figures 20 and 21).

MPNSTs are characterized by other signaling pathways (Figure 29). Most of them are the purpose of preclinical tests (Farid *et al.*, 2014; Walker & Upadhyaya, 2018; Prudner *et al.*, 2020; Somatilaka *et al.*, 2022). For instance, the **Wnt-**β **catenin pathway** is responsible for the neural crest differentiation (Section 1.3), but also proliferation and migration. Its over-activation is described in most cancers (Somatilaka *et al.*, 2022). The **hypoxia pathway**, through HIF1A and HIF1B, is notably activated during the nerve regeneration process to promote functional recovery via the angiogenesis (Section 1.4.2). In cancer, it is implied in tumor growth and metastasis, also through angiogenesis. Finally, the **Hippo pathway** is implied in the limitation of cell proliferation and promotion of apoptosis. When the pathway is inactivated, the proteins YAP or TAZ are not phosphorylated and can translocate to the nucleus (Figure 26). They activate the TEAD transcription factors to regulate proliferation and death. When the pathway is activated, YAP or TAZ are phosphorylated and cannot translocate. The over-activation of Hippo pathway promotes the progression of high grade MPNSTs (Wu *et al.*, 2018).

2.4.4.2 Sox9 is biomarker of MPNSTs

A **biomarker** is "a naturally occurring molecule or characteristic by which a particular pathological or physiological process, disease or disease characteristic can be identified" (Tripathi *et al.*, 2022). Identification of biomarkers can help to better stratify tumors, thereby improving diagnosis and patient care. Many approaches can be considered to identify biomarkers.

By comparing RNA-Seq profiles of normal SCs, pNFs and MPNSTs human cell lines, Miller *et al.* (2009) identified *SOX9* as a biomarker and potential therapeutic target in MPNSTs (Box 1).

2.4.4.3 Paucity of MPNST micro-environment

In addition to tumor cells, which consists of the main cellular component in MPNSTs, the **tumor micro-environment** (TME) populations is restricted to macrophages and endothelial cells (White & Rhodes, 2024). Few mast cells, T cells and fibroblasts remain for benign and pre-malignant stages. This composition is recapitulated in MPNSTs developed in mouse models, as identified by scRNA-Seq (Wu *et al.*, 2022a).

Since the PD-1/PDL-1 axis has been identified as a therapeutic target in several cancers, it has been investigated in MPNSTs. Expression of the PDL-1 membrane receptor allows tumor cells to escape cell death induced by PD-1 expressing **T cells** and is a reason for tumor progression. Blocking PDL-1 is a therapeutic option to restore the anti-tumorigenic role of the immune system. Indeed, Shurell *et al.* (2016) showed that PDL-1 staining is increased in MPNSTs compared to benign tumors. However, T cells are excluded from MPNSTs. Such a treatment requires to call back them inside the tumors, which is possible (Somatilaka *et al.*, 2024). Indeed, in MPNSTs, immunotherapy may require a two-steps treatment, by first re-activating the immune system in MPNSTs, and then targeting them to kill tumor cells.

In MPNSTs, **M2 macrophages** outnumber M1 type in a 6:1 ratio. Preclinical studies and clinical trials showed that targeting the reduction of M2 macrophages in MPNSTs resulted in a reduction of MPNST cell proliferation and clinical benefit. These results suggest the pro-tumorigenic role of macrophages in MPNST (White & Rhodes, 2024).

Finally, since MPNSTs develop in both NF1 and sporadic contexts, but pre-malignant stages are detected only in NF1 patients, one may suspect the implication of ***NF1* haploinsufficiency in TME**, in tumor progression. Indeed, mouse models on $NF1^{+/+}$ background, which better reflect sporadic context, are more prone to malignant transformation compared to $NF1^{+/-}$ background (White & Rhodes, 2024). This may explain why pre-malignant states are not detected in sporadic cases.

Although the **extra-cellular matrix** (ECM) is the major component of pNF, the evolution of its composition during malignant transformation has not yet been characterized. However, a recent study highlighted that ECM remodeling is a hallmark of tumor progression, particularly prior to metastasis (Brockman *et al.*, 2022). The authors therefore investigated the correlation between loss of PRC2 activity, ECM remodeling and metastatic potential. Using cell lines derived from different mouse models, they showed that loss of PRC2, and thus of H3K27me3, increases the expression and activity of matrix metalloproteases (encoded by *MMP** genes) and lysyl oxidases (encoded by *LOX** genes) and further increases metastasis.

2.4.5 Cell lines and genetically engineered animal models for MPNSTs

To hypothesize on the cellular origin of MPNSTs, understand their mechanisms of initiation and progression, identify therapeutic targets, and assess drug efficacy prior to clinical trials, modeling is required (Somatilaka *et al.*, 2022).

Cell lines are either (i) immortalized tumor cell lines derived from primary patient tumors or (ii) induced pluripotent stem cells (iPSCs), also derived from patient cells. They can be used as 2D or 3D *in vitro* or *in vivo* models (Mazuelas *et al.*, 2022). Although such a model does not recapitulate the tumor micro-environment, cell lines are better suited for drug screening and *in vitro* manipulation compared to animal models. In addition, since they are derived from patients, each cell line reflects some of the genetic diversity found in NF1 patients. However, immortalized cell lines require in-depth characterization to ensure that they truly reflect MPNSTs. Indeed, Magallón-Lorenz *et al.* (2023) provided a complete genomic and epigenetic characterization of human MPNST cell lines, which can serve as a resource for diagnosis and cell line selection for drug screening.

Primary tumor cells from patients can be xenografted into mice, called **PDX model**. They are suitable for personalized medicine because each model is specific to a given patient (Creus-Bachiller *et al.*, 2023).

Knowledge of genetic alterations in MPNSTs allows the development of animal models. Currently, **genetically engineered animal models** include fly, zebrafish and mouse (Somatilaka *et al.*, 2022). Under the induced genetic changes, they can develop symptoms of the disease. *Nf1*-mutant mouse models develop pNFs that do not undergo malignant transformation without additional genetic mutations (Section 2.2). Current mouse models that develop MPNST are targeted for conditional additional mutations in either *Trp53*, *Cdkn2a* (encoding INK4A and ARF proteins, Box 3), *Pten*, or *Eed*, all four of which are described in human MPNSTs (Mohamad *et al.*, 2021; Somatilaka *et al.*, 2022). This leads to *de novo* and rapid MPNST development, preventing any study of pre-malignant states. Furthermore, animal models suffer from limited genetic diversity (Dehner *et al.*, 2021). For example, *Trp53*$^{-/-}$-based mouse models reflect only the fraction of human MPNSTs that are altered for *TP53*.

An animal model that recapitulates the malignant transformation of benign neurofibromas would be of great use to uncover the mechanisms driving MPNSTs development and identify therapeutic targets to prevent their development.

2.5 A *Nf1*-KO mouse model recapitulates several features of NF1

Boundary cap cells derivatives give rise to SCs and are located from nerve roots to nerve endings, two locations where neurofibromas develop in NF1 disease (Section 1.3.2). To investigate the propensity of such cells to form neurofibromas, Radomska *et al.* (2019) targeted the conditional bi-allelic inactivation of *Nf1* in *Prss56* cells and their derivatives by definitive excision of exons 31 and 32, encoding the protein region responsible for RAS inactivation (Figure 19). Inactivation of *Nf1* is accompanied by the constitutive expression of the *Tomato* reporter gene, allowing the identification of mutant cells even if *Prss56* is no longer expressed (Figure 30, top).

Authors showed that male and female mice develop two subtypes of plexiform neurofibromas and small cutaneous neurofibromas (cNFs), in both $Nf1^{+/+}$ and $Nf1^{+/-}$ genetic backgrounds (Figure 30). A skin injury accelerates the development and increase the size of cNFs. Based on this observation, mice are kept together in the same cages in the animal facility. As a result, only male mice are subjected to chronic skin trauma due to their natural aggressive behavior. In the results chapter, we will show that, in this mouse model, neurofibromas progress into MPNSTs.

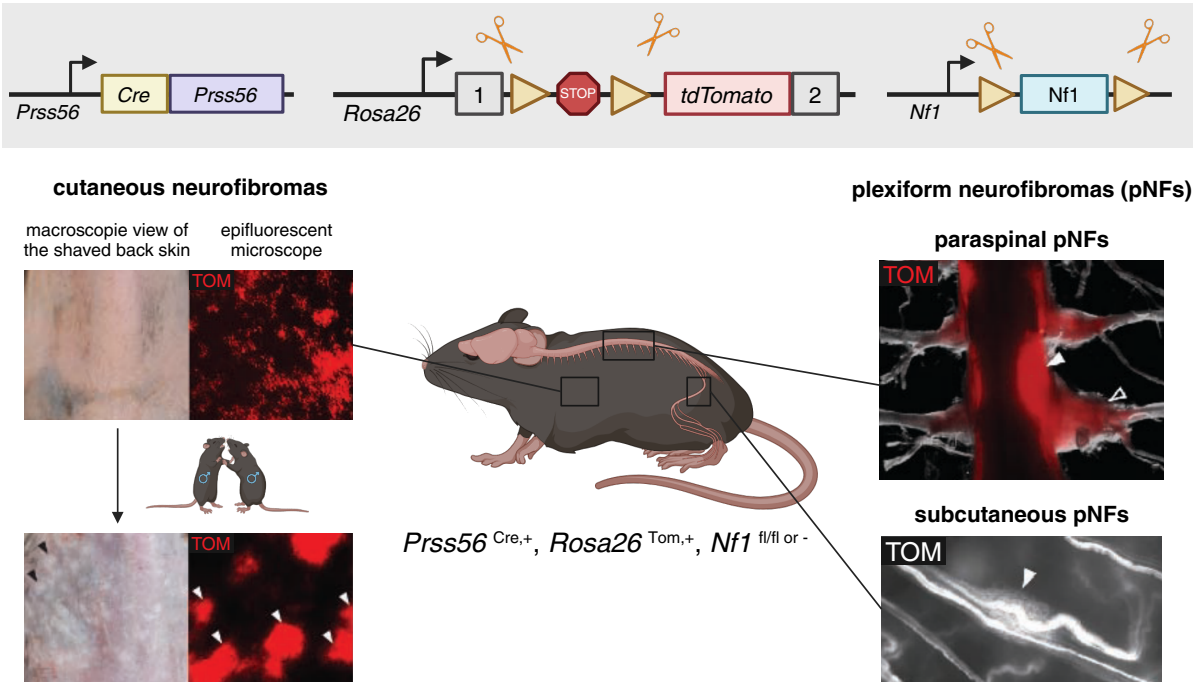


Figure 30 – *Prss56*^{Cre} *Nf1*-KO mouse model recapitulates several features of NF1. The construction of the mouse model is schematized on top. See Box 2 for further details on the Cre recombinase system. The mice develop cutaneous neurofibromas, that increase of size when male are housed together (left panel) and various types of plexiform neurofibromas (right panel).

2.6 Overview of the chapter

Neurofibromatosis Type 1 (NF1) is a monogenic disease characterized by a germinal haploinsufficiency in neurofibromin protein, encoded by *NF1*: patients are *NF1*$^{+/-}$. Neurofibromin is a multi-domain protein with many roles described. Among them, NF1 inactivates RAS proteins, whose downstream pathways are related to the metabolism, cell shape and cell survival or death.

NF1 patients can develop peripheral nerve sheath tumors (PNSTs). Particularly, **plexiform neurofibromas** (pNFs) arise from Schwann cells having lost the second allele of *NF1*: they are *NF1*$^{-/-}$. These benign tumors arise in early childhood. They are well characterized and share a similar cellular composition with peripheral nerves, in addition of fibrosis. The inflammation has been shown as a major trigger in pNFs development, but the early mechanisms driving their initiation remain unknown.

In NF1 patients, pNFs can transform into malignant PNSTs (MPNSTs), which are aggressive tumors prone to metastasis and with few therapeutic options. The malignant transformation is preceded by a transient intermediate stage (**dysplastic neurofibromas**, dyNFs). The first genetic event driving the pNF to dyNF transition is the genetic loss of *Cdkn2a*, which normally holds cells during their cell cycle for DNA integrity checking. Unlike pNFs, dyNFs are characterized by an infiltration of immune cells. Few additional features are known. Although a plethora of mouse models has been develop to study pNFs, the malignant transformation was reported once and not investigated.

While pNFs and dyNFs are almost specific to NF1 patients, half of MPNSTs are sporadic. **MPNSTs** are characterized by a genetic burden, notably amplifications in the chromosomic regions encoding activators of RAS, loss of functions in P53, implied in the maintenance of genome integrity, and in the proteins forming the PRC2, a complex acting in the epigenetic repression of gene expression. Unlike pNFs and dyNFs, in which *Nf1*$^{-/-}$ tumor cells have a glial identity, MPNSTs have lost this Schwann cells identity in favor of a mesenchymal profile. The micro-environment is further restricted to macrophages and endothelial cells. Similarly to pNFs, many models have been developed to study MPNSTs. All animal models are engineered to carry MPNSTs-associated mutations and do not recapitulate the benign-to-malignant transition. Thus, the mechanisms driving malignant transformation are poorly described.

The last section of this chapter introduced the *Nf1*-KO mouse model developed in the lab and presented in 2019. Mice recapitulate several features of NF1 patients, including pNFs. In the Results chapter, we will took advantage of this model to study the early mechanisms driving pNFs, and show they stepwise transform into MPNSTs.

3 Sequencing breakthroughs revolutionize biology

Cells' structure, functions and fate are determined by their genome and by their local environment. The genome is stored as DNA (Figure 31). Based on chromatin accessibility, histone and DNA modifications and the action of regulatory elements, the genome expresses itself.

About 75% of the human genome can be transcribed into RNA ("Decoding noncoding RNAs" 2022). The remaining 25% includes regulatory regions, recognized by regulatory elements, or yet to be characterized. The 75% of RNA-coding DNA is segmented into genes, delimited by a transcription start site (TSS) and transcription termination site (TTS) (Figure 32A). Only 4% RNA-coding DNA gives messenger RNA (mRNA), which can be further translated into proteins (Figure 32B).

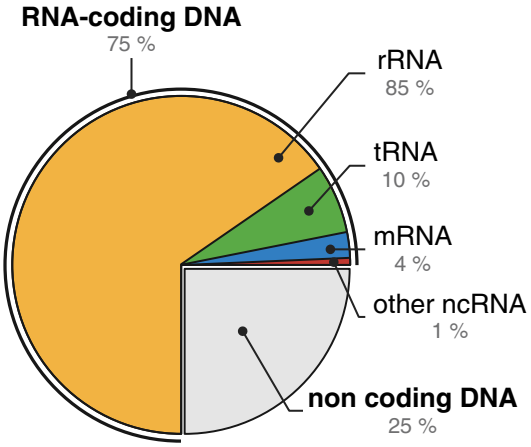


Figure 31 – **Human genome composition**. Human genome consists of non coding and RNA-coding DNA. The majority of RNA-coding DNA sequences encodes ribosomal RNA (rRNA) and transfer RNA (tRNA), used for the translation of messenger RNA (mRNA) into proteins. Other non-protein-coding RNA (ncRNA) include long non-coding RNA, small non-coding RNA and microRNA. Proportions are given in weight of DNA (Palazzo & Lee, 2015).

3.1 Overview of genomic approaches

To characterize and understand biological tissues, several genomic levels can be studied: epigenetic, genetic, transcriptomic and proteomic. In this section, I address two sequencing approaches, applied in the laboratory (Figure 32C).

Proteins consist of the last level of DNA processing, and are directly responsible for most of cell activity. They are encoded by exonic regions of genes. Therefore, mutations in exonic regions can lead to non functional proteins. The set of exons in the genome is called the exome. Exome sequencing (**Exome-Seq**) can be used to detect such genetic abnormalities, including single-nucleotide variations (SNVs) and copy number variations (CNVs). While SNVs correspond to nucleotide deletion, insertion or modification, CNVs are amplifications or losses of large chromosomal regions. Based on the nucleotides they affect, SNVs can lead to aberrant protein function. CNVs cause alterations in the quantity of proteins produced.

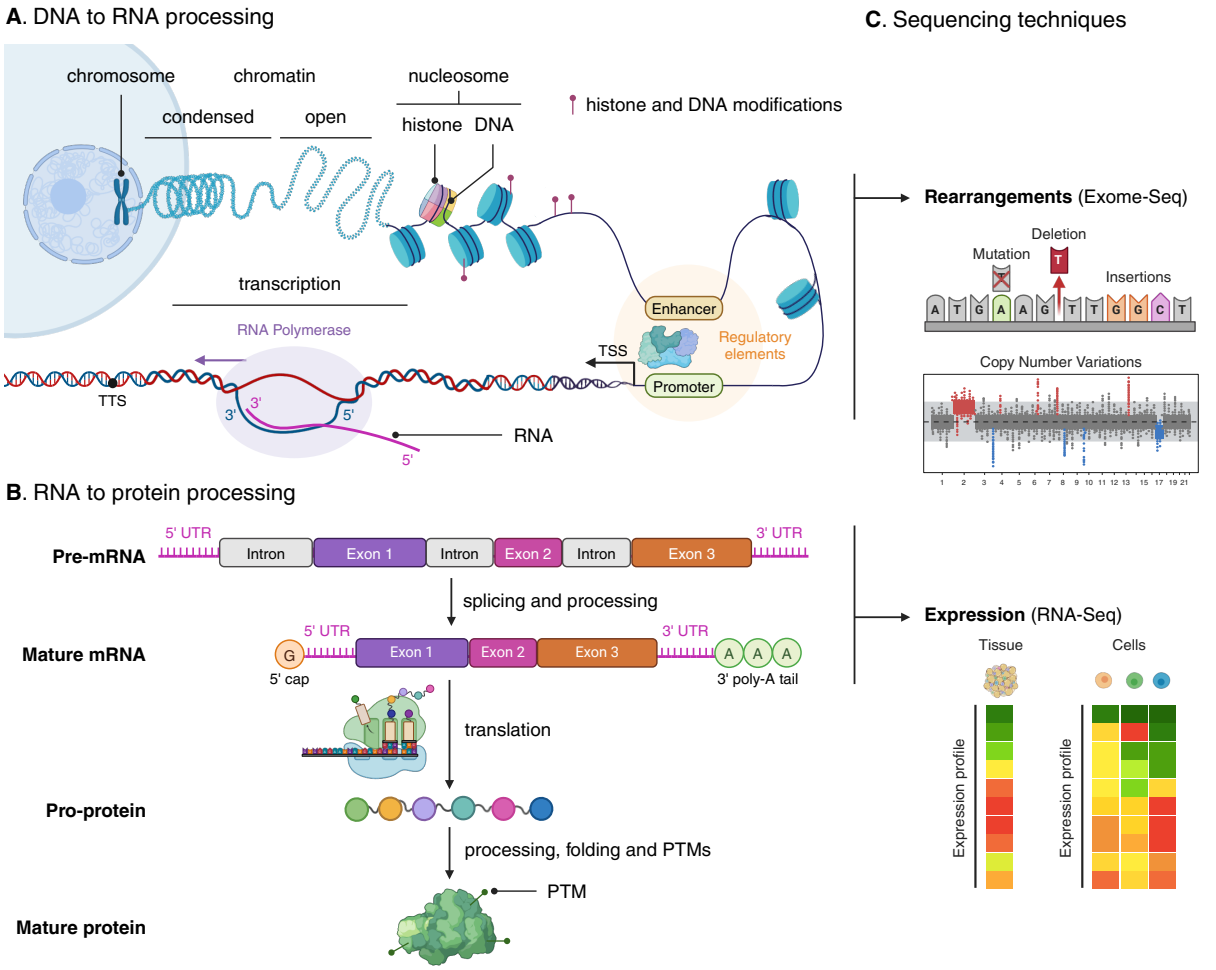


Figure 32 – DNA to protein processing and genomic approaches of interest. **A**. Chromosomes are made of condensed or open chromatin, which consists of nucleosomes composed of 8 histones surrounded by a double-stranded DNA molecule. Both histones and DNA can carry modifications that affect their accessibility. Genes correspond to transcribed DNA sequences starting with a transcription start site (TSS) and ending at a transcription termination site (TTS). RNA polymerase transcribes genes into RNA by reading the coding strand of DNA. **B**. Pre-messenger RNA (pre-mRNA) is a single-stranded molecule containing exonic and intronic sequences. Protein-coding RNA maturation involves splicing to remove intronic sequences, as well as further processing, including the addition of a poly-adenylation (poly-A tail) to the 3' untranslated region (UTR). Mature mRNA is exported from the nucleus and translated into proteins by the ribosome. The resulting protein precursor (pro-protein) undergoes editing, such as cleavage followed by three-dimensional folding and post-translational modifications (PTMs), such as phosphorylation. **C**. Sequencing techniques provide detailed information about the genetic material and molecules in a sample. RNA sequencing (RNA-Seq) can assess the presence and amount of RNA at the tissue or single-cell level. Exome sequencing (Exome-Seq) can detect genomic abnormalities in exonic regions. Adapted from "Regulation of Transcription in Eukaryotic Cells" and "RNA Processing in Eukaryotes", by BioRender.com (2023). Retrieved from `https://app.biorender.com/biorender-templates`.

Due to the three-dimensional structure of proteins and the challenge of extracting and characterizing the entire proteome from a sample (Aslam *et al.*, 2017; Timp & Timp, 2020), it is easier to quantify the penultimate processing step before protein, the messenger RNA (mRNA). It is possible to extract all RNA molecules and sequence them through RNA sequencing (**RNA-Seq**) approaches. RNA-Seq provides information regarding the amount of transcripts present in a tissue, the different types of RNA molecules (Figure 31), as well as splicing isoforms. Since all transcripts are captured blindly, RNA-Seq can also be used to discover novel transcripts specific to a tissue.

There are two main kinds of RNA-Seq. At a tissue level, the technique is known as **bulk RNA-Seq**. However, tissues contain various cells. Single-cell level RNA-Seq (**scRNA-Seq**) can help to characterize these diverse cell populations.

The main challenge in scRNA-Seq is to isolate single cells while tracing the cellular origin of transcripts. The first analysis of gene expression in single cells was performed by Eberwine *et al.* (1992). Cells were isolated manually and transcript capture was targeted to specific genes. Since then, Tang *et al.* (2009) improved the RNA capture method to bypass specific transcript targets. This was the first scRNA-Seq experiment dealing with transcriptomic scale data. However, only a few cells were characterized through this technique, as single-cell isolation was done manually. Since 2013, new technologies have appeared to reduce the production costs and automatically trace single cells (Svensson *et al.*, 2018). This allows for the characterization of the transcriptome of hundreds of cells simultaneously.

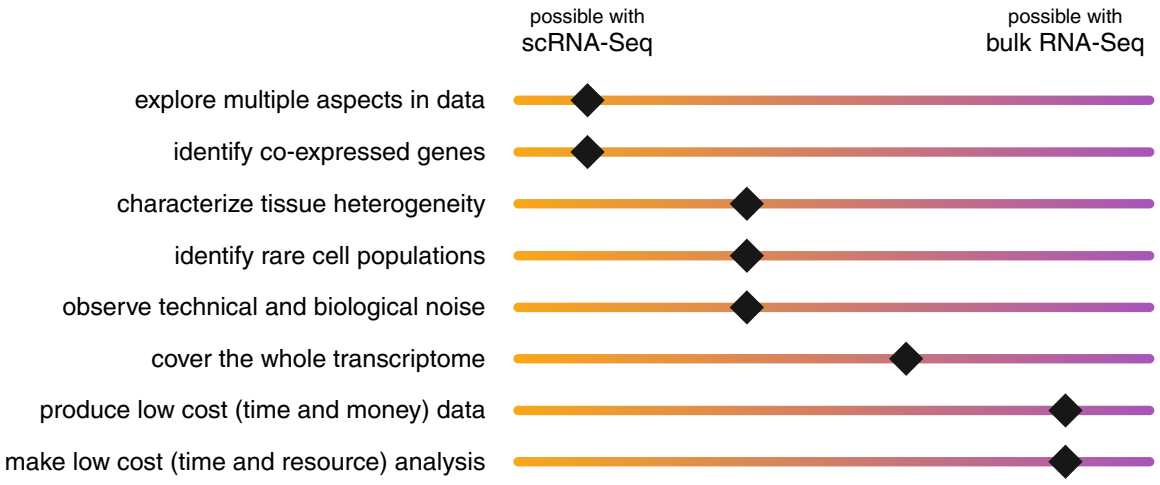


Figure 33 – **Advantages and limitations of scRNA-Seq over bulk RNA-Seq**. Each line represents a criterion. The cursor delineates whether the criterion is met with scRNA-Seq data or bulk RNA-Seq data, although it can vary according to the approach used.

The advantages and limitations of today's scRNA-Seq approaches compared to bulk RNA-Seq are illustrated in Figure 33. The high dimensionality of scRNA-Seq data enable to explore multiple aspects in data (Section 3.3.2.2). For instance, identifying co-expressed genes within a cell is only possible at single-cell resolution. However, and despite some scRNA-Seq approaches aiming to overcome this limit, the detection rate of transcripts is lower in scRNA-Seq compared to bulk RNA-Seq, thus transcriptome coverage is better with bulk strategies. Among the limitations, scRNA-Seq data are noisy. Due to single cell preparation, including dissociation and sorting, cells may suffer from stress, which represents a technical noise. Biological noise further arise from stochastic gene expression and cell-to-cell variability, which is averaged in bulk RNA-Seq. Additionally, producing scRNA-Seq is much more expensive and time-consuming compared to bulk RNA-Seq due to the cell isolation step, specific library preparation and intensive sequencing. Finally, scRNA-Seq data analysis involves more steps and more memory resources.

3.2 scRNA-Seq enables single-cell level analysis of tissue

Since the scRNA-Seq data is the endpoint of a multiple-steps processing, understanding each step is necessary to avoid over-interpretation of the data. The purpose of this section is to explain the generation of scRNA-Seq data, the inherent bias they contain, and the content of the input data for downstream computation analysis. Since all protocols are based on single cell labeling, the first step consists of preparing a single cells suspension (Section 3.2.1). The second step involves capturing RNA molecules while keeping track of their cellular origin (Section 3.2.2). The last step concerns the preparation of libraries and their sequencing (Section 3.2.3).

3.2.1 Preparation of the single cells suspension

This section highlights the key steps of the preparation of the single cells preparation (Figure 34) and its impact on final data.

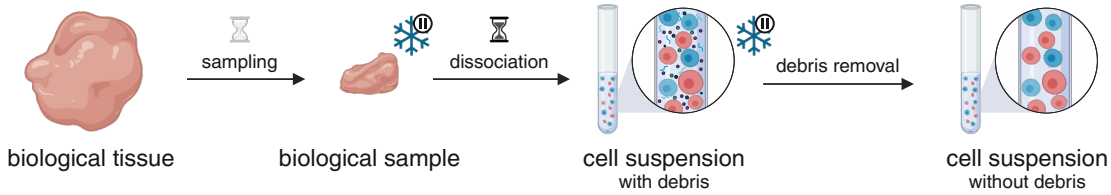


Figure 34 – **Single cells suspension preparation**. The four main steps of single cells preparation are illustrated. Light hourglass indicates that sampling may take time for very small or difficult-to-access tissues. Snowflake indicates that the sample can be frozen.

Cells of interest are contained in a biological tissue, which can be of large size or precious. Therefore, the entire tissue is often not used for scRNA-Seq preparation and requires a **sampling**. Due to tissue spatial variation, this biological sample may not be representative of the original whole tissue (Figure 34, first step).

The biological sample can vary regarding its extra-cellular matrix (ECM) composition and cell adhesion. A pivotal step consists of **dissociating the sample** (Figure 34, second step). Dissociation involves depleting the ECM and breaking junctions between cells, using mechanical and/or enzymatic methods. This step is highly stressful for cells. Additionally, setting up the dissociation process involves identifying a good balance between excessive tissue degradation, which leads to cell damage, and insufficient dissociation, which leads to a waste of large tissue pieces. Consequently, dissociation can damage cells or be incomplete, thereby biasing the composition of the samples.

After dissociation, or if the cells suspension has been frozen, one should **remove debris** (Figure 34, third step). Debris include non-dissociated pieces of tissue, poor-quality cells, such as dead cells, and intracellular content released after cell membrane disruption. While large debris are removed through successive sieves of decreasing mesh sizes, small debris are often removed through FACS (fluorescence-activated cell sorting) sorting. FACS sorting requires to set thresholds for particle size and granularity, because cells are expected to have a certain size and to have a round shape. FACS sorting has limitations. Firstly, setting thresholds may remove cells with cell-type specific features. Secondly, FACS outputs cells in a tube and poor *landing* can lead to membrane disruption and RNA release. Finally, FACS sorting can be used to filter cells based on the presence of relevant markers. For instance, all immune cells (CD45+) can be filtered, by setting a threshold on the measured level of this marker.

Debris removal can also be performed using columns. For instance, when dissociating nerves, myelin sheath forms small lipid structures that can be mistaken as cells by the FACS. Myelin can be removed through filtering columns. In publications presenting scRNA-Seq data from sciatic nerves, MACS (magnetic activated cell sorting) column-based cell separators were used (Carr *et al.*, 2018; Toma *et al.*, 2020; Zhao *et al.*, 2022). Briefly, cell suspension pass through a column, which retains myelin. However, this method can retain myelinating cells, thus depleting them from the analysis. To overcome this bias, Gerber *et al.* (2021) used Smart-Seq protocols to generate their data, and Yim *et al.* (2022) used snRNA-Seq technologies to directly extract nuclei and have a better representativness of cell types.

3.2.2 Tracing the cellular origin of transcripts

In bulk RNA-Seq, all transcripts are extracted at a tissue level, and pooled together. To produce scRNA-Seq data, the main challenges was to determine the cellular origin of each transcript. Most techniques are based on the attribution of a cell-specific DNA barcode to each transcript. Such attribution requires to physically isolate single cells, and many approaches have been set up (Svensson *et al.*, 2018). While manual capture of cells enable the analysis of a limited number of cells, with integrated fluidic circuits, hundreds of cells can be isolated.

3.2.2.1 The 10X Genomics technology

Our laboratory chose the 10X Genomics technology to produce scRNA-Seq data. Reasons include its reasonable experimental duration and financial cost, the possibility to analyze a large amount of cells, and its ease of use at a laboratory scale.

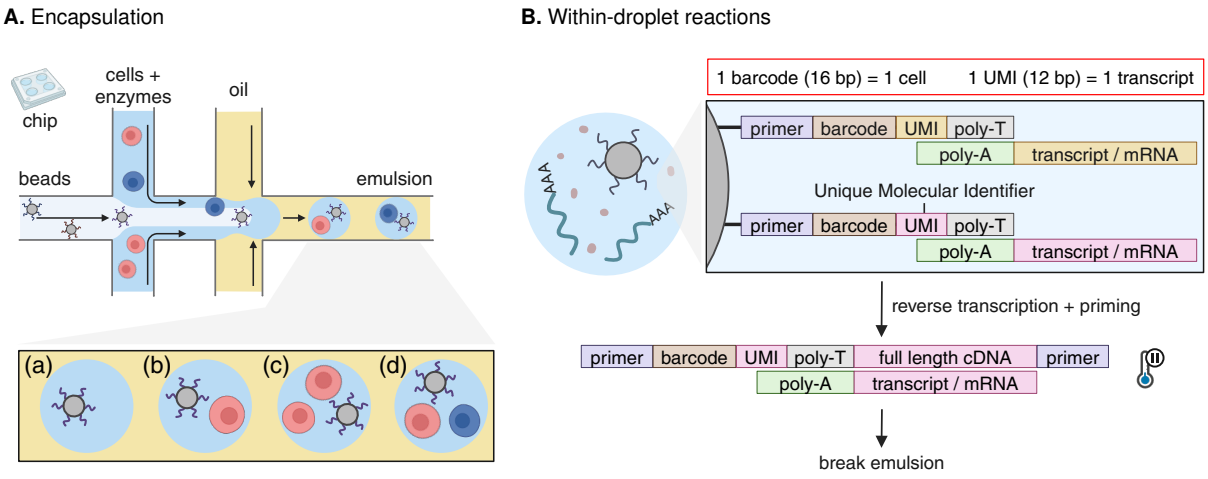


Figure 35 – Tracing the cellular origin of transcripts with barcoded droplets. **A.** Principle of the microfluidic system using by 10X Genomics to physically isolate single cells in an emulsion containing aqueous droplets. The aqueous phase is depicted in blue while oil is yellow. The possible composition of droplets is indicated in (a) to (d). Cells are represented in blue or red, depicted two distinct cell types. Beads are represented as gray shape. **B.** Illustration of the molecular reactions within each droplet. On top, zoom is made on the droplet, after lysis of the cell membrane. On a given bead, each capture sequence share the same cell barcode, but differ by a unique molecular identified (UMI). The thermometer indicates that the sequences can be stored at 4 ℃ after the reverse transcription and priming step, and before breaking the emulsion.

The 10X Genomics protocol starts from the single cell suspension (Figure 34). Thousands of cells, enzymes and millions of gel-composed beads are deposited onto a chip in a water-based fluid, along with an oil-based fluid, and loaded onto the

microfluidic platform (Figure 35A). Cells and enzymes on one side, and hydrogel beads on the other side, are conveyed up to the oil-based fluid, which, through its flow rate, cuts the water flow. This results in an emulsion, where each water bubble is called a "**droplet**". Due to the large amount of beads, majority of droplets is empty of cells (Figure 35Aa). The droplets of interest are those containing one cell and one bead (Figure 35Ab). In the cases of incomplete dissociation, highly adhesive cells or loading too many cells in the machine, it is possible that the droplet contains more than one cell. This configuration is called "**doublets**" (Figure 35Ac-d). Doublet cells can be filtered using computational methods, once sequencing is done (Xi & Li, 2021). The bead volume is approximately 80% of the droplet, so no bead doublets are observable.

Once the emulsion is ready, the cell membrane is lysed by the enzymes. RNA molecules are captured by the probe sequences on the bead (Figure 35B, top). With this technology, only 30% of transcripts are captured, because there are many more transcripts in each cell compared to the number of probes on the bead. Therefore, rare transcripts may be missed. In scRNA-Seq data, missing transcripts can be due to non-expression or capture issue, the latest being called "**drop-out events**".

RNA capture is based on a poly-T sequence, binding to large repeats of adenine in RNA molecule. Such repeats mostly concern the poly-A tail of mRNA (Box 5). Other techniques, including Smart-Seq, capture transcripts based on random sequences, avoiding to limit the analysis to only poly-adenylated molecules.

Finally, an amplification of the molecules is required to improve their detection (Figure 35B, bottom). With the enzymes already present in the droplet, RNA molecules undergo a reverse transcription, converting them into complementary DNA (cDNA), and primers are added to enable amplification. Since mRNA molecules are tagged according to the cell they originate from, the emulsion is broken and all sequences are pooled together. They are ready to prepare the sequencing library.

3.2.2.2 Differences with Drop-Seq and Smart-Seq

During my thesis, I retrieved publicly available data obtained with 10X Genomics protocols, but also with Drop-Seq and Smart-Seq data. Many articles compared the different protocols to generate scRNA-Seq data (Zhang *et al.*, 2019; Ding *et al.*, 2020). They showed that **Drop-Seq** is very similar to 10X Genomics. Cells are individually isolated in droplets, with beads carrying similar barcode and capture sequence. However, 10X Genomics outperforms Drop-Seq in terms of barcode sequence quality, RNA capture sensitivity and the presence of technical noise.

Box 5: RNA transcription and poly-adenylation

Figure 36 highlights the temporal organization of each processing step during RNA transcription. Despite initiation, elongation and termination being sequential events, RNA processing co-occurs with transcription elongation. Particularly, splicing starts right after elongation, once 5' capping is added. Furthermore, 3' poly-adenylation is a marker of transcription termination, but it is not required for splicing initiation. Thus, it is possible to capture unspliced transcripts in scRNA-Seq data.

Proteins are translated from mRNA molecules carrying a poly-A tail. However, a subset on non-coding RNA can also be poly-adenylated (Xu *et al.*, 2022). Indeed, scRNA-Seq data, based on poly-A capture of RNA, also gives access to some non-coding RNA.

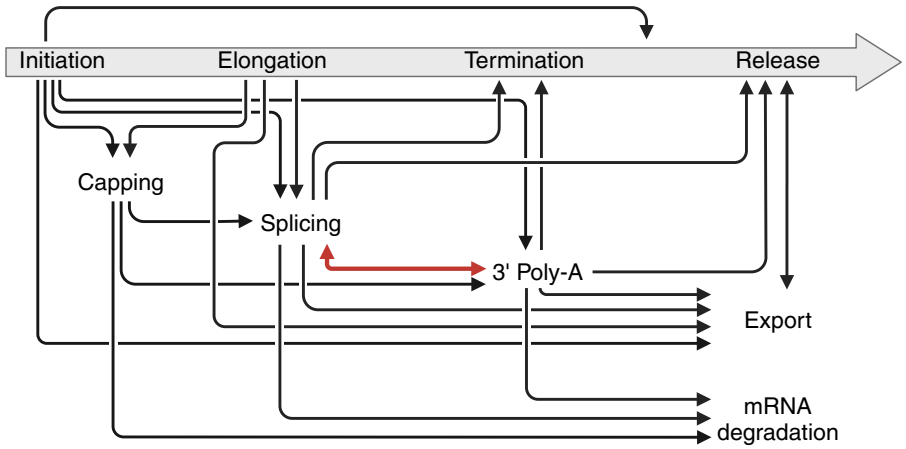


Figure 36 – **RNA transcription and processing**. The major steps of transcription are shown on the top arrow. The steps in RNA processing are arranged below. The arrows denote temporal organization of steps, and physical dependence. Adapted from Maniatis and Reed (2002).

Transcripts are barcoded at one of their extremities. To retrieve both the mRNA and the barcode sequences, either a long-read sequencing or a 3' or 5'-biased short-read sequencing should be considered. Furthermore, barcode-based techniques only capture a subset of cell's transcriptome, 30% in the case of the 10X Genomics technology. Lastly, due to the poly-T capture sequence, only poly-adenylated transcripts are captured. Overall, only specific transcripts are captured and full transcripts sequences are challenging to obtain.

To overcome these limitations, **Smart-Seq** protocols are based on the sorting of individual cells in well plates, and thus cell-specific sequencing library construction. This drastically increases the cost of scRNA-Seq production, but leads to bulk-comparable transcriptome coverage and full-length transcript sequencing.

This technology is used to answer distinct biological questions, compared to droplet-based methods. Firstly, since cells are individually sorted into wells, the number of cells studied is limited. Secondly, because RNA capture does not depend on a small number of probes on a bead, potentially all transcripts can be detected. Additionally, RNA can be captured with random probes, regardless of their poly-adenylation status, enabling the capture of more RNA families. Thus, this method is chosen to study a small group of cells in great detail, including the characterization of rare transcripts and the study of isoform usage.

3.2.3 Library preparation and sequencing with Illumina

The library preparation depends on the sequencing technique. In the laboratory, only the Illumina sequencing technique was used. Once the emulsion is broken, sequences are amplified to improve their detection (Figure 37A).

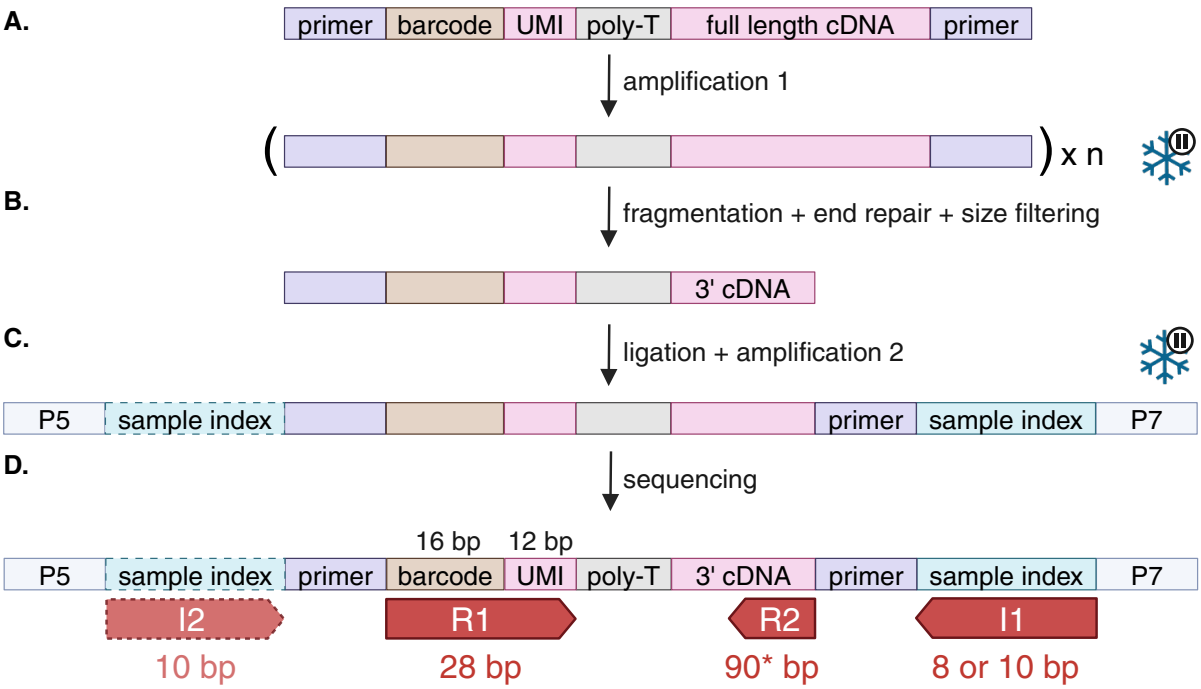


Figure 37 – Illumina-specific library preparation and sequencing. Snowflake indicates that the preparation can be frozen. **A.** Sequences amplification. **B.** Fragmentation and size filtering. **C.** Index and sequencing primers ligation. **D.** Sequencing of reads (R) and indices (I). Dashed box indicates optional sequencing. The size of sequences is positioned along the library. The index 1 (I1) is 8 or 10 bp depending on the protocol version. The star (*) denotes that 90 bp size is a recommendation. bp: base pairs. Adapted from *Single Cell 3' v3.1 Dual Index Gene Expression* (2023).

Illumina sequencer requires sequences sizing between 200 and 700 base pairs (bp). Thus, the library preparation involves a fragmentation step (Figure 37B). The full-length

cDNA is randomly cut and only the 3' region remains. Molecules are further filtered based on their size. As a consequence, very small RNA or splicing variants cannot be studied. Splicing variant analysis requires the sequencing of full-length cDNA, which is possible with Nanopore or PacBio techniques (De Paoli-Iseppi *et al.*, 2021).

Sample indices and sequencing primers (P5 and P7) are then flanked to the sequences (Figure 37C). The sample index is a fixed size DNA tag with known sequences for a given library. In the **single index** protocol, one sample index is a quartet of 8 bp sequences added on the P7 side of the molecule. In the **dual index** protocol, one sample index consists in a pair of two distinct 10 bp sequences added on both P5 and P7 sides. To decrease the sequencing costs, several sequencing libraries can be sequenced together. Since the sample index is sequenced and accessible in the **index 1** (I1) files, it can be used to demultiplex data and retrieve sample-specific signal. To demultiplex data more precisely, it is also possible to sequence the P5 side sample index, thus generating an **index 2** (I2) files, but increasing the sequencing cost.

Sequencing consists of reading each sequence at three locations, which leads the sequencer to build three types of matching files (Figure 37D, in red). I1 and I2 files have just been described. **Reads 1** (R1) file corresponds to the barcode-UMI sequences. Since their length is fixed and known, the barcode and UMI can be further separated. **Reads 2** (R2) file contains the biological signal and corresponds to the cDNA sequence on the most 5' side. Illumina sequencer builds Illumina-specific formatted data (BCL), stored in a folder, which is the input format for computational analyses (Section 3.3). I1, R1 and R2 files can be retrieved this folder. By abuse of language, this folder is called "**BCL folder**" because it includes *binary base call* (BCL) files.

3.3 Computational processing of scRNA-Seq

The computational processing of scRNA-Seq can be divided into three main steps (Figure 38). It starts with the pre-processing, which transforms the raw data into a cells-by-genes count matrix (Section 3.3.1). The processing filters the count matrix (Section 3.3.2.1). The dataset is then ready for downstream analyses (Section 3.3.2.2). Processing and analyses are supported by many visualization steps (Section 3.3.2.3).

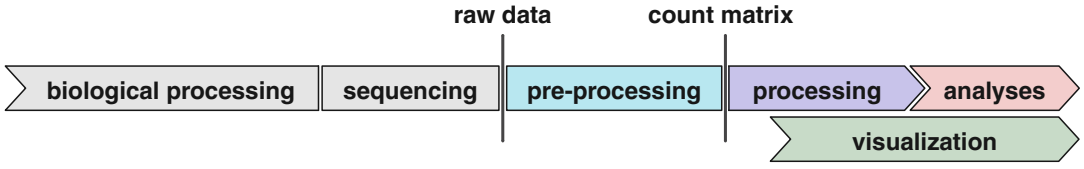


Figure 38 – Overview of scRNA-Seq data processing. Experimental procedures are in gray while computation processing is depicted as colored boxes.

3.3.1 From the reads to a count matrix

The pipeline to obtain the scRNA-Seq count matrix from the sequencing data is standardized (Figure 39). First, index (I1 and I2) files are used for demultiplexing. Reads 2 (R2) files, containing the biological information, are mapped on a reference transcriptome. Reference transcriptome can be homemade or retrieved from public repositories online. The output BAM files can be used for further visualizations. Lastly, reads 1 (R1) files are used to build the cells-by-genes count matrix.

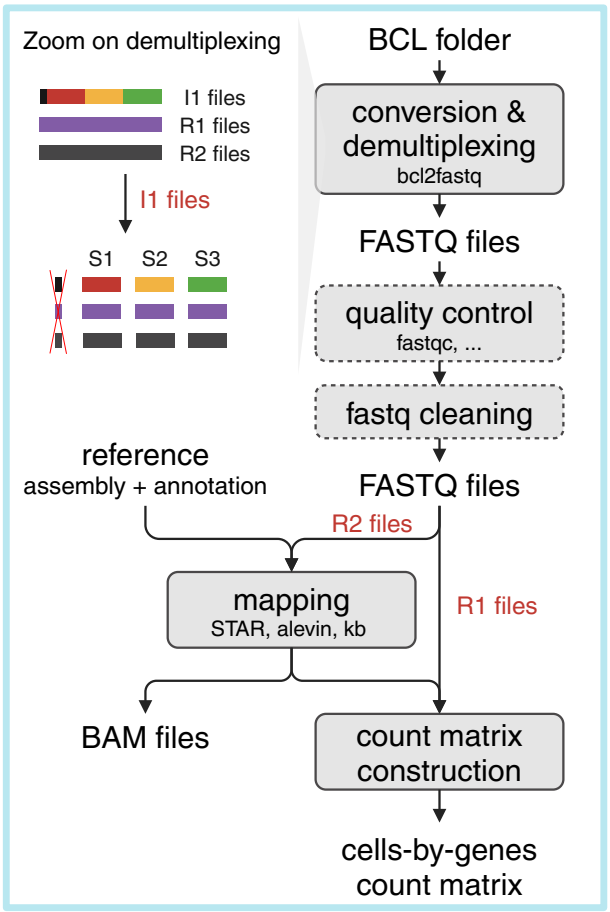


Figure 39 – **Main processing steps for reads to count matrix**. This figure details the pre-processing step shown on Figure 38. Boxes delineate a transformation of the data while arrows denote their temporal arrangements. Each element is detailed in next sections.

3.3.1.1 Conversion, demultiplexing and quality control

First, the BCL folder is converted to FASTQ standard format using the bcl2fastq tool proposed by Illumina. It builds FASTQ files for each R1, R2 and I1 group of reads (Figure 37, left). Based on the I1 file, reads are then split by sample. Each

sample corresponds to a given index (Figure 37, colored boxes). This step is called demultiplexing. Unassigned indices and their matching reads are discarded.

The **quality control** step further summarizes the quality metrics of all sequences. It can be followed by an optional filtering step, which consists of removing low-quality sequences.

3.3.1.2 Retrieving the transcriptomic origin of transcripts

The next step is the alignment, or **mapping, of R2 FASTQ files**, onto a reference transcriptome. It allows to retrieve the genomic origin of reads, and is necessary to attribute reads to genes. Mapping requires a **reference**. With scRNA-Seq data from mouse, the reference corresponds to the whole sequence of the mouse transcriptome assembly and the sequence annotation. In my thesis, I analyzed scRNA-Seq data generated from *Homo sapiens* and *Mus musculus*. While the first genomes were published 20 years ago[3], both assembly and annotation are updated regularly. This leads to changes in gene names and start and stop positions. Fortunately, genes are characterized by a unique and definitive identifier, making it possible to trace genes across transcriptome (assembly + annotation) versions.

Multiple strategies exist to map reads onto a reference. Because reads from RNA transcripts can straddle two exons that are not consecutive in the genome, splice-aware aligners have been developed and include STAR (Dobin *et al.*, 2013) and hisat2 (Kim *et al.*, 2019). Both aligners output a BAM (*Binary Alignment Map*) file, containing the genomic coordinates of the reference sequence on which each R2 read most probably originates from. Alternatively, kallisto (Bray *et al.*, 2016) and salmon (Patro *et al.*, 2017) are "pseudo-aligners", which do not directly map reads onto the reference, but go through an intermediary step. They are used to generate the count matrix in a more resource-efficient way compared to splice-aware aligners, as highlighted by Brüning *et al.* (2022), but do not generate BAM files. All four tools cited above are free of rights. Their source code is accessible and many parameters can be adjusted.

10X Genomics has developed its own private solution, called Cell Ranger, to wrap existing tools in a user-friendly manner (Zheng *et al.*, 2017). Few parameters are adjustable and the source code is private. Since 10X Genomics data are often sequenced by Illumina sequencers, Cell Ranger includes bcl2fastq to convert BCL file to FASTQ files. Mapping is performed with a wrapped STAR tool. Cell Ranger performs similarly to STAR, but in a slower way (Brüning *et al.*, 2022).

[3]in July 2003 for *Homo sapiens*, called hg16, and March 2005 *Mus musculus*, called mm6.

Some reads are not successfully mapped to the reference. First, bad quality reads, due to uncertain sequencing, lead poorly confident alignments. Second, when reads come from a transcript that is not in the reference, there is no mapping. It can occur when those reads are issued from a biological contamination, or when the reference does not contain any information about this transcript. The latter can occur for non-model organisms, for which the construction of complete consensus reference is still under investigation (Lehmann, 2021). Finally, small reads may map to several locations on the reference and are then classified as ambiguous.

3.3.1.3 Construction of the UMI count matrix

The last step is the **construction of the count matrix** (Figure 40). R2 reads that have been mapped onto the transcriptome are associated with genic or intergenic regions based on the reference annotation. Only the R1-R2-I1 triplet of reads with **R2 mapping in genic regions** are kept to build the count matrix (Figure 40, step 1). Before Cell Ranger version 7, only reads that mapped to exonic regions were taken into account for downstream processing. From version 7, both exonic and intronic mapping reads are used. These reads are associated with the gene they belong to. Such genes define the **columns** of the count matrix.

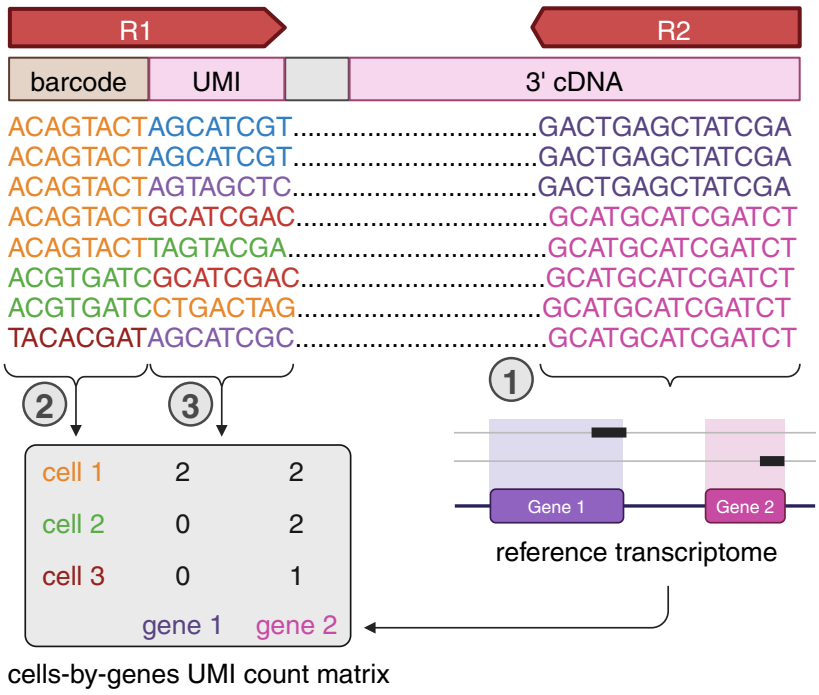


Figure 40 – **Build the count matrix from reads**. R1 and R2 reads are illustrated on top of the sequence they represent. Steps are depicted by circled numbers: mapping (1), identifying cells (2) and counting UMI (3). UMI: unique molecular identifier. Inspired from Lafzi *et al.* (2018).

Next, Cell Ranger looks at the R1 reads and splits cell barcodes from UMIs. Barcodes are compared to the whitelist. When possible, sequencing errors are corrected in the barcode sequence, in line with the whitelist. The **correct barcodes** are used to define the **rows** of the count matrix (Figure 40, step 2). Some R1-R2-I1 reads are discarded due to ambiguous cell barcodes.

Finally, **UMIs** are counted, for each cell barcode and gene, to **fill the count matrix** (Figure 40, step 3). In the example in the figure, cell 1 has three reads for gene 1. Two of them correspond to amplicons of the same transcript because they share the same UMI. The third one corresponds to another transcript, delineated by a specific UMI.

3.3.2 The count matrix as a wealth of information

This section draws the main focus of my thesis: the computational analyses of scRNA-Seq data. Two processing languages are mainly used: Python and R. Many modules (Python) or packages (R) have been developed to process data, each targeting a specific aspect of the data. A need was to make all tools compatible together, through data format standardization. Thus, the scanpy module and the AnnData format were set up in Python language (Wolf *et al.*, 2018), while the Seurat package and the Seurat object format are used in R language (Stuart *et al.*, 2019).

3.3.2.1 Standardized processing of the count matrix

Processing the count matrix up to downstream analysis is standardized (Luecken and Theis (2019), Figure 41). First, the **quality control** consists in evaluating cells or genes, to decide for their **filtering**. Firstly, cells with a low number of **detected genes** (column) or **transcripts** (UMI) will not be of interest for downstream analysis and can impair the processing by adding noise in the data. In fact, cells with a very low number of transcripts can be empty droplets, where ambient RNA has been captured by probes (Figure 35Aa). Secondly, because they do not contain enough diversity of RNA, cells expressing too much **ribosomal RNA** or genes from the **mitochondrial genome** are considered of poor quality. Such a profile can result from technical issues, including capture or amplification bias, or biological stress. Thirdly, several cells may share the same cell barcode, due to encapsulation within the same droplets (Figure 35Ac-d). Such **doublet cells** can impair downstream analyses. Homotypic doublet cells, two similar cells, will present inconsistently high expression of genes. Heterotypic doublet cells, two cells from two distinct cell types, will show unusual gene co-expression. The principle of the **doublet removal** tools is to create numerous artificial doublets by combining the expression profile of two cells, and to calculate the distance between

gene expression from each cell and each artificial doublet (Xi & Li, 2021). Cells showing significant proximity with such doublets will be marked as doublet cells. Finally, scRNA-Seq data are highly sparse. The count matrix contains more than 90% of zeros due to the so-called **drop-out events** (Section 3.2.2). To decrease computer resource usage, at the gene level, it is common to remove genes detected in too few cells.

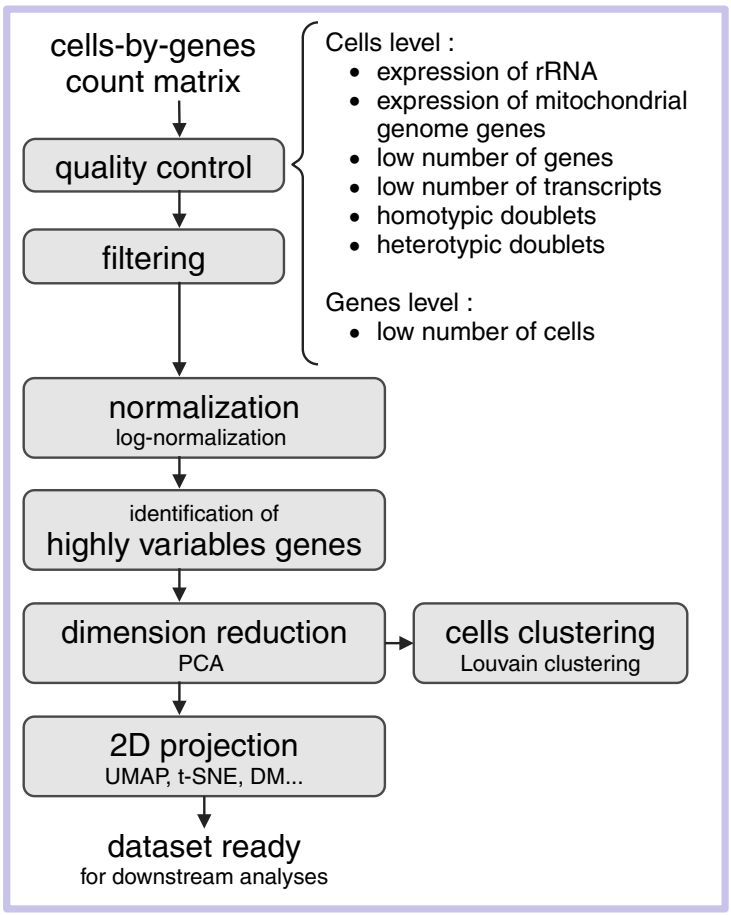


Figure 41 – **Standard pipeline to prepare the count matrix for analysis**. This figure details the processing step shown on Figure 38. Boxes delineate a transformation of the data while arrows denote their temporal arrangements. rRNA: RNA encoding ribosome subunits.

Once the definitive count matrix dimension is set, UMI count should be **normalized** to compare gene expression between cells. Many methods exist and have been compared by Ahlmann-Eltze and Huber (2023). Sophisticated methods are not always the best performing, and the authors showed that the logarithmic transformation is among the best, in terms of computational cost and biological interpretation.

In order to visualize cells in two dimensions and for some downstream analyses, the next steps involve a dimension reduction. The challenge of 2D visualization is to arrange high-dimensional data such that it depicts the sample cellular heterogeneity.

Indeed, the representation should reflect a distance, in terms of gene expression proximity, between all cells. Thus, the first step in dimension reduction consists in selecting the most "**highly variably** expressed **genes**" (HVG). Such genes are identified by computing a measure of dispersion of expression across all cells. Indeed, a gene expressed at the same intensity in all cells is not of interest to distinguish cells.

Once 1,000 to 5,000 HVGs are identified, their expression is scaled to ensure that they all vary equally and will be treated with equal importance. Then, all HVGs are pairwise compared. From this comparison, 50 to 100 new components are defined, by associating several HVGs together and attributing them weights. Thus, such components are linear combinations of genes expressions. Each component is also associated with a single value, depicting to what extent the component represents the variability between cells. These values enable raking the components by decreasing importance, and allow for the selection of the principal components that capture the most variability in the data. This process is known as **principal component analysis** (PCA). The number of principal components selected is often determined by the amount of variability one aims to preserve ; the remaining variability is considered as noise or not informative enough.

PCA is currently the most used method to reduce the dimensionality of scRNA-Seq data, as it preserves well the biological information (Sun *et al.*, 2019). PCA reduces the data to few dozens of dimensions. To then project cells in a two-dimensional space, two main methods are mainly applied: Uniform Manifold Approximation and Projection (**UMAP**) and t-distributed stochastic neighbor embedding (**tSNE**) (Section 3.3.2.3).

Besides using PCA to project cells into two dimensions, principal components can be used to compute distances between cells, and to group cells in clusters with similar transcriptomic profiles. Multiple **clustering** methods exist for high-dimensional data. Yu *et al.* (2022) compared 14 implementations of clustering approaches and concludes that Louvain clustering implemented in Seurat package is ranked at the fourth position in terms of speed, memory, and its ability to retrieve a consistent number of cell types.

Once the number of cells is defined and visualization is available, data is ready for downstream analyses. Based on the output of each processing step, the understanding of data may change, leading to going back to earlier steps of the process to modify tools or parameters. Analyzing scRNA-Seq data is not a straightforward process.

3.3.2.2 Analysing scRNA-Seq data

Based on the biological questions, scRNA-Seq data can be explored in multiple directions (Figure 42).

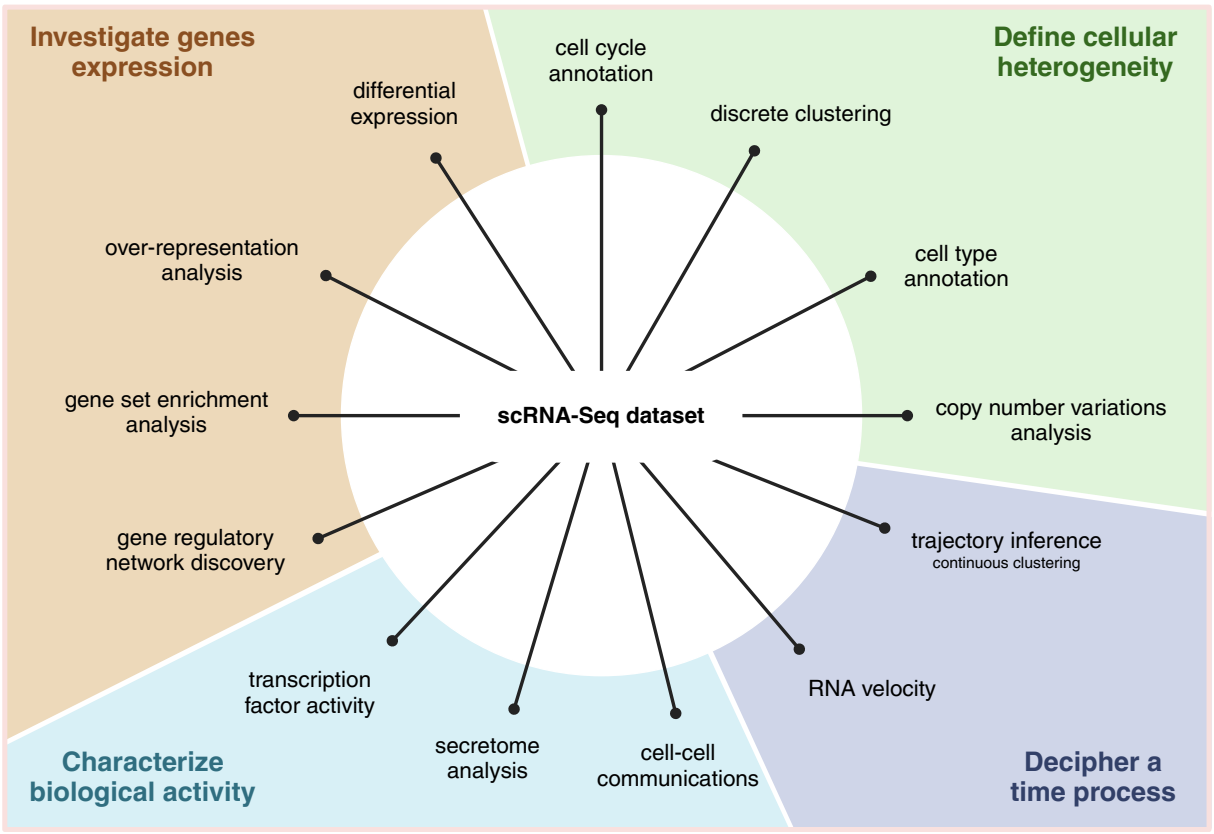


Figure 42 – **Multiple directions make sense of scRNA-Seq data**. This figure details the analysis step shown on Figure 38. Based on the biological questions, scRNA-Seq data can be explored in multiple directions. Analysis types are arranged in four main categories, despite some types may be shared by several categories.

Since scRNA-Seq enables to investigate the cellular heterogeneity underlying biological samples (Figure 42, green corner), an important step is the **cell-type annotation**. Several datasets can be merged to build biological tissue **atlas**, meaning a map of all the possible cellular transcriptomic profiles making a tissue. The 10X Genomics technology is particularly suited to build atlases (Mereu *et al.*, 2020).

The most common way to investigate **gene expression** (Figure 42, brown corner) is to perform differential expression analysis (DEA) between groups of cells, for instance defined by discrete clustering (Soneson & Robinson, 2018; Nguyen *et al.*, 2023). Co-expression analysis methods can also be considered (Chowdhury *et al.*, 2020). SCENIC package can be used to discover gene regulatory networks (GRN) among

data (Van de Sande *et al.*, 2020). While SCENIC infer new GRN, one may want to use existing and peer-reviewed GRN database to characterize the biological activity of its cells (Figure 42, blue corner). To take advantage of transcription factors / targets or pathways databases, several mathematical models can be applied. Badia-i-Mompel *et al.* (2022) published the decoupleR package, which make these models interoperable. To investigate cell-cell communications, many methods have been developed (Armingol *et al.*, 2021).

Biological activity can be considered as a **time process** (Figure 42, purple corner). Two kinds of approaches can be considered. Based on the expected trajectory in the data, a plethora of methods have been published (Saelens *et al.*, 2019). Authors made a considerable effort to make more than 50 trajectory inference methods interoperable.

Analyzing data often consists of **comparing** several approaches, or **combining** them. For instance, trajectory inference can be applied on a single cell population to identify gene expression variations across a progression. This cell population is accompanied by micro-environmental cells in the dataset. Cell-cell communications can be inferred to decipher cross-talk between the population of interest and other cells. Combining trajectory and cell-cell communication to investigate the contribution of the micro-environment to the progression of a population of interest is the purpose of the ENTRAIN package (Kyaw *et al.*, 2023).

While studying multiple conditions, for instance several tumor types, it is necessary to merge several scRNA-Seq datasets. This combination is known a **integration**. However, each scRNA-Seq dataset is biased, which is called batch-effect (Section 3.2). To promote the biological information in the data, integration should **correct batch-effect**. Once again, a plethora of methods have been proposed and regularly compared (Tran *et al.*, 2020; Luecken *et al.*, 2022). Indeed, some methods perform best on large or small, homogeneous or heterogeneous datasets. They generate either a corrected count matrix or a dimensionality reduced space, such as PCA. For instance, the method implemented in harmony package takes as input a PCA space, generated from the combined dataset, and correct, "harmonize", the PCA to remove the individual effect of each dataset (Korsunsky *et al.*, 2019).

This is an overview of different kinds of analysis for scRNA-Seq data, but such data can take advantage of other -omics data. Combining data from various methods is known as **multimodal integration** (Argelaguet *et al.*, 2021; Heumos *et al.*, 2023).

3.3.2.3 Visualizing scRNA-Seq data

Each analysis is accompanied by its own visual representations, such as distributions for quality control step, or heatmap for differential expression analysis. This section addresses the specificity of the 2D representation of single cells. Because each projection displays specific aspects of the data, Chari and Pachter (2023) classified the visualization approaches into three categories, based on three aspects one wants to visualize in the data: their global structure, their local structure, or the distances between cells. One should not be attached to a single representation (Figure 43).

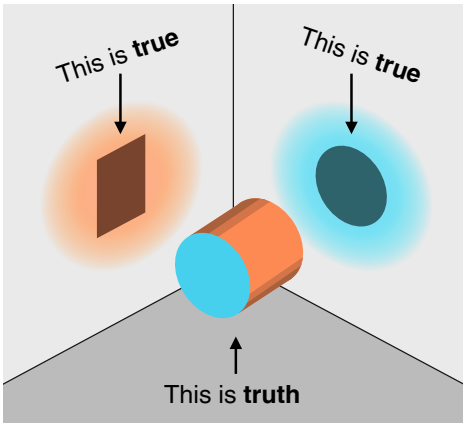


Figure 43 – Representations are a glimpse of high-dimensional data. While visualizing cells from scRNA-Seq data on two dimensions, one should not be be attached to the precise arrangements of cells within the plan. Indeed, 2D representations summarize thus simplify the data. Two representations of the same input data may highly differ, however none of them is false. Adapted from `https://9gag.com/gag/a4YnVGp` accessed 09/03/2024

scRNA-Seq data are commonly represented through UMAP or tSNE. The UMAP algorithm was established in 2018 to overcome the high running time and variability of tSNE (McInnes *et al.*, 2018). Despite Becht *et al.* (2019) showed that UMAP exhibit better performance than tSNE, both methods are recurrently found in the literature.

While UMAP preserves the **global distance** between cells, tSNE focuses more on **local heterogeneity**. Indeed on a UMAP, the empty space between cells is meaningful, while on tSNE, one should pay attention to the local arrangement of cells. Thereby, on a UMAP, all immune cells will be arranged in the same area of the projection, with dendritic cells closer to macrophages and NK cells closer to T cells, because they share common markers. On a tSNE, there is no guarantee for NK cells to be close to T cells because, despite sharing some features, many genes are specific to each cell type. Furthermore, tSNE will enable to visualize distinguish the different subpopulations within T cells, compared to UMAP.

Other visualization methods are applied to scRNA-Seq data, including diffusion map (Coifman *et al.*, 2005). Diffusion map is used to arrange cells by **gradients of gene expression** and is applied to trajectory-related questions (Angerer *et al.*, 2016).

3.4 Outlook and feedback

In this chapter, I presented the complexity of sequencing data, particularly focusing on scRNA-Seq data. Many biases are introduced in the data, during the experimental procedure. The sequencing data comes from biological tissue, and corresponds to a snapshot of a living tissue. Cells for which data is available ("sequenced cells") are not necessarily representative of the biological tissue. Sequenced transcripts correspond to a subset of what was available in the cell, due to RNA capture efficiency, library preparation and sequencing depth. Sequenced transcripts do not always reflect the RNA produced in the tissue, due to the cell dissociation method, which can stress cells and make them express stress-related genes.

Therefore, scRNA-Seq data that have not been generated by the same people, using the same protocol from sample preparation to sequencing, should be compared sparingly, considering their inherent differences.

Once sequenced, scRNA-Seq data requires multiple consecutive processing steps. Furthermore, there are many directions that could be taken when exploring the data. To analyze the data, many tools have been developed to answer similar initial biological questions. Therefore, a question remains: How to choose a method ? Here are some criteria that might help in making a decision:

- The package in which the method is implemented is easily installable.
- The package is easy to use, there is a good documentation to use it and their is an active community, on forum, using it.
- The method is explainable, i.e. customizing the parameters is intuitive, and variations in the results can be explained by the parameter setting.
- The tool has good default setting, meaning that results are consistent without modifying the default parameter values.
- The method runs in a resource-efficient manner. This is important since scRNA-Seq analysis yields many results, and visual interpretation is a pivotal step in validating the method. It is preferable for the method to run on a local rather than a remote computing cluster where graphical visualization is limited.
- The method is trusted by the community and/or is performing, compared to other similar tools, within benchmark articles. However, benchmark of tools published by authors who are at the originators of one of the tools should be considered with great care, because such a benchmark is not always objective.

4 Reproducibility challenges in computational biology

The reproducibility has many definitions based on its field of application. At a global scale, the reproducibility is the **ability to produce again**, consistently, something, for instance data, methods or results. The purpose of this chapter is to motivate the need for reproducible research in science (Section 4.1) and to highlight the current context of reproducibility crisis (Section 4.2). To overcome the crisis, various solutions exist, especially in the field of computational biology (Section 4.3).

When not explicitly supported by scientific articles, the content of this section is based on the 2023 French Reproducible Research Day[4], the course "Ethics, good scientific conduct and plagiarism" offered to Master students in 2023 at École Normale Supérieure in Paris, the 52nd ORAP forum about reproducibility in the context of high-performance computing[5], and personal communications.

4.1 Why making reproducible research ?

Scientific research is a team effort, at both the laboratory level and the global scale. In biology, research aims to expand our understanding of living organisms and answer fundamental questions. This has many implications, including progress in addressing public health problems and advancing medicine. As a worldwide team, basic biological research can be thought of as a jigsaw puzzle with millions of pieces. At the laboratory level, each team is responsible for a single small piece. Scientists build their research on their piece of the puzzle based on the neighboring pieces published in the literature. What happens if the existing pieces are not reliable ? There is no puzzle model to check if a piece is correctly placed. The **reliability** of existing knowledge is a key to build on it, be more efficient, re-use what have been suggested, combined different projects and avoiding expensive repetitions.

Furthermore, biological research has a direct impact on non-scientists because of its health implications. The public should have confidence and trust in science, and "any corruption of the scientific process affects the perceived credibility of important contributions to knowledge."[6]. When the **credibility** of science is undermined, the public may be skeptical, including about vaccine discovery and global change (Schulz *et al.*, 2016; Casadevall *et al.*, 2016). **Transparency** of the research process is crucial to make research results credible.

[4]French Reproducible Research Network: `https://www.recherche-reproductible.fr/`
[5]52nd ORAP forum: `http://orap.irisa.fr/52ieme-forum-reproductibilite/`
[6]From `https://www3.weforum.org/docs/WEF_Code_of_Ethics.pdf`, accessed in February 2024

Since research is a self-regulated endeavor, individuals should establish a moral and social framework, an ethic, for conducting reliable research. In the context of research quality, making the research reproducible has been proposed (Leonelli, 2018). Reproducibility has different meanings depending on scientific practice. At the various reproducibility events I have attended, each speaker has started his presentation with his own terminology, albeit slightly different.

Computational biology is the science of analyzing biological data using computers. In the scRNA-Seq context, the plethora of analysis pipelines, both due to many successive processing steps and existing tools, leads a single dataset to output a myriad of results (Vieth *et al.*, 2019). As pipelines can produce different results with distinct biological implications, it is essential for computational biologists to properly manage their pipelines. In the field of computational biology, I thereby define reproducibility as the **ability to produce again exactly the same results from the same input data and the same methods on a new machine**, which is compatible with the taxonomy from Heroux *et al.* (2018).

4.2 Reproducibility crisis

In 2016, 90% of the 1,500 scientists who responded to a survey agreed that there is a reproducibility crisis in all fields of research (Baker, 2016). The survey found that more than 60% of respondents from the fields of chemistry and biology had experienced a failure to reproduce someone else's results, as well as their own. This failure to reproduce published results brought to a crisis of confidence.

Several studies report concrete examples of reproducibility failures. For example, Samuel and Mietchen (2024) deal with the reproducibility of analyses performed by 20,000 Python computational notebooks. The authors were able to install tools for half of them, but successfully ran only 1,200 of them without errors, of which 900 gave identical results to those initially reported.

4.2.1 Global reasons for the reproducibility crisis

Many factors contribute to the irreproducible research (Figure 44). Due to competition, selective reporting, and the fact that obtaining grants to continue research is based on publications, the scientific integrity might be altered (Edwards & Roy, 2017). Indeed, the **pressure to publish** comes from funding agencies that reward laboratories for publishing frequently and in prestigious journals. The **selective reporting** concerns the publication of novel, positive and original results, and can negatively influence the research process toward false discovery.

False discoveries can result from a **poor design**. For example, a low number of replicates leads to the application of statistical methods to data that do not satisfy the statistical hypothesis. It can conduct to an over-interpretation of the results (Button *et al.*, 2013). Poor design can also be the lack of positive and negative controls, leading to non-robust results or false positives. Such false positive results from intrinsic variability in data collection or analysis. The result may not be statistically significant when the number of replicates is increased. Finally, poor design can result from competition and pressure, but also from lack of time or money to conduct a study properly.

HARKing, which stands for hypothesizing after results are known, was introduced in the field of social science (Kerr, 1998). HARKing is the failure to distinguish between hypothesis testing and hypothesis generating research, the latter being exploratory research. It can lead to the misuse of statistical tests, also known as **P-hacking** (Gelman & Loken, 2019). P-hacking also consists of deliberately using the statistical test that results in the lowest p-value.

Lastly, once results are published, the lack of reproducibility can come from the **unavailability of data**. Reasons for not sharing data include competition and fear of sharing a source of new results, but also fraud behind the production of data. Similarly, **method unavailability** can result from poor traceability of methods within the laboratory, leading to inability to explain how the data were collected.

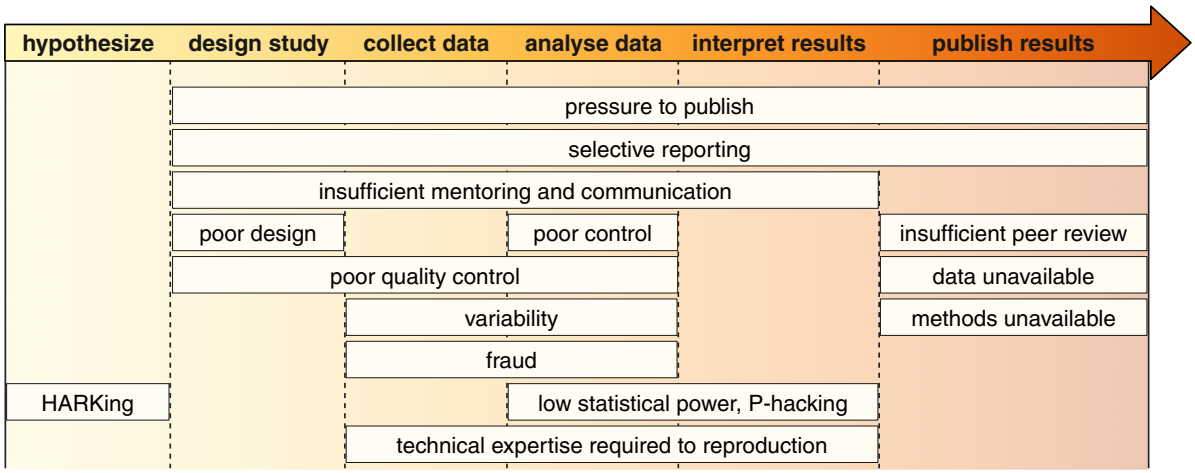


Figure 44 – **Factors contributing to irreproducible research**. A timeline of the research process is schematized. Elements contributing to irreproducible research are indicated below each step concerned. The figure is a graphical interpretation of Baker (2016).

All of these factors can also result from a lack of training (Mandl *et al.*, 2024) or mentoring, **communication**, and knowledge, because they require technical expertise.

4.2.2 The case of computational biology

This section highlights the importance of reproducibility issues in computational biology. Firstly, computational biology suffers from some stereotypes. Data are stored numerically, so there should be no variability in the results they produce. The code is there, so the analysis should work. In fact, the factors that hinder reproducibility fall into two categories, relative to the **data** and to the **methods**.

Since a few decades, journals require that raw **data** be provided for each publication. However, this requirement is not always fulfilled. Moreover, due to the plethora of ways to perform analyses, having the raw data does not mean that the results will be reproducible.

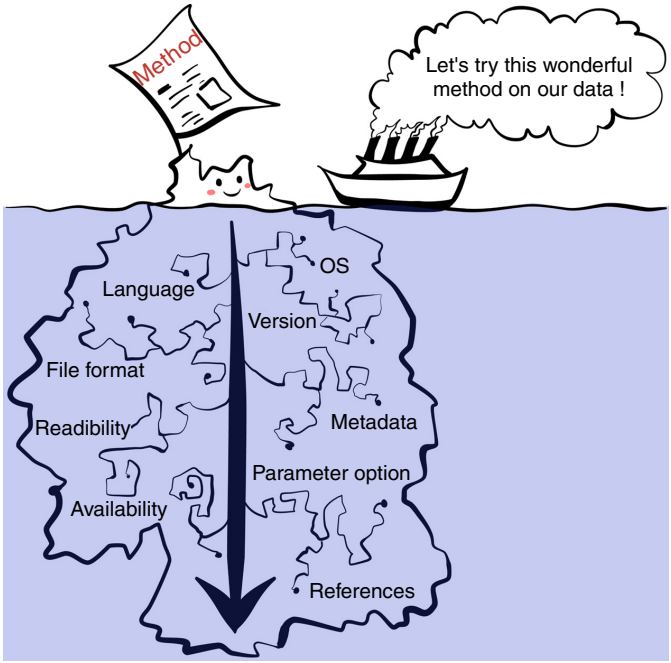


Figure 45 – **Reproducibility issues in computational biology**. The availability of the code or the method does not guarantee its usability. Many requirements and configuration are necessary to reuse a code. Adapted from Kim *et al.* (2018).

The Materials and **Methods** section of any publication is largely inadequate for reproducing results because of the hidden underlying settings (Figure 45, Simoneau *et al.* (2021)). Its purpose is to understand how the data were processed and how the results were obtained, rather than to be able to repeat the analysis. Additionally, unintentional and inconsistent data handling can lead to irreproducible results, because mistakes are untraceable if the code is not present. Therefore, it is necessary for authors to provide the full code along with their manuscript, which is currently not required in all journals.

Processing sequencing data involves many sequential steps and a trial-and-error method can be followed to perform the analysis. A poor traceability during the research process will lead to the absence of the full code and thus to irreproducible results. Although this is rare in the computational literature, the lack of code is very frequent in the field of biology. For example, the following statements in the "Data and Code Availability" section can be found in recent articles:

- "This study did not generate any unique code", in a publication with original ATAC-Seq, ChIP-Seq, RNA-Seq, CUT&RUN and scRNA-Seq data ;
- "No special code was used in this study, and code for all figures in the study are available for research purposes from the corresponding authors on request.", in a publication with original methylation, Exome-Seq, RNA-Seq and snRNA-Seq data.

Finally, the full code can be mentioned as *available upon request*, but this is not a sustainable way to the traceability, nor to ensure the reproducibility of analyses and results. Even if the code is available and accessible, it may be impossible to reuse because of insufficient documentation of the code or the methodology, or inability to install some dependencies (Samuel & Mietchen, 2024).

4.3 Promote reproducible research

To address the questions raised above, the section is divided into the global solutions (Section 4.3.1) and specific solutions for computational biology (Section 4.3.2).

As of today, nothing is mandatory and reproducible research consists in recommendations raised at different levels, from research laboratories, to research organisms and stakeholders, including funding agencies and journals. The progress toward reproducible science has to be progressive for people to adapt their practices, until it becomes a standard way of working.

4.3.1 Global solutions to overcome the crisis

In the same survey that raised the issues of reproducibility, Baker (2016) proposes solutions to promote reproducible research (Figure 46).

At the institutional and funding levels, **rewarding** open science processes and considering reproducible results and replication studies can foster for reproducible research, and make incentives for making reproducible research as a method and not as a result. For example, the pharmaceutical company Merck & Co. requires full or partial reimbursement if research is not conducted according to their good practice

guidelines. Rewarding open science project can also consist in stamping articles with **badges**. While submitting articles to conference or journals, it is possible to ask for evaluation of the computational reproducibility of the study. Examples include the Super Computing conference[7] and the Computer Aided Geometric Design journal.

Among other solutions, **recording method** aims to overcome HARKing. The preregistration of study before data collection begins ensure that the methods will not be modify after the results are known. **Adding control data** includes separating data into exploration and replication sets, or adding of positive and negative controls. Results can also be validated by independent approaches. Finally, **sharing** data and methods is promoted by the Open Science movement to increase confidence in the published results (Tay, 2024). Collaboration, communication and training can also be considered to promote reproducible research (Kohrs *et al.*, 2023). In bioinformatics, reprohackathons can be considered (Cokelaer *et al.*, 2023). In France, courses have been developed to train people to apply good practices in bioinformatics. These include a MOOC (Massive Open Online Course) entitled "Reproducible Research: Methodological principles for transparent science"[8].

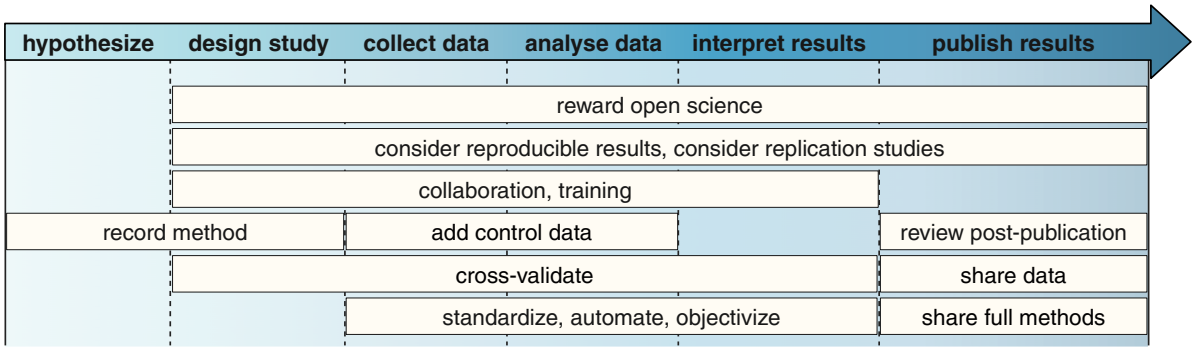


Figure 46 – Solutions promoting reproducible research. A timeline of the research process is schematized. Similarly to Figure 44, main solutions promoting reproducible research are indicated below each step concerned. The figure is a graphical interpretation of Baker (2016) and a presentation done by Dorothy V. M. Bishop entitled "What is the reproducibility crisis, and what can be done about it?" at the Incredible Research Conference 2020.

In the current context of reproducibility crisis, Nature journal encourages scientists to trust what have been published, but to verify anyway ("Trust but verify" 2024). However, such **replication studies** can fail to evidence the published results (Brodeur *et al.*, 2023). Therefore, authors highlighted that replicators should be encouraged and protected, or authorized to stay anonymous.

[7]Super Computing conference 2023: `https://sc23.supercomputing.org/program/papers/reproducibility-appendices-badges/`

[8]The MOOC is accessible at `https://www.fun-mooc.fr/en/courses/reproducible-research-methodological-principles-transparent-scie/`

4.3.2 Specific solutions for computational biology

In computational biology, the reproducibility rests on five pillars: the **data** used for the analysis, the **code** used to obtain results from the data, the **version** associated with the code, the computational **environment** in which the code runs, and the **documentation** that comprehensively describes all facets of the analysis, ensuring its reusability (Ziemann *et al.* (2023), Figure 47). These building blocks enable to assess the repeatability of the process and ensure the stability of the results. Then, exporting the analysis pipeline to a new machine make it reproducible, and the results portable. Finally, by varying the data (replicability) or the methods (robustness), the results become reliable.

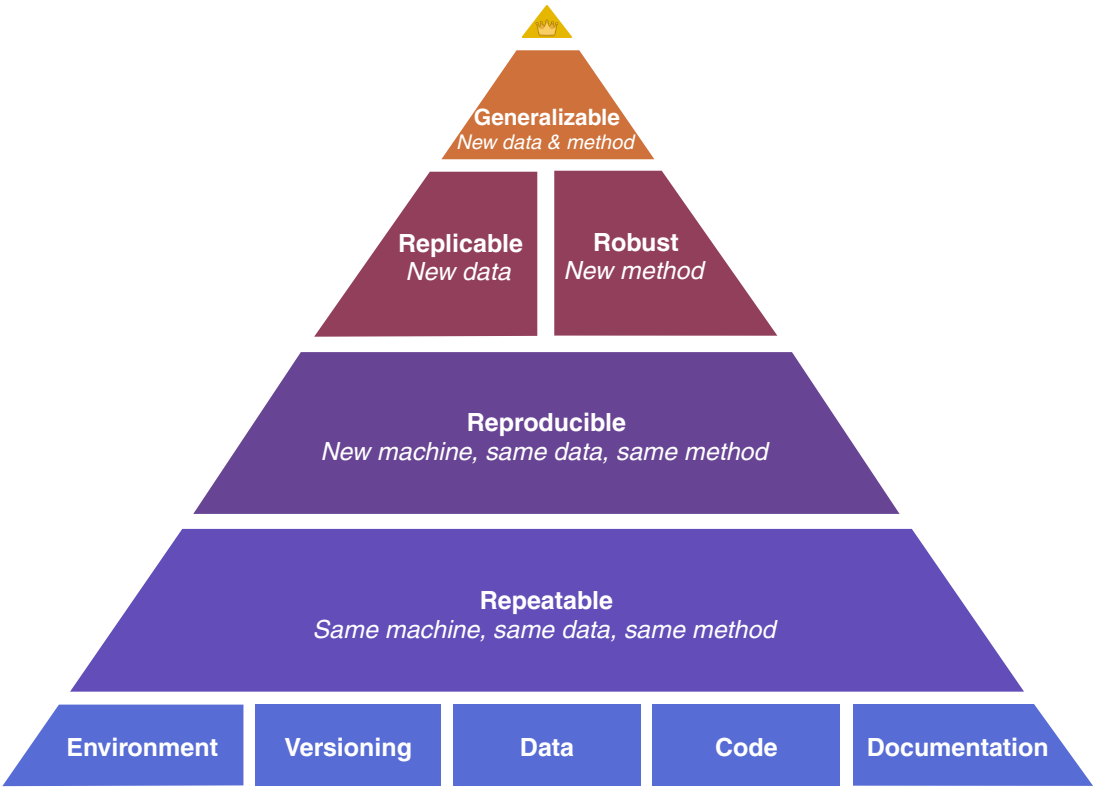


Figure 47 – **Reproducible computational biology relies on five pillars**. The five pillars are indicated at the bottom. They define solid basis to build up, from repeatability to reproducibility and finally to reach generalized statements. Inspired by Ziemann *et al.* (2023), Keefe *et al.* (2023), and the online support of a course taught in 2023 by the French Institute of Bioinformatics, called "FAIR-bioinfo"[9].

Existing tools for managing these five pillars are the subject of the next sections. Section 4.3.2.1 concerns the traceability of analyses, which is code-dependent. Traceable analyses are repeatable and portable if the environment in which they run is controlled (Section 4.3.2.2). Finally, reusable analyses relies on documentation and **f**indable, **a**ccessible, **i**nteroperable and **r**eusable (FAIR) data (Section 4.3.2.3).

4.3.2.1 Traceability with notebooks and workflow managers

The traceability is a prerequisite for the repeatability. For this purpose, two kinds of tools have been developed: **computational notebooks** and **workflow manager**.

The principle of **computational notebooks** is to interleave programming code with natural statements. They are used to (i) make the code human-readable, because it is documented, and (ii) precisely trace how the analyses were performed and what they output, because notebooks can store the graphical results of the analyses. Perfect traceability makes it possible to identify any unintentional errors in the process that would otherwise be undetectable. Indeed, notebooks can be published as scientific articles, as done in the Environmental Data Science book[10].

For scRNA-Seq data analysis, two types of notebooks can be used, depending on the programming language used: Jupyter notebooks in Python, and R Markdown in R language. Although they share a similar structure, they also have peculiarities. While Jupyter notebooks enable to save the code output, R Markdown notebook require compilation to build HTML or PDF immutable documents. Computational notebooks tend to be used for data analysis with many visualization steps, and are not well suited for long-running processes, parallelizable processes, and command line tools.

Their counterparts for data (pre-)processing are **workflow managers**. Workflow managers are used to (i) arrange multiple steps that could be written in different programming languages, (ii) automate redundant and standardized analyses that do not require much visualization, and (iii) parallelize the processing of multiple input data with the same format (Wratten *et al.*, 2021). Depending on the field, hundreds of workflow systems are available (Amstutz *et al.*, 2024). Among them, two main solutions are used in bioinformatics: Nextflow (Tommaso *et al.*, 2017) and snakemake (Köster & Rahmann, 2012). They share the same concept, namely process (or rule) and channel. While processes define what to do with data, channels organize the flow of data through processes.

Notebooks and workflow managers allow full traceability of data processing. However, they do not guarantee repeatability or reproducibility (Cohen-Boulakia *et al.*, 2017; Samuel & Mietchen, 2024). In fact, they can suffer from randomness, computational sensitivity, error propagation, and they must be carefully written to ensure their stability. A trial-and-error method may be used to modify their content and fix the variability. Once the analysis is stable, the environment in which it runs should be controlled to

[10]EDS book: `https://edsbook.org/`, accessed in March 2024

make it portable. The environment consists of the tools needed to process the analysis, as well as the other tools that enable the notebook or workflow manager to run.

4.3.2.2 Portability with environment control

An analysis is based on different tools, under specific versions. Each tool further relies on its dependencies, with compatible versions. The tools, their dependencies and all the version make the environment. Managing environments makes it easier to install, update and distribution them. Two sets of solutions exist to manage environment: **package managers**, such as Conda[11] and **containerization solutions**, such as Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017). Their respective advantages and limitations are summarized in the Table 2.

	package manager	**containerization solution**
	Conda	Docker, Singularity
level	no OS-level dependencies included	encapsulate entire software along with their OS-level dependencies
usability	easy and intuitive setup and usage	can be hard to setup and use
edibility	simplifies package installation and updating	not easily editable
compatibility	only access to a defined set of pre-compiled packages	everything installable
portability	through YAML specification files	directly portable container

Table 2 – Comparison of Conda with Docker and Singularity. OS: operating system.

While Docker and Singularity containers are easily shared via a single operating system (OS)-independent file, a **Conda** environment consists of a folder containing multiple files, which runnability depends on OS-specific dependencies. Thus, transferring conda environments across different operating systems can be challenging. In fact, conda environments are shared via a description file, which is OS-specific. Installing the environment from this file won't give exactly the same environment in a computer with different OS. Also, a package or a specific version may not exist or be compatible across OS[12]. These issues are more limiting for environments that contain a large number of packages.

Although they are developed for the same purpose, **Docker** and **Singularity** present some differences. By default, with Singularity, environment variables from the hosting system, and home directory, are shared. Thus, one should pay attention that elements

[11] Conda: `https://github.com/conda/conda`
[12] Conda across operating systems: `https://stackoverflow.com/questions/55554431`

running within Singularity container are not in fact using the settings of the hosting system. Controlling the container parameters and using two distinct machines should help to prevent reproducibility issues.

Docker and Singularity containers are built on a given machine. Programs requiring compilation are thus installed using the technical parameters of the machine, for instance the processors. Compilation on a low performing machine can prevent the performance, including speed, of the programs, once the containers are used on high performing computing clusters. **Guix** is specifically designed to overcome the limitations of choosing between reproducibility and performance (Courtès, 2013). It is a package manager that enables reproducible deployment of software, because, contrary to Conda, OS-level dependencies are taken into account (Vallet *et al.*, 2022).

4.3.2.3 Re-usability with code and data documentation

A documentation of the environment, code and data is essential for reusing pipelines. Documenting **code** is possible at several levels. At the **analysis level**, computational notebooks (see Section 4.3.2.1) are developed to facilitate the addition of natural language around lines of code, although it is also possible to make comments at the code level using specific comment characters. At the **development level**, for redundant parts of code, it is standard to write functions, which include original way to handle data, or wrappers of pre-existing functions. In R language, several functions can be arranged in a package, and documented using the Oxygen method, with roxygen2 package. Oxygen provides standardized and formatted comments that, once filled, allow to generate automatic documentation in an easy to read format.

To systematically inform users of the version of custom functions, **version control system** such as Git can be used[13]. A git file can be initialized in a folder containing multiple code files. Git allows to track all changes in these files. One can define a version at a time when the files are sufficiently stabilized and usable. Git will tag all the files for that version. Then, installing a package from a particular version means fetching the files associated with the version tag and installing them.

Beyond documenting the code, one should document the reasons why the specified code was written, what has been learn from it, and what should be done next, i.e., the **research process**. This meta documentation can be seen as part of an experimental notebook, and can help someone else to follow what was already done.

[13]Git: `https://www.git-scm.com/`

Last but not least, data should be documented with what is known as **metadata**. Metadata are used to describe and annotate data. While data may not be human-readable, metadata should allow for the understanding of its content. Similarly, workflows be annotated by ontology terms to be automatically compared to other similar workflows (Cohen-Boulakia *et al.*, 2017).

To improve the reusability of data, **FAIR principles** have been defined (Wilkinson *et al.*, 2016). FAIR data should be **findable**, meaning that metadata are sufficient to find and understand data; **accessible**, meaning that metadata are available and associate data with a unique identifier, even if data themselves are not accessible; **interoperable**, meaning that both data and metadata use a standard format, or that there is sufficient information to read (meta)data; and finally **reusable**.

4.4 Outlook and feedback

In this chapter, I presented the current scientific context of reproducibility crisis, and highlighted specificities in the computational biology field. Over the past decade, many solutions have been proposed to overcome the crisis and apply best practices. Although setting up a daily framework for doing reproducible research is time consuming, there are many long term benefits.

During my thesis, I was faced with the following problems related to reproducible research. First, I started my thesis after a Master student had already developed functions and scripts to analyze data. There was **no documentation** about the code, making it challenging to read, debug and re-use. Next, I faced issues regarding the **repeatability** to produce figures, within the same day, using the same code, input data and computing environment. I was involved in several projects entailing similar analyses, which required to generalize the code. The **re-usability** of pipelines was a key to progress effectively on several projects. Furthermore, the laboratory bought me a new computer in the middle of my thesis journey. I subsequently needed to transfer all the analysis framework on this computer to continue the analyses, while ensuring that it gives the exact same final figures on both computers. Then, I was able to retrieve published data, but the associated **metadata** was difficult to find, even though its content was mentioned in the publication. It was challenging to make the link between retrieved data and their representations in published articles. Finally, after retrieving data, I was not able to **reproduce published results**, either because the code was shared but incompletely or not sufficiently documented to understand its content and be able to re-use it, or the methods were incompletely described.

II Aim of the thesis

Malignant peripheral nerve sheath tumors (MPNSTs) pose a significant lifetime risk for individuals with Neurofibromatosis Type 1 (NF1). MPNSTs and their benign precursors, plexiform neurofibromas (pNFs), are characterized at multiple biological levels. However, a lack of understanding remains regarding the mechanisms driving the initiation of pNFs and their malignant transformation. To decipher these mechanisms, the laboratory has developed a mouse model that unprecedentedly recapitulate several features of NF1 disease, including the development of MPNSTs from benign tumors.

Using this mouse model, the biological purpose of my PhD thesis was to **reconstruct the natural history of tumor development, from asymptomatic nerves to pNF development, and then to malignant progression**. This project takes part in a broader research project in the laboratory. While I took advantage of the computational analysis of single-cell RNA Sequencing (scRNA-Seq) data, my colleagues were in charge of the experimental procedures underlying the project.

The project was split among the team members into two parts, focusing on these complementary topics: the initiation of pNF and the malignant transformation. However from a computational point of view, both parts of the project rely on very similar processes. Thus, the computational purpose of my PhD thesis was methodological and consisted in **setting up a computational framework, adaptable to various projects, and fulfilling the reproducible research criteria**. The latter implies the traceability of the whole analysis, the portability to new machines, and the ability to reproduce the exact same results from the same input data on a new machine. Working in close collaboration with biologists, the framework should also be readable by other team members.

In the next chapter, I present the datasets and the computational methods used to answer the biological questions. They are more contextualized compared the Materials and Methods section of associated manuscripts. In the Results chapter, I first present the computational framework set up for reproducible scRNA-Seqdata analysis (Section 1). Secondly, I present our manuscript relative to the identification of the events underlying pNF initiation in our mouse model (Section 2). Thirdly, I present the main contribution of my PhD thesis, namely the characterization of the cellular and molecular mechanisms driving malignant progression of pNF into MPNST (Section 3). Finally, I applied the computational framework to an independent biological project, focusing on the characterization of hair follicles from patients suffering from *Hidradenitis suppurativa* (Section 4).

III Materials and Methods

1 Biological sample presentation

In this section, the scRNA-Seq data analyzed during my thesis are presented. They come from mice or from patients, generated in the laboratory or downloaded from public databases. Up to the sequencing, I participated to discussion relative to data generation. From tissue collection to sequencing, experiments were conducted by my colleagues.

1.1 Mice data produced in the lab

1.1.1 Presentation of the mouse models

Two mouse models (***Nf1*-KO**) have been generated by our laboratory and used in the studies reported in this thesis. In the *Prss56*$^{Cre,+}$;*Nf1*$^{fl/fl}$;*Rosa26*$^{Tom,+}$ (***Nf1*-KO^{wt}**) model, the micro-environment is wild-type (wt) for *Nf1*, while it is haploinsufficient (heterozygous, het) in the *Prss56*$^{Cre,+}$;*Nf1*$^{fl/-}$;*Rosa26*$^{Tom,+}$ (***Nf1*-KO^{het}**) model (Table 3).

mouse model	abbreviation	*Prss56*+ lineage	ME	Model for
Prss56$^{Cre,+}$;*Nf1*$^{fl/fl}$;*Rosa26*$^{Tom,+}$	*Nf1*-KO^{wt}	• *Nf1*$^{-/-}$	• *Nf1*$^{+/+}$	sporadic MPNST
Prss56$^{Cre,+}$;*Nf1*$^{fl/-}$;*Rosa26*$^{Tom,+}$	*Nf1*-KO^{het}	• *Nf1*$^{-/-}$	• *Nf1*$^{+/-}$	NF1-related MPNST
Prss56$^{Cre,+}$;*Nf1*$^{+/+}$;*Rosa26*$^{Tom,+}$	control	• *Nf1*$^{+/+}$	• *Nf1*$^{+/+}$	control

Table 3 – Mouse models of interest in the laboratory. Mouse model corresponds to full model name. Associated abbreviations correspond only to the *Nf1* gene status. *Prss56* lineage and micro-environment (ME) indicate the *Nf1* gene status of cells. The "-" sign means that exons 31 and 32 are deleted on one allele. Depending on the micro-environment *Nf1* status, a model is better suited to study sporadic MPNST or NF1-related MPNST.

1.1.2 Experimental designs

Spontaneous tumors develop in *Nf1*-KO male mice housed together. Mice are monitored at the animal facility on a weekly basis (Figure 48).

Enlarged hyperplastic nerves (hN) and plexiform neurofibromas (pNFs) were invisible on mouse skin. Mice were sacrificed at a specific age to collect them. Advanced tumors, dyNFs and MPNSTs, are visible, either as bumps under the skin, or as fast growing tumors with large size. Differential diagnosis between them was either made by an experimented histologist, or by immunostaining, or after investigating gene expression in scRNA-Seq data (Article 2).

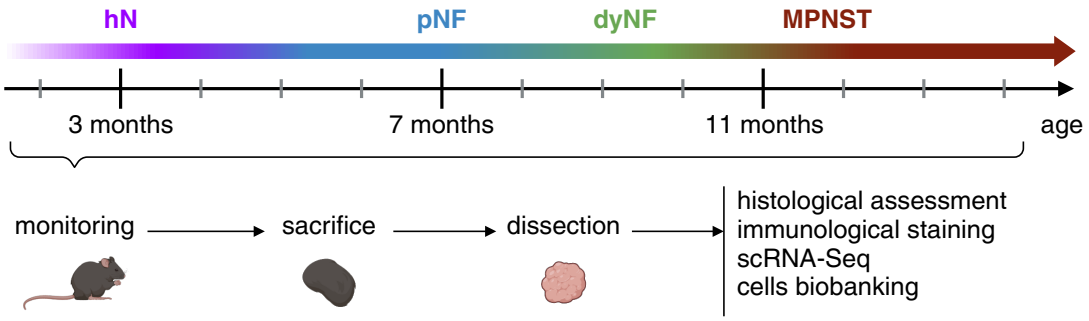


Figure 48 – Management of spontaneous tumors in *Nf1*-KO male mice. The timeline schematizes the spontaneous tumor development in male mice housed together. After sacrifice, tumors are dissected for histological assessment, immunological staining and sequencing, including scRNA-Seq. Cells are also stored in biobank for further investigation. hN: hyperplastic nerve, pNF: plexiform neurofibroma, dyNF: dysplastic neurofibroma, MPNST: malignant peripheral nerve sheath tumor.

Dissociated cells from primary tumors were collected. They can be allografted without additional treatments (Figure 49). To monitor tumor development, dissociated cells were directly transplanted in recipient mice (allografting). To evaluate specific treatment, primary tumor cells underwent transduction with viral particles, in 2D plate culture. Viral particles contained either an empty vector (control) or a vector carrying a short hairpin RNA (shRNA). The shRNA specifically recognizes the mRNA of a given gene. By binding to this mRNA, it prevents its translation into proteins and promotes its degradation. Control and treated tumor cells were injected in the right and left flank of recipient mice, respectively. All transplanted mice were immunocompromised "Nude" female mice, genetically depleted from T cells, aged between 3 and 6 months. They were sacrificed 2 weeks after transplantation to characterize tumors.

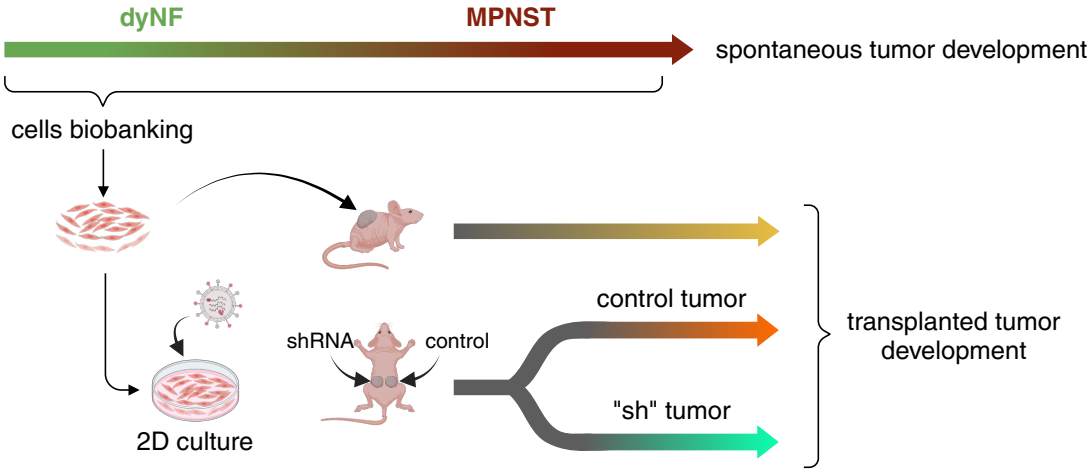


Figure 49 – Management of biobanked cells from *Nf1*-KO male mice. The timeline schematizes the spontaneous tumor development in male mice housed together. Dissociated cells can be directly transplanted in nude recipient or treated in cell culture before transplantation.

1.1.3 Available data overview

Figure 50 presents the scRNA-Seq mice data produced in the laboratory and discussed in this thesis. All libraries were sequenced on a Illumina sequencer. The term "**sample**" is used to address biological sample, while "**dataset**" is about the computational dataset, after the sequencing.

Datasets are associated with two manuscripts, or presented as complementary results (**results** row). In the Article 1 ("Mansour"), we compared hN from individually (alone) or group (together) housed male mice (**condition** row). In the Article 2 ("Radomska"), a majority of datasets were already analyzed when I joined the laboratory. I reanalyzed them from the raw count matrices (CM) (**traceability** row). For compatibility with this project, new data were mapped to the mm10 transcriptome. For novel projects, the mm39 transcriptome was used (**mapping** row). The numbers of cells in each dataset are indicated (**nb_cells** row). They correspond to the final individual datasets, after quality-based filtering.

Datasets are associated to four main sample types (**sample_type** row). Note that hN and pNF datasets were obtained from several mice (**nb_mice** row), because they represent small biological samples. The two pooled pNF datasets took part in both manuscripts (**results** row). Thus, they were mapped to mm10 and mm39 transcriptomes to ensure compatibility. However, they still differ from hN datasets by the experimental protocol applied to generate single cells suspension (**protocol** row). For hN and pooled pNF samples respectively, dissociation occurs in a bain-marie or an incubator, the latter being more stressful for cells. Furthermore, dissociated hN cells were passed through a column to remove myelin, an essential step in nervous tissue processing. Therefore, hN and pNF datasets were carefully compared.

The two pNF datasets from 2020 originated from pooled pNFs, with either nodular or diffuse nature. One of them (dataset 2020_17) further mixed mice aged from 7 to 20 months (**age** row). In 2024, two new pNF datasets, obtained from 7.4 month-old mice from the same litter, were generated to compare nodular and diffuse pNFs (**sample_type** row). All samples were individual sequenced, except these two new pNF datasets, that were multiplexed. They thus contain less reads that single samples, a point to be keep in mind during analysis.

Finally, datasets originate from 4 genotypes (**genotype** row). Control mice corresponds to *Prss56*$^{Cre,+}$;*Nf1*$^{+/+}$;*Rosa26*$^{Tom,+}$ male. All Nude mice were female mice (**gender** row). Some datasets are derived from the same primary mouse, resulting from allografting or differences in experimental procedures (**mouse_association** row).

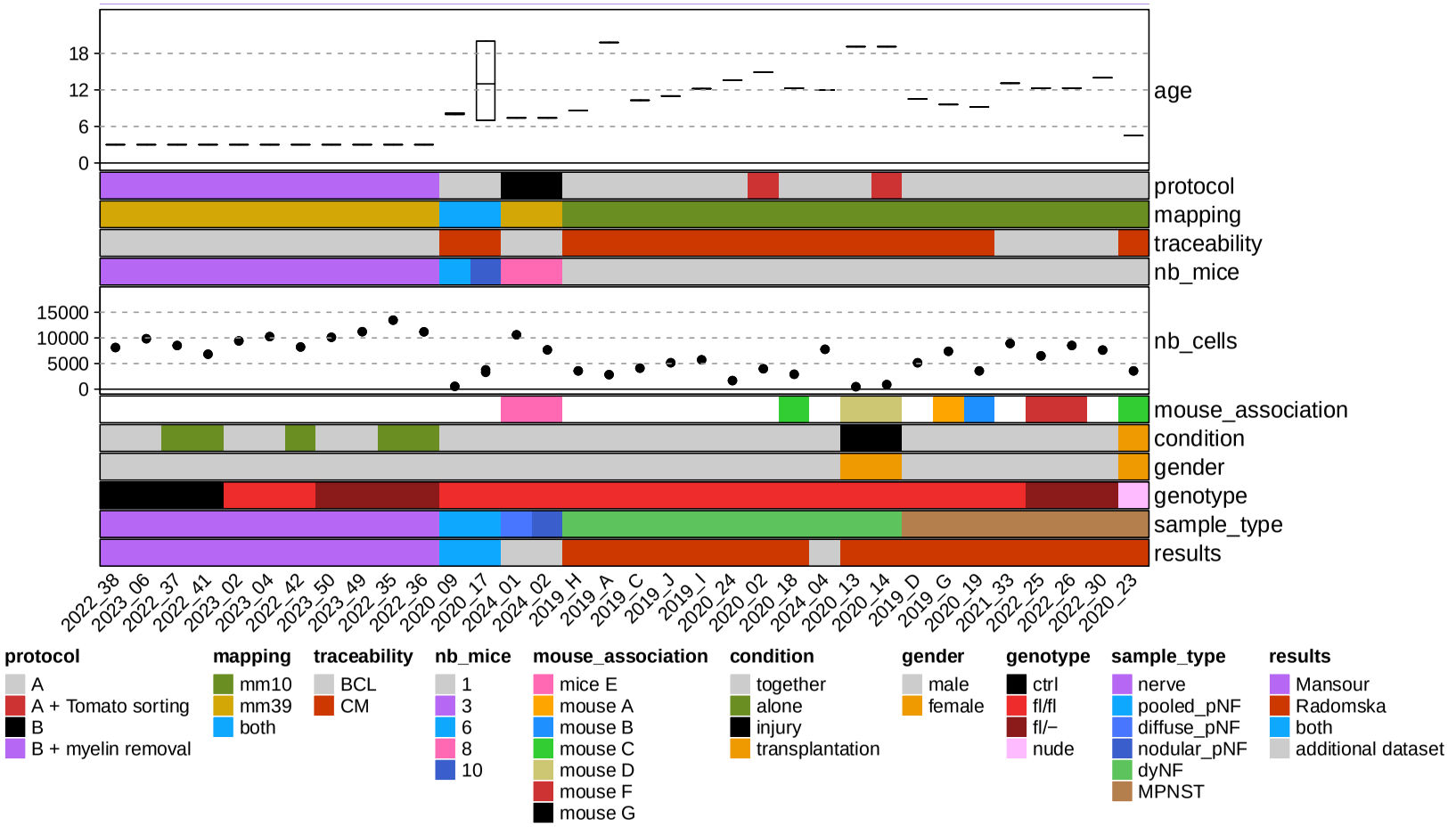


Figure 50 – **Metadata of scRNA-Seq from mice, generated in the laboratory**. Each column corresponds to an individual scRNA-Seq dataset and rows to metadata. Rows content is explained in the text. Figure made with the ComplexHeatmap R package (Gu *et al.*, 2016).

1.2 Patients data

For the Article 2 ("Radomska"), **NF1 patients** data were produced to validate findings from mouse data. Patient samples resulted from surgical biopsies. A first piece was used for histological-based diagnosis, by an experimented histologist. A second piece underwent preparation to produce scRNA-Seq data. The reads were already mapped on the hg19 transcriptome when I joined the lab. The datasets were analyzed from the count matrices. To enrich our datasets, public scRNA-Seq data were retrieved from Gene Expression Omnibus (GEO) public repositories (Figure 51, left panel). They were analyzed from the filtered count matrices.

For the Article 3 ("Hue"), hair follicles (HFs) from healthy donor or *Hidradenitis suppurativa* patients were sampled from discarded plastic surgery, and processed to produce scRNA-Seq data. To validate our findings, public scRNA-Seq data from healthy donors were retrieved as FASTQ files from the biosino public repository (Figure 51, right panel). For all the datasets in this project, the traceability is ensured from the FASTQ files. Reads were mapped onto the hg19 transcriptome using Cell Ranger version 3.

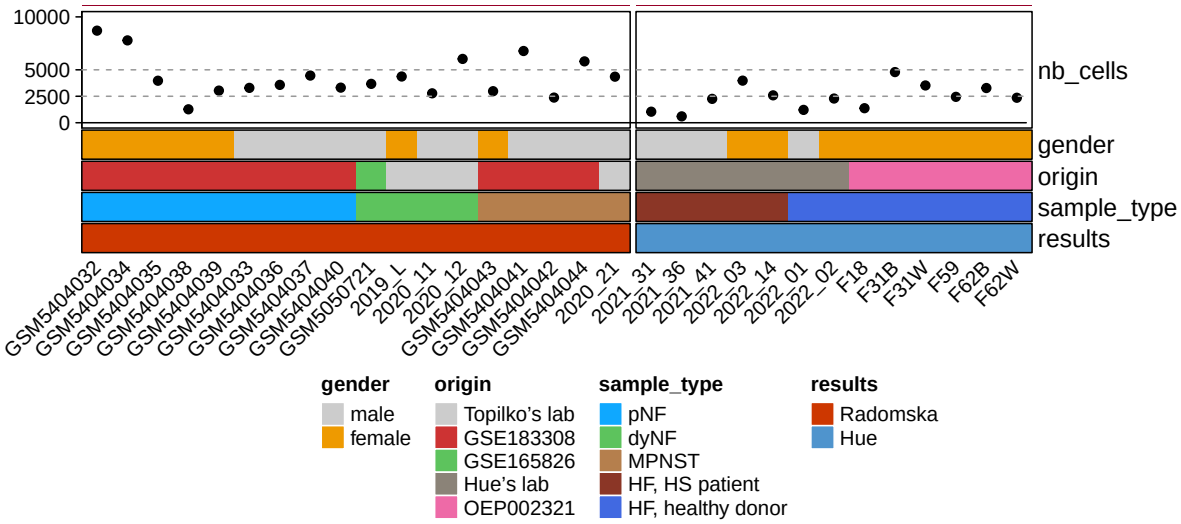


Figure 51 – Information regarding all patients data presented in this thesis. Each column corresponds to an individual scRNA-Seq dataset, and rows are metadata. Topilko's lab corresponds to our lab. Hue's lab is associated with the Article 3. pNF: plexiform neurofibroma, dyNF: dysplastic neurofibroma, MPNST: malignant peripheral nerve sheath tumors, HF: hair follicle, HS: Hidradenitis suppurativa. Figure made with the ComplexHeatmap R package (Gu *et al.*, 2016).

1.3 Data from public repositories

The public datasets re-analyzed during my thesis are summarized in Table 4. Datasets from group A were used to compare hN and pNF tumor cells to the diversity of wild-type Schwann cells. Datasets from group B were used to enrich our atlas of NF1 patients data. Finally, datasets from group C were used to validate our findings.

	name	content	accession	#	reference
A	Zhao	uninjured and injured (cut) mouse sciatic nerve	GSE198582	4	Zhao *et al.* (2022)
	Carr	uninjured and injured (crush) mouse sciatic nerve	GSE120678	3	Carr *et al.* (2018)
B	Wu	pNFs and MPNSTs from NF1 patients	GSE179043	13	Wu *et al.* (2022a)
	Sun	dyNF from a NF1 patient	GSE165826	1	Sun *et al.* (2021)
C	Wu	hair follicles from healthy donors	OEP002321	6	Wu *et al.* (2022b)

Table 4 – Overview of the reanalyzed published scRNA-Seq data. Each row corresponds to the publication documenting the datasets. Accession number and number of datasets (#) retrieved from the repository are indicated. The capital letter in the first column groups data by project.

Data from OEP002321 accession number were retrieved as FASTQ files on the biosino repository. All the other data were downloaded from public repositories as pre-filtered count matrices. The experimental procedure between public datasets and our datasets are inherently different. This introduces much more biases in the data compared to the differences related to read mapping. Furthermore, both human and mouse transcriptomes are very well annotated. Excepting gene names, there are no major changes in transcriptome annotations across recent versions.

2 Reproducibility of the computational analyses

This section presents the tools used to reach reproducible computational analyses, from raw data (FASTQ files) to the figures presented in publications. Figure 52 shows the global framework set up during my thesis.

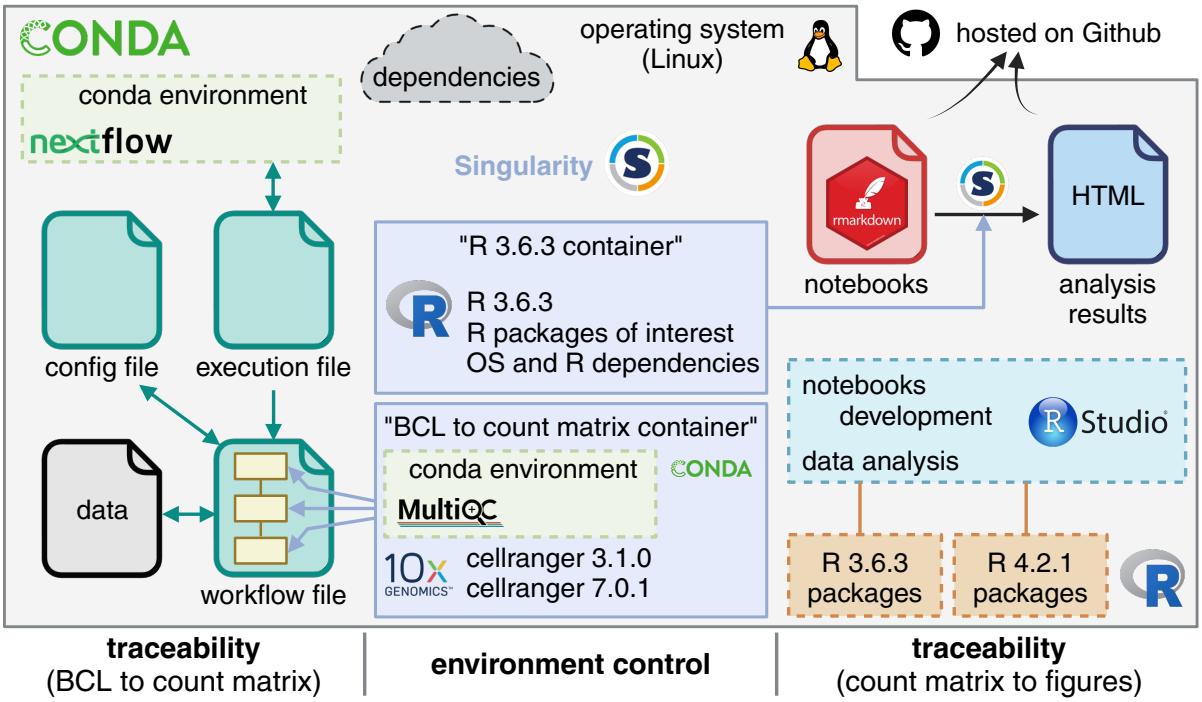


Figure 52 – Framework for reproducible computational analyses. The gray frame corresponds to the operating system (OS) on the local machine, which includes many dependencies (cloud shape). Dashed boxes depict elements that require dependencies from the OS. Hard line boxes are OS-independent: Singularity containers and files. The left part represents the framework for Nextflow pipelines (BCL) to count matrix), while the right part corresponds to the analyses after count matrices.

2.1 Traceability

All the code files produced during my thesis are stored in private repositories, on **Github**, to track code changes. A **dedicated repository** is attached in the Code availability section of each manuscript, for future publication.

The computational processing of scRNA-Seq data is divided into two main parts: before and after the count matrix. The traceability of the processing from raw data to count matrix was ensured using the version 20.10.0 of the **Nextflow** workflow manager (Tommaso *et al.*, 2017), while **R Markdown** notebooks were developed to process the count matrices.

Because the R Markdown notebooks do not contain the results of the analyses, they require compilation. Executing the code and converting the file to an HTML format is made using the knitr[1] package, and is called "**compilation**". It successively converts a R Markdown file into a Markdown file, and finally a HTML file. Each format has its own specificities (Figure 53). The HTML format is particularly useful for sharing and ease the access to the results.

	R Markdown	**Markdown** (temporary file)	**HTML** (self contained)
readable	✓	✗	✓
editable	✓	not supposed	not supposed
code (programming language)	✓	✓	✓
comments (nature language)	✓	✓	✓
text output	✗	✓	✓
graphical output	✗	encoded	✓

Figure 53 – From R Markdown notebook to their HTML compiled version. The R Markdown notebook is the working file to build the analyses. The compilation consists in converting the notebook in a Markdown temporary file, a full text file which is not supposed to be modified. Finally, the Markdown file is converted into an HTML file. By default, it is self-contained, meaning that the definition of the layout is included inside the file. The HTML file is not supposed to be altered and contain the code, comments, text and graphical outputs of an analysis.

An HTML index page would ease the navigation across the notebooks. To wrap all compiled notebooks in this **integrated page**, several R packages were tested (Table 5). Although **knitr** was used to generate each individual HTML file, it does not allow to integrate several HTML files together. All other tools require to build again the HTML pages. We still evaluated their functionalities with simple tasks, to assess if we could arrange the HTML pages *a posteriori*. The **pkgdown**[2] and **blogdown**[3] packages were quickly dismissed because they are more suited for other use, respectively package website or books that do not contain heavy analyses. The **bookdown**[4] package

[1]knitr website: `https://yihui.org/knitr/`
[2]pkgdown website: `https://pkgdown.r-lib.org/`
[3]blogdown website: `https://bookdown.org/yihui/blogdown/`
[4]bookdown website: `https://bookdown.org/`

considers that all R Markdown files are child files of a main parent. Thus, they should not contain first-level titles (large titles) and must share a common header. These requirements were too restrictive. The **gitbook**[5] tool was the most promising. It is independent from R language and takes as input the Markdown files. Using those generated by knitr, the HTML pages obtained with gitbook were raw, empty of styles and lacking interactivity. We tried to embed the HTML files generated by knitr into empty Markdown files before using gitbook. But, significant Java-related errors made debugging challenging.

	input	output	issues
knitr	single Rmd	single HTML	conversion of a single file (Rmd $\longrightarrow$ md $\longrightarrow$ html)
pkgdown	md	R package website	suited for the folder tree of R package
blogdown	Rmd	website	global layout is generated with hugo (`gohugo.io`)
bookdown	Rmd	website	requires Rmd files with common header (YML), and no title of level 1
gitbook	md	website	using knitr's md files, fails to retrieve style (CSS) and interactivity (Java)

Table 5 – Issues with existing solutions to build HTML pages from (R) Markdown files. The tools are listed in the first column. Input and output formats are indicated: R Markdown (Rmd), Markdown (md) or HTML. Issues are briefly stated and are detailed in the main text.

Finally, a custom HTML page was developed to connect all the compiled notebooks, without needing a new compilation (Results, Section 1.4.4).

2.2 R package

Many custom R functions, mainly wrappers, were also developed. To ensure portability, documentation and re-usability, they were included in a **R package**, called **aquarius**[6], build with a user-friendly tutorial[7]. The successive versions of aquarius are tracked and stored on Github (Results, Section 1.3). A website was built using the pkgdown package[8].

[5]gitbook documentation: `https://gitbook-ng.github.io/`

[6]Why aquarius ? Because this is my astrological sign, and before my arrival, functions were loaded using the "source" function, again a water-related name.

[7]Tutorial to create a R package: `https://thinkr.fr/creer-package-r-quelques-minutes/` (accessed from 2021 to 2024, in French)

[8]pkgdown website: `https://pkgdown.r-lib.org/`

2.3 Computational environment control

It is not reasonable to always use the last versions of all tools, for many reasons. Firstly, scientific projects spans over several months, if not years. To ensure traceability, it is better to keep the same environment for the whole project. Secondly, R packages have many dependencies. During my thesis, around 30 packages were used. Including their dependencies, the environment involves 437 packages. Updating one package require many other updates for compatibility reasons. Additionally, across package versions, some functions may evolve in their implemented methods, parameter names can change, as well as default parameter settings. So, the environment has been fixed.

Four kinds of **Conda** environments were created (Table 6). One contains Nextflow. Nextflow was not installed in a Singularity container because launching the workflow on a computing cluster in a container does not work smoothly. To infer RNA velocity on single cells, two tools are necessary (velocity and scvelo) but they do not require the same dependencies. Therefore, two environments were developed. The versions are detailed in the Appendix (Section 3).

main tool	running environment	purpose
Nextflow	local machine and HPC clusters	run Nextflow workflows
fastqc and multiqc	Singularity container	reads quality control
velocyto	local machine (terminal)	generate spliced and unspliced matrices from BAM file
scvelo	local machine (RStudio)	infer RNA velocity from spliced and unspliced matrices

Table 6 – **Content of the Conda environments**. Each row corresponds to a Conda environment. It has been successfully used in the specified running environment. HPC: high-performance computing.

I met some issues with Conda environments when too many tools were included. This lead to poor portability between machines, notably deployment failure when changing the operating system. Furthermore, only a subset of tools can be included in a Conda environment. To overcome these limitations, a **Singularity** containerization solution was used. Each Singularity container is stored as a single file. They are easy to share between machines, including high-performance computing (HPC) cluster.

Two containers were built (Figure 52, middle). The first contains all the tools for BCL to count matrix processing, including the Conda environment with the quality control tools (Table 6). The second container contains R version 3.6.3 and all the packages of interest, including aquarius. All packages and their versions are listed in

the Appendix (Section 2). Installing all these packages along with their dependencies was challenging. The strategy set up to develop this container is the purpose of an independent section (Results, Section 1.4.3).

2.4 Reproducibility evaluation

To evaluate the reproducibility of the computational analyses, various systematic verification tasks were done.

For the **BCL to count matrix processing**, the pipeline was run using the same input data (dataset 2024_04), with same Nextflow workflow and Singularity container, on two distinct high-performance computing (HPC) cluster: the **IBENS HPC cluster**, dedicated to members of the Institute of Biology of the École Normale Supérieure (IBENS), where my thesis project was co-supervised, and the **IFB HPC cluster**, publicly available from the French Institute of Bioinformatics (IFB). The reproducibility was assessed if (i) the same FASTQ files are obtained, based on their size (octets), the same numerical metrics are the web_summary.html files, the last file output by Cell Ranger and a perfect overlapping between the resulting tSNE plots, in this file.

For the **dataset construction**, including the individual datasets and the dataset integration, the same R Markdown notebooks were compiled using the same input data (hN datasets) and Singularity container, on three distinct machines, one under Mac OS and two under Linux OS. The reproducibility was assessed if (i) the same number of cells before and after QC-based filtering and (ii) the same tSNE plots were obtained.

Finally, for the **investigation of gene expression**, the notebook were compiled using the same input data (Schwann cells from hN datasets) and Singularity container, on two Linux machines. The reproducibility was assessed if the same fold change and p-value for top differentially expressed genes were obtained, as well as same curves and normalized enrichment scores from the gene set enrichment analyses.

3 Construction of the datasets

This section explains the preparation of datasets, starting from the sequencing files. The **package::function** notation is used to cite a specific function from a given R package. The same framework was used for all projects. The same individual processing was applied to each dataset (Section 3.1). Then, all or part of the individual datasets were integrated into a combined dataset (Section 3.2). The analysis framework takes advantage of the Seurat object (Stuart *et al.*, 2019), whose structure is detailed in the Appendix (Section 4.1)

For the analysis of scRNA-Seq from the count matrices, a tree of R Markdown notebooks, connected with a directory tree, was set up (Results, Section 1.4.2).

3.1 Individual processing

3.1.1 Obtaining the individual count matrices

The processing of the reads up to the count matrix was already done for some datasets (Figure 1.1.3). For other datasets, a Nextflow pipeline was developed to automate the process (Results, Section 1.2). The BCL files were de-multiplexed into FASTQ files using the **bcl2fastq** tool, developed by Illumina. Read quality check was performed with **fastqc**[9] and files were assembled with **multiqc** (Ewels *et al.*, 2016).

The mapping of the reads on a reference transcriptome and the construction of the count matrix was conducted using **cellranger count** command, using the **Cell Ranger** software version 3.1.0 or 7.0.1. Cell Ranger generates the raw and a pre-filtered UMI count matrix. The web summary generated from the second matrix was used to check the reads mapping quality and the global composition of datasets: number and quality of cells and cell populations. The raw count matrix was used for downstream analyses.

A reference transcriptome compatible with the Cell Ranger software has to be prepared. For patients data or some mouse projects, both hg19 and mm10 reference transcriptomes were already built before I joined the laboratory. Their origin was not traceable. For the mm39 mouse reference transcriptome, a Nextflow pipeline was developed to generate a Cell Ranger compatible reference (Results, Section 1.1). It includes the addition of custom sequences, such that of the *Tomato* reporter gene. The reference sequence file Mus_musculus.GRCm39.dna.primary_assembly.fa and the annotation file Mus_musculus.GRCm39.108.gtf from release 108 were obtained from the Ensembl repository.

[9] `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/` (Accessed in 2024)

3.1.2 Gene names and Ensembl ID

By default, in the version 3, Seurat object contains gene names. However, gene identifiers are also contained in the raw count matrix files generated by Cell Ranger. Gene names are useful for interpretation (human-level), but not for interaction with published datasets or public databases (computer-level). Thus, gene names were kept as default information, but the correspondence between gene names and identifiers (**Ensembl ID**) were added in the Seurat object. Ensembl IDs were used to interact with various databases, including for gene set enrichment analysis (Section 4.1.2).

Datasets from patients were not obtained using the same reference transcriptome annotation. To combine individual datasets into a all-patients dataset, a consensus matrix was built using the intersection of Ensembl IDs of each individual dataset.

3.1.3 Count matrix to annotated and visualisable Seurat object

task	**function**
add the gene Ensembl ID	base functions
remove cells expressing few genes	base functions
remove cells having few UMI	base functions
annotate and remove doublet cells	scds (Bais & Kostka, 2020) and scDblFinder (Germain *et al.*, 2022)
remove cells with too many UMI related to ribosomal protein coding genes	Seurat::PercentageFeatureSet
remove cells with too many UMI related to genes from the mitochondrial genome	Seurat::PercentageFeatureSet
annotate cells for cell types	own method (Results, Section 1.3.2), based on Seurat::AddModuleScore
annotate cells for cell cycle phase	Seurat::AddCellCycle
log-normalize gene expression matrix	Seurat::NormalizeData
identify highly variable genes (HVGs)	Seurat::FindVariableFeatures
scale the HVGs	Seurat::ScaleData
generate a PCA from the HVG scaled matrix	Seurat::RunPCA
cluster cells	Seurat::FindNeighbors and Seurat::FindClusters
project cells in a tSNE from the PCA	Seurat::RunTSNE
project cells in a UMAP from the PCA	Seurat::RunUMAP

Table 7 – Content of the individual Seurat object. The first column lists the functionalities of the Seurat object, and the second column explains the tool used to achieve this.

The raw count matrix was loaded in RStudio using the Seurat::Read10X function. The package EmptyDrops was used to detect and filter out empty droplets from the matrix (Lun *et al.*, 2019). The filtered count matrix was then used to initialize an individual Seurat object, through the Seurat::CreateSeuratObject function. After initialization, the Seurat object was processed as listed in Table 7. The Seurat object is then ready for further analyses or for integration with other datasets.

3.2 Atlas and cell population datasets

To address questions requiring a comparison between several datasets, individual datasets need to be merged. It is possible to merge all the raw count matrices, and analyze them as a single dataset. The same mathematical models, including estimated parameters, are then applied to all cells (doublet detection, cell scoring...). However, there are two limits. If an individual dataset appears in two different combined datasets, it will be processed twice but will not contain the exact same number of cells, which is confusing. Furthermore, the doublet detection is time and memory consuming, particularly for large datasets. For these reasons, each dataset were individually analyzed up to the final number of cells and annotated for all metrics of interest. Next, the combine datasets were generated by combining cells from individual datasets. That way, there is no need to re-process the individual datasets. The method to build a combined dataset consists in the following steps.

1. Choose the **cells of interest** in each individual dataset, using filters. Filters applied in this work are cells (i) belonging to specific clusters, (ii) annotated for a specific cell type, or set of cell types, (iii) expressing a gene, or having a score of expression, above or below a specific threshold.
2. **Merge datasets** using the **base::merge** function, which has been adapted by Seurat developers team for the Seurat object format.
3. **Filter the genes** expressed in less than 5 cells. This step is never done at the individual dataset level because some genes may appear to be of interest once several datasets have been combined.
4. **Erase** the normalization and projections generated at a single-dataset level.
5. **Normalize** the combined UMI count matrix.
6. Generate one or several **representations** or projections.
7. Remove dataset-specific effect (**batch-effect correction**) on the representations.

As for individual datasets, processing consists in log-normalization, selection of HVGs, then PCA and 2D projection using UMAP or tSNE (Table 7). To generate the 2D projection, the number of PC from the PCA was selected as the number of dimensions accounting for a fixed proportion of the total variability, based on the elbow plot. In the case depicted on the Figure 54, I chose the number of dimensions that recapitulates 65% of the variability between cells, 100% corresponding to all 100 components. Here, 65% corresponds to 43 dimensions.

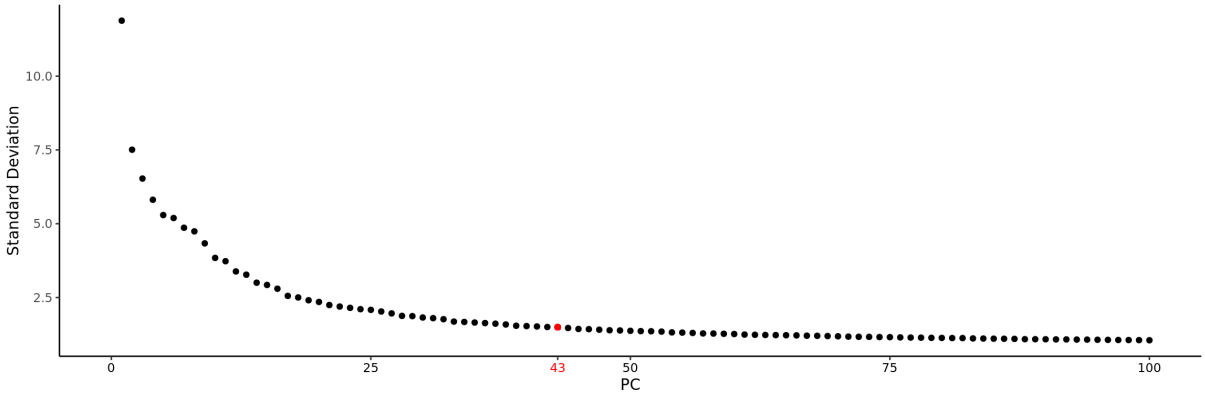


Figure 54 – **Selection of the number of dimensions to generate 2D representation**. The elbow plot represents the standard deviation (y-axis) of each principal component (x-axis) issued from the PCA. This figure is extracted from the compiled notebook **1_make_combined_all.html** from the Article 1.

Two kinds of combined datasets were produced. The **atlas** datasets gather all cells from several individual datasets of interest. The atlases were mainly used as a way to visualize all datasets of interest. Since the cell projections are specific to each dataset and take into account all the cells with the same weight, they were not used to further investigate subpopulations. The atlas datasets were not used for downstream analyses. Indeed, the atlas representation is not suited to identify rare populations. On the opposite, the **population** datasets gather specific cell populations from several individual datasets. They are used to investigate their underlying heterogeneity.

The 2D projection obtained from the PCA contains a **batch-effect**. To remove it, several popular packages were tested, without going too far from default parameters setting. The package scanorama (Hie *et al.*, 2019) was time consuming (hours) and mixed tumor cells (*Tom*+) and fibroblasts (*Tom*-) on the projection. The method in the Seurat package and the ComBat method (Johnson *et al.*, 2007; Leek *et al.*, 2012) led to memory error. The **FastMNN** method, implemented in the batchelor package (Haghverdi *et al.*, 2018), also tends to mix tumor cells and fibroblasts. However, on the tumor cells dataset, it corrected very well the batch-effect, while preserving the differences between tumor types. FastMNN was thus used to remove batch-effect in **population** specific datasets. Conversely, the **Harmony** method (Korsunsky *et al.*,

2019), performs very quickly (few minutes) and successfully removes batch-effect, while still keeping tumor cells and fibroblasts away on the 2D projection. However, it tends to mix tumor types in population-specific datasets. Harmony was thus used to remove batch-effect in all **atlas**-level datasets (Table 8).

Both FastMNN and Harmony methods produce a batch-corrected reduced space, respectively called mnn and harmony. To generate 2D projection, the number of dimensions was defined from the PCA (Figure 54) and is tracked in the notebooks.

projection	dataset
RNA - PCA - tSNE	individual datasets
RNA - PCA - harmony - tSNE	atlas datasets
RNA - mnn - UMAP	primary/transplanted tumor cells and human tumor cells
RNA - mnn - tSNE	mice tumor cells (sub-populations)
RNA - mnn - DM - UMAP	mice tumor cells (trajectory interpretable)

Table 8 – Two-dimension projections to visualize cells. The method depends on the type of dataset.

Diffusion maps (DM) were used to project cells by reflecting gene expression levels gradient. The method implemented in destiny package (Angerer *et al.*, 2016) was applied on the batch-effect corrected space from FastMNN, to generate 50 components. The usual representation consists in visualizing cells on the two first components (Figure 55A). In the case of tumor cells from our mouse model, this is not sufficient to apprehend their heterogeneity. Originally, a UMAP was made from several components (Figure 55B-F). Except the panel B, all other panels are trajectory-interpretable because they reflect the pNF-to-dyNF-to-MPNST biological progression.

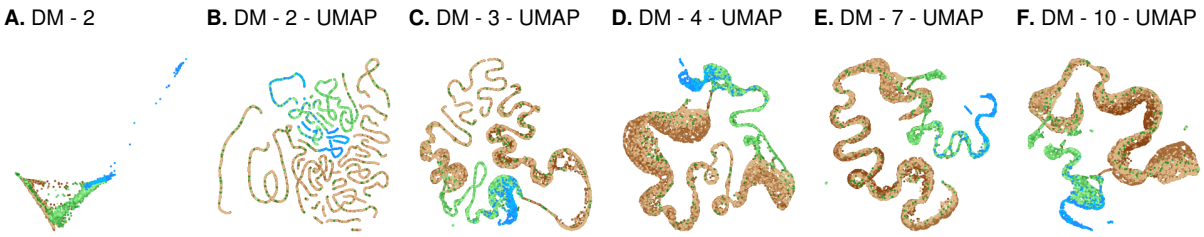


Figure 55 – Influence of the number of diffusion components on the UMAP appearance. The diffusion map (DM) was made from the FastMNN reduced space, in the tumor cells dataset presented in the Article 2. Tumor cells are colored according to tumor types: pNF (blue), dyNF (green) or MPNST (brown), with a shading corresponding to individual datasets. **A.** Cells are represented in the two first diffusion components. The cell dispersion is mainly related to particularity in a subset of pNFs cells (upper right corner), which are myelinating Schwann cells (see Article 2). **B-F.** 2D UMAPs were was made from the DM, using a specific number of components, indicating on top of each panel. The arrangement in **E** has been selected for the Article 2. Figures have been made with the notebook **1_tumor_cells_dataset** from the Article 2.

4 Analysis of the datasets

The exploration of scRNA-Seq data focused on differential expression analysis and functional analysis (Section 4.1), trajectory inference (Section 4.2), and inference of genomic alterations (Section 4.3).

4.1 Gene expression analyses

4.1.1 Differential expression analysis

The Seurat::FindMarkers function was used with the default setting. For each gene between two cell populations of interest, the statistical difference between gene expression levels is determined through a **Wilcoxon** test. More computational details are given in the Appendix (Section 4.3).

In the Article 1, the **DESeq2 package** was also used (Love *et al.*, 2014). It was originally developed for bulk RNA-Seq data, but the tutorial suggests also parameters to apply the method on scRNA-Seq data[10]. Unlike Wilcoxon test, DESeq2 is able to identify differentially expressed genes even when these genes are expressed in very few cells.

4.1.2 Enrichment analysis

The differential expression analysis outputs a list of differentially expressed (DE) genes. Enrichment analysis was used to attribute a biological meaning to this list. Two methods exist: over-representation analysis and gene set enrichment analysis. These methods are implemented in the very popular **clusterProfiler** package (Yu *et al.*, 2012).

These two kinds of analysis evaluate and rank the enrichment of specific gene sets. We used the **Molecular Signature Database** (MSigDB), which stores gene sets from many other databases. MSigDB is available and versioned in the **msigdbr** R package. For all analyses, the interrogated databases were Hallmark, KEGG pathways, WikiPathways, Reactome, Protein Interaction Database (PID), Biocarta and Gene Ontology related to Cellular Components (GOCC), Molecular Function (GOMF) and Biological Processes (GOBP). For compatibility between the databases and the count matrices, **Ensembl IDs** were used instead of gene names (Section 3.1.2).

[10]DESeq2 recommendations for scRNA-Seq data: `https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#recommendations-for-single-cell-analysis`

The **over-representation analysis** was done using the clusterProfiler::enrichr function. It attributes the DE genes to the gene sets they probably belong to. However, this method does not take into account the fold change in gene expression levels. Gene set enrichment analysis (GSEA) was applied to get more precision in the biological significance behind the DE genes. This method was particularly used in the Article 3.

The **GSEA** was performed using all genes in the count matrix, even if they were not significantly differentially expressed between the populations of interest. However, the Seurat::FindMarkers did not output the fold change for all the genes available in the count matrix. Thus, the formula provided on 10X website[11] was used to compute such fold changes. The formula below is implemented in the aquarius::run_foldchange function, using R base functions:

$$\mathsf{FC}_g = log\left(\frac{1+\sum_{c\in p_1}\mathsf{UMI}_{g,c}}{1+\mathsf{sf}_1}\right) - log\left(\frac{1+\sum_{c\in p_2}\mathsf{UMI}_{g,c}}{1+\mathsf{sf}_2}\right)$$

where:

- FC_g is the fold change for gene g ;
- g is a gene, c is a cell, p_i is the population i with $i \in \{1, 2\}$;
- sf_p is the size factor for population p: $\mathsf{sf}_p = \frac{\sum_g \mathsf{UMI}_{g,p}}{\text{median}\left(\sum_g \mathsf{UMI}_{g,p}\right)}$;
- $\mathsf{UMI}_{g,c}$ is the number of UMI in the count matrix, for cell c and gene g ;
- $\mathsf{UMI}_{g,p}$ is the sum of UMI for g is all cells from p: $\mathsf{UMI}_{g,p} = \sum_{c\in p}\mathsf{UMI}_{g,c}$.

The clusterProfiler::GSEA function outputs a list of candidate gene sets, with p-value of enrichment. Enriched gene sets were selected using a p-value threshold of 0.05. Then, word cloud representations were built using the ggwordcloud R package (Le Pennec & Slowikowski, 2023), to evaluate the genes responsible for the enrichment. Indeed, some genes are present in many gene sets, leading to irrelevant enrichment. Finally, gene sets were retained when the genes responsible for enrichment were not find in too many other unrelated gene sets. This curation was difficult to automate, and thus remains a manual step.

GSEA is performed between two populations, and thus does not reflect the variability between cells within each population. To evaluate this variability, the Seurat::AddModuleScore function was used to attribute, to each cell, a score of expression

[11] Fold change formula: `https://kb.10xgenomics.com/hc/en-us/articles/360007388751-How-is-Log2-fold-change-calculated`

for all genes in a gene set. Visualizing these expression scores on a tSNE or an UMAP allows to check if the gene set is expressed according to some specific pattern. Such a visualization is proposed in the Figure 5 from the Article 1.

4.2 Trajectory inference and RNA velocity

The scRNA-Seq technology is a powerful but destructive technique that captures a snapshot of the transcriptomic profiles of individual cells. However, it may capture closely related cells that may have originated from the same parent cell. To identify such connections, trajectory inference and RNA velocity were used.

4.2.1 Trajectory inference from a topology

The 2D projection is a very simplified view of cells heterogeneity, and is highly variable, depending on the input parameters. Consequently, trajectory inference was performed on the 50 dimensions from the batch-effect corrected space obtained with FastMNN (Article 2) or Harmony (Article 3).

Two independent methods were used to evaluate linear and branching trajectories: **TInGa**, for trajectory inference with growing neural gas algorithm (Todorov *et al.*, 2020), and **slingshot** (Street *et al.*, 2018). Both methods output a pseudo-temporal ordering of cells starting from 0, corresponding to the root cell. The root cell was always manually set in the population biologically known to be at the origin of the others. In this sense, trajectory inference was never use to discover a new trajectory, but rather to order cells in a biologically meaningful order. This is useful to build heatmaps or to make progressive differential expression analysis.

4.2.2 RNA velocity

The RNA velocity method relies on the spliced and unspliced UMI count matrices. Most methods, such as kallisto-bustools, STARsolo and alevin-fry require to map again the read on the reference transcriptome. However, the information of intronic or exonic mapping is available in the BAM file output by Cell Ranger after mapping, combined with the GTF file containing gene annotation.

The **velocyto** tool was used to generate spliced, unspliced, and ambiguous UMI count matrices from the BAM and GTF files (La Manno *et al.*, 2018). Although this tool is no more maintained, it is, to our knowledge, the only one that build these matrices from the BAM files, without need to map again the reads. The tool was installed through a Conda environment (Table 15 in Appendix). The GTF file is part of the reference

transcriptome and the BAM file was obtained at the end of the reads processing pipeline. Consequently, while trajectory inference was performed on population-level combined datasets, RNA velocity was applied at the level of individual datasets, using the scvelo Python module with default setting (Bergen *et al.*, 2020). The outputs are presented as complementary results of the Article 2 (Section 3.3.3). It turned out difficult to install a working set of scvelo with its dependencies, because of a lack of documentation. By trials and errors, a Conda environment was developed (Table 16 in Appendix).

4.2.3 Expression pattern along pseudotime

Trajectory inference orders cells by pseudotime values, which is of great interest to visualize gene expression changes. These pseudo-temporal changes were visualized on heatmaps, made using the very rich ComplexHeatmap package, enabling extensive customization (Gu *et al.*, 2016).

For the Article 2, we used the dynfeature::calculate_overall_feature_importance function to identify genes whose expression changed along pseudotime (Saelens *et al.*, 2019). This function outputs genes by importance metric, an important gene being expressed variably along the pseudotime. However, some "important" genes were expressed only by dozens of cells, and not relevant among thousands cells. The output genes list was thus manually curated to filter genes with high amplitude of expression during pseudotime, and high expression maintained in several consecutive cell.

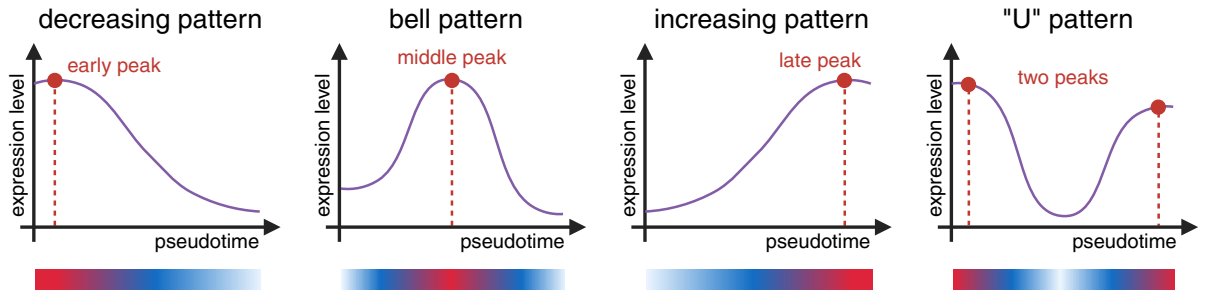


Figure 56 – **Gene expression patterns along pseudotime**. The gene expression level is represented as a function of pseudotime. Four patterns are expected, depending on the location of the peak of expression, along pseudotime. The corresponding expression profiles, as they may appear on a heatmap, are depicted by the blue (low expression) to red (high expression) shape on the bottom.

Along pseudotime, features, either genes or gene sets scores, have different profiles (Figure 56). We would like to automatically order features in the heatmap, based on their maximum of expression. To order features, the aquarius::find_features_order

function was created. It takes as input a matrix of cells-by-features, and the pseudotime ordering of cells. For each feature, the function fits a curve along pseudotime, using the stats::loess and stats::predict R base functions. While the first returns a function, the second evaluates the function along the pseudotime. Then, aquarius::find_extrema_curve identifies all the maxima in the curve, corresponding to zero point in the derivative. It outputs several metrics (Figure 57).

Finally, features were ranked by the pseudotime corresponding to their maximum of expression (max_x on Figure 57). To decide between features peaking at the same pseudotime, four other metrics were used for ranking (numbered on Figure 57). Such ranking enables to visualize transient expression first in the heatmap, and lasting expression afterwards.

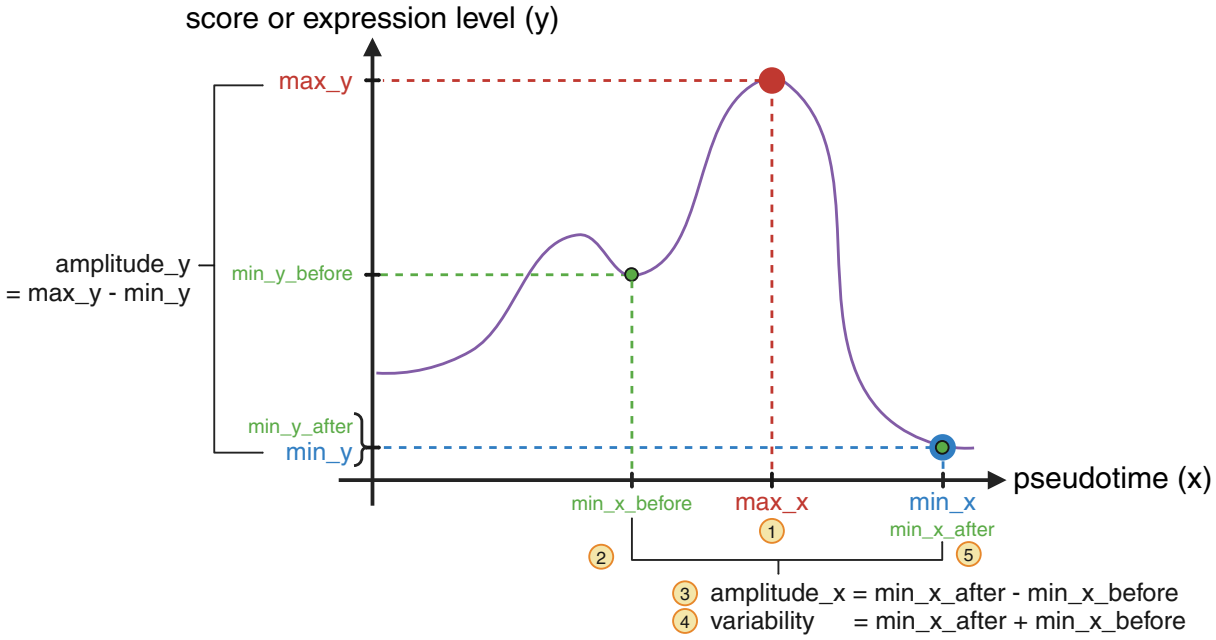


Figure 57 – Metrics used to order features on the heatmap. The feature score or expression level (y) is represented as a function of pseudotime (x), as in Figure 56. Metrics of interest are annotated. Genes are ordered based on the numbered metrics.

4.3 Inference of CNV from scRNA-Seq data

During malignant transformation, tumor cells are expected to acquire chromosomic aberrations. Since it is possible to infer genomic copy number variations (CNV) from scRNA-Seq data, we wanted to explore two aspects.

While *Tomato* expression helps to annotate tumor cells in datasets from our mouse model, the **differential annotation** between tumor cells and fibroblasts is challenging

in patients datasets. We expect tumor cells to carry more anomalies compared to fibroblasts. Based on this hypothesis, we validated the differential annotation between tumor cells and fibroblasts, in NF1 patients data (Article 2).

Then, we wanted to evaluate the **clonality** of the tumor cell population in MPNSTs (Results, Section 3.3.4). Such a study is difficult to make from the Whole Exome Sequencing data, because they were not generated at a single-cell resolution.

The infercnv package[12] was used to infer CNV in the individual scRNA-Seq datasets. This tool relies on the hypothesis that amplified chromosomic regions lead to over-expression of many genes, and conversely for lost regions, compared to normal diploid cells. The tool smooths the expression level of genes, aggregated by bins along the genome, and compare the pattern to a reference cell population. It then estimates the CNV of each bin. The reference was set in the micro-environment, either in all immune cells, macrophages, mDC or endothelial cells. Cells to infer CNV on were those annotated as tumor cells (mice datasets), with eventual addition of fibroblasts (patients datasets).

We did not interpret loss of copies at the beginning of the human chromosome 6 or in the middle of the mouse chromosome 17. These sequences contain several consecutive genes, coding for the major histocompatibility complex, which were strongly expressed in most immune cells in the context of NF1. Their absence of expression in fibroblasts and tumor cells was therefore interpreted as a loss of copies by infercnv. This behavior has already been described[13].

The infercnv function outputs many intermediate files, including a genes-by-cells table containing the CNV values (**infercnv.observations.txt**). This file was visualized as a heatmap using the ComplexHeatmap package (Gu *et al.*, 2016). It was further added as an additional assay in the Seurat object. Unlike the standard processing, this assay was not normalized. After identification of 3,000 HVGs, scaling and PCA, we generated a tSNE to visually evaluate the clones. The underlying structure of the Seurat object is illustrated in the Appendix (Section 4.2).

[12] infercnv package: `https://github.com/broadinstitute/inferCNV`

[13] Issues in using immune cells as reference for non immune cells while running infercnv: `https://github.com/broadinstitute/infercnv/issues/222` and `https://github.com/broadinstitute/infercnv/issues/225`.

5 Tools to write this thesis

This document was written in LaTeX language, using Overleaf online editor and additional tools (Table 9). The raw files behind this PDF document are hosted on Github. The repository will be publicly accessible once this thesis will be public.

name	url	purpose
Overleaf	`https://www.overleaf.com/`	layout
Reverso	`https://www.reverso.net/`	translation
Thesaurus	`https://www.thesaurus.com/`	English synonym
DeepL write	`https://www.deepl.com/write`	syntax correction
ChatGPT	`https://chatgpt.com/`	rephrasing
Stack Overflow	`https://stackoverflow.com/`	LaTeX layout help
Google Scholar	`https://scholar.google.com/`	scientific literature mining
Mendeley	`https://www.mendeley.com/`	references management
Adobe InDesign	computer software	figure layout (scientific publications)
BioRender	`https://app.biorender.com/`	scientific illustrations (this thesis)

Table 9 – **Tools to write this thesis**. The tools are indicated along with their purpose.

IV Results

During my thesis, I contributed to the following projects:

- **Article 1: Exploring the mechanisms governing the initiation of plexiform neurofibromas using Prss56Cre, *Nf1*-KO mouse model at a single-cell resolution**
  Myriam Mansour†, Audrey Onfroy†, Fanny Coulpier, Katarzyna J. Radomska, Pierre Wolkenstein and Piotr Topilko. †co-first authors
  This article is submitted to the *Journal of Investigative Dermatology* and currently under revision (Section 2).
- **Article 2: A mouse model uncovers mechanisms of malignant transformation in Neurofibromatosis Type 1**
  Katarzyna J. Radomska, Audrey Onfroy, Laure Lecerf, Bastien Job, Aurélien Beaude, Laura Sesma Sanz, Tatiana El Jalkh, Denis Thieffry, Patrick Charnay, Pierre Wolkenstein, Nicolas Ortonne, Fanny Coulpier and Piotr Topilko.
  This article is submitted (Section 3).
- **Article 3: The molecular signature of hair follicles classifies *Hidradenitis Suppurativa* patients into inflammatory, keratinizing, and mixed endotypes**
  Audrey Onfroy, Francette Jean-Louis, Philippe Lecorvoisier, Fanny Coulpier, Raphaele Arrouasse, Caroline Boucle, Salwa Abid, Émilie Sbidian, Christina Bergqvist, Pierre Wolkenstein, Véronique Godot, Yves Lévy, Piotr Topilko, Étienne Audureau and Sophie Hüe.
  This article is submitted (Section 4).

The following articles are in preparation, and are not included in my thesis manuscript due to a minor contribution.

- **Article 4: Mapping mouse cutaneous neurofibromas development with single-cell transcriptomics**
  Pernelle Pulh, Fanny Coulpier, Audrey Onfroy, Layna Oubrou, Wanzhen Zhang, Léa Toledano, Élie Abou Zougheib, Laura Fertitta, Katarzyna J. Radomska and Piotr Topilko.
  We conceived of mouse model recapitulating the development of cutaneous neurofibromas (cNFs). The aim of this study was to decipher the cellular and molecular events underlying the initiation, progression and maturation of cNFs. The study relies on immunohistochemistry and scRNA-Seq. My contribution was technical and methodological and consisted in assisting the first author to analyze her scRNA-Seq data and set up the analysis framework on her computer.
  This article is in preparation.

- **Article 5: Characterization immune cells in dysplastic neurofibromas: identification of PD1 and the Tim-3/Galectine 9 pathway as potential targets for immunotherapy**
  Anaïs Brunet, Laetitia Lacroix, Fanny Coulpier, Audrey Onfroy, Katarzyna J. Radomska, Valérie Aftimos, Audrey Briand-Suleau, Pascale Maille, Oana Hermeziu, Ingrid Laurendeau, Eric Pasmant, Pierre Wolkenstein, Piotr Topilko and Nicolas Ortonne.
  An observation at the origin of this study was the massive and transient immune infiltration in dysplastic neurofibromas (dyNFs), the intermediate step between benign and malignant tumors in NF1 disease. The aim of this study was to characterize the immune infiltrate in dyNFs in NF1 patients. The methodology included immunohistochemistry, bulk transcriptomic and scRNA-Seq. My contribution was to assess the findings from the two first methods, in the scRNA-Seq data. The computational analyses of these data was very similar to that reported in the Article 2.
  This article is in preparation.
- **Article 6: A benchmark study of bioinformatics methods for single cell RNA-seq Nanopore long reads data**
  Ali Hamraoui, Sophie Lemoine, Catherine Senamaud-Beaufort, Audrey Onfroy, Fanny Coulpier, Piotr Topilko, Stéphane Le Crom and Morgane Thomas-Chollier.
  Majority of scRNA-Seq data consists in end-biased short reads, precluding the analysis of splice isoforms. Current long-read scRNA-Seq data suffer from small library sizes, sequencing errors and poor isoform annotation, requiring distinct computational tool. The recent expansion of long-read scRNA-Seq, using the Nanopore technology, was accompanied by new tools to build the isoform count matrix. The aim of this study was to compare the performance of these tools, using synthetic and experimental scRNA-Seq data. The experimental data combined long-read and short-read paired datasets from our NF1 mouse model. My contribution consisted in providing the cell type annotation method, explaining the biological context behind data, and helping in the design of R Markdown notebooks for the differential analysis of isoforms.
  This article is in preparation.

Since the scRNA-Seq data analysis of these projects rely on the same analysis framework, the next section details the computational development carried out during my thesis.

1 Computational results

This chapter is divided into four parts, corresponding to the construction of the reference transcriptome for reads mapping (Section 1.1), the improvement of a pre-existing pipeline to build the count matrix from reads (Section 1.2), the development of an R package (Section 1.3), and the framework for reproducible computational analyses, from the count matrices to the final figures (Section 1.4).

1.1 Automation of the custom reference construction

The raw input data for computational analysis are the reads, which should be mapped to a reference transcriptome to retrieve their biological meaning. When using the Cell Ranger software, the reference transcriptome must fit a particular format. Additionally, our mouse model is genetically engineered to express the reporter gene *Tomato*, and its sequence must be added to the reference transcriptome.

In the laboratory, the pre-existing transcriptome was based on the mm10 version of *Mus musculus* (house mouse), with the addition of sequences encoding Cre recombinase (CRE), Yellow Fluorescent Protein (YFP), and *Tomato*. For the novel projects, we wanted to update the transcriptome to the mm39 version. New projects not presented in this thesis required adding sequences encoding puromycin resistance and Green Fluorescent Protein (GFP). Therefore, we decided to write a **Nextflow pipeline** to generate the reference transcriptomes (Figure 58). In the long term, this will enable easy updates of future transcriptomes.

The pipeline requires two types of input. The FASTA sequence of the reference transcriptome and the corresponding GTF annotation file can be obtained from Ensembl[1]. The second input is a folder containing one FASTA file per custom sequence of interest. Indeed, our transcriptome contains the mRNA sequences corresponding to *CRE*, *YFP*, *GFP*, puromycin resistance, and *Tomato* genes.

The main step of the pipeline is the concatenation of the FASTA (sequence) and GTF (annotation) files from the reference and the custom sequences (Figure 58, "merge" steps). Since no annotation exists for the custom sequences, the pipeline generates a single-row annotation file for each sequence. Each custom sequence is defined as a coding sequence, with one exon, of the same size as the sequence in the FASTA file (Figure 58, orange steps). All custom FASTA and GTF files are then merged together, and subsequently with the downloaded reference files.

[1] Ensembl files repository: `https://ftp.ensembl.org/pub/`

As suggested in the Cell Ranger documentation, the GTF file is filtered to remove pseudogenes and prevent multi-mapping. Finally, the custom GTF and FASTA files are arranged in a reference folder following the format required by Cell Ranger, using the cellranger mkref command (Figure 58, processing steps).

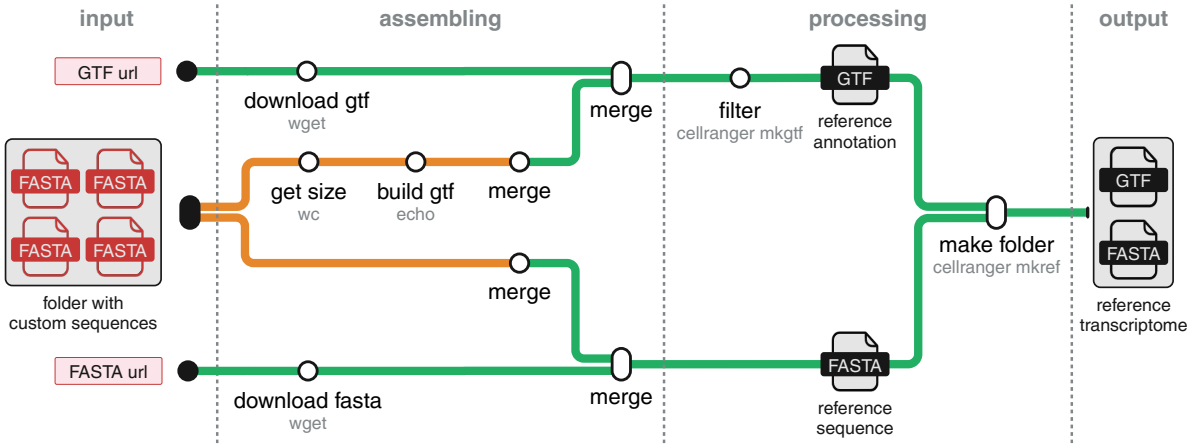


Figure 58 – Nextflow workflow structure to build the custom reference. Red elements represent customizable input files, and new files are in black. The white dots represent the main processes. During an execution, the green branches are run only once, while the orange branches are run as many times as there are input FASTA files.

This pipeline runs in **15-20 minutes** to build the mm39 transcriptome, using 4 CPUs and 20 GB of memory. It ensures the traceability of the construction of the reference transcriptome, particularly the exact version of the gene annotation file. Although it has not been tested with other input files, this pipeline should work on other organisms.

1.2 Improvement of the pre-existing pipeline for reads processing

1.2.1 Motivations and global features

The pre-existing pipeline was developed by a Master student. Written in Python, it could be run on a local machine, but suffered from a poor version control of the various tools (Table 10). The **improvement** involved using the Nextflow workflow manager to enable more modularity, parallelization of processes, and easier use of Singularity containers. Except for the Nextflow version, which is maintained through a Conda environment, all other tools are included in a Singularity container. It contains the Cell Ranger software version 3.1.0 (for old projects) or 7.0.1 (for novel projects). The version of Cell Ranger can be chosen through the pipeline configuration file. This environment control enables the pipeline to be exported to two high-performance computing (HPC) clusters, freeing resources on the local machine and shortening runtime.

	pre-existing pipeline	new pipeline
main language	Python	Nextflow
V/E control of the main language	none	Conda
V/E control of reads QC tools	Conda	Conda in Singularity
V/E control of Cell Ranger	none	Singularity
runnability	local machine	local machine + HPC clusters
documentation for re-usability	poor	yes

Table 10 – Features underlying the improvement of the reads-to-count matrix pipeline. V/E: version and environment, QC: control quality; HPC: high-performance computing.

An additional feature compared to the pre-existing pipeline is the **demultiplexing of multiple samples**, relying on distinct sample indices. As a result, the directory tree slightly changes between the old and new pipelines. The new directory tree contains more nested folders, corresponding to each individual sample. It is not shown in this thesis, but illustrated and explained in a documentation.

A **documentation** was written for other bioinformatic-aware members of the laboratory to reuse the pipeline on the IFB core cluster. This documentation explains how to connect to the cluster, import and export data, understand the directory tree and the pipeline, and set up, run, monitor, and clean the pipeline.

1.2.2 Architecture of the pipeline

The pipeline to obtain the count matrices is presented in Figure 59. It takes as recurrent input a reference transcriptome and the index sequences for demultiplexing. These index sequences are stored in the Singularity container along with the Cell Ranger software. The customizable inputs are the run folder containing the BCL files, and character strings corresponding to sample names and their respective indices names for demultiplexing.

The **pre-processing** step involves constructing a sample sheet compatible with the bcl2fastq tool. The **demultiplexing** step uses the sample sheet to split the run folder by sample and convert BCL files to FASTQ files. The reads that are not associated with the given indices are discarded in files called "undetermined". Originally, the input file containing the index sequences is reused to check whether an index is over-represented in the undetermined files, suggesting an error in the input (Figure 59, "check index" process). The **quality control** further applies two standard tools to all FASTQ files and generates a single HTML report. Finally, for each sample, the **mapping and counting** step build the count matrices and the web summary report.

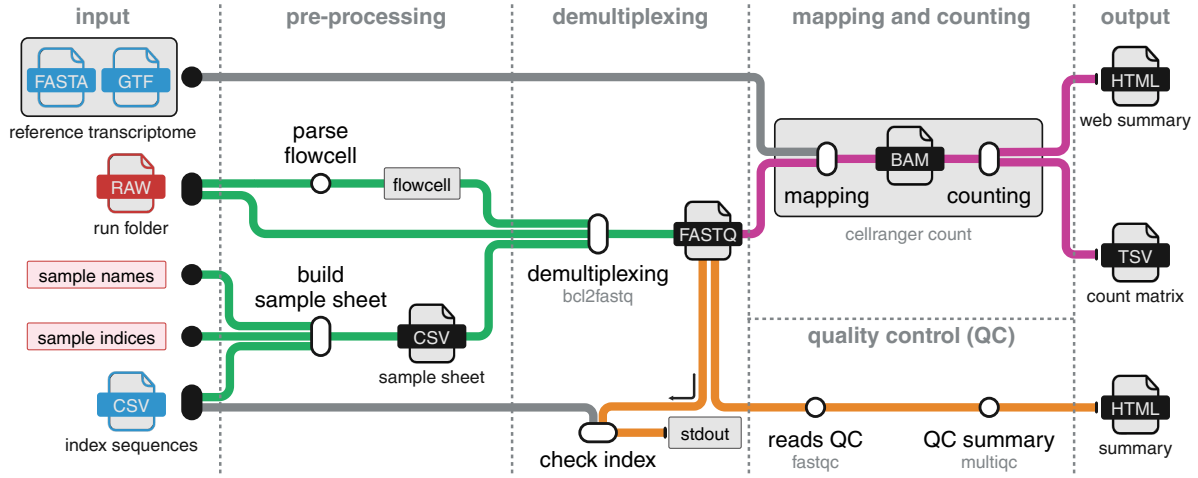


Figure 59 – **Nextflow workflow structure to obtain the count matrix**. Blue elements are recurrent input files and red depicts the input that may change regularly. The white dots represent the processes. During an execution, the green branches run only once, while the magenta branches run as times as there are samples, using the same input files from the gray branches. The orange branches run one time more compared to the magenta ones, because it includes the files with undetermined reads.

This pipeline is automatic and runs on HPC clusters. It enables reproducible and fast construction of the count matrices. Highly variable running times were observed, depending on the queue on the HPC clusters and the run folder size: between 2 and 10 hours, with an average of **6 hours**, using 32 CPUs and 100 GB of memory

1.3 The aquarius package: a wrapper for simplified analyses

To systematically document the usage and parameters of custom functions, the aquarius R package was developed. It is associated with a website but still lacks a comprehensive documentation[2]. The version 0.1.5 of the package contains over 100 functions, including 23 that were written by a Master student during a previous internship (Table 11).

origin	total number	essential	specific usage	not used
Master student internship	23	9	/	14
PhD thesis	83	32	18	33

Table 11 – **Content of the aquarius package, version 0.1.5**. The total number of functions is indicated. Essential functions are routinely used in the R Markdown notebooks. Functions with specific usage correspond to visualization functions applied to build ready-to-publish figures. Remaining functions are not used and should be removed from the package.

[2]Website associated with aquarius: `https://audrey-onfroy.github.io/aquarius/`

1.3.1 Example of functions

The Seurat package was primarily used to process the data (Stuart *et al.*, 2019). It provides a data format object and many functions to analyze the data. However, methods such as batch-effect correction, trajectory inference, and diffusion maps are implemented in other packages and are not directly compatible with the Seurat object format. Taken advantage of the seurat-wrappers package[3], additional wrappers were built to make the Seurat object compatible with independant packages (Table 12).

purpose	aquarius function	external function
generate a diffusion map	run_diffusion_map	destiny::DiffusionMap
remove batch effect using FastMNN	integration_fastmnn	batchelor::fastMNN
infer trajectory using TInGa	traj_tinga	TInGa::gng_param2

Table 12 – **Example of wrappers in the aquarius package**. The name of wrapper functions are given along with the function they wrap (Angerer *et al.*, 2016; Haghverdi *et al.*, 2018; Todorov *et al.*, 2020).

Most of the other essential functions are related to visualization (Table 13).

aquarius function	purpose	example in the Article 2
plot_piechart	piechart using as input a cell-by-annotation dataframe	/
plot_qc_facslike	scatter plot with marginal distributions	/
plot_split_dimred	2D projection of cells split according to an annotation, with a colored background	Figure 2C
plot_dotplot	dotplot with custom theme	Figure S2A
plot_barplot	barplot using as input a cell-by-annotation dataframe	Figure S2C
gsea_plot	barplot to visualize GSEA results	Figure S3B

Table 13 – **Example of aquarius functions for visualization**. These wrappers are based on ggplot and Seurat packages. The two first functions are used in the notebooks only for the quality control.

Some functions contain original implementations, such as the run_foldchange, find_features_order and find_extrema_curve functions, described in the Materials and Methods (Sections 4.1.2 and 4.2.3). Another function was developed to ease the construction of Singularity container containing all the packages of interest (Section 1.4.3).

[3]Github page of the seurat-wrappers package: `https://github.com/satijalab/seurat-wrappers`

1.3.2 A marker-based, single-cell level, cell type annotation

Cell type annotation can be done using a reference dataset. A reference corresponds to a dataset containing well-annotated cell populations and can be used to annotate a new dataset when one expect it to contain similar or lower cellular heterogeneity. Such a reference does not exist in the context of NF1 disease. Without a reference, most annotation methods involve attributing a cell type score to cell clusters. But, the clustering is not robust, and this affects the annotation.

The method described here is implemented in the aquarius::cell_annot_custom function and was presented at the JOBIM 2022 event [4] and the MaDICS symposium[5], with a poster provided in the Appendix (Section 6, Poster 1).

1.3.2.1 Identification of cell-type specific markers

During an internship, a Master student identified a set of markers for each cell type of interest. His method involved manually annotating cell clusters based on the markers they expressed. This set of markers served as the initial basis for our annotation and continues to be used to assess annotations on dot plots (eg. Article 1, Figure S2D).

Due to the high rate of drop-out events in scRNA-Seq, a few markers are not sufficient to robustly annotate single cells. To enrich the set of markers, we performed the manual cluster annotation, conducted differential expression analysis, and then selected the best markers. An ideal marker should be expressed across the entire cell population of interest (**representativeness**) and not in any other cells (**selectivity**). These metrics are quantified numerically as pct.1 and pct.2 values, output from differential expression analysis using the Seurat package (see Appendix, Section 4.3).

This process is manual and challenging to generalize. It was repeated for each project, either involving mouse nerves (Article 1), mouse peripheral nerve sheath tumors (Article 2), or human hair follicles (Article 3). For each project, the marker set was defined using three datasets depicting a high cellular heterogeneity.

1.3.2.2 Annotation method with positive and negative scores

Following the proposal of a colleague (Lehmann, 2021), the Seurat::AddModuleScore function was used to assign a score for each cell type, based on the expression levels of the provided marker genes, to each single cell (Figure 60). Each score ranges from negative to positive values. While negative values correspond to an under-expression,

[4]JOBIM 2022 event: `https://jobim2022.sciencesconf.org/`
[5]MaDICS symposium 2022: `https://www.madics.fr/event/symposium-madics-4/` (in French)

positive values indicate a global enrichment of the markers. The value "0" is thus meaningful and must be considered if a scaling is applied.

Two score matrices were defined: one for **positive markers** and another for **negative markers**. For a given cell type, the negative markers comprise the union of positive markers from all other cell types (Figure 60, "complement" process). Shared markers are included only once. The negative marker scores were consistently higher than the positive marker scores due to differences in gene numbers between the two sets. To ensure comparability, both matrices were scaled for each cell type to range between -1 and 1, with 0 maintained at 0 (Figure 60, "scaling" process). The final score matrix represents the difference between the two matrices. Cells are ultimately annotated based on the cell type corresponding to the highest score.

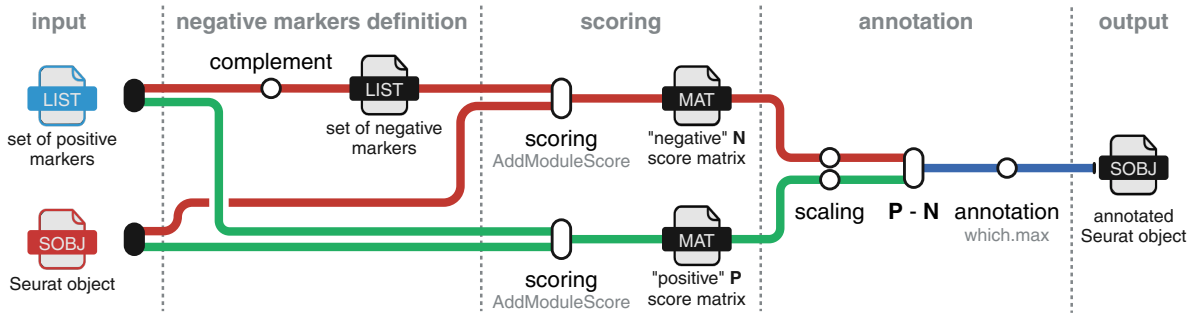


Figure 60 – Cell type annotation method in aquarius::cell_annot_custom function. R data formats indicated: list, matrix, or Seurat objects. The function takes as input the Seurat object to annotate and the set of positive markers (R list). The red (resp. green) lines depict the process to obtain a negative (resp. positive) cell-by-cell type score matrix. The blue line corresponds to the final annotation process after merging the two matrices. The function outputs the Seurat object annotated with cell types.

The Seurat::AddModuleScore function runs with default parameters and a seed to fix the variability. The scores can be visualized on 2D cell projections to evaluate potential **annotation conflicts**.

Cell type annotation was visually assessed using dot plots or 2D cell projections. On the projection, cells were colored according to their cell types. An effective annotation shows cells from a given population colored with a single color. Conversely, a poor annotation results in a "rainbow" coloring.

Single-cell level annotation is not perfect: a small proportion of cells may remain mis-annotated. However, our method enables the retrieval of rare cell types within individual datasets. When combining multiple datasets, rare cell types become more apparent and can be identified through clustering. In combined datasets, for visual clarity purpose, **annotation smoothing** was applied by selecting the most represented cell type within each cluster using the base::table and base::prop.table functions.

1.3.2.3 Validation of the single-cell level method and the set of markers

The robustness of the sets of markers was assessed by applying our annotation method to each novel dataset. Furthermore, the annotation method was successfully applied to public datasets, independent of our lab (Table 13).

set of markers	# pop.	independent validation datasets	figure
nerves	13	sciatic nerves atlas from Carr *et al.* (2018), Toma *et al.* (2020) and Zhao *et al.* (2022)	not shown
PNSTs	14	NF1 patients PNSTs from Sun *et al.* (2021) and Wu *et al.* (2022a)	Article 2, Figure S4
HFs	15	HFs atlas from Wu *et al.* (2022b)	Article 3, Figure S2

Table 14 – Validation datasets for the set of markers used in cellular annotation. The sets of markers, number of cell types (# pop.) they represent, and independent validation datasets are indicated. Results can be visualized in the figures from the manuscript. HFs: hair follicles, PNSTs: peripheral nerve sheath tumors.

1.4 Establishment of a framework for reproducible analyses

From the count matrices, the analysis framework consists of four main elements:

- Write R Markdown notebooks to ensure **traceability** of the analyses.
- **Generalize** the analysis across multiple projects using a **modular** notebook tree and associated directory structure.
- Compile the notebooks into a container to manage the **environment**.
- Organize the compiled notebooks into a website to enhance **navigability** and **facilitate sharing**.

1.4.1 Use of R Markdown notebooks to ensure traceability

From the count matrices to the figures presented in scientific publications, only R Markdown notebooks were used. Existing best practices were enhanced and applied while writing notebooks (Figure 61). Each notebook was divided into three sections.

The **layout** section includes the header, defining the title, date, and other document metadata. Global options can be adjusted to specify figure saving locations or define functions to time the analysis or add custom buttons (Figure 61, green).

The **environment** section is placed around the main content (Figure 61, blue). In the part where libraries are loaded, all R packages for graphics and data handling should be included. It is important to use ".libPaths()" to retrieve the directory of packages on the machine. At the end, "sessionInfo()" lists all the packages required in the notebook, along with their dependencies and versions. This fosters reusability or understanding the issues if the replication of the results fails.

The **main** section is divided into three parts (Figure 61, red). The global settings define directory names and parameter values. To ensure portability of the notebook, absolute paths were avoided, maintaining a consistent directory structure across projects. Notebooks are designed to rely only on the files located in this directory structure. After defining the parameters, all relevant data are loaded. The second part involves a series of data handling scripts, the outputs of which are captured after the notebook is compiled. Data can then be saved. RDS and CSV formats are preferred over the RDA format, which is challenging to manage. Finally, comments and headers are used to enhance the notebook readability. Functions are always called explicitly by their originating package (package::function).

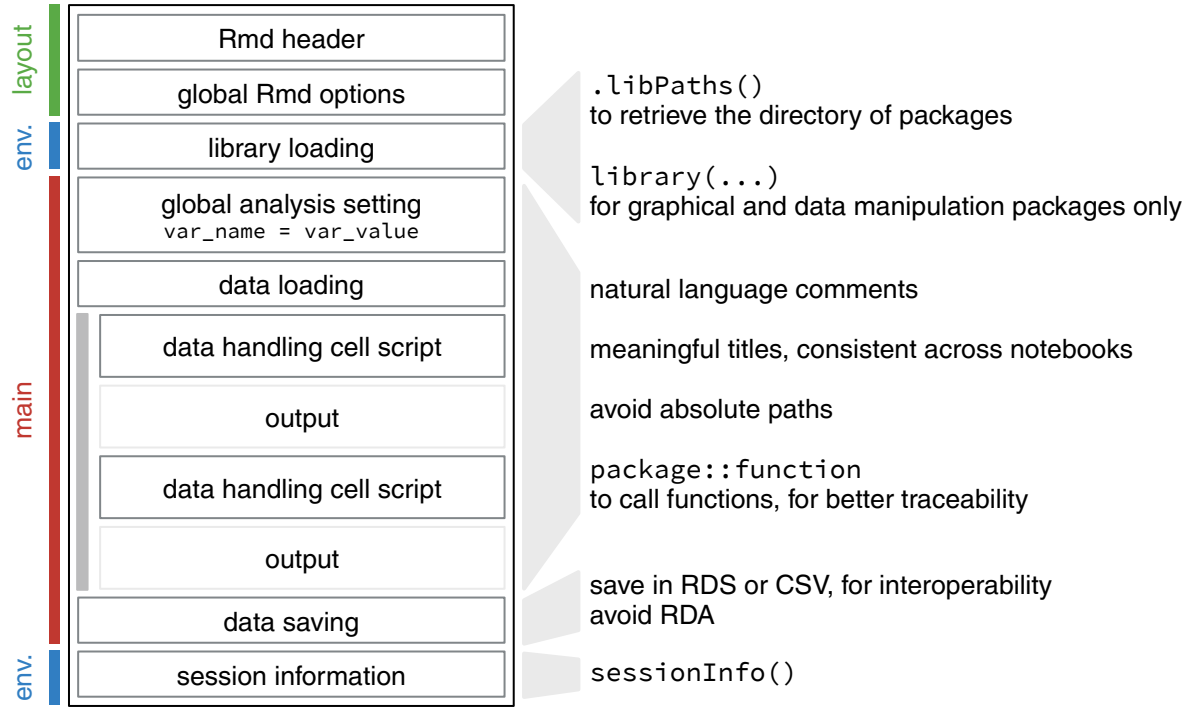


Figure 61 – Best practices in the development of R Markdown notebooks. The standard structure of an R Markdown notebook is illustrated on the left. It is divided into three sections (green, blue, red), each containing specific components. The horizontal gray line in the "main" section represents an undefined number of data handling scripts. Their output, shown in light gray, is captured only once the notebook has been compiled (see Figure 53). On the right, specific best practices are highlighted.

1.4.2 Notebooks tree for flexible analyses

The scRNA-Seq data processing consists in many steps and can be seen as an tree. To avoid compiling heavy notebooks, the code was split into three types of notebooks, for building **datasets**, running **analyses** and creating **figures**. These notebooks are interconnected, forming a notebook tree (Figure 63). It aligns with a directory tree, specific to each project (example in Figure 62).

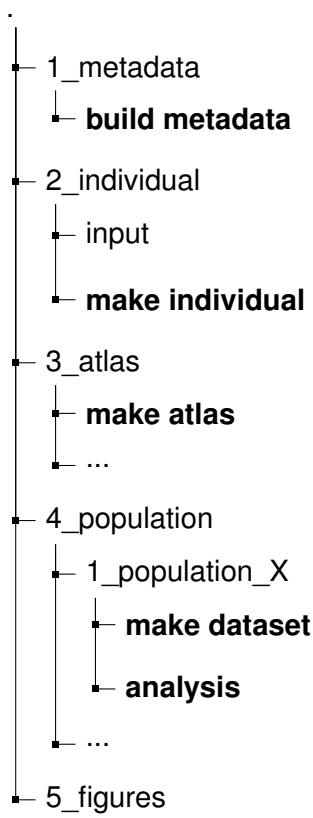


Figure 62 – **Directory tree**. The notebooks names are written in bold.

For each project, the first step builds the metadata, which includes the sample information (Figure 50), a set of markers for cell type annotation, and a color palette for visualizing cell types (Figure 63, "build metadata"). In the second folder (Figure 62, "2_individual"), the input directory contains the count matrices, represented as symbolic links. A single notebook is used to generate individual datasets from each count matrix, utilizing the metadata (Figure 63, "make individual"). These datasets can further be merged into an atlas or population-specific datasets using appropriate notebooks (Figure 63, "make atlas/population" and Figure 62, "3_atlas" and "4_population"). Each population-specific dataset resides in its own folder, alongside the associated analyses. In the last folder, no analyses are made. The notebooks are there used to build the final figures from the manuscript (Figure 62, "5_figures").

Numerous analyses are feasible (Figure 63). Except the cell-cell communications analysis, results are detailed in subsequent sections: gene expression analysis (Article 2, Figure 3), RNA velocity (Results, Section 3.3.3), copy number variations (CNV) analysis (Results, Section 3.3.4), transcription factor (TF) activity (Discussion, Section 3.7.3), extra-cellular matrix (ECM) analysis (Poster 3 in Appendix, Section 6).

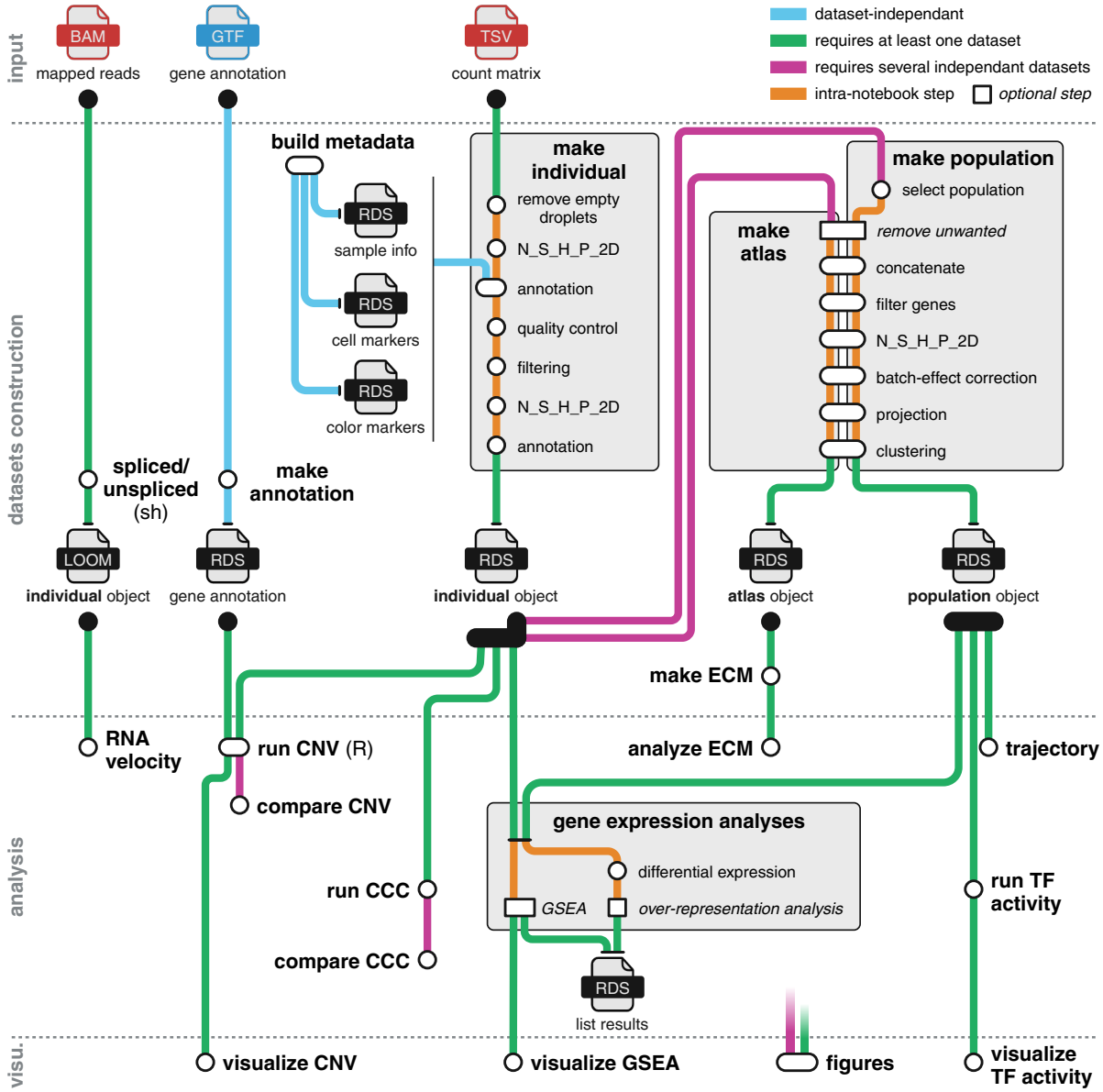


Figure 63 – **Notebook tree**. Sample-specific files are depicted in red, while other input files are in blue. Black dots represent start points. Individual R Markdown notebooks are written in bold, except for the spliced/unspliced script (bash file) and run CNV script (R file). The dependency relationships between notebooks are symbolized by lines, color-coded according to the legend in the upper right corner. The tree is composed of four parts corresponding to input data, dataset construction, analysis, and visualization. N_S_H_P_2D: normalization, scaling, identification of highly variably expressed genes, PCA and 2D projection, CCC: cell-cell communications, CNV: copy number variations, ECM: extra-cellular matrix, GSEA: gene set enrichment analysis, TF: transcription factors.

1.4.3 Containerized compilation ensures reproducible analyses

In the efforts to fully develop all analyses within a controlled environment, I encountered issues to containerize the RStudio graphical interface in a software container or to call specific tools from the container, within RStudio. Consequently, the notebooks were developed in RStudio on a local machine, but were **compiled using a Singularity container**. To maintain consistency, the goal was to deploy the same set of packages and versions, both locally and within the container.

With R version 3.6.3, our analysis panel relies on 437 packages. The primary challenge involved installing all R packages and their dependencies within a container, necessitating a precise installation order based on the dependencies.

Since we could not find an existing method to establish this order, the function aquarius::repro_dependency_tree was written. Given a set of locally installed packages of interest, this function utilizes tools::package_dependencies to list all dependencies. It then iterates through these packages and their recursive dependencies alphabetically, incrementing a table ordered by dependencies. If the dependencies of a package are already in the table, the package name and version are added. Otherwise, it proceeds to the next package. This process continues until all packages are included. Subsequently, the corresponding archive files for each package version were retrieved and downloaded on BioConductor[6], CRAN[7] or Github[8].

The definition file of the Singularity container is then straightforward. First, it installs the specified version of the R language and executes an R script to install packages based on the dependency-ordered table. Some R packages require system-level dependencies, which must be installed beforehand. A trial-and-error approach was followed to determine which packages necessitated these dependencies. Ultimately, constructing the Singularity container takes **6 hours** and is done only once.

1.4.4 Index page enables easy navigation through the compiled notebooks

In support of a scientific project or publication, numerous R Markdown notebooks and their compiled HTML versions are generated. To facilitate navigation and enhance shareability, a custom solution was set up (Figure 64). The resulting index.html page has the same structure for all projects. The code to generate the index.html page is documented and shared through a Github repository[9].

[6]BioConductor website: `https://www.bioconductor.org/`
[7]CRAN website: `https://cran.r-project.org/`
[8]Github website: `https://github.com/`
[9]Github repository: `https://github.com/audrey-onfroy/organize_html`

Utilizing the **iframe** functionality enables embedding external content in the page, specifically the compiled HTML files containing our analyses. The left-hand menu permits navigation across these HTML files. Selecting an item updates the iframe content. The top banner dynamically updates with a link to the embedded HTML file, and opens the iframe content in a new browser tab. It also enables easy tracking within the directory structure.

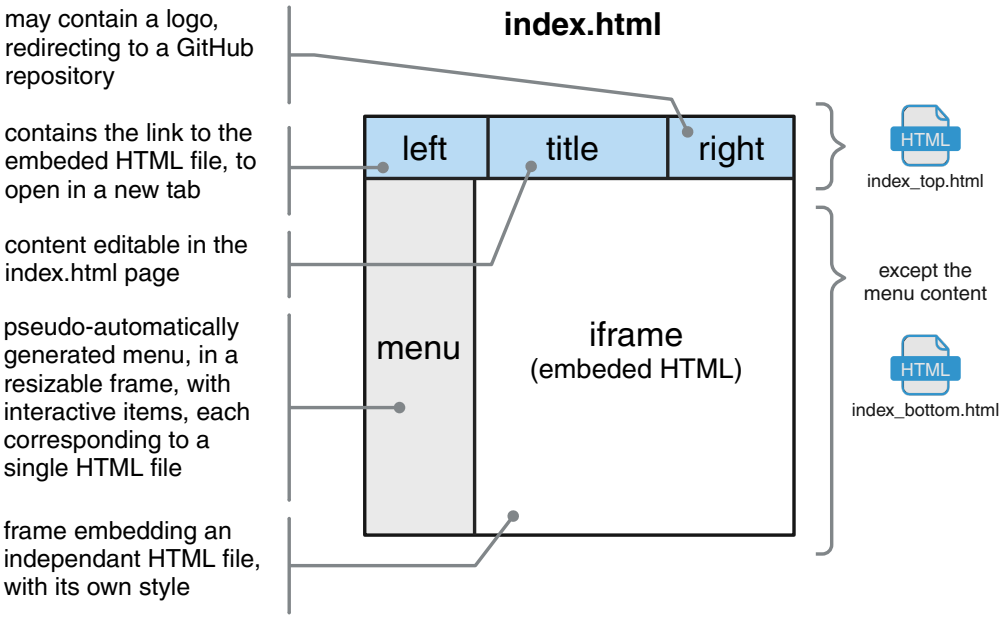


Figure 64 – **Structure of the index.html page**. The index page is structured in five blocks, with their functionalities indicated on the left. The index.html page content is pre-defined by the index_top.html and index_bottom.html fragments, excepting the menu content (see main text).

Two HTML fragments help to build this page: one for the top banner and another for the iframe, complemented by custom styles for the page layout (Figure 64, right). These two fragments need to be integrated with the menu, positioned in the middle of the final index.html file.

The compiled HTML files are organized in nested directories (Figure 62), which varies for each project. Therefore, a **semi-automated** approach generates the menu and the index.html page (Figure 65). Firstly (orange line), a function recursively identifies all HTML files in a specified directory, generating a menu (site-index.html) implemented in the HTML language. Secondly (green lines), the index_top.html file is concatenated with the menu and the index_bottom.html files. Thirdly (magenta line), the files encoding the graphical style of the index page are copied is the directory of interest. Finally, the index.html can be manually adjusted to customize the title, content of the upper right corner, menu tree structure, and item names. By default, the latter correspond to the names of the individual HTML files. Styles can also be customized.

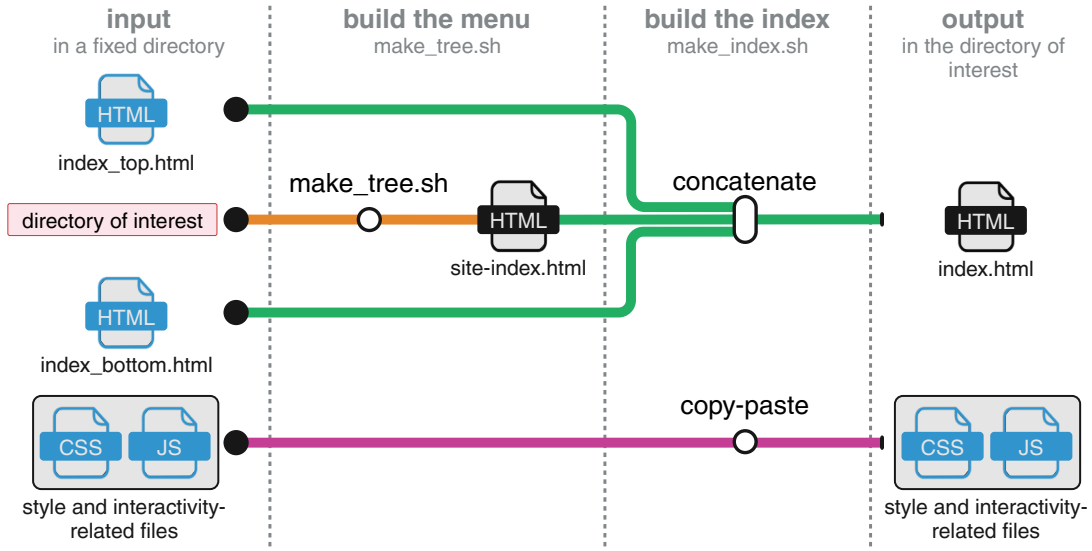


Figure 65 – Three-step pipeline to generate the index.html page. Blue elements denote input files, red indicates the customizable input, and new files are depicted in black. Main processes are represented by white dots. The steps are color-coded branches. This pipeline runs from the command line.

Together with the notebook tree, the index.html and associated files can be included in a Github repository and integrated as Github pages. A concrete example is provided in Section 2.3.1.

1.5 Summary

A dedicated pipeline has been developed to ensure the traceability of the reference transcriptome construction and ease the future update. Then, our existing pipeline to process reads and generate the count matrix was enhanced. It was executed successfully on two distinct HPC clusters, enabling to ensure reproducible processing. This pipeline is supported by a documentation to be reused.

A significant enhancement lies in the processing of the count matrix. A R package has been developed to document and track the code changes in custom functions. This package includes a robust method for annotating single cells. An analysis framework has been established to ensure the reproducibility of the count matrix processing, up to the figures associated with publications. This framework uses a modular R Markdown notebook tree connected to a directory structure. The notebooks are compiled with a Singularity container. An index page facilitates navigation among all compiled notebooks. This index page is included on a Github page alongside each publication. The method to build this index page is publicly available and documented to be reused.

The lab designed an *Nf1*-KO mouse model in which bi-allelic loss of *Nf1* and permanent expression of a fluorescent reporter were targeted in boundary cap cells and their derivatives. Mutant mice develop hyperplastic nerves and skin trauma accelerates their progression into pNFs (Article 1) and ultimately MPNSTs (Article 2).

2 Article 1

Plexiform neurofibromas (pNFs) are slow growing peripheral nerve sheath tumors that can evolve in MPNSTs. To enable early detection and develop preventive therapeutic solutions, a characterization of their initiation is required.

2.1 Overview of the study

The aim of this study was to identify the cellular and molecular events that precede pNF formation using our mouse model. We used immunohistochemistry and scRNA-Seq methods to understand the impact *Nf1* loss in Schwann cells (SCs) alone or in combination with *Nf1* haploinsufficiency in the micro-environment and skin trauma, on subcutaneous nerves. My contribution was to analyze the scRNA-Seq data, design the figures and participate in the formalization of the results and write the manuscript.

In wild-type mice, the skin trauma is not sufficient to induce macroscopic, cellular or transcriptomic changes in nerves. The loss of *Nf1* in SCs alone leads to a local increase of the nerve thickness (hyperplasia). This was related to an increased proportion of mutant SCs, further depicting a pro-fibrotic profile. The pro-fibrotic factor periostin is proposed as a therapeutic target to prevent fibrosis and nerve thickening.

The additional context of skin trauma or $Nf1^{+/-}$ background leads to a more severe macroscopic, cellular and transcriptomic phenotypes in nerves (Figure 66). This suggests that both triggers play similar roles. We then focused on SCs, fibroblasts and immune cells by combining all scRNA-Seq datasets in a unprecedented transcriptomic atlas. Compared to the controls or mutant condition without trauma, the proportion of **pro-fibrotic *Nf1*$^{-/-}$ (*Tom*+) SCs** increased at the expense of *Tom*- SCs. As the increased proliferative activity of mutant SCs was not confirmed, we suggest either slow proliferation or rapid and transient proliferation preceding nerve hyperplasia. Among fibroblasts, the main feature was the **expansion of endoneurial fibroblasts** with over-expression of the TNF signaling pathway. In immune cells, the **expansion of *Cx3cr1+* endoneurial macrophages** was stronger in the context of trauma compared to *Nf1* haploinsufficiency. These three features (in bold) are maintained in pNFs.

To conclude, we have mapped the transcriptomic heterogeneity of subcutaneous nerves from wild-type and *Nf1*-KO mice for the first time. Hyperplastic nerves are thicker nerves that precede the development of pNFs. The skin trauma or *Nf1* haploinsufficiency accelerates their progression into pNFs (Figure 66).

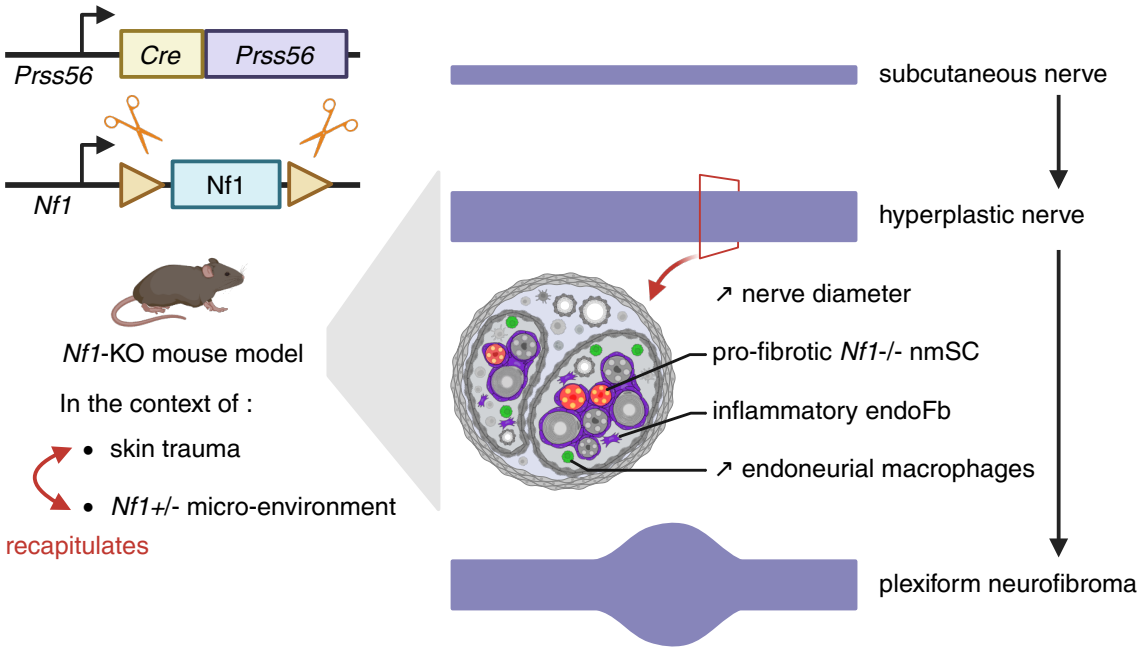


Figure 66 – Graphical abstract of the Article 1. *Nf1*-KO mice develop hyperplastic nerves characterized by the expansion of tumor SCs with a pro-fibrotic profile. Skin trauma or, equivalently, *Nf1* haploinsufficiency in the nerve micro-environment, triggers the expansion if (i) endoneurial fibroblasts (endoFb) with over-expression of TNF signaling and (ii) endoneurial macrophages. They are thought to be the first events preceding plexiform neurofibroma formation.

Note: The manuscript has been submitted to the *Journal of Investigative Dermatology* and is under revision. The Github link to make code available is broken because the repository is private. It will be made public once the manuscript has been accepted.

The layout of the manuscript has been improved using a customized version of the Oxford University Press free layout, under the LaTeX Project Public License.

MANUSCRIPT UNDER REVISION, IN THE JOURNAL OF INVESTIGATIVE DERMATOLOGY

Exploring the mechanisms governing the initiation of plexiform neurofibromas using Prss56Cre, *Nf1*-KO mouse model at a single-cell resolution

Myriam Mansour[1,†], Audrey Onfroy[1,†], Fanny Coulpier[1], Katarzyna J. Radomska[1], Pierre Wolkenstein[2] and Piotr Topilko[1,*]

[1]Institut Mondor de Recherche Biomédicale, Inserm U955 Team 9, Créteil, France and [2]Department of Dermatology, Reference Center for Neurofibromatoses, Henri-Mondor hospital, AP-HP, Créteil, France
[†]These authors contributed equally to this work[*]Corresponding author: piotr.topilko@inserm.fr

Abstract

About half of patients with the genetic disease neurofibromatosis type 1 (NF1) develop benign nerve sheath tumors, called plexiform neurofibromas (pNFs). Despite important advances in understanding the pathogenesis of pNFs, mainly due to the plethora of dedicated mouse models, the mechanisms responsible for the initiation of this process remain poorly understood. Here, we used a *Nf1*-KO mouse model targeting biallelic loss of Nf1 in boundary cap cells on a wild type and heterozygous Nf1 background to explore the early events driving pNFs development. All mutants develop subcutaneous hyperplastic nerves with some progressing into pNFs from one year of age. We discovered that skin trauma accelerates this process, highlighting the role of inflammation. While skin trauma has no effect on control subcutaneous nerves, in the mutant nerves on a wild-type background, we observed an expansion of mutant Schwann cells expressing a profibrotic program. Unexpectedly, a similar response was observed in the mutant nerves on a heterozygous background without skin trauma, highlighting the impact of Nf1 heterozygosity in the nerve micro-environment on the development of pNFs.

Key words: neurofibromatosis type 1, plexiform neurofibroma, subcutaneous nerves, mouse model, Schwann cells

Introduction

Neurofibromatosis type 1 (NF1) is a complex monogenetic disease due to mutations in the *NF1* gene encoding a negative regulator of the RAS signaling pathway [1]. NF1 affects 1 in 3000 people worldwide and is characterized by a broad set of symptoms beginning in childhood and including café-au-lait macules, bone and eye lesions, learning disabilities, and neurofibromas (NFs), among many others [2]. Neurofibromas are benign nerve sheath tumors developing at the level of nerve roots/deep nerve bundles (plexiform neurofibromas; pNFs) and nerve terminals of skin (cutaneous neurofibromas; cNFs). Both types of tumors originate from biallelic loss of NF1 in Schwann cells (SCs) lineage [3]. In addition to mutant SCs, NFs are composed of fibroblasts, a complex immune landscape, and are densely vascularized and innervated [4–6]. Despite their common glial origin and benign character, pNFs and cNFs show differences. PNFs are developing prenatally or in early childhood in about 50% of NF1 patients while cNFs emerge in almost all patients starting from puberty [7]. PNFs growth is slow but continuous and their number per patient is low while the number of cNFs may reach thousands and their growth is transient [2]. Finally,

about 10% of pNFs progress into malignant peripheral nerve sheath tumors (MPNSTs) while the malignant progression of cNFs was never reported [7, 8].

Important progress, mostly due to a plethora of *Nf1* genetically engineered mouse models, was recently achieved in decoding cellular and molecular mechanisms driving the development of pNFs. In those models, biallelic inactivation of *Nf1* was targeted into the SCs lineage at successive stages of their development [9–12]. Interestingly, while loss of *Nf1* in SCs precursors, immature and non-myelinating SCs leads to the development of pNFs, inactivation of *Nf1* in myelinating SCs remains asymptomatic [13]. Such an observation suggests that SCs in their immature state constitute the population at the origin of pNFs. Interestingly, nerve trauma in mice with *Nf1* loss in myelinating SCs promotes their dedifferentiation and development of pNFs-like tumors further supporting the immature/inflammatory state of SCs as a prerequisite for pNFs development [14, 15]. Recent studies showed that mutant SCs orchestrate the development of pNFs through their interaction with the complex nerve micro-environment, mostly composed of fibroblasts, immune cells, axons, and blood vessels [8, 16]. One of the flagship molecules is stem cell factor (SCF) secreted by mutant SCs that binds to the receptor kit on mast cells, which in turn stimulates collagen deposition through activation of fibroblasts by the transforming growth factor β (TGFβ) [17, 18].

Despite such progress, the mechanisms governing the initiation of pNFs tumorigenesis remain to be decoded. In this study, we used the *Nf1*-KO mouse model, carrying biallelic *Nf1* loss and expression of fluorescent reporter Tomato targeted to a subpopulation of SCs originating from boundary cap (BC) cells, to address the early events and triggers responsible for pNFs development. We discovered that skin trauma and *Nf1* heterozygosity in the SCs micro-environment promote subcutaneous nerve enlargement defined as hyperplasia. While in the mutant nerves on the wild type (wt) background, the skin trauma is necessary to promote mutant SCs proliferation, *Nf1* heterozygosity in the nerve micro-environment in the absence of trauma is sufficient to mimic such effect. Finally, we observed that mutant SCs express pro-fibrotic marker periostin whose role is broadly known in the regulation of fibrosis and inflammation in different situations including cancer [19, 20].

Results

Skin trauma promotes subcutaneous nerve hyperplasia in Nf1 mutant mice

We have conceived a *Nf1* mutant mouse model (*Nf1*-KO) carrying simultaneous biallelic *Nf1* inactivation and permanent expression of fluorescent reporter Tomato (Tom) into *Prss56*-expressing BC cells and their derivatives, *Prss56* being a specific marker of BCs. BCs form clusters of stem-like cells transiently located during embryonic development at the dorsal root entry zone and ventral exit point of cranial and spinal nerves [21]. Genetic fate mapping of BCs using the *Prss56*^{Cre} line identified derivatives migrating into the nerve roots and along spinal nerves into nerve endings of skin where they differentiate into myelinated and non-myelinated SCs [22]. While in the nerve roots, almost all SCs originate from BCs, in the distal segments of spinal nerves (hypodermis, dermis, and epidermis) BC-derived SCs represent about 30% of the glia, the remaining cells originating from the neural crest [22]. We have previously reported that *Nf1*-KO mice on wt; *Nf1*-KO^{wt} (*Prss56*^{Cre}, *Nf1*$^{fl/fl}$, *R26*^{Tom}) and Nf1 heterozygous; *Nf1*-KO^{het} (Prss56^{Cre}, *Nf1*$^{fl/-}$, R26^{Tom}) background develop cNFs and pNFs starting from one year. PNFs are of two types, paraspinal, emerging at the level of dorsal root ganglia (DRG), and subcutaneous, developing along nerves in the hypodermis (subcutaneous nerves; snerves) (**Figure 1A**) [23]. It's well known that the grouped house males, but not females, are naturally aggressive and have numerous skin bites. In the grouped house mutant males with skin trauma, we observed accelerated development of cNFs and subcutaneous but not of paraspinal pNFs [2].

Since, in this model, *Nf1* inactivation takes place during embryogenesis and the first symptoms appear several months after birth, we

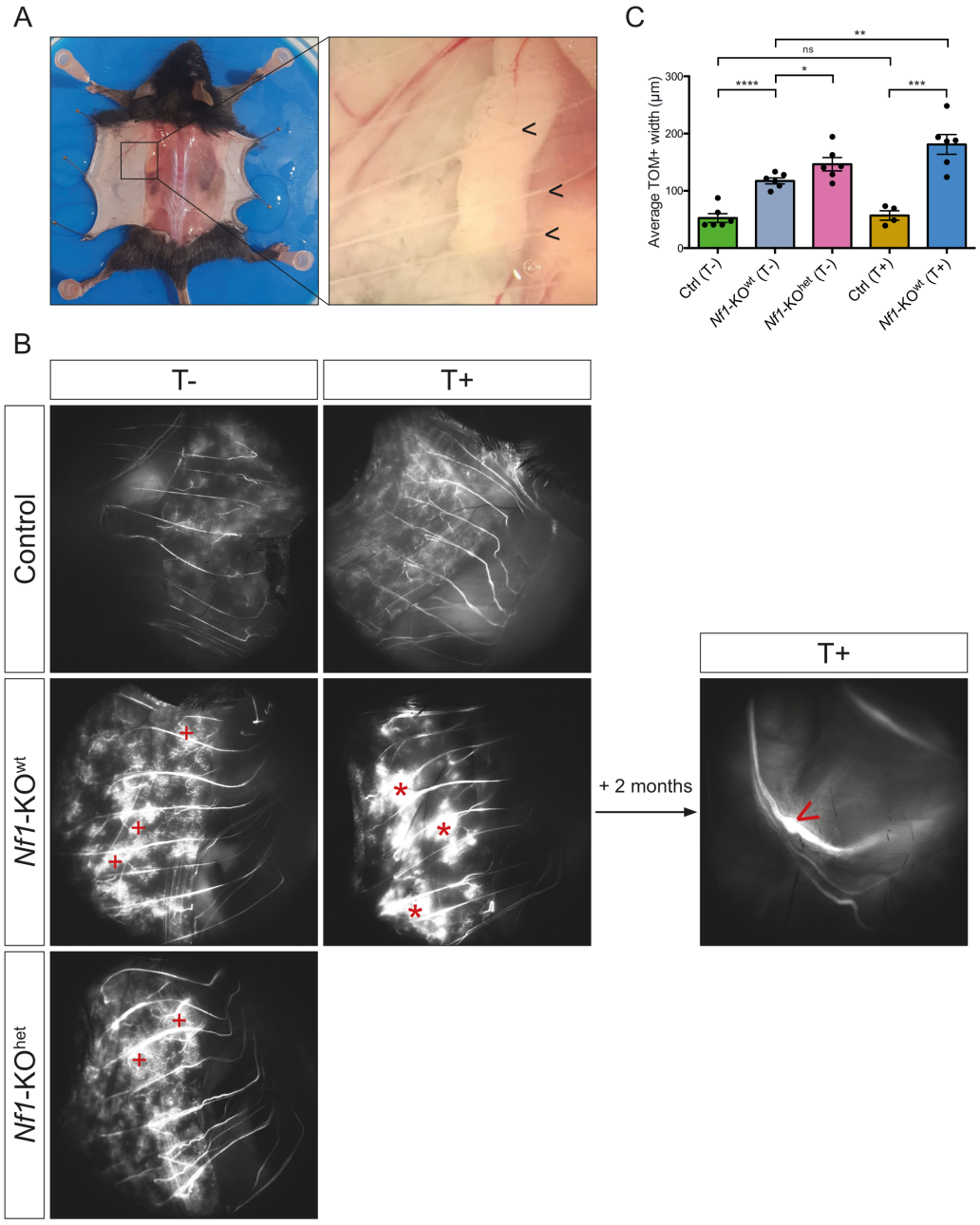


Figure 1. Macroscopic visualization of subcutaneous nerves from *Nf1*-KO mice. (A) Metameric organization of snerves (marked by <). (B) Macroscopic epifluorescent (Tom+) imaging of snerves from control, *Nf1*-KO^{wt} and *Nf1*-KO^{het} 3-month-old male mice, kept separately (without trauma; T-) or grouped house (with trauma; T+). Mutant snerves are enlarged with an increased proportion of Tom+ cells. Note the presence of numerous cutaneous neurofibromas (*) in grouped house mutant males. In the context of trauma, progression into pNF (marked by <) has been observed in 5-month-old *Nf1*-KO^{wt} mice. (C) Average TOM+ snerve width, as mean by mouse +/- SEM (n = 6 nerves, 3-5 mice/experimental condition). Non-parametric t-test p-value: ns: non-significant, *: pvalue < 0.05, **: p-value < 0.01, ***: p-value < 0.001, ****: p-value < 0.0001.

characterized snerves from control and 3-month-old mutant males kept separately (without trauma; T-) or grouped house (with trauma; T+). Finally, we explored the role of *Nf1* heterozygosity in the

nerve micro-environment by analyzing snerves from *Nf1*-KO^{het} males. All results presented in this study are based on analyses of at least 20 snerves/animal, distributed bilaterally and issued from 3 animals.

First, observing snerves under the epifluorescent macroscopic showed no differences in the thickness of snerve segments populated by Tom+ cells across T- and T+ control mice, suggesting that skin trauma has no major impact on Tom+ SCs (**Figures 1B and C**). To validate such observations at the cellular level, we performed immunostaining of snerves from T- and T+ control mice with markers of all SCs (SOX10), all fibroblasts (PDGFRα), and subpopulations (SOX9, CLDN1, ZO-1), all immune cells (CD45) and macrophages (IBA1), the latter considered as a major immune component of injured nerves (**Figure 2A**) [5]. No significant differences were observed in the cellular composition of control snerves supporting results from macroscopic analysis (**Figure 2B**).

Then, we showed that the thickness of snerve segments populated by Tom+ cells is increased by 216% in T- *Nf1*-KO^{wt}, by 334% in T+ *Nf1*-KO^{wt}, and by 270% in T- *Nf1*-KO^{het} compared to control snerves (**Figures 1B and C**). We defined this local increase in snerve thickness as hyperplasia. Immunostaining of mutant and control snerves without trauma with a marker of proliferation (Ki67) did not identify any Tom+ dividing cells suggesting either rapid but transient proliferation of mutant SCs at the earlier period or low but constant proliferative activity, as previously reported [9]. Immunostaining of mutant snerves (*Nf1*-KO^{wt} T-, *Nf1*-KO^{wt} T+ and *Nf1*-KO^{het} T-) with markers of SCs (TOM, SOX10) showed an increase of the Tom+ SCs proportion (**Figures 1B and C**). Such results suggest that while skin trauma in grouped house males has no effect on Tom+ SCs from control snerves the same trigger promotes expansion of Tom+ mutant SCs in mutant nerves. Finally, in some *Nf1*-KO^{wt} males grouped house for more than 2-3 months, we observed the development of encapsulated subcutaneous pNFs characterized by numerous tumor SCs detached from axons with abnormal multipolar morphology and decompacted perineurium, suggesting that hyperplastic nerves are at their origin (**Figures 1B and S1**). Overall, our results suggest that in the *Nf1*-KO mice, skin trauma induced by bites and *Nf1* heterozygosity in the nerve micro-environment promotes local nerve hyperplasia with some progressing into pNFs.

Skin trauma does not modify the transcriptomic signature of subcutaneous control nerves

To explore the changes in cell composition and transcriptomic activities that take place in *Nf1* mutant snerves, we performed scRNA transcriptomic analysis of control and mutant snerves using 10x Genomics technology. We focused our analyses on the role of (i) *Nf1* loss in Tom+ SCs, (ii) *Nf1* heterozygosity in the snerve micro-environment, and (ii) skin trauma in grouped house males. To do so, snerves from control and *Nf1*-KO^{wt} and *Nf1*-KO^{het} males with and without skin trauma were collected, dissociated and isolated cells were processed for scRNA-Seq. Two replicates/experimental conditions, except for T- *Nf1*-KO^{wt} (n=1), using 20 nerves/animal (3 animals/experimental condition, 60 nerves in total) were used and analyzed. Each biological replicate included between 6,000 and 14,000 cells. We added, to this set, two recently generated datasets from scRNA-Seq of subcutaneous pNFs issued from grouped house *Nf1*-KO^{wt} males (Radomska *et al.*, manuscript in preparation). All datasets were merged to form a transcriptomic atlas of snerves of 90,159 cells (**Figure 3A**). Since the cellular composition and transcriptomic signature for each cell type were recently established for the adult sciatic nerve [24, 25], we used this dataset as a reference to compare with our scRNA-Seq results from T- and T+ control snerves.

Both sciatic and control snerves include 13 distinct cell populations based on a cell specific markers based annotation (**Figure S2D**). The predominant population is made up of three types of nerve fibroblasts: epineurial (epiFb: Clec3b+, 70.8%), perineurial (periFb: Tjp1+, 0.8%), and endoneurial (endoFb: Lumhigh+, 3.1%). The second major population corresponds to immune cells (Ptprc+) and includes macrophages (16.9%), dendritic cells (2.2%), T lymphocytes (0.6%), NK cells (0.8%), mast cells (0.2%),

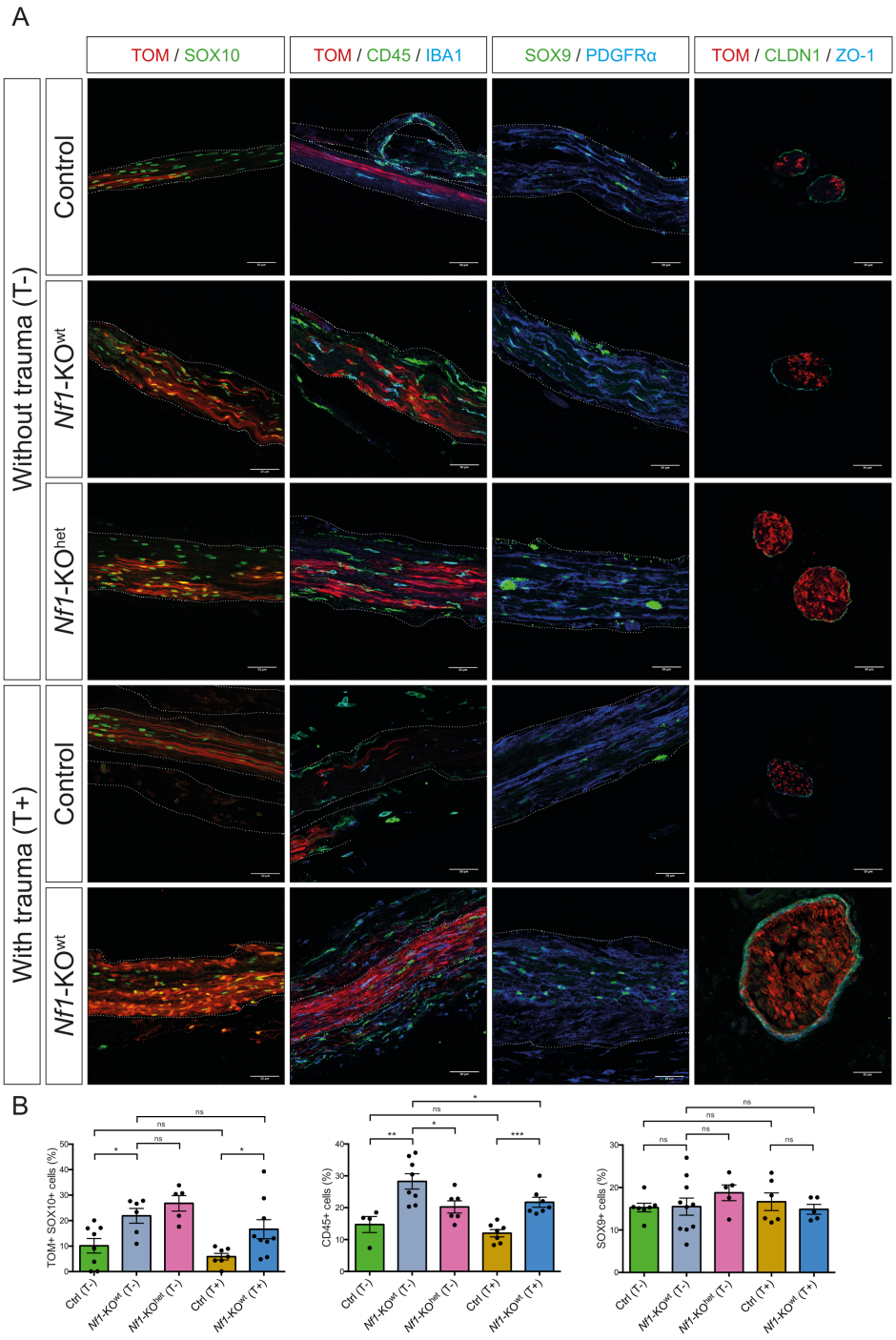


Figure 2. Immunohistochemical characterization of subcutaneous nerves. (A) Immunofluorescence analysis of longitudinal and transverse sections of snerves from 3-month-old control, *Nf1*-KO^{wt} and *Nf1*-KO^{het} mice, kept separately (T-) or grouped house (T+), with markers of BC-derived Schwann cells (TOM), all Schwann cells (SOX10), all immune cells (CD45), macrophages (IBA1), all nerve fibroblasts (PDGFRα), endoneurial (SOX9) and perineurial (CLDN1 and ZO-1) cells. Scale bar: 50 µm. (B) Quantification of TOM+ SOX10+ Schwann cells, CD45+ immune cells, and SOX9+ fibroblasts in each experimental condition, over the total number of DAPI+ cells. Non-parametric t-test p-value: ns: non-significant, *: p-value < 0.05, **: p-value < 0.01, ***: p-value < 0.001, ****: p-value < 0.0001.

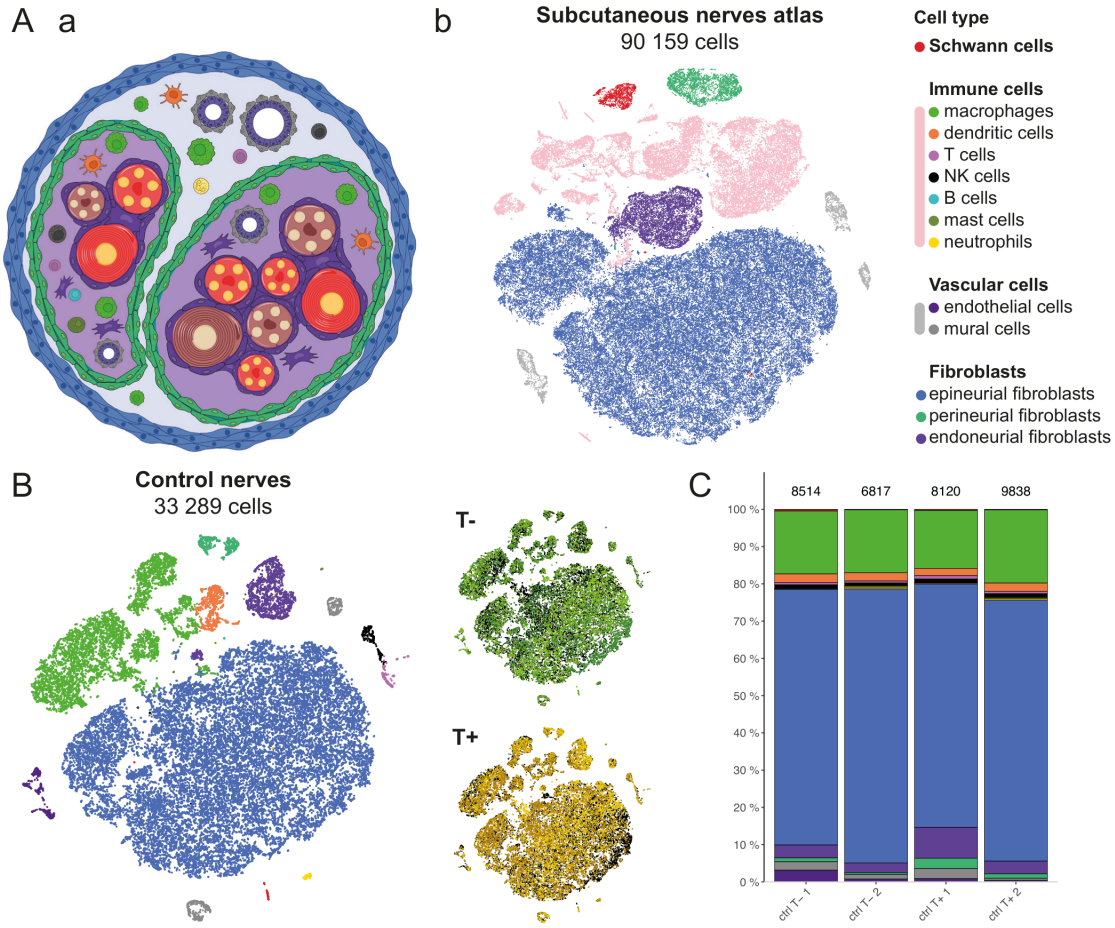


Figure 3. Transcriptomic atlas of control and *Nf1*-KO subcutaneous nerves. (A) Single-cell snerve transcriptomic atlas (n = 11 independent datasets). (a) Schematic representation of nerve architecture including cell types (designed using BioRender), (b) tSNE visualization of 90,159 cells, colored according to cell type. Immune and vascular cells were colored in pink and gray respectively. See **Figure S2A** to visualize cell populations. (B) tSNE presentation of 33,289 cells from control datasets (n = 4 independent datasets), colored according to cell types (left) or by experimental condition (right). In the right panel, each replicate (n = 2) within each experimental condition (T- or T+) is delineated as colored cells, while the background, corresponding to the other experimental condition, is represented as a black background. This tSNE is not corrected for sample effect. The right panel shows that all 4 samples (n = 2/experimental condition) are perfectly overlapping, ming there are no populations with unique transcriptomic signatures. (C) Barplot showing cell populations representativeness in each control dataset. The text above bars corresponds to the number of cells in each dataset.

B lymphocytes (0.1%) and neutrophils (0.3%) (**Figure S2C**). The remaining populations correspond to endothelial cells (Egfl7+, 2.1%), mural cells (Des+, 1.8%), and rare SCs (Plp1+, 0.3%). A subpopulation of SCs expresses Tom and corresponds to BCs-derived glia (**Figure S3C**). Such a low proportion of SCs is likely due to the incomplete snerves dissociation since we observed numerous Tom+ cells attached to myelin debris that were eliminated during the cell isolation process. All 13 cell populations were identified in the sciatic nerve indicating that the global cell composition of adult snerves and sciatic nerve is the same. Interestingly, our results demonstrate that the snerves are still enveloped by the epineurium since this highly vascularized and composed of the rich connective tissue mantle nerve layer is no longer present in the distal part of the snerves entering the dermis.

Next, we explored the impact of skin trauma in control mice on cellular composition and transcripto-mic signature of snerves (**Figures 3B-C**). To do so, we extracted, from the global dataset, 33,289 cells corresponding to snerves isolated from T+ and T- control mice. Then, we analyzed the intra-experimental condition heterogeneity by comparing two replicates issued from the same experimental condition and finally, we compared results across experimental conditions. Barplot for individual datasets showed that intra-experimental condition heterogeneity is slightly higher versus interexperimental condition differences (**Figure 3C**) revealing no major differences in cell composition between snerves isolated from T+ and T- control males. Immunolabelling on T+ and T- snerves with specific markers of SCs (TOM, SOX10), all immune cells (CD45), macrophages (IBA1), fibroblasts (PDGFRα, SOX9), perineurial fibroblasts (CLDN1, ZO-1) did not identify differences (**Figures 2A and B**).

Next, we compared the transcriptomic signature of each cell population between T+ and T- control snerves. First, we performed principal component analysis (PCA) and project cells using tSNE (**Figure 3B**), without applying sample-effect correction on PCA space. Thus, tSNE was used to identify any new cell population based on overlapping between each dataset. No condition-specific population was identified showing that transcriptomic signatures for each cell population across T+ and T- control snerves are similar (**Figure 3B**, right panel). Overall, our observations reveal that skin trauma induced by bites is probably too superficial to promote snerve injury and inflammation, in control males.

Pro-fibrotic signature is a hallmark of Nf1-/- Schwann cells from hyperplastic snerves

Since *Nf1* mutant SCs are at the origin of the hyperplastic snerve phenotype, our initial focus was on this cell population. We extracted 1,046 SCs from the atlas and projected them using UMAP (**Figure 4A**). The dataset was annotated using a specific panel of SCs markers of myelinating (mSCs) and non-myelinating (nmSCs) subpopulations (**Figure S3B**). The majority of SCs issued from the snerves correspond to non-myelinating glia, which is consistent with the architecture of snerves, primarily composed of thin, poorly, or unmyelinated sensory axons (**Figures 4Aa and C**). Transcriptomic analyses led us to the following observations. First, there were no significant differences in cellular composition and transcriptomic signature between Tom+ and Tom- SCs from the control nerves, showing that despite their distinct, BCs and neural crest, origin, they share a common molecular signature (**Figures 4A, C, and S3A, C**). Second, in the control snerves, the proportion of Tom+ SCs and their transcriptomic signature remained unchanged in the presence of trauma, confirming that the skin bites do not affect wt Tom+ SCs (**Figures 4B-C and S3A, C**). Third, no differences in transcriptomic signature were observed between wt and Nf1+/- SCs, demonstrating that Nf1 heterozygosity in SCs in the absence of trauma does not affect their transcriptomic activities (**Figures 4A, C, and S3A**).

Next, we compared *Nf1* mutant SCs isolated from T- snerves with those issued from T+ snerves and snerves on Nf1+/- genetic background. We observed that while Tom+ SCs represent 25% of all SCs in the control snerves (T+ and T-), this proportion is almost double (47%) in the T- mutant snerves and even higher (70%) in the presence of trauma or an Nf1+/- background (**Figure 4B**). This suggests that *Nf1* loss in SCs affects their proliferative activity, which is further enhanced by trauma and the Nf1+/- background. Bioinformatic analysis revealed that SCs from snerves are grouped in two subclusters (**Figure S3D**). Wt and Nf1+/- SCs grouped in subcluster 1 and were separated from subcluster 2 formed by mutant SCs issued from T+ and Nf1+/- background snerves supporting differences in their transcriptomic signature. Unfortunately, the small number of Tom+ mutant SCs from T- snerves on wt background (8 cells) does not allow us to classify them into one of the two subclusters. To further identify transcriptional differences between those 2 subclusters, we performed a differential expression (DE) analysis. Using two distinct DE methods, we identified 715 up-regulated and 27 down-regulated genes in subcluster 2, the most significant were

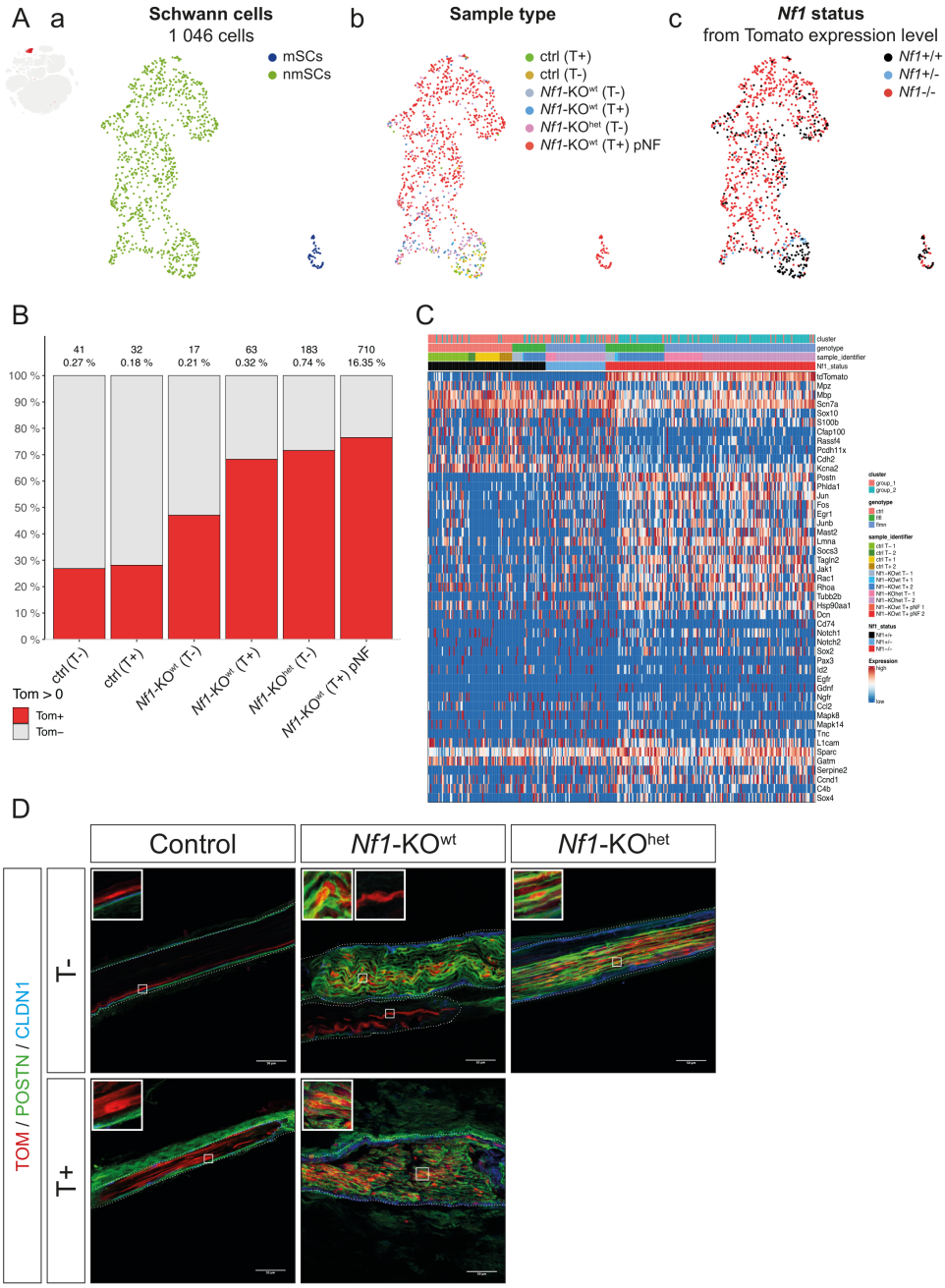


Figure 4. Transcriptomic signature of control and mutant Schwann cells from subcutaneous nerves. (A) Inset on the left represents all SCs, in red, extracted from the atlas. The 1,046 SCs are then represented on a UMAP colored by: (a) subtype, (b) experimental condition and (c) *Nf1* status, inferred from Tomato expression level (**Figure S3C**) and genotype. (B) Barplot showing ratio of BC-derived SCs (Tom+) versus neural crest derived SCs (Tom-) in each experimental condition. The number and proportion of SCs in the population among all cells within the experimental condition are indicated. (C) Heatmap showing expression of genes in SCs, across experimental conditions. Gene expression is log-normalized, then scaled across cells. Cells from pNFs are not depicted in this heatmap. (D) Immunostaining of longitudinal sections of control and mutant snerves with markers of (a) BC-derived SCs (TOM), periostin (POSTN), perineurial cells (CLDN1) and (b) myelin binding protein (MBP). Scale bar: 50 µm. Abbreviation: mSCs, myelinating SCs; nmSCs, non-myelinating SCs.

represented on a Heatmap (**Figure 4C**). Such imbalance in the proportion of DE genes is likely due to globally higher transcriptional activity in *Nf1* mutant SCs. Among the top upregulated genes, we have identified a profibrotic factor periostin (Postn), Phlda1, and Sox2 (**Figures 4C and S3E**). Postn expression in control and mutant snerves was analyzed by IHC. In T- and T+ control snerves, Postn expression is restricted to perineural cells (CLDN1+) and absent from Tom+ SCs. In mutant snerves, in addition to perineural cells, we observed a subpopulation of Tom+ SCs expressing high levels of Postn (**Figure 4Da**). Finally, we showed that some mutant Tom+ SCs assembled in small clusters remain Postn-. To differentiate between myelinating and non-myelinating SCs, we performed IHC with MBP, a marker for compact myelin. We observed that the Tom+, Postn+ population corresponded to non-myelinating glial cells, while the Tom+, Postn- expressed high levels of MBP and corresponded to myelinating SCs (**Figure 4Db**). These observations demonstrate that loss of NF1 in BC-derived SCs specifically activates expression of the pro-fibrotic marker Postn in non-myelinating SCs without affecting myelinating SCs. Interestingly, a high level of Postn expression has also been observed in mutant SCs from pNFs, making this molecule a potential biomarker of the early development of pNFs (**Figures S1d and S3E**).

Phlda1 encodes Pleckstrin Homology-Like Domain family A member 1 and emerges in many studies as a critical mediator of resistance to FGFR inhibition in cancer [26]. Interestingly, Phlda1 and Sox2 expressions are also maintained in mutant SCs from pNFs (**Figure S3E**). In addition to Postn, Phlda1, and Sox2, we observed an upregulation in mutant SCs from cluster 2 of several immediate early genes (IEGs) including Fos, Jun, Junb, and Egr1 (**Figure 4C**). Since the expression of this set of genes was not observed in SCs from the control condition, we believed that such signature is rather associated with *Nf1* loss alone or in combination with trauma than with stress associated with tissue dissociation procedure [27]. Moreover, overexpression of the same panel of IEGs in *Nf1* mutant SCs was also reported in pNFs from our model, further supporting their implication in the pathogenesis of those tumors. Interestingly, overexpression of Postn, Phlda1, Sox2, and the IEGs was also reported in SCs from the injured sciatic nerve and is associated with repairing SCs phenotype suggesting that *Nf1* loss in SCs in the presence of Nf1+/- background partly mimics such phenotype.

Endoneurial fibroblasts from hyperplastic nerves overactivated the TNF signaling pathway

Peripheral nerves are protected by nerve fibroblasts, which are considered the primary nerve component and can be classified into three distinct populations: epiFb, periFb, and endoFb, based on their location in the nerve and functions. Since in pNFs, the nerve fibroblasts play a role in secreting the molecules of the extracellular matrix, we performed a comparative analysis of the proportions and transcriptomic activities of each type of fibroblasts between control and *Nf1* mutant snerves.

First, we focused on epiFb, which constitutes 87% of all nerve fibroblasts in our nerve atlas and forms a peripheral nerve mantle layer. 56,957 epiFb cells were extracted from the atlas (**Figure 5Aa**). The population of epiFb could be further divided into two subpopulations; A (orange) and B (green), distinguished by their specific gene expression profiles. The epiFb cells from subpopulation A express Nav3, Calclr, Pla1a, Limch1, and Sema3e, while those from subpopulation B are characterized by the expression of Gdf10, Cxcl12 and Cxcl14 (**Figure 5Ba**). Interestingly, such two types of epiFb have not been described in the adult sciatic nerve [24]. In the mutant versus control snerves comparison, we did not observe any differences in the proportion of epiFb and their transcriptomic signature, irrespective of the experimental conditions (**Figure 5Bb**). However, in pNFs from our mouse model, we mainly identified Gdf10+ epiFb belonging to the subpopulation B (**Figure 5Bb**). This suggests either enhanced proliferation of the B subpopulation, or A to B switch in transcriptomic signature of the subpopulation A.

PeriFb forms a protective barrier around nerve bundles, known as the perineurium. Its primary function is to protect nerves against

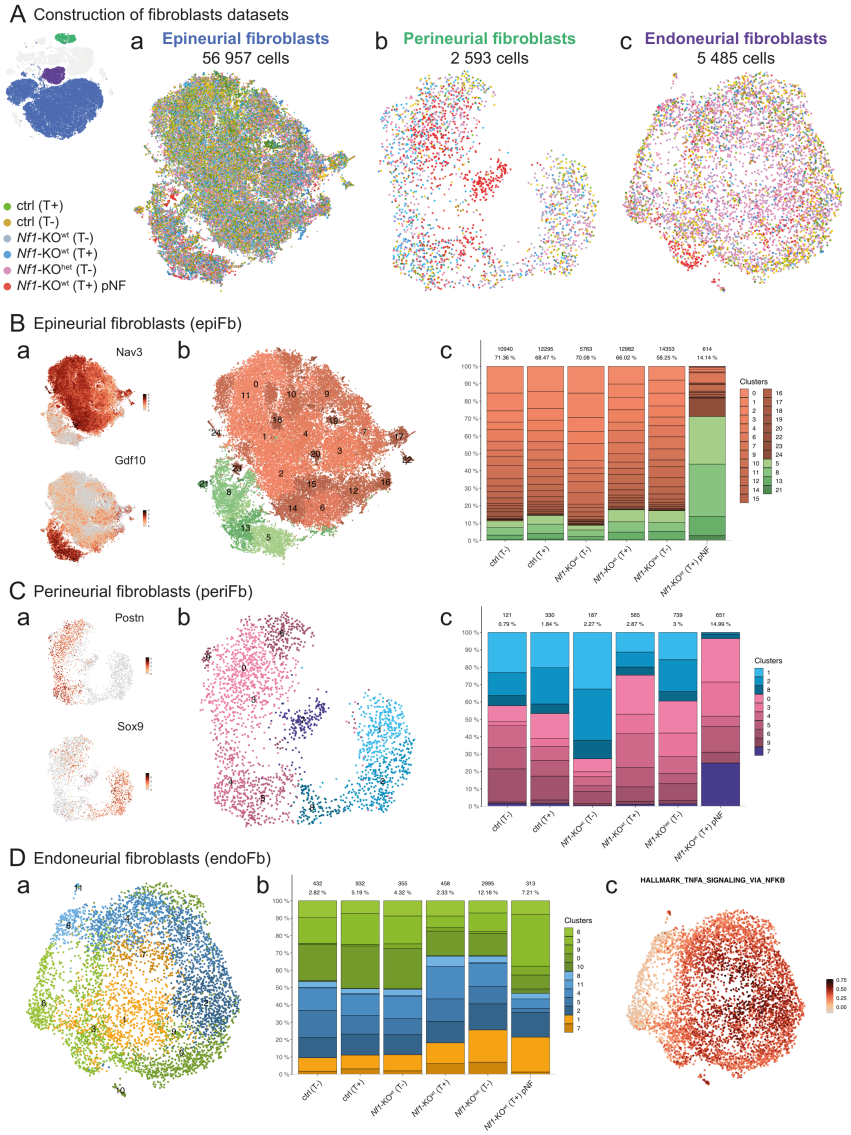


Figure 5. Transcriptomic signature of subcutaneous nerves derived fibroblasts. (A) Inset on the left represents all fibroblasts from the atlas. 2D representation of each type of snerve fibroblasts including: (a) 56,957 epineurial fibroblasts (epiFb), (b) 2,593 perineurial fibroblasts (periFb) and (c) 5,485 endoneurial fibroblasts (endoFb) colored by experimental condition. (B) Transcriptomic signature of epiFb (a) log-normalized gene expression levels identifying two subpopulations among epiFb. The orange population is Nav3+ while the green population is Gdf10+. (b) Barplot showing clusters representativeness in each experimental condition. The number and proportion they represent among all cells within the experimental condition are indicated. (C) Transcriptomic signature of periFb (a) log-normalized gene expression level of two subpopulations of perineurial Fb. The blue population is Sox9+ while the pink population is Postn+. Purple subpopulation is specific to the pNFs dataset. (b) Barplot showing clusters representativeness in each experimental condition. The number and proportion they represent among all cells within the experimental condition are indicated. (D) Transcriptomic signature of endoFb (a) endoFb clustering grouped into three subpopulations based on gene expression levels gradient. (b) Barplot showing clusters representativeness in each experimental condition. The number and proportion they represent among all cells within the experimental condition are indicated. (c) Representation of HALLMARK_TNFA_SIGNALING_VIA_NFKB genes set score, computed using Seurat's AddModuleScore, showing score gradient from the surrounding cells to the middle yellow subpopulation.

pathogens and toxic compounds. We isolated and analyzed 2,593 periFb from our atlas, which accounted for 4% of the total fibroblasts (**Figure 5Ab**). In control nerves, periFb can be further divided into two subpopulations: A (blue) and B (pink). Subpopulation A is characterized by the expression of transcription factor Sox9, while subpopulation B expresses the profibrotic marker Postn (**Figures 5Ca**). In the mutant snerves, we observed an increased proportion of both subpopulations of periFb compared to controls (**Figure 5Cb**). However, we did not observe differences in their transcriptomic signature, regardless of the experimental conditions suggesting that molecular characteristics of periFb are not affected in mutant snerves. IHC analysis of control and mutant snerves using ZO-1 and CLDN1, both markers of compact perineurium, revealed no differences further supporting transcriptomic data (**Figure 2A**). Interestingly, in the pNFs from our model only the population expressing Postn was present (**Figure 5Cb**). This observation suggests either a transcriptional switch from Sox9+ to the Postn+ population of periFb or an expansion of the Postn-expressing Fb.

The endoFb infiltrates the nerve bundles and is the only population of fibroblasts in direct contact with SCs. To examine endoFb, we isolated and analyzed 5,485 endoFb from our atlas, accounting for 9% of all fibroblasts (**Figure 5Ac**). Notably, all endoFb are grouped in one cluster. However, we identified three distinct populations connected by partly overlapping transcriptomic signatures (**Figure 5Da**). Population A (blue) expresses higher levels of Dbi, Gng5, Serf2, and ribosome-protein coding genes while population B (green) shows higher expression of Nsd2, Neat1, Rian, Cdk8, and Lars2. Population C (yellow) over-expresses several IEGs in addition to Tnfaip1-6 and Nfkbia, both hallmarks of the TNF signaling pathway (**Figure 5Dc**). While in T- and T+ control snerves, endoFb is composed of 45% of cells from population A, 45% of cells from population B, and 10% cells from population C, we observed expansion of the population C to 25% in T+ mutant snerves (**Figure 5Db**). Interestingly, such expansion of endoFb with TNF signature was also observed in pNFs from our model.

Subcutaneous nerves from control and mutant mice share the complex immune landscape

Immune cells represent the second most abundant cell type, after fibroblasts, of the snerves. To further analyze the transcriptomic signature of each immune cell type, we isolated and analyzed 19,922 immune cells from our atlas, corresponding to approximately 21% of all cells (**Figure 6A**). Each class of immune cell was identified based on the expression of a panel of specific markers (**Figure 6B**) and was present in the snerves across all experimental conditions (**Figure 6C**). Among the immune cells, the predominant group corresponded to M1 (Plac8+, Chil3+) and M2 (Pf4+, Mrc1+) macrophages. The M2 population comprised epineurial and endoneurial macrophages, each type depicted by a specific transcriptomic profile and location within the nerve [28] (**Figure 6B**). Epineurial macrophages express Lyve1, Retnla and Clec10a and are present in the epineurium. Endoneurial macrophages, expressing Bcl2a1b, Dst, Creb5, and Dock4, are in close contact with SCs and endoneurial fibroblasts. This latter population can be separated into M2 Cx3cr1+ and Cx3cr1- macrophages. While the overall proportion of macrophages remains constant under the different experimental conditions, we observed a significant increase in the proportion of M2 Cx3cr1+ macrophages in T+ mutant snerves and T- mutant snerves on Nf1+/- background (**Figures 6C-D**) but not in the mutant T- snerves on wt background. While Cx3cr1+ and Cx3cr1- macrophages co-exist within mutant snerves, in pNFs, all macrophages express Cx3cr1+ (**Figure 6D**). Despite the important complexity of the other immune cell types populating snerves, we did not identify major differences in proportion or transcriptome signature between control and mutant snerves.

Discussion

Plexiform neurofibromas develop in about half of NF1 patients, but the mechanisms responsible for their initiation remain poorly characterized.

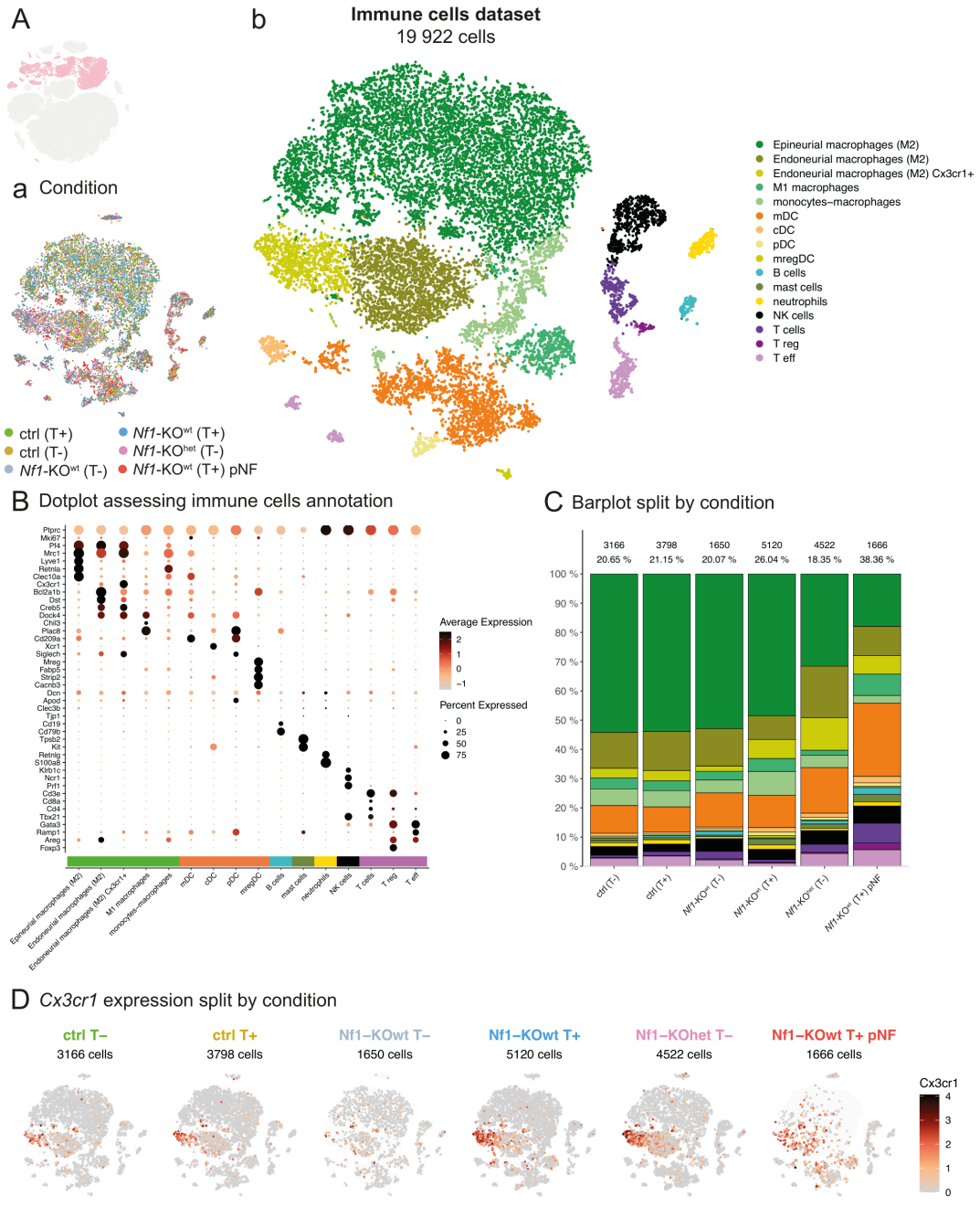


Figure 6. Transcriptomic signature of immune cells isolated from subcutaneous nerves. (A) Inset on the left represents immune cells from the atlas. All 19,922 immune cells are represented on a tSNE colored by experimental condition (a) or by immune cell type (b). (B) Dotplot assessing immune cell populations annotation. Dot size represents the proportion of cells expressing the corresponding marker. The dot color represents average gene expression levels for cells concerned by the dot. (C) Barplot showing immune cell type representativeness in each experimental condition. The number and proportion they represent among all cells within the experimental condition are indicated. (D) tSNE representations of Cx3cr1 expression levels, for each experimental condition. The expression level is log normalized and ranges from gray to black. A light gray background represents cells from the other experimental conditions. Note that in pNFs datasets, the majority of macrophages and a subpopulation of dendritic cells expressed Cx3cr1.

In this study, we explored an *Nf1*-KO mouse model that exhibits *Nf1* bi-allelic loss restricted to BC cells and their derivatives. Adult *Nf1*-KO mice develop paraspinal and subcutaneous plexiform neurofibromas. We have shown that the development of subcutaneous plexiform neurofibromas (pNFs) is preceded by the presence of hyperplastic snerves characterized by the local accumulation of Tom+ mutant SCs expressing high levels of Postn. Finally, we discovered that skin trauma in group-housed mutant males and *Nf1*+/- genetic background in the nerve micro-environment without trauma accelerated the development of pNFs but remained ineffective on paraspinal pNFs, highlighting the role of local triggers and *Nf1* heterozygosity in the nerve micro-environment as important players.

Our observations support the hyperplastic snerves origin of pNFs. First, hyperplastic snerves and pNFs share some commonalities, including an increased proportion of mutant SCs expressing Postn. Second, hyperplastic snerves are localized precisely where pNFs develop. Interestingly, the presence of numerous hyperplastic snerves but only a few pNFs suggests that their progression to pNFs is rare and probably requires additional local stimuli.

Postn expression is detected in mutant snerves and persists in pNFs, suggesting the role in neurofibromas pathogenesis. Recently, Postn is one of the most upregulated extracellular matrix proteins in paraspinal pNFs, with a 7,38 fold change compared to wt DRGs [29]. Known to contribute to immature SCs migration during embryogenesis [30], this protein is also expressed in repair SCs during nerve injury [31, 32]. This highlights the immature character of mutant SCs in hyperplastic nerves and pNFs.

scRNA-Seq analysis of all SCs isolated from control snerves enabled us to compare for the first time the transcriptomic signature of Tom+ BC-derived versus Tom- neural crest-derived glial cells. We showed that both populations include mSCs and nmSCs, but no differences in transcriptomic signature were identified for these populations. This demonstrates that despite their distinct origins, BC-derived and NC-derived SCs share similar molecular characteristics.

Unexpectedly, we have shown that in control males, bite-induced skin trauma has no impact on the transcriptomic signature of snerves. No signs of mechanical nerve trauma, including nerve defasciculation and SC proliferation, were observed further supporting this observation. Interestingly, under the same experimental conditions, mutant snerves behave differently since we observed an expansion in the proportion of Tom+ mutant SCs expressing Postn and a panel of IEGs. Since Prss56- expressing BCs give rise to nmSCs and mSCs in snerves, we wondered whether both types of mutant SCs responded to bite-induced skin trauma. We observed that despite their coexistence in the same snerve, only mutant nmSCs respond to skin trauma by proliferation and Postn expression. Our observations suggest that mutant but not control nmSCs or mutant mSCs, respond to inflammation present in the injured dermis. Moreover, the fact that the transcriptomic signature of immune cells in mutant snerves upon skin trauma remains unchanged suggests that mutant nmSCs respond directly to such inflammatory stimuli.

In our model, hyperplastic snerves as well as paraspinal pNFs were often observed beneath cNFs, which develop in the dermis, reinforcing the role of local inflammation and suggesting that the mechanisms governing their initiation may be the same. Detailed comparison of the transcriptomic signatures of mutant SCs from pNFs and cNFs might help identify such mechanisms and associated candidate genes. By comparing the transcriptome of mutant nmSCs isolated from T+ and T- snerves, we hoped to identify genes whose expression is directly affected by *Nf1* loss. Unfortunately, the small number of SCs isolated from the T- mutant snerves made such an analysis impossible. However, Postn expression in T- mutant snerves but not in control snerves strongly suggest that expression of this profibrotic molecule is under NF1 control.

We observed that *Nf1* heterozygosity in the micro-environment of mutant snerves mimics the impact of trauma. This result was unexpected as we did not observe the increased incidence of pNFs in mutant males on Nf1+/- background. These results suggest that the Nf1+/- background might favors the initiation

phase of nerve hyperplasia but is not sufficient for their progression into pNFs.

Materials and Methods

Animals

The following mouse lines were used and genotyped as described in the original publications [16, 22, 23]. The mice with the following genotype were used: *Prss56*$^{Cre/+}$, *R26*$^{tdTom/+}$, *Nf1*$^{flox/flox}$ (*Nf1*-KO^{wt}), *Prss56*$^{Cre/+}$, *R26*$^{tdTom/+}$, *Nf1*$^{flox/-}$ (*Nf1*-KO^{het}) and *Prss56*$^{Cre/+}$, *R26*$^{tdTom/+}$ (control). The handling of the animals was conducted in agreement with French and European regulations. All experimentations on mice were approved by the French Ethical Committee (APAFIS #35296-2022020816327611). All mice were housed in a temperature- and humidity-controlled vivarium with a 12-hour dark-light cycle. Mice were genotyped by the PCR method.

Sacrifice, dissection, and processing of subcutaneous nerves

Mice (males) were euthanized. Their skin was shaved, disinfected with 70% ethanol, and removed. The subcutaneous nerves were imaged under a Leica epifluorescence microscope. Pictures were taken using LaX software and images were analyzed using ImageJ software. Subcutaneous nerves were dissected and fixed by immersion in 4% paraformaldehyde (PFA) for 2 hours at 4°C on a shaking plate. After fixation, nerves were rinsed with 1X PBS before being cryoprotected by overnight immersion in 30% sucrose in PBS at 4°C. Nerves were placed in a Tissue-Tek OCT compound embedding medium (Sakura, ref: 4583) and were frozen at -20°C with dry ice and isopentane. The resulting blocks were stored at -80°C.

Immunofluorescence

10 µm frozen sections were blocked with 4% BSA in 0.3% TritonX-100 in 1X PBS for 1 hour, then stained with primary antibodies overnight at 4°C in the blocking solution. Secondary antibodies were added to the blocking solution for 2 hours at room temperature. Cell nuclei were labeled with Hoechst 33342 (Life technologies, ref: H3570). Sections were washed three times with 1X PBS between each step. Slides were mounted using Fluoromount G (Southern Biotec, ref: 0100-01). The following antibodies were used:

Antibodies	Species	Type	Reference	Dilution
TdTomato	Rat	Kerafast	EST203	1/1000
mCherry	Goat	Sicgen	AB0040-200	1/1000
RFP (DsRed)	Rabbit	Rockland	600-401-379	1/1000
SOX10	Rabbit	Bio SB	BSB2584	1/500
IBA1	Rabbit	Wako	019-19741	1/400
PDGFRα	Goat	R&D systems	AF1062	1/100
CD45	Goat	BioTechne	AF114-SP	1/100
Ki67	Rabbit	Abcam	ab15580	1/200
ZO-1	Rat	ThermoFisher	14-9776-82	1/100
CLDN1	Rabbit	Thermofisher	71-7800	1/100
PGP9.5	Rabbit	Abcam	ab108986	1/200
SOX9	Rabbit	Abcam	ab185230	1/2000
POSTN	Goat	BioTechne	AF2955	1/400
TUJ1	Chicken	Synaptic Systems	302306	1/500
CD49f	Rabbit	ThermoFisher	710209	1/300
S100	Rabbit	DAKO	Z0311	1/500
MBP	Rat	Millipore	MAB386	1/100

Imaging

All images were acquired using the Leica TCS SP5 confocal microscope equipped with 10X, 25X, and 63X objectives. Images were saved in .tif format. All analyses (measurements, nerve delimitation, quantification) were done manually with ImageJ software. GraphPad Prism6 software was used for graphs. t-tests and nonparametric tests were used for statistical analyses. Significant p values are indicated as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Single-cell RNA sequencing

Dissection and digestion

20 nerves per mouse were dissected and incubated in cold HBSS (Sigma, ref: H6648-500ML). After mechanical dissociation, nerves were transferred into a mix of enzymes (0.15% Trypsin (Sigma #T9201), 0.04% Hyaluronidase (Bioconcept #LS005474), 0.3% Collagenase type 2 (Bioconcept #LS004174) and DNAse I (25 µg/mL)) in HBSS, during 2 hours in a water bath. After digestion, nerves were tritured with a 1000 pipette, and a solution of 20% FBS in RPMI (Gibco, ref: 42401018) was added to

stop digestion. Tubes were centrifuged and the supernatant was removed.

Cell sorting

Single viable cells were sorted (BD InfluxTM Cell Sorter) using a nuclear labeling DRAQ7TM Far-Red Fluorescent Live-Cell Impermeant DNA Dye (Abcam, ref: ab109202). Cells were collected in 20% FBS-RPMI solution.

Myelin depletion

Cells were centrifuged, the supernatant removed, and myelin depletion was performed by using anti-myelin beads (Miltenyi Biotec, ref: 130-096-733) in a 0.5% MACS solution (Miltenyi Biotec, ref: 5130-091-376). Cells were collected after the passage through the magnetic separator.

Single-cell RNA sequencing

Around 25,000 cells were loaded into one channel of the Chromium system using the V3.1 single-cell reagent kit (10X Genomics) to generate single-cell GEMs. Following capture and lysis, cDNAs were synthesized, and then amplified by PCR for 12 cycles as per the manufacturer's protocol (10X Genomics). The amplified cDNAs were used to generate Illumina sequencing libraries that were each sequenced on one flow cell Nextseq500 Illumina.

Bioinformatic analyses

BCL files were processed using 10x Genomics Cell Ranger 7.0.1. Reads were mapped onto a custom mouse transcriptome based on GRCm39 release 108 transcriptome, in which we added sequences such as tdTomato. For both custom transcriptome construction and BCL to count matrix pipeline, we designed two pipelines using Nextflow [33] and Singularity containerization [34]. Count matrices were processed from raw matrices to the figures using RStudio and Rmarkdown to generate fully traceable notebook for each analysis. To ensure version stability, a Singularity container containing R version 3.6.3 and all specific versions for all the packages of interest was developed and used to compile notebooks.

Each dataset, corresponding to one biological replicate from one experimental condition, was analyzed individually from its raw count matrix using Seurat V3 package [35]. Cells having less than 6 log number of UMI or less than 600 genes were filtered out. Doublet cells were removed using scDblFinder tool and scds in hybrid mode. Then, cells having more than 20% of UMI related to mitochondrial genes or more than 30% related to ribosomal genes were filtered out. UMI count matrix for remaining cells was normalized using LogNormalize method implemented in Seurat V3 package. From the normalized count matrix, we used AddModuleScore to annotate cells for cell type using the following marker sets : Sox10, Plp1, tdTomato, Nkain2, Ank3 and Kcna1 for Schwann cells ; Clec3b, Gdf10, Dpt, Gpx3, Slit3, Aebp1, Prrx1, Ly6c1, Pcolce2 and Cxcl14 for epineurial fibroblasts (epiFb) ; Lum, Smoc2, Spp1 and Gpc3 for endoneurial fibroblasts (endoFb) ; Cldn1, Tjp1, Itga6, Ildr2, Zbtb7c, Cttnbp2, Adamtsl3, Fxyd6 and Itgb4 for perineurial fibroblasts (periFb) ; Adgre1, Lyz2, Cd68, Pf4, Csf1r and Apoe for macrophages ; Cd209a, Mgl2, Tnip3, Mcemp1, Cfp, Xcr1 and Napsa for dendritic cells ; Cd3g, Cd3d, Cd3e, Cd8a, Cd4, Foxp3, Gata3 and Areg for T lymphocytes ; Nkg7, Klrk1, Klri2, Klrc2, Klrb1c, Ncr1 and Klrb1f for Natural Killer cells (NK cells) ; Cd79a, Ly6d, Cd79b and Cd19b for B lymphocytes ; Tpsb2, Tpsab1, Cpa3 and Kit for mast cells ; S100a8 and S100a9 for neutrophils ; Egfl7, Flt1, Ptprb, Shank3 and Mecom endothelial cells and Des, (Rgs5, Notch3, Gucy1a1), (Chodl, Nppc, Gal and Myf5) for mural cells. Note that two subpopulations of mural cells were identified, corresponding to the two sets of genes in brackets. Finally, cells were clustered using Seurat's FindClusters function with a resolution of 2, on 20 principal components of a 100-dimension PCA obtained with Seurat's RunPCA function from 3,000 highly variable features.

Same pipeline was used to either combine all 4 control datasets or all 11 datasets. Datasets of interest were merged using base's merge function. Genes expressed in less than 5 cells were removed. As for individual datasets, we generated a PCA. Sample-effect was removed on the PCA using RunHarmony function from harmony package [36]. To generate UMAP or tSNE, we used the implementation for Seurat

package: RunUMAP or RunTSNE. Cells were clustered on the harmonized PCA. To smoothen eventual mis-annotation, cell type annotation was grouped by cluster.

To build the 5 cell types-related datasets, same procedure was applied. Cells were extracted from each individual datasets based on single cell or cluster smoothen annotation. After first processing as for the atlas, eventual contaminating cells were automatically removed through a clustering and marker genes. A second processing was applied on the selected cells. For Schwann cells and fibroblasts datasets, sample-effect was removed from the normalized count matrix using fastMNN function [37] from batchelor package. For immune cells, we used harmony. Schwann cells projection was obtained using diffusion map, implemented in destiny package [38], on the corrected space, then UMAP. UMAP or tSNE from Seurat package were used for fibroblasts and immune cells datasets projection. Cells were clustered on the corrected space. Schwann cells were re-annotated using the following genes: Scn7a, Plac9, Ifitm3, Prkg1, Ptn and Ncam1 for nmSC and Ncmap, Cldn19, Gfra1, Mlip, Fxyd6 and Mpz for mSC. Annotation was smoothed at a cluster level. Immune cells dataset was investigated for subpopulations based on clustering, differential expression and known markers from literature. For differential expression, both the methods implemented in DESeq2 package [39] and Wilcoxon test from Seurat's FindMarkers function were used. Differentially expressed genes were filtered based on p-value adjusted, lower than 0.05.

Code and data availability

Single cell RNA sequencing data are available as FASTQ files and raw count matrices in ArrayExpress repository, under the accession number E-MTAB-13334. All scripts and complete methods used to process raw count matrices up to the figures are shared through a Github repository (`https://github.com/audrey-onfroy/Mansour_et_al`).

Authors contribution

Conceptualization (AO, FC, PW, PT), Data Curation (AO), Formal Analysis (MM, AO), Funding Acquisition (PT), Investigation (MM, FC, KR), Project Administration (AO, FC, PT), Supervision (PW, PT), Validation (MM, AO, FC, PT), Visualization (MM, AO, FC, PT), Writing - Original Draft Preparation (MM, AO, FC, PT), Writing - Review & Editing (AO, PT)

Acknowledgments

This work was supported by the Institut National de la Santé et de la Recherche Médicale (INSERM), and by a grant from the Institut Mondor de Recherche Biomédicale (IMRB), Fondation pour la Recherche Médicale (FRM), Fondation Maladies Rares and the Association Neurofibromatose et Recklinghausen. All authors have disclosed any financial or personal relationships with organizations that could influence the results.

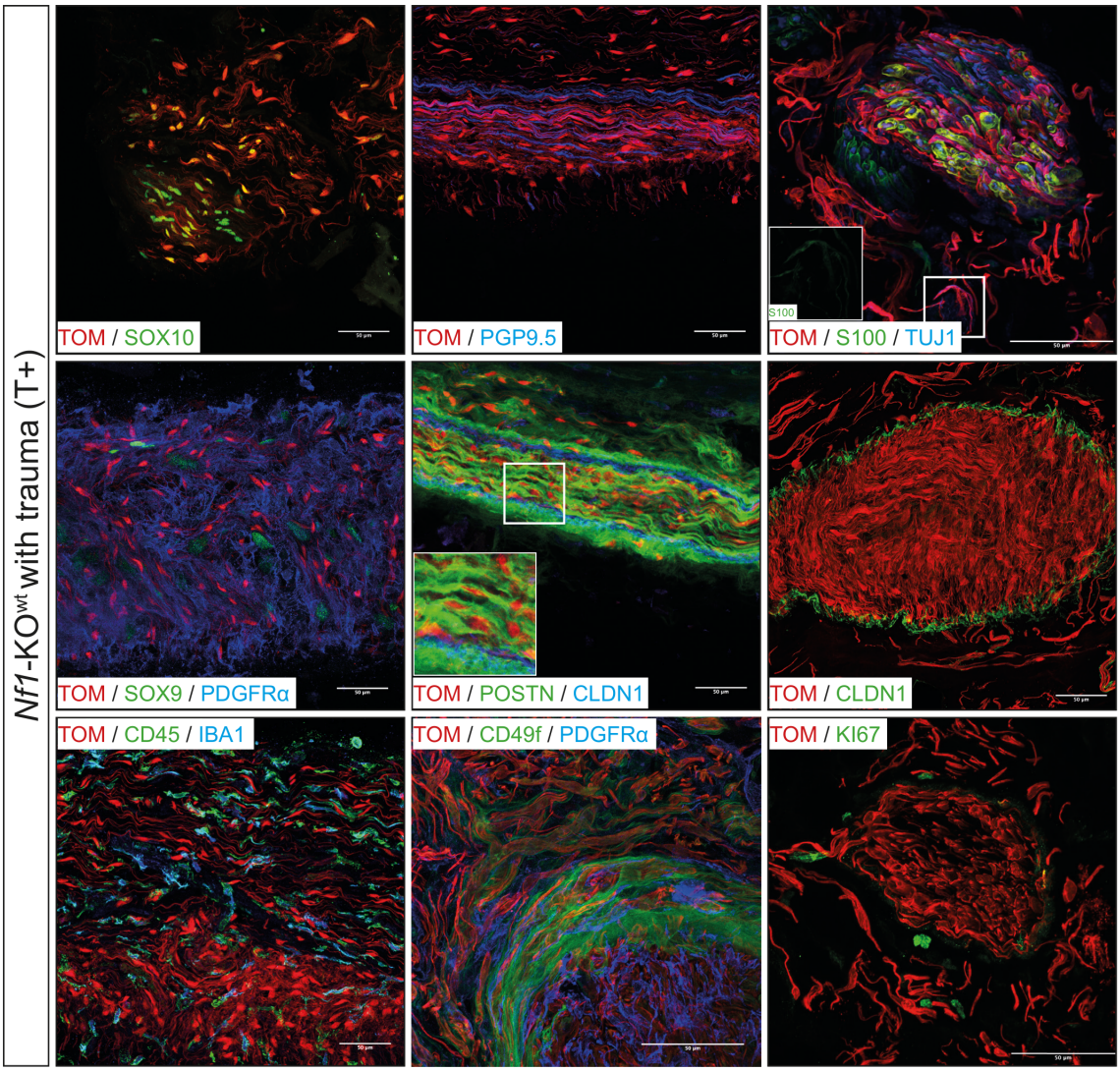
References

1. Mohammed Bergoug, Michel Doudeau, Fabienne Godin, Christine Mosrin, Béatrice Vallée, and Hélène Bénédetti. Neurofibromin Structure, Functions and Regulation. *Cells*, 9(11):2365, 10 2020.
2. Rosalie E. Ferner and David H. Gutmann. Neurofibromatosis type 1 (NF1): diagnosis and management. In *Handbook of clinical neurology*, pages 939–955. Elsevier, 2013.
3. Stephen Li, Zhiguo Chen, and Lu Q Le. New insights into the neurofibroma tumor cells of origin. *Neuro-Oncology Advances*, 2:i13–i22, 6 2020.
4. Fanny Coulpier, Pernelle Pulh, Layna Oubrou, Julie Naudet, Laura Fertitta, Jean-Marc Gregoire, Arnaud Bocquet, Anne-Marie Schmitt, Pierre Wolkenstein, Katarzyna J. Radomska, and Piotr Topilko. Topical delivery of mitogen-activated protein kinase inhibitor binimetinib prevents the development of cutaneous neurofibromas in neurofibromatosis type 1 mutant mice. *Translational Research*, 261:16–27, 11 2023.
5. Lu Q. Le and Luis F. Parada. Tumor microenvironment and neurofibromatosis type I: connecting the GAPs. *Oncogene*, 26(32):4609–4616, 7 2007.
6. Chunhui Jiang, Renee M. McKay, and Lu Q. Le. Tumorigenesis in neurofibromatosis type 1: role of the

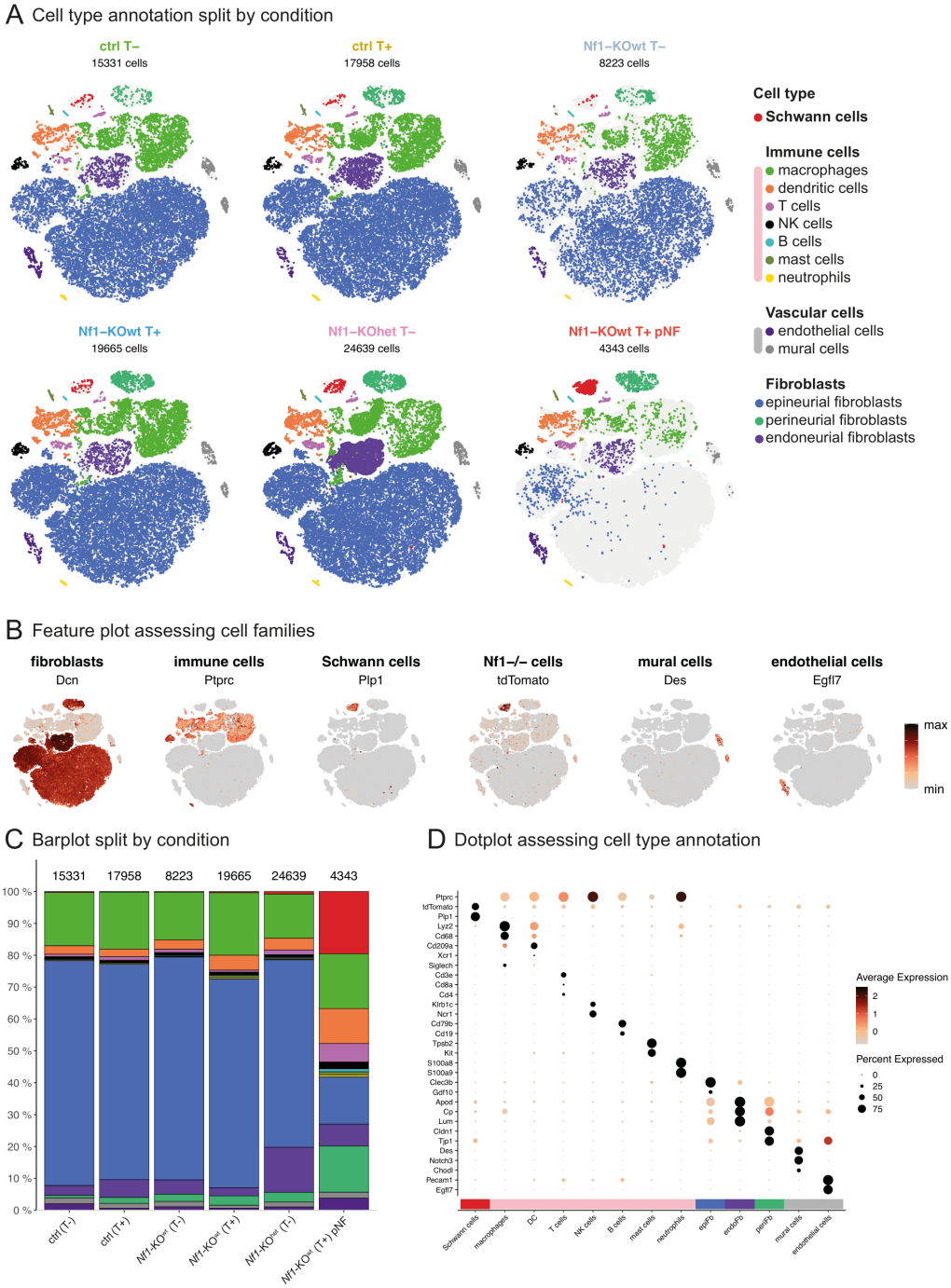
microenvironment. *Oncogene*, 40(39):5781–5787, 9 2021.

7. Nicolas Ortonne, Pierre Wolkenstein, Jaishri O. Blakeley, Bruce Korf, Scott R. Plotkin, Vincent M. Riccardi, Douglas C. Miller, Susan Huson, Juha Peltonen, Andrew Rosenberg, Steven L. Carroll, Sharad K. Verma, Victor Mautner, Meena Upadhyaya, and Anat Stemmer-Rachamimov. Cutaneous neurofibromas: Current clinical and pathologic issues. *Neurology*, 91(2_Supplement_1), 7 2018.
8. Ashley Bui, Chunhui Jiang, Renee M. McKay, Laura J. Klesse, and Lu Q. Le. Insights into the Pathogenesis of NF1-Associated Neoplasms. *JID Innovations*, 1(3):100044, 9 2021.
9. Nicole M. Brossier and Steven L. Carroll. Genetically engineered mouse models shed new light on the pathogenesis of neurofibromatosis type I-related neoplasms of the peripheral nervous system. *Brain Research Bulletin*, 88(1):58–71, 5 2012.
10. Rui M. Costa and Alcino J. Silva. Mouse models of neurofibromatosis type I: bridging the GAP. *Trends in Molecular Medicine*, 9(1):19–23, 1 2003.
11. Nancy M. Joseph, Jack T. Mosher, Johanna Buchstaller, Paige Snider, Paul E. McKeever, Megan Lim, Simon J. Conway, Luis F. Parada, Yuan Zhu, and Sean J. Morrison. The Loss of Nf1 Transiently Promotes Self-Renewal but Not Tumorigenesis by Neural Crest Stem Cells. *Cancer Cell*, 13(2):129–140, 2 2008.
12. Jianqiang Wu, Jon P. Williams, Tilat A. Rizvi, Jennifer J. Kordich, David Witte, Dies Meijer, Anat O. Stemmer-Rachamimov, Jose A. Cancelas, and Nancy Ratner. Plexiform and Dermal Neurofibromas and Pigmentation Are Caused by Nf1 Loss in Desert Hedgehog-Expressing Cells. *Cancer Cell*, 13(2):105–116, 2 2008.
13. Melanie P. Clements, Elizabeth Byrne, Luis F. Camarillo Guerrero, Anne-Laure Cattin, Leila Zakka, Azhaar Ashraf, Jemima J. Burden, Sanjay Khadayate, Alison C. Lloyd, Samuel Marguerat, and Simona Parrinello. The Wound Microenvironment Reprograms Schwann Cells to Invasive Mesenchymal-like Cells to Drive Peripheral Nerve Regeneration. *Neuron*, 96(1):98–114, 9 2017.
14. Sara Ribeiro, Ilaria Napoli, Ian J. White, Simona Parrinello, Adrienne M. Flanagan, Ueli Suter, Luis F. Parada, and Alison C. Lloyd. Injury Signals Cooperate with Nf1 Loss to Relieve the Tumor-Suppressive Environment of Adult Peripheral Nerve. *Cell Reports*, 5(1):126–136, 10 2013.
15. Huarui Zheng, Lou Chang, Neha Patel, Jiong Yang, Lori Lowe, Dennis K. Burns, and Yuan Zhu. Induction of Abnormal Proliferation by Nonmyelinating Schwann Cells Triggers Neurofibroma Formation. *Cancer Cell*, 13(2):117–128, 2 2008.
16. Yuan Zhu, Pritam Ghosh, Patrick Charnay, Dennis K. Burns, and Luis F. Parada. Neurofibromas in NF1: Schwann cell origin and role of tumor environment. *Science*, 296(5569):920–922, 5 2002.
17. Chung Ping Liao, Reid C. Booker, Jean Philippe Brosseau, Zhiguo Chen, Juan Mo, Edem Tchegnon, Yong Wang, D. Wade Clapp, and Lu Q. Le. Contributions of inflammation and tumor microenvironment to neurofibroma tumorigenesis. *Journal of Clinical Investigation*, 128(7):2848–2861, 7 2018.
18. Feng-Chun Yang, Shi Chen, Travis Clegg, Xiaohong Li, Trent Morgan, Selina A. Estwick, Jin Yuan, Waleed Khalaf, Sarah Burgin, Jeff Travers, Luis F. Parada, David A. Ingram, and D. Wade Clapp. Nf1+/- mast cells induce neurofibroma like phenotypes through secreted TGF-b signaling. *Human Molecular Genetics*, 15(16):2421–2437, 8 2006.
19. Laura González-González and Javier Alonso. Periostin: A Matricellular Protein With Multiple Functions in Cancer Development and Progression. *Frontiers in Oncology*, 8, 6 2018.
20. Shima Dorafshan, Mahdieh Razmi, Sadegh Safaei, Erica Gentilin, Zahra Madjd, and Roya Ghods. Periostin: biology and function in cancer. *Cancer Cell International*, 22(1):315, 10 2022.
21. Piotr Topilko, Sylvie Schneider-Maunoury, Giovanni Levi, Anne Baron-Van Evercooren, Amina Ben Younes Chennoufi, Tania Seitanidou, Charles Babinet, and Patrick Charnay. Krox-20 controls myelination in the peripheral nervous system. *Nature*, 371(6500):796–799, 10 1994.
22. Aurélie Gresset, Fanny Coulpier, Gaspard Gerschenfeld, Alexandre Jourdon, Graziella Matesic, Laurence Richard, Jean Michel Vallat, Patrick Charnay, and Piotr Topilko. Boundary Caps Give Rise to Neurogenic Stem Cells and Terminal Glia in the Skin. *Stem Cell Reports*, 5(2):278–290, 2015.
23. Katarzyna J. Radomska, Fanny Coulpier, Aurélie Gresset, Alain Schmitt, Amal Debbiche, Sophie Lemoine, Pierre Wolkenstein, Jean Michel Vallat, Patrick Charnay, and Piotr Topilko. Cellular origin, tumor progression, and pathogenic mechanisms of cutaneous neurofibromas revealed by mice with Nf1 knockout in boundary cap cells. *Cancer Discovery*, 9(1):130–147, 1 2019.

24. Daniel Gerber, Jorge A. Pereira, Joanne Gerber, Ge Tan, Slavica Dimitrieva, Emilio Yángüez, and Ueli Suter. Transcriptional profiling of mouse peripheral nerves to the single-cell level to build a sciatic nerve atlas (Snat). *eLife*, 10, 2021.
25. Jolien Wolbert, Xiaolin Li, Michael Heming, Anne K. Mausberg, Dagmar Akkermann, Clara Frydrychowicz, Robert Fledrich, Linda Groeneweg, Christian Schulz, Mark Stettner, Noelia Alonso Gonzalez, Heinz Wiendl, Ruth Stassart, and Gerd Meyer Zu Hörste. Redefining the heterogeneity of peripheral nerve cells in health and autoimmunity. *PNAS*, 117(17):9466–9476, 2020.
26. Abbie E. Fearon, Edward P. Carter, Natasha S. Clayton, Edmund H. Wilkes, Ann-Marie Baker, Ekaterina Kapitonova, Bakhouche A. Bakhouche, Yasmine Tanner, Jun Wang, Emanuela Gadaleta, Claude Chelala, Kate M. Moore, John F. Marshall, Juliette Chupin, Peter Schmid, J. Louise Jones, Michelle Lockley, Pedro R. Cutillas, and Richard P. Grose. PHLDA1 Mediates Drug Resistance in Receptor Tyrosine Kinase-Driven Cancer. *Cell Reports*, 22(9):2469–2481, 2 2018.
27. Léo Machado, Perla Geara, Jordi Camps, Matthieu Dos Santos, Fatima Teixeira-Clerc, Jens Van Herck, Hugo Varet, Rachel Legendre, Jean-Michel Pawlotsky, Maurilio Sampaolesi, Thierry Voet, Pascal Maire, Frederic Relaix, and Philippos Mourikis. Tissue damage induces a conserved stress response that initiates quiescent muscle stem cell activation. *Cell Stem Cell*, 28(6):1125–1135, 6 2021.
28. Bing Chen, Matthew C. Banton, Lolita Singh, David B. Parkinson, and Xin Peng Dun. Single Cell Transcriptome Data Analysis Defines the Heterogeneity of Peripheral Nerve Cells in Homeostasis and Regeneration. *Frontiers in Cellular Neuroscience*, 15, 3 2021.
29. Chunhui Jiang, Ashwani Kumar, Ze Yu, Tracey Shipman, Yong Wang, Renee M. McKay, Chao Xing, and Lu Q. Le. Basement membrane proteins in extracellular matrix characterize NF1 neurofibroma development and response to MEK inhibitor. *Journal of Clinical Investigation*, 133(12), 6 2023.
30. Eva Sonnenberg-Riethmacher, Michaela Miehe, and Dieter Riethmacher. Promotion of periostin expression contributes to the migration of Schwann cells. *Journal of Cell Science*, 1 2015.
31. Jeremy S. Toma, Konstantina Karamboulas, Matthew J. Carr, Adelaida Kolaj, Scott A. Yuzwa, Neemat Mahmud, Mekayla A. Storer, David R. Kaplan, and Freda D. Miller. Peripheral nerve single-cell analysis identifies mesenchymal ligands that promote axonal growth. *eNeuro*, 7(3):1–70, 5 2020.
32. Xiao-Feng Zhao, Lucas D. Huffman, Hannah Hafner, Mitre Athaiya, Matthew C. Finneran, Ashley L. Kalinski, Rafi Kohen, Corey Flynn, Ryan Passino, Craig N. Johnson, David Kohrman, Riki Kawaguchi, Lynda J.S. Yang, Jeffery L. Twiss, Daniel H. Geschwind, Gabriel Corfas, and Roman J. Giger. The injured sciatic nerve atlas (iSNAT), insights into the cellular and molecular basis of neural tissue degeneration and regeneration. *eLife*, 11, 12 2022.
33. Paolo Di Tommaso, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Palumbo; Emilio, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, 2017.
34. Gregory M. Kurtzer, Vanessa Sochat, and Michael W. Bauer. Singularity: Scientific containers for mobility of compute. *PLOS ONE*, 12(5):e0177459, 5 2017.
35. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902, 6 2019.
36. Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 12 2019.
37. Laleh Haghverdi, Aaron T.L. Lun, Michael D. Morgan, and John C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 5 2018.
38. Philipp Angerer, Laleh Haghverdi, Maren Büttner, Fabian J. Theis, Carsten Marr, and Florian Buettner. destiny : diffusion maps for large-scale single-cell data in R. *Bioinformatics*, 32(8):1241–1243, 4 2016.
39. Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 12 2014.



Supplementary Figure 1. *Nf1*-KO^{wt} grouped house males develop subcutaneous plexiform neurofibromas. Immunofluorescence analyses of longitudinal and transverse sections of pNFs show accumulation of mutant SCs (TOM) that are detaching from axons (PGP9.5) (a). When going out of nerve fascicles, SCs express CD49f (also known as ITGA6), which is also expressed by periFb. Once out, CD49F expression shut down (b). Mutant SCs still keep their glial identity (SOX10) (c) and maintain expression of (POSTN), which is also found around perineurial cells (CLDN1) (d). Perineurial layer is decompacted (CLDN1) (e). pNF shows accumulation of fibroblasts (PDGFRα) including endoneurial (SOX9) fibroblasts. However, almost all cells are either quiescent or slowly proliferating (KI67). pNF are infiltrated by immune cells (CD45) including macrophages (IBA1) and mastocytes (c-KIT). Scale bar: 50 µm.



Supplementary Figure 2. Transcriptomic atlas of control and *Nf1*-KO subcutaneous nerves. (A) tSNE visualization of the main dataset split by experimental condition and colored according to cell types. A light grey background corresponds to cells from other conditions. Except for *Nf1*-KO^{wt} T- (n = 1), each experimental condition consists of two independent replicates. Number of cells in each condition is indicated. (B) tSNE representation of gene expression levels to assess cell type assignation. Expression levels are log-normalized and scaled within the same range for all genes. (C) Barplot showing cell type representativeness in each experimental condition. Number of cells in each condition is indicated. (D) Dotplot assessing cell population annotation. Dot size represents the proportion of cells expressing the corresponding marker. The dot color represents average gene expression levels for cells concerned by the dot.

Supplementary Figure 3. Transcriptomic signature of control and mutant Schwann cells from subcutaneous nerves. (A) UMAP visualization of SCs dataset split by experimental condition and colored according to *Nf1* status. A light grey background corresponds to cells from other conditions. Except for *Nf1*-KO^{wt} T- (n = 1), each experimental condition consists of two independent replicates. Number of cells in each condition is indicated. (B) Dotplot assessing cell population annotation. Dot size represents the proportion of cells expressing the corresponding marker. The dot color represents average gene expression levels for cells concerned by the dot. (C) UMAP representation of Tomato expression level. expression level is log normalized. Parameter order = TRUE was set to represent cells with strong expression in the foreground. (D) Schwann cells clustering. (a) UMAP visualization of SCs clustering. SCs were grouped into 3 groups. Group 1 (red) contains mainly Tom- cells while group 2 (blue shades) contains Tom+ cells, both isolated from snerves. Group 3 contains Tom+ SCs isolated from pNFs and was not used for differential expression analysis. (b) Barplot showing the distribution of each experimental condition across groups 1 and 2. Among each experimental condition, Tom+ SCs are depicted as hatched colors. Note that some *Nf1*-/- Tom+ cells have similar transcriptomic profiles as control (*Nf1*+/+) SCs. Number of cells in each group is indicated. (E) Violin plot representing log-normalized expression for Postn, Phlda1, and Sox2, split by experimental condition. The proportion of SCs with not null expression levels is indicated.

2.3 Current and future directions

2.3.1 Code availability

The code used to process the count matrices up to the figures is shared through via a Github repository. Since the manuscript is under revision, the repository is private. Once public, the link will redirect to the repository page (Figure 67). The repository contains the R Markdown notebooks and their associated compiled HTML versions. In the "About" section in the top right corner, there is a direct link to the main index.html (red shape), presented in previous Section 1.4.4.

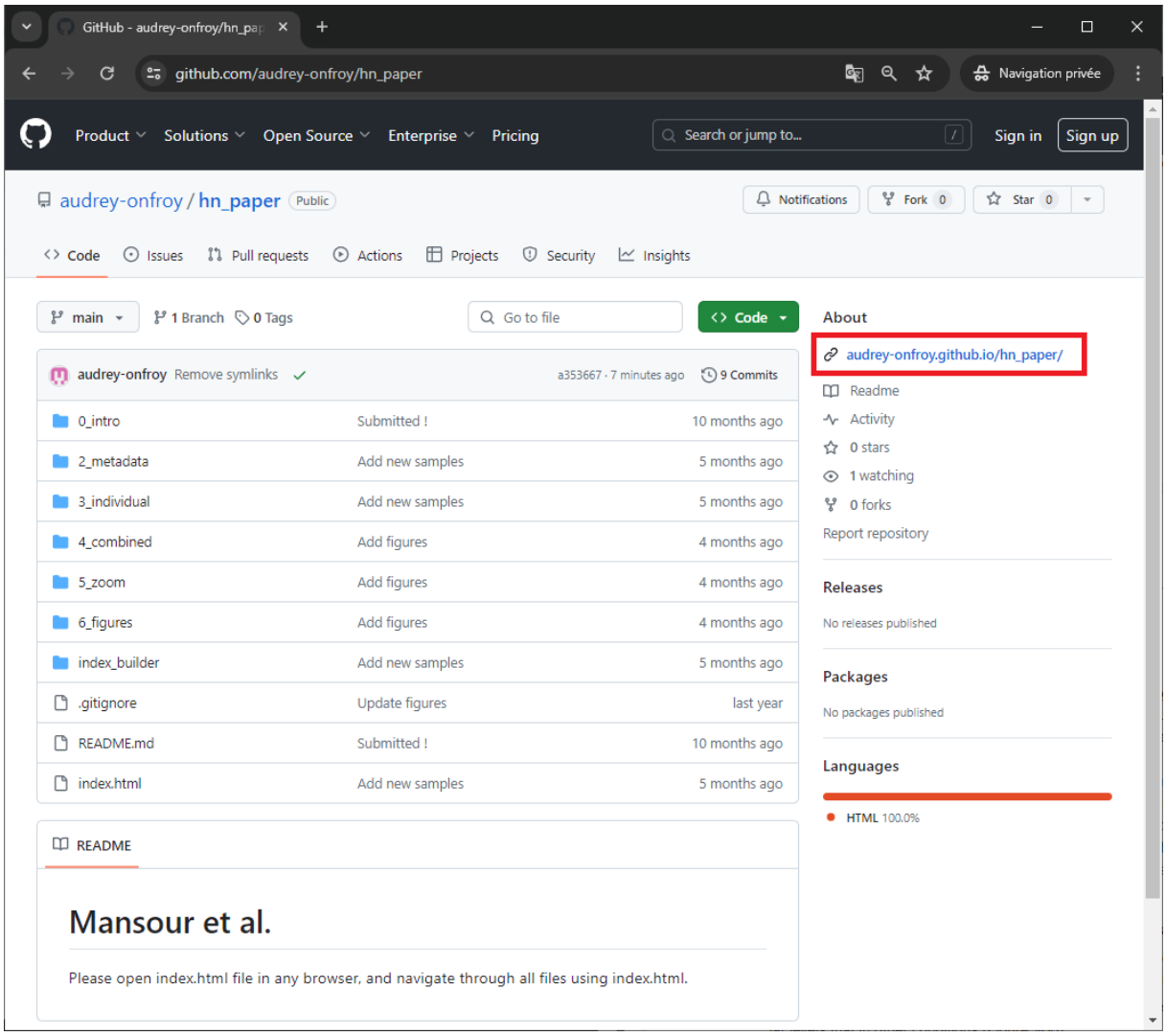


Figure 67 – Screenshot of the repository page associated with the Article 1. The repository name will change to be more informative. The screenshot was taken after temporarily changing the repository visibility to public.

Once the link is clicked, the reader is redirected to the index.html page (Figure 68). All the compiled notebooks can be accessed from the menu on the left. After selecting an item in the menu, it appears in red and the HTML notebook appears in the main frame. Its own menu is navigable. In the upper right corner, a button allows to directly download the R Markdown file behind the compiled version.

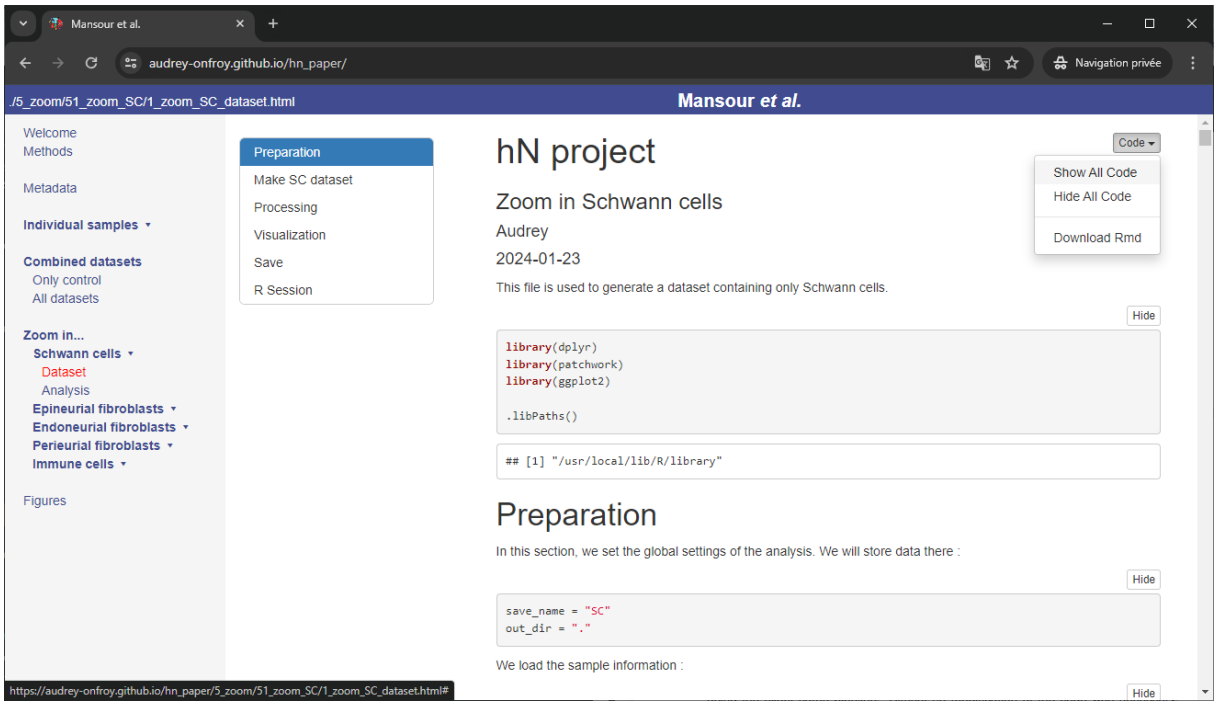


Figure 68 – Screenshot of a page from the website associated with the Article 1. The blue header and the left menu column are fixed in the source code of the page. The menu is navigable. After clicking on an item, it appears in red. The associated page, corresponding to the compiled notebook, is embedded in the main body of the page. The presented notebook was used to generate the Schwann cells dataset. The screenshot was taken after temporarily changing the repository visibility to public.

2.3.2 Combination of skin trauma and *Nf1* haploinsufficiency in nerves

The main criticism from the reviewers is the incomplete experimental design, specifically the lack of subcutaneous nerve samples from mutant mice with an *Nf1*$^{+/-}$ background and skin trauma (*Nf1*-KO^{het} T+ condition). We conducted two independent scRNA-Seq experiments for this condition using the same procedures and computational pipeline. Almost no modification of the code was necessary, except for the new original comparisons underlying the addition of this new condition.

Subcutaneous nerves from *Nf1*-KO^{het} T+ mice showed similar cellular heterogeneity to other conditions but exhibited changes in cellular proportions. However, no increased proliferative activity was noticed, suggesting a slow process.

The proportion of **Schwann cells** doubled compared to *Nf1*-KO^{het} T- nerves and increased ten-fold compared to *Nf1*-KO^{wt} T- nerves. Among them, the proportion of tumor (*Tom*+) SCs rose from 72% in *Nf1*-KO^{het} T- nerves to nearly 85% with skin trauma. Among tumor SCs, more cells expressed *Postn*, *Phlda1*, and *Sox2*. Finally, as in pNF datasets, both *Tom+* and *Tom-* SCs had similar transcriptomic profiles in *Nf1*-KO^{het} T+ nerves. These results suggest that the combination of trauma and *Nf1* haploinsufficiency induces a faster progression of the pro-fibrotic tumor SCs.

Endoneurial fibroblasts (endoFb) with a TNF signaling profile increased from 30% in *Nf1*-KO^{het} T- nerves to 40% in the context of skin trauma. It highlights a progressive acquisition of a pro-inflammatory profile in the endoneurium.

In *Nf1*-KO^{het} T+ nerves, the *Nav3*+ population of **epineurial fibroblasts** (epiFb) also expressed *Gdf10*. In other conditions, the two markers enabled to clearly identified two populations. This observation suggests a transformation of *Nav3*+ cells into *Gdf10*+ cells during pNF formation.

On the cells projection, the overlapping between the previous and new conditions shows that the endoneurial macrophages that were *Cx3cr1*- in other mutant conditions start to express this marker in *Nf1*-KO^{het} T+ nerves. While 28% of macrophages express *Cx3cr1* in *Nf1*-KO^{het} T- nerves, they are 41% in the new condition. This suggests a progressive activation of endoneurial macrophages, promoted by the combination of trauma and *Nf1* haploinsufficiency.

We evaluated the potent cells, expressing the ligand of CX3CR1 : *Cx3cl1*. In control and mutant mice, some vascular and dendritic cells are *Cx3cl1*+. In *Nf1*-KO^{het} T- nerves, endoFb and a few SCs could also receive macrophage signals. Finally, in *Nf1*-KO^{het} T+ nerves, most SCs, periFb, and *Gdf10*+ epiFb cells are potent. This emphasizes the *Gdf10*+ epiFb cells as new active components in pNF initiation.

New observations extend the scope of our study, showing more **vascular cells** and **mast cells** in the new datasets. Vascular cells, which support nerve populations and participate in immune infiltration, are further responsive to *Cx3cr1*+ macrophages.

To conclude, we observed a more severe phenotype in *Nf1*-KO^{het} T+ nerves compared to all other conditions. The scRNA-Seq analysis concludes that trauma and *Nf1* haploinsufficiency in 3-month-old male mice accelerate pNF development. Immunohistochemistry validations are ongoing to complete the study.

3 Article 2

In the context of NF1, malignant peripheral nerve sheath tumors (MPNSTs) arise from plexiform neurofibromas (pNFs). They are aggressive tumors and lack of effective therapeutic options. Better characterization of the mechanisms driving the pNF-to-MPNST transformation is crucial. The poor characterization of this transition results from a lack of animal models recapitulating the malignant transformation.

3.1 Overview of the study

The purpose of this study was to decipher the molecular mechanisms driving the malignant transformation, using our *Nf1*-KO mouse model. The methodology involves histology, immunohistochemistry, scRNA-Seq, whole exome sequencing, biochemistry and drug screening. My contribution has been to analyze the scRNA-Seq data, participate to formalize the results and design the figures.

When *Nf1*-KO male mice are housed together, their pNFs stepwise progress into dysplastic neurofibromas (dyNFs) and MPNSTs with a high penetrance. This highlights the active role of skin trauma in malignant transformation, in a mouse model engineered for the mere loss of function of NF1 protein. We further showed shared histological, genetic and transcriptomic features between MPNSTs from patients and our mouse model. To characterize the cellular heterogeneity of mouse pNFs, dyNFs and MPNSTs, we built a transcriptomic atlas. It could serve as a resource for the NF1 community. As in patients, the malignant transformation is preceded by an immune infiltration and then a glial-to-mesenchymal transition in tumor cells, i.e. the extinction of glial markers (eg. *Sox10*) and activation of mesenchymal markers (eg. *Sox9*).

As in patients, mouse MPNSTs are characterized by the genetic loss of *Cdkn2a*. Tumors can be propagated by successive allografting, and these allografted tumors recapitulate more genetic and epigenetic features described in patients MPNSTs.

Given that *Sox9* is a biomarker of MPNSTs, we proposed it as a therapeutic target. Its inhibition in mesenchymal tumor cells from mouse, followed by their transplantation into immunocompromised mice, delayed tumor growth. A screening of drugs lowering *Sox9* expression identified compounds related to the inhibition of Ras signaling pathways. This suggests Ras signaling as a therapeutic target in MPNSTs.

This study presented the first model recapitulating the malignant progression of neurofibromas described in patients, while targeting only the loss of *Nf1*. It is a suited tool to understand the malignant transformation and develop therapeutic strategies.

This manuscript was the subject of a poster at the 2024 Global NF Conference, which is provided in the Appendix (Section 6, Poster 4).

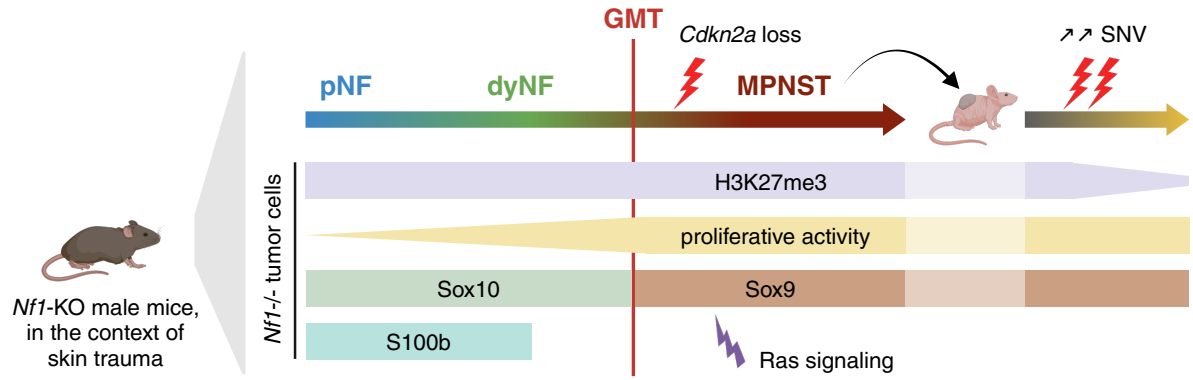


Figure 69 – **Graphical abstract of the Article 2**. *Nf1*-KO male mice recapitulate the development of neurofibromas and their malignant transformation. As in their human counterparts, malignant tumors lack *Cdkn2a*. During tumor progression, mice tumor cells undergo a glial-to-mesenchymal transition, characterized by a progressive loss of glial markers (*S100b* and *Sox10*) and acquisition of mesenchymal markers (*Sox9*). The activation of Ras signaling is related to the sustained expression of *Sox9* is MPNST and can be a therapeutic target. MPNSTs can be propagated through successive allogeneic transplantations.

Note: The manuscript is submitted. The Github link to make code available is broken and points to a private repository. It will be made public once the manuscript has been accepted.

The layout of the manuscript has been improved using a customized version of the Oxford University Press free layout, under the LaTeX Project Public License.

MANUSCRIPT SUBMITTED

A mouse model uncovers mechanisms of malignant transformation in Neurofibromatosis Type 1

Katarzyna J. Radomska[1], Audrey Onfroy[1,4], Laure Lecerf[1], Bastien Job[2], Aurélien Beaude[1], Laura Sesma Sanz[3], Tatiana El Jalkh[1], Denis Thieffry[4], Patrick Charnay[4], Pierre Wolkenstein[1,5], Nicolas Ortonne[1,6], Fanny Coulpier[1] and Piotr Topilko[1,*]

[1]Mondor Institute for Biomedical Research (IMRB), INSERM, Créteil, France, [2]Gustave Roussy Institute of Oncology, Villejuif, France, [3]National Phenotypic Screening Center, School of Life Sciences, University of Dundee, Dundee, United Kingdom, [4]Ecole Normale Supérieure, PSL Research University, CNRS, INSERM, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Paris, France, [5]Department of Dermatology, Reference Center for Neurofibromatoses, Henri-Mondor Hospital, AP-HP, Créteil, France and [6]Department of Pathology, Henri-Mondor Hospital, AP-HP, Créteil, France
[*]Corresponding author: piotr.topilko@inserm.fr

Abstract

There is currently no effective treatment for malignant peripheral nerve sheath tumors (MPNSTs), half of which result from malignant progression of neurofibromas (NFs) in patients with Neurofibromatosis Type 1 (NF1). NFs are due to biallelic loss-of-function of NF1, which negatively regulates the Ras pathway, in the Schwann cell lineage. We generated a conditional *Nf1*-mutant mouse model where NFs spontaneously transform into MPNSTs, faithfully recapitulating the human situation. Single-cell transcriptomic profiling demonstrated progression of NFs into MPNSTs, with a glial-to-mesenchymal transition. *Sox9* was identified as a marker of this transition and key player in tumor growth. The transition is followed by a loss of the tumor suppressor gene (TSG) *Cdkn2a* and acquisition of pathogenic variants of other TSGs. Finally, a proof-of-concept for screening drugs able to lower *Sox9* expression in tumor cells identified 12 FDA-approved compounds targeting the Ras signaling cascade, whose multi-targeted inhibition might thus be promising in fighting MPNSTs.

Key words: NF1, neurofibroma, dysplastic NF, ANNUBP, MPNST, malignant progression, glial-to-mesenchymal transition, Ras pathway inhibitors, *Cdkn2a* loss, *Sox9*

Introduction

Neurofibromatosis Type 1 (NF1) is caused by loss-of-function (LOF) mutations in the NF1 gene encoding a negative regulator of the Ras signaling pathway. Approximately half of NF1 patients develop plexiform neurofibromas (pNFs) along the nerve roots and bundles in soft tissues. pNFs are complex tumors composed of Schwann cells (SCs) mixed with other nerve fiber elements, including axons, fibroblasts, blood vessels, mastocytes, and macrophages. Despite their cellular heterogeneity, pNFs arise from SC lineage that lose their remaining functional NF1 (loss-of-heterozygosity, LOH) copy, with permanent hyperactivation of RAS targets, including the mitogen-activated protein kinase (MAPK) pathway [1]. pNFs are typically diagnosed in childhood and can transform

into malignant peripheral nerve sheath tumors (MPNSTs), without effective therapy. MPNSTs account for 5-10% of soft tissue sarcomas, affecting adults between 20 and 60 years of age. About half of MPNSTs occur in NF1 patients, the remaining being radiotherapy-induced (10%) or sporadic (40%) [1]. Malignant transformation of pNFs is preceded by a ”premalignant” state, described as ”dysplastic” or ”atypical” neurofibroma (NF) [2], exhibiting abnormal SC morphology or increased cellular density. Recently, experts proposed a new terminology, classifying tumors into three categories: atypical NF (only atypical SCs), cellular NF (increased cellularity), and atypical neurofibromatous neoplasms of uncertain biologic potential (ANNUBP) (at least two atypical features like cellular atypia, hypercellularity, loss of pNF architecture, and frequent mitoses) [3]. Despite this terminology being now a reference, these tumors will herein be designed as dysplastic NFs (dyNFs). Although dyNFs are premalignant lesions of MPNSTs [2], definite evidence of them being an early stage of pNF transformation into MPNST is lacking.

MPNST cells harbor complex genomic rearrangements, with NF1 invalidation necessary yet insufficient for MPNST development. MPNST development likely requires amplification of *PDGFRA*, *EGFR*, or *MET* proto-oncogenes, LOF of *CDKN2A/B* or *TP53* tumor suppressor genes (TSGs) or genes encoding components of PRC2 epigenetic regulator, *SUZ12* and *EED* [1]. The genomic landscape of ANNUBP revealed a common loss of *CDKN2A/B*, without alterations in other genes mutated in MPNST [4]. Despite seminal advances in molecular players and pathways controlling MPNST growth, the mechanisms governing stepwise progression of pNFs into dysplastic and malignant tumors remain poorly understood. This can be accounted for by interpatient molecular heterogeneity of MPNSTs and lack of corresponding animal models.

Over the past decade, numerous genetic tools have been developed to study NF1-associated MPNSTs, including genetically engineered mouse models carrying targeted inactivation of both *Nf1* and several aforementioned tumor suppressors. None of them recapitulated the evolution of pNFs into either MPNSTs or a transitory dysplastic stage. As previously observed, boundary cap (BC) cells, representing a neural crest-derived cell population transiently located at the dorsal root entry zone and ventral exit points of cranial and spinal nerves, give rise to SCs that migrate into the nerve roots and nerves innervating the skin [5]. These are privileged sites where NFs develop in NF1 patients. Using a conditional mutation in *Nf1*, we specifically inactivated the gene in BC cells, showing that mutant mice develop cutaneous NFs and pNFs, which faithfully recapitulate molecular and cellular aspects of human pathology [6, 7]. Upon characterizing this animal model, we discovered that pNFs in mutant mice often transform into MPNSTs while this process undergoes a dysplastic stage, offering a robust system for characterizing dyNFs. Utilizing combined genomics, single-cell transcriptomics, immunohistochemistry, and drug screening approaches, we found that malignant progression of pNFs is initiated by a fate transition of tumor SCs from a glial to a mesenchymal identity, followed by genetic alterations in TSGs. Drug screening studies using this model identified novel compounds targeting the RAS pathway, thereby underscoring its pivotal role.

Results

Spontaneous progression of plexiform neurofibromas to dysplastic and malignant tumors in ***Prss56***^{Cre}, ***Nf1***$^{fl/fl}$ mutant mice

Our mouse model relies on the expression of Prss56 gene, labeling a BC cell subpopulation that generates nerve root SCs and can elicit NFs [5]. It consists of mice carrying the following allele combinations: *Prss56*^{Cre}, *R26*^{tdTom}, and *Nf1*$^{fl/fl}$ or *Prss56*^{Cre}, *R26*^{tdTom}, and *Nf1*$^{fl/-}$ (referred to as *Prss56*^{Cre}, *Nf1*$^{fl/fl}$ or *Prss56*^{Cre}, *Nf1*$^{fl/-}$ mice, respectively). The *R26*^{tdTom} allele enabled identifying cells derived from progenitors having expressed Prss56, which permanently inactivated *Nf1*, yet activated tomato expression (TOM+). As reported, animals aged between 6 and 20 months develop various benign peripheral nerve sheath tumors (PNSTs), including cutaneous NFs (cNFs), and paraspinal and intraneural

(subcutaneous) NFs, both defined as pNFs, with high penetrance [6]. Regardless their location, PNST types display a similar cellular composition with *Nf1*-null (TOM+) SCs, mixed with immune cells, nerve fibroblasts, and endothelial cells, within a dense extracellular matrix (ECM) [6].

We followed mutant animals for longer periods, observing that some *Prss56*^{Cre}, *Nf1*$^{fl/fl}$ and *Prss56*^{Cre}, *Nf1*$^{fl/-}$ older mice also develop large, rapidly growing superficial TOM+ tumors in dorsum, limb, and cranial regions. Due to tumor size/ulceration or movement impairment, these animals require euthanasia at around 13.1±3.2 months (**Figure 1A**, **Figure S1A**). Fast-growing tumors were only observed in group-housed mutant males with numerous bite marks resulting from their naturally aggressive behavior. Lack of tumors in males from the control *Prss56*^{Cre}, *Nf1*+/+ cohort, and in mutant males and females housed individually, suggested their development promoted by skin trauma or inflammation. Fast-growing tumors were never observed in nerve roots where paraspinal pNFs developed.

Gross histological evaluation of lesions (**Figure 1B**, **Figure S1B**), using criteria established for human PNSTs [3], revealed them to display characteristics of either MPNSTs (pronounced hypercellularity and nuclear atypia, frequent mitotic figures [$>4/mm^2$], apoptotic or necrotic cells, and cellular debris) or dyNFs, rather than of cellular NF and ANNUBP types (increased cellularity with fascicular growth pattern; loss of NF architecture; nuclear atypia; rare mitotic figures [$\leq 4/mm^2$]). Numerous lesions exhibited a mixed architecture with varying proportions of both dyNF- and MPNST-like areas, reflecting an ongoing evolution towards malignity (**Figure 1B**, **Figure S1B**, **Table S1**). Among MPNST-like tumors, both low-grade (LG) (moderate hypercellularity and high mitotic index [>4 - ≤ 8]) and high-grade (HG) variants (marked hypercellularity; high mitotic index [>8 - ≤ 30]) were identified, with rare necrosis, in contrast to human MPNSTs (**Figure 1B**, **Figure S1B**). The diagnostic delineation correlated with variations in immune infiltrates. Mast cells were profuse in benign lesions, but scarce/absent in dyNF- and MPNST-like tumors, with other mononuclear cells like macrophages and T lymphocytes being abundant in dyNFs versus both pNFs and MPNSTs (**Figure 1B**, **Figure S1B**, **Table S1**). In dyNF-like tumors, T cells were aggregated around blood vessels (**Figure 1B**, **Figure S1B**). Similar T cell-infiltrated dyNFs were observed in NF1 patients [4, 8]. Details on tumor phenotypes, gender, age at sacrifice, and genetic background of tumor-bearing mice are provided in **Table S1**.

Immunohistochemical (IHC) analyses of pNFs and dyNF- and MPNST-like tumors (**Figure 1C**) confirmed increased density and transformed (sarcomatous) morphology of *Nf1*-mutant (TOM+) cells, with augmented mitotic activity (PH3+), robust macrophages (IBA1), and scarce T cell (CD3+) infiltration in malignant versus benign/dysplastic lesions, as typical for human MPNSTs [3]. Whereas most tumor SCs co-expressed the SC markers S100B and SOX10 in pNFs, these markers were not detected in MPNST-like tumors (**Figure 1C**). Considering dyNF-like tumors, we identified two dyNF types, where tumor cells (TCs) were S100B+/SOX10+ (dyNF S100B+ type) or S100B-/SOX10+ (dyNF S100B- type) (**Figure 1C**). Having never observed TCs with a S100B+/SOX10- signature, we hypothesized *S100b* silencing to precede that of *Sox10*, being possibly the earliest indication of malignant progression.

Moreover, dyNFs and MPNSTs similarly developed in both *Nf1* compound heterozygous (*Prss56*^{Cre}, *Nf1*$^{fl/-}$; 19/48 males = 39.6%) and floxed *Nf1* homozygous (*Prss56*^{Cre},*Nf1*$^{fl/fl}$; 58/81 males = 71.6%) backgrounds. Tumor localization, growth rate and histological characteristics were similar in both genotypes (**Figure 1**, **Figure S1** and **Table S1**). Of note, 30% of *Prss56*^{Cre}, *Nf1*$^{fl/-}$ males, within an age spectrum of 5 to 18 months, were found dead in cages or required euthanasia due to paralysis secondary to paraspinal NFs. Upon excluding premature deaths, the dyNF/MPNST incidence surged to 57.6%. The average age at sacrifice, dictated by tumor burden, was 12.6±3.2 months and 14.1±2.8 months for *Prss56*^{Cre}, *Nf1*$^{fl/fl}$ and *Prss56*^{Cre}, *Nf1*$^{fl/-}$ mutants, suggesting *Nf1* inactivation by Cre recombinase to be very efficient, inactivation of one or two alleles making no much difference.

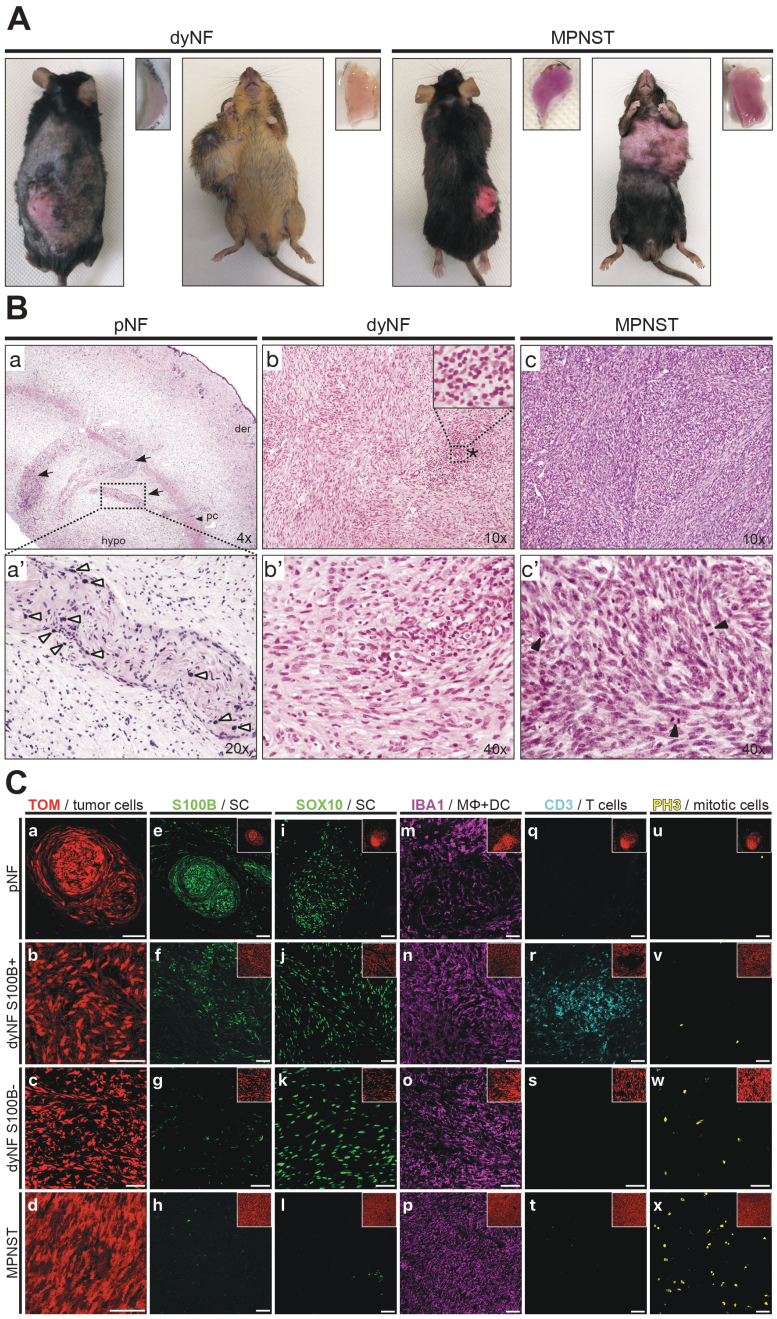


Figure 1. Spectrum of peripheral nerve sheath tumors in *Prss56*^{Cre},*Nf1*$^{fl/fl}$ mice. (A) Gross anatomy of dyNF- and MPNST-bearing males. Dissected tumors are shown on the right. Note enhanced red shading of MPNSTs as compared to dyNFs, reflecting augmented densities of TOM+ neoplastic cells. (B) H&E staining of benign subcutaneous NFs (a, arrows), dyNFs (b), and MPNSTs (c) from *Prss56*^{Cre}, *Nf1*$^{fl/fl}$ mice (for H&E staining of MPNSTs in *Prss56*^{Cre}, *Nf1*$^{fl/-}$ mice, refer to **Figure S1**). High magnification images are shown in the lower panel. Note increasing cellular densities and cellular/nuclear atypia in dyNF and MPNST versus pNF samples. Open arrowheads in (a') indicate numerous mast cells infiltrating pNFs, but absent in dyNFs and MPNSTs. An asterisk in (b) indicates a cluster of mononuclear cells typically found in dyNFs (higher magnification is shown in the inset). Closed arrowheads in (c') indicate frequent mitotic cells in MPNST. der: dermis, hypo: hypodermis, pc: panniculus carnosus muscle. (C) Sections through pNF, dyNF, and MPNST immunolabeled for TOM (tumor cells) (a-d), S100B (e-h) and SOX10 (i-l) (SC lineage), IBA1 (macrophages, dendritic cells, DC) (m-p), CD3 (T cells) (q-t), and PH3 (mitotic cells) (u-x). Insets show TOM staining of corresponding regions. Scale bar = 50µm.

All dyNFs and MPNSTs were located in the hypodermis, often infiltrating subcutaneous fat/muscle and extending into the dermis, raising the question of their precursor lesion's nature. We identified two subcutaneous NF types, "nodular" and "diffuse", in young mutant mice (**Figure S1C**). Nodular NFs corresponded to a well-demarcated accumulation of TOM+ SCs and tumor microenvironment, contributing to a local nerve enlargement (**Figure 1B-C**, **Figure S1C**). Diffuse NFs showed nerve defasciculation with a "diffuse" growth pattern of TOM+ SCs (**Figure S1C**). As they could not be analyzed in detail, we were unable to assess whether the diffuse form corresponded to a more advanced-stage tumor development or distinct NF type. These two types of subcutaneous NF were reminiscent of those observed in human PNSTs [9].

Conclusively, in our mouse model, subcutaneous NFs likely progress to dyNFs and MPNSTs, whose cellular characteristics faithfully recapitulate fast-growing PNSTs in NF1 patients.

Single-cell dissection of cellular and molecular landscapes in benign, dysplastic, and malignant tumors

To decipher how tumor cellular composition, biological activities, and their microenvironment evolve during malignant transformation, we performed single-cell transcriptomic profiling of pNFs, dyNFs, and MPNSTs, issued from *Prss56*^{Cre},*Nf1*$^{fl/fl}$ mice, using the 10x Chromium platform (see Materials and Methods [M&M]). We assembled all tumor samples (2 pNF datasets, each containing a pool of 10-15 subcutaneous pNFs; 9 dyNF and 4 MPNST datasets, each corresponding to individual tumors) into a comprehensive peripheral nerve sheath tumor (pan-PNST) transcriptomic atlas containing 60,126 cells. We used t-distributed stochastic neighbor embedding (tSNE) to visualize all cells in this unified dataset (**Figure 2A**). We employed a gene expression score-based cell-type annotation method, enabling us to identify 12 distinct cell populations. *Tom* expression was a TC marker (**Figure 2A-B**, **Figure S2A**). TCs (*Tom*+) were the dominant group. Since *Prss56* reactivation did not occur during pNF development and malignant progression (**Figure 2B**), all *Tom*+ tumor SCs were considered originating from BC cells. The second group were immune cells (*Ptprc*+) comprising the myeloid lineage, including macrophages/monocytes (Mø/Mo) (*Adgre1*+/*Cd68*+), dendritic cells (DCs), neutrophils (*S100a9*+), and mast cells (*Tpsb2+*), and the lymphoid lineage, with T cells (*Cd3e*+), natural killer (NK) cells (*Ncr1*+), and B cells (*Cd79a*+). DCs exhibited heterogeneity, segregating into monocyte-derived DCs (mDCs) (*Cd209a*+/*Mgl2*+), classical DCs (cDCs) (*Xcr1*+), and plasmacytoid DCs (pDCs) (*Siglech*+). The remaining cells encompassed four distinct cell types: fibroblasts (*Lum*+/*Serping1*+), endothelial cells (*Pecam1*+), mural cells (*Pdgfrb*+/*Rgs5*+), with vascular smooth muscle cells and pericytes, and skeletal muscle cells (*Myl1*+).

We next split up the pan-PNST atlas by tumor type (**Figure 2C-D**) or dataset (**Figure S2C**) to explore how cellular composition and transcriptomic signatures evolve over the transformation steps. All cell populations were shared between all datasets (**Figure S2C**), suggesting that they are not tumor type-specific, although their proportions varied among tumor types (**Figure 2C-D**). All cell types exhibited notable dynamics during transition. Unsurprisingly, the TC proportion progressively increased from the benign to the malignant stage, consistent with their relative proliferative activity (**Figure 1C**, **Figure S2B**), whereas the proportion of immune cells peaked in the dysplastic phase versus both pNF and MPNST, the proportion of all other cell types decreasing with malignity (**Figure 2D**).

The fibroblast (Fb) population was divided into distinct, non-overlapping subsets in the pNF versus dyNF and MPNST datasets (**Figure 3C**). To analyze the Fb population's diversity, the Uniform Manifold Approximation and Projection (UMAP) technique was employed. UMAP delineated four distinct subsets (**Figure 3E**), three of them corresponding to classical peripheral nerve sheath Fb subtypes, namely epineurial Fbs (epiFb) (enriched for *Clec3b* and *Gpnbm2*), perineurial Fbs (periFb) (*Itga6*, *Itga4*), and endoneurial Fbs (endoFb) (*Apod*, *Cp*) [10]. The fourth subset occurred only

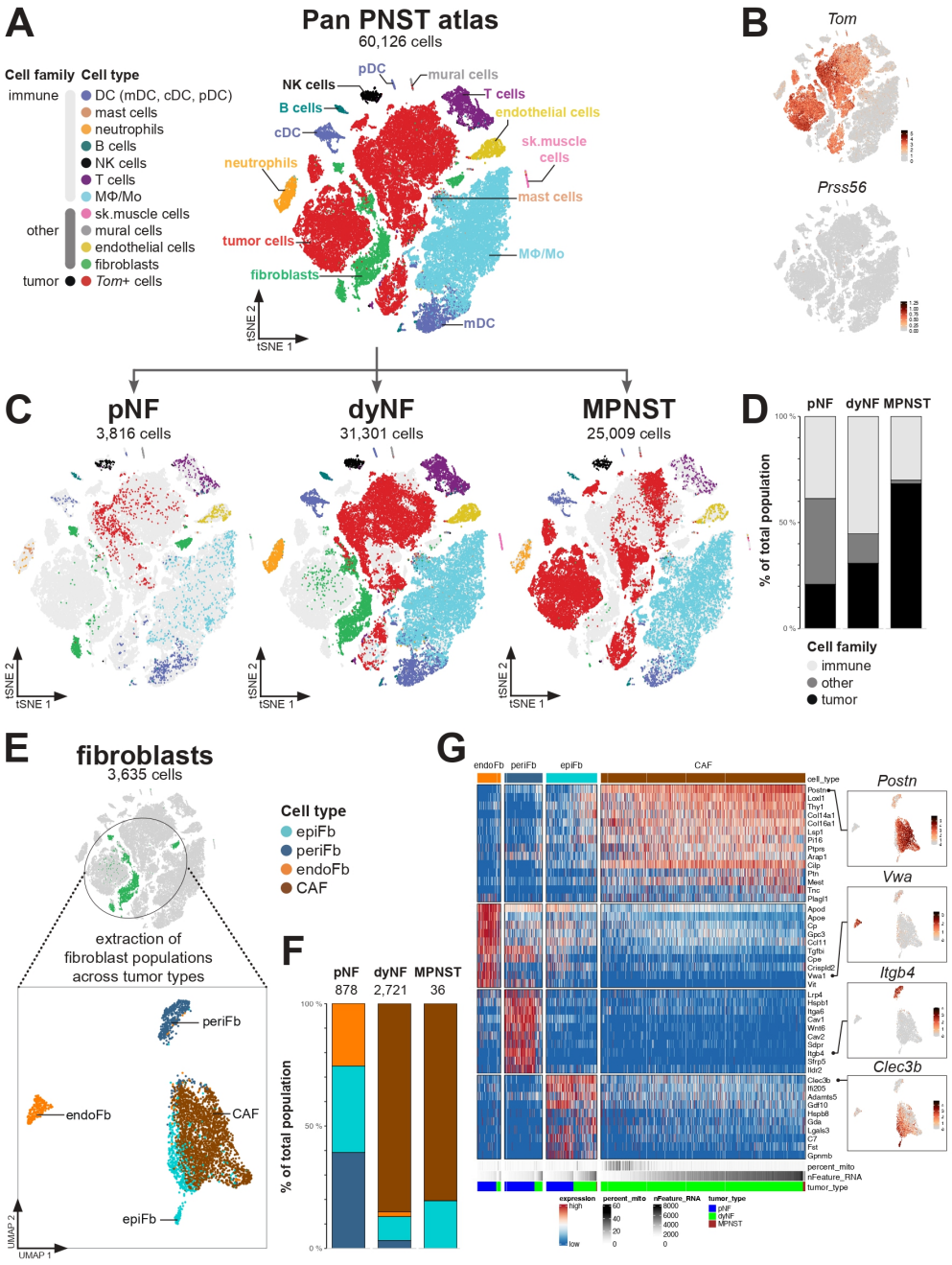


Figure 2. Single-cell profiling of *Prss56*^{Cre}, *Nf1*$^{fl/fl}$ - derived tumors. (A) Pan-PNST transcriptomic atlas assembled from *Prss56*^{Cre}, *Nf1*$^{fl/fl}$ - derived pNF, dyNF, and MPNST datasets. Data are presented as tSNE projection to visualize cellular heterogeneity. Expression score-based annotation identified 12 cell populations. (B) Overlay of the pan-PNST atlas with expression of *Tom* and Prss56. Cells from the complete atlas are shown as light gray background. Cells are color-coded according to gene expression level. (C) Overlay of the pan-PNST atlas with transcriptomic data from each tumor type. Note the distinct locations of fibroblasts and tumor cells between different tumor types. (D) Dynamics of major cell populations throughout pNF-to-MPNST progression. (E) Heterogeneity of the fibroblast population, extracted from the pan-PNST atlas and visualized as UMAP plot. (F) Frequencies of fibroblast subtypes, according to tumor type. Cell numbers are shown on each bar's top. (G) Heatmap representing differentially expressed genes for each fibroblast subtype. The expression of selected transcripts is shown on UMAP, on the right. CAF: cancer-associated fibroblasts, endoFb: endoneurial fibroblasts, epiFb: epineurial fibroblasts, periFb: perineurial fibroblasts.

in dysplastic/malignant stages (**Figure 2E-F**), leading us to annotate them as cancer-associated fibroblasts (CAFs). Using differential gene expression analysis between classical Fb subtypes and CAFs, we identified CAF-enriched transcripts, including *Postn*, *Cilp*, *Loxl1*, *Thy1*, and *Pi16*, previously associated with CAF identity in various tumor types [11, 12] (**Figure 2G**). As CAF transcriptomic signature shared similarities among all three nerve Fb types, we could not determine from which specific type they originated. Missing nerve Fb in dyNF and MPNST can be interpreted as their robust reprogramming or dilution impacting MPNST cellular and transcriptomic landscapes. We compared single-cell transcriptomes of MPNSTs derived from *Prss56*^{Cre}, *Nf1*$^{fl/fl}$ (n=4) and *Prss56*^{Cre}, *Nf1*$^{fl/-}$ (n=3) mutant mice, creating a combined MPNST atlas comprising 47,653 cells (**Figure S2D-E**). When analyzed by genotype, the cell types and transcriptomic profiles showed minimal variability. However, the immune (mainly macrophages/monocytes) fraction was underrepresented in tumors from *Prss56*^{Cre}, *Nf1*$^{fl/-}$ mutants (**Figure S2F-G**). Accordingly, the *Nf1* genetic background in the tumor microenvironment did not significantly impact the cellular composition and transcriptomic activity of MPNSTs, underscoring the similarities between sporadic and NF1-associated MPNSTs and demonstrating that both models can effectively decipher the mechanisms governing their development.

Transcriptional dynamics of *Nf1*-deficient tumor cells during benign-to-malignant progression

We sought to explore the transcriptomic heterogeneity within the TC population extracted from the pan-PNST atlas. *Tom*+ TCs were processed to remove low-quality cells. After batch-effect correction using FastMNN, 16,965 cells were projected by combining diffusion map and UMAP (**Figure 3A**). Diffusion maps helped organizing cells by gradient of gene expression. Subsequent UMAP arranged almost all cells in a connected string, with dyNF cells among pNF and MPNST cells (**Figure 3Aa**), pointing towards a progressive transformation recapitulating the global histological evolution. Only a small *Tom*+ cell subset originating from pNF datasets remained unconnected. These cells were myelinating SCs, as suggested by the expression of myelin-related transcripts like *Mbp*, *Mpz*, or *Pmp22* (**Figure S3A**). Individual sample contributions are shown in **Figure 3Ab**. Note the gradual cell distribution shift, supporting a progressive transformation process. Consistent with our previous findings, most cycling cells originated from MPNST datasets (**Figure 3Ac**).

To identify possible transition states, these cells were subjected to pseudo-temporal (pseudotime) ordering, using the Slingshot trajectory inference package on the batch-effect corrected space [13, 14]. Cells were organized along a linear pseudotime trajectory, passing through successive pNF, dyNF, and MPNST samples (**Figure 3Aa,d**). We employed the dynfeature package to identify genes whose expression varied according to pseudotime. We explored the expression patterns of top-ranked 2000 genes, visualizing on a heatmap 120 representative genes selected based on their high-level expression in specific (pNF, dyNF, or MPNST) tumor types, with cells ordered according to pseudotime (**Figure 3B**). This analysis identified three main cell groups (**Figure 3Ae**). To explore their underlying biology, we performed gene set enrichment analysis (GSEA) between each pair of groups [15, 16] (**Figure S3B**).

Group 1 set was preferentially expressed in TCs from pNF and dyNF samples (early pseudotime) and enriched for roles in gliogenesis (**Figure S3Ba**). This set comprised numerous classical SC-related markers like Plp1, *Sox10*, *S100b*, *Mcam*, *L1cam*, *Ngfr*, *Cdh19*, and *Foxd3*. Group 2 set (mid-pseudotime) was enriched in functional categories related to cell cycle regulation, exemplified by numerous proliferation-related transcripts like *Cenpa*, *Mki67*, *Top2a*, or *Birc5*. Group 2 proliferative cells exhibited high ribosomal RNA content, mainly derived from MPNST samples, yet additionally including a small dyNF cell subset (**Figure 3Aa,c**, **Figure S3Bb**).

Group 3 set was upregulated in the outermost trajectory (late pseudotime), featuring part of the MPNST cells and being enriched with

mesenchymal-like signature. Enrichment was observed in categories linked to ECM degradation/reorganization, cellular migration, hypoxia, and apoptosis (**Figure S3Bc**): ECM and adhesion molecules (*Dcn*, *Mgp*, *Bgn*, *Vcan*, *Cd34*, *Col3a1/2*, *Fn1*, *Mmp2*, *Cdh11/13*, *Vcam1*), growth factors and receptors (*Pdgfra/b*, *Fgfr1*, *Igfbp3/4/6*), stress and metabolism mediators (*Gas1*, *Nupr1*, *Errfi1*), and transcription factors, including *Prrx1*, *Twist1*, *Sox9*, and *Snai1* as known drivers of epithelial-to-mesenchymal transition (EMT) [17] (**Figure 3B**). EMT was previously implicated in malignant progression in epithelial cancers by promoting acquisition of invasive cell phenotype, metastasis, and resistance to chemo- and immunotherapy [18]. According to our findings, a similar mechanism, which we define as glial-to-mesenchymal transition (GMT), was shown to drive malignant progression of dyNFs, supported by the morphological transition of TOM+ cells from spindle-shaped, bipolar SCs in pNFs to large-size, multipolar, fibroblast-like TCs in MPNSTs (**Figure 1Ca-d**). Validation of the down- or up-regulation of selected markers during pNF-to-MPNST progression is presented in **Figure 3C**.

We analyzed in detail the transcriptomic signature of TCs from dyNFs to identify characteristic markers (**Figure 3B**). In all dyNF cells, we observed the extinction of myelinating SC specific gene expression, along with decreasing expression of mature SC-associated markers (*S100b*, *Kcna1*, *Cdh19*). In contrast, SC lineage markers (*Sox10*, *Erbb3*, *Plp1*) were still expressed. Unexpectedly, we identified a group of genes whose expression was restricted to dyNFs (*Moxd1*, *Uchl1*, *Tagln*, *Acta2*). Their biological significance remains unknown. No mesenchymal markers (*Crabp1*, *Sox9*, *Spp1*) were detected in TCs from dyNFs. Using two markers, *Sox10* and *Sox9*, was thus sufficient to distinguish between dysplastic (*Sox10*+/*Sox9*-) and malignant (*Sox10*-/*Sox9*+) TCs. The additional use of *S100b* rendered it likely possible to discriminate between early-(*S100b*+) and late-(*S100b*-) stage dyNFs (**Figure 1C**). Recently, the biallelic loss of *Cdkn2a*, a negative cell cycle regulator, was considered an early event driving MPNST development in NF1 patients4. Accordingly, while *Cdkn2a* was expressed in TCs from pNFs and almost all dyNFs, complete expression loss occurred in MPNSTs (**Figure 3B**).

Accordingly, the notion of cellular malignant progression from pNF to MPNST appears robust, being characterized by gradual glial marker losses, along with robust activation of mesenchymal markers and increased proliferation. The transcription factors *Sox10* and *Sox9* proved sufficient to classify TCs as dysplastic or malignant.

Nf1-KO mice and NF1 patients display similar transcriptome landscapes during malignant progression

We investigated if transcriptional signatures observed upon pNF-to-MPNST progression reflected those of NF1 patients. Therefore, we generated a single-cell transcriptomic atlas from patient-derived dyNFs (N=3) and MPNST (N=1) (**Table S2**), integrating it into publicly available datasets of dyNF (N=1) (Acc. No. GSE16582619), pNFs (N=9), and MPNSTs (N=4) (Acc. No. GSE17903320), creating a unified tSNE representation (pan-PNST atlas) comprising 78,518 cells (**Figure S4A**). Significant heterogeneity in cell composition and transcriptomic signatures across patient samples was observed, possibly accounted for by several factors: (i) several-hour delay between surgical tumor removal and RNA extraction; (ii) frequent incomplete tissue dissociation; (iii) inter-patient variations (tumor size, location, and developmental stage, including HG-MPNSTs). Cells were annotated for cell types using the score-based cell annotation established in our mouse PNST atlas. This strategy effectively separated TCs from other cell types. Since both MPNST TCs and Fbs display high-level expression of mesenchymal markers, we further evaluated their differences using the infercnv package. TCs exhibited more inferred large-scale chromosomal copy-number variations (CNVs) versus Fbs, validating our gene-based annotation.

Once cell types were successfully annotated, the global cellular compositions of pNFs, dyNFs,

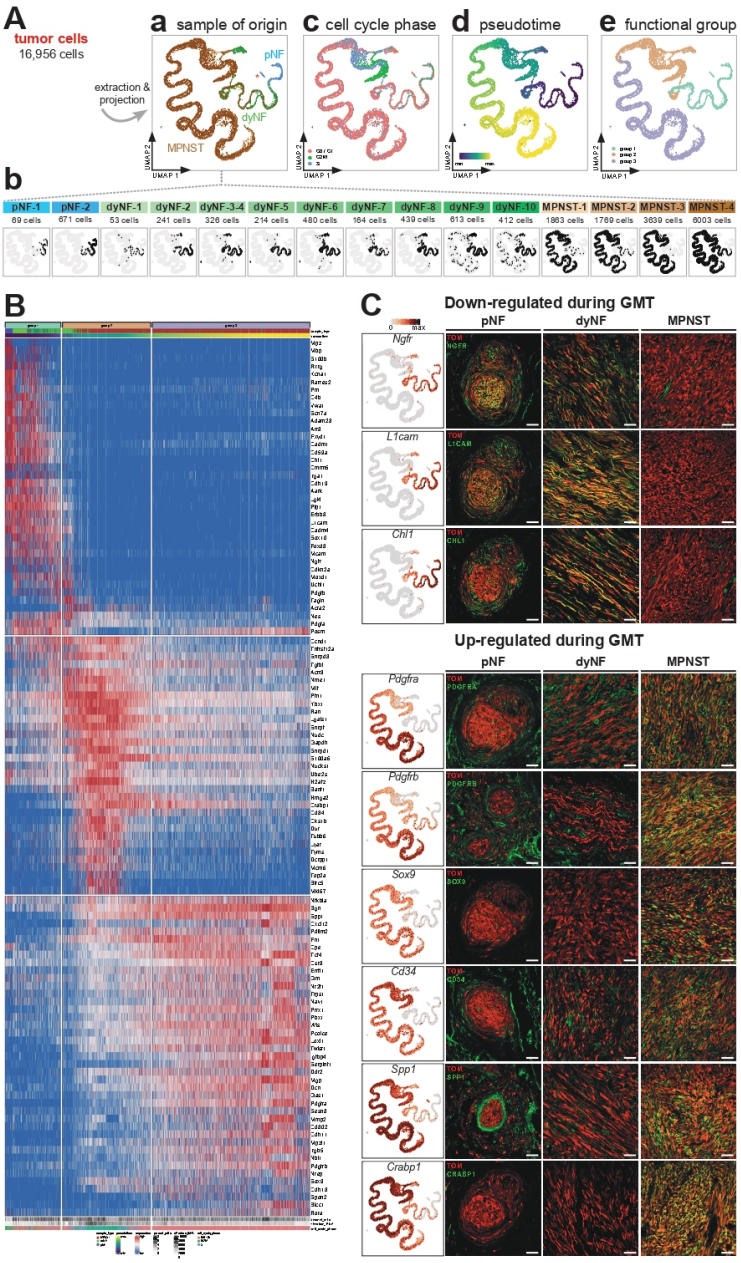


Figure 3. Transcriptional dynamics of tumor cells along malignant progression. (A) Tumor cells were extracted and projected by combining diffusion map and UMAP. On the resulting projection, cells were colored based on sample of origin (a), with individual contribution of pNF (n=2), dyNF (n=9), and MPNST (n=4) datasets (b), cell cycle phase annotation (c), pseudotime inferred with slingshot (d), and functional group (e). In (a), the red star annotates mSCs from pNF samples. See details in **Figure S3A**. In (b), datasets dyNF_3 and dyNF_4 were merged because they originated from the same primary tumor. (B) Heatmap showing transcriptomic changes in tumor cells during pNF-to-MPNST progression. Cells were ordered by increasing pseudotime values. Top annotation shows functional groups, tumor type, and pseudotime. Bottom annotation shows cell cycle phase (cell_cycle_phase) and quality metrics: proportion of expression level related to mitochondrial genes (percent.mito) and total number of genes detected by cell (nFeature_RNA). Genes were divided into three groups: glial signature, cell cycle signature, and mesenchymal signature. (C) Single-cell gene expression levels (left column) and protein detection of selected markers during glial-to-mesenchymal transition. Note that glial markers are co-expressed with TOM+ (tumor) cells in pNF and dyNF, but downregulated in MPNST. Conversely, mesenchymal markers are expressed by non-tumor cells in pNF and dyNF and activated by tumor cells in MPNST.

and MPNSTs in patients resembled those of our mouse model, including an increased TC proportion, decreased Fb number, and macrophage abundance (**Figure S4A**). We then extracted TCs to generate a UMAP (**Figure S4B**). A mouse-like transition in *SOX10*/*SOX9* expression was observed among pNF, dyNF, and MPNST samples (**Figure S4C-D**). In TCs from pNFs and dyNFs, *SOX10* was co-expressed with other glial markers like *S100B*, *SCN7A*, *CDH19*, *CHL1*, and *L1CAM*. Conversely, *SOX9* was activated in MPNSTs along with markers like *CD34*, *PDGFRA*, and *PDGFRB*.

Overall, cell type annotations across both species underscore the robustness of our mouse model, which accurately reproduces the transcriptomic profiles of human PNSTs. An interesting difference emerged: differential gene expression analysis between patient TC and Fb populations identified delta-like non-canonical Notch ligand 1 (*DLK1*) as the most selective marker for MPNST TCs. *DLK1* was not expressed in MPNST-associated Fbs (**Figure S4A**) or in TCs from pNFs and dyNFs. In contrast, *DLK1* expression was not detected in our mouse pan-PNST atlas. As *DLK1* encodes a transmembrane protein, we propose using it to identify and separate TCs from Fbs in human MPNSTs.

Dysplastic NF allografts progress into MPNSTs

To demonstrate that MPNSTs originated from dyNFs through malignant progression, we transplanted TCs isolated from two dyNFs, diagnosed by their histological and transcriptomic signatures, into nude recipient mice (**Figure S5**). In both primary tumors, nearly all *Tom*+ TCs expressed *Sox10* and *Cdkn2a*, but not *Sox9* (data not shown). Grafted cells generated fast-growing tumors collected 8 weeks later. Their transcriptomes, analyzed using scRNA-Seq, were compared with those of primary tumors (**Figure S5Aa-c**). In each grafted tumor, we observed a mixture of *Sox10*+/*Sox9*- and *Sox10*-/*Sox9*+ TCs, suggesting GMT within the tumor population (**Figure S5B**). Using the same methods as for TC profiling from the pan-PNST atlas, both primary and transplanted TCs were ordered by pseudotime values (**Figure S5Ad**), with gene expression profiles on a heatmap (**Figure S5C**). The transcriptomic signature of transplanted TOM+ TCs resembled that of primary MPNSTs (**Figures 3B, S5C**). Whereas *Cdkn2a* was maintained in *Sox10*+ cells, almost all *Sox9*+ TCs were *Cdkn2a*, and all *Cdkn2a*- cells were *Sox9*+, suggesting *Sox9* activation to occur before *Cdkn2a* loss.

These data support the preserved malignant transformation process in our xenograft model (**Figure S5D**), underlining the hypotheses of a GMT and dyNF origin of MPNST cells.

Progression of neurofibromas into MPNSTs is associated with biallelic ***Cdkn2a*** loss in tumor cells

To investigate whether malignant NF progression correlates with acquiring specific genomic alterations, we performed whole-exome sequencing (WES), and copy number variant (CNV) and single nucleotide variant (SNV) analyses, on tumor biopsies from *Prss56*^{Cre}, *Nf1*$^{fl/fl}$ (n=7) and *Prss56*^{Cre}, *Nf1*$^{fl/-}$ (n=6) mutant mice. Spleen samples from the same individuals served as internal controls. All 13 tumors were classified as MPNSTs, using histological evaluation by experienced pathologists and TC scRNA-Seq signature, with detailed characteristics provided in **Table S1**.

Based on CNV analysis, 11/13 tumors presented deletions of portions of chromosome 4 covering *Cdkn2a* (**Figure 4A**) [3]. For 10 of these 11 tumors, deletion extended to neighboring regions containing Cdkn2b, Mtap and Ifn(s) loci, whose co-deletion with *Cdkn2a* was reported in patient MPNSTs [19]. Other rare genomic rearrangements comprised amplifications of portions of chromosomes 10, 15, and 19, including *Ptprk*, *Cdk1/4*, *Pten*, *Acta2*, also described in MPNSTs from NF1 patients [20, 21]. SNV across all tumors exhibited a low number of variants (**Figure 4B**). Neither CNV nor SNV analyses detected mutations in known TSGs like *Suz12*, *Eed*, and *Trp53*, commonly linked with sporadic and NF1-associated HG-MPNSTs [4]. CNV and SNV analyses of the remaining two MPNSTs (both *Prss56*^{Cre}, *Nf1*$^{fl/-}$) revealed no genetic alterations in *Cdkn2a* or any other MPNST-associated TSG, with a low number of variants. Strikingly, 3/13 tumors exhibited

pathogenic variants in *Ptpn11* (encoding SHP-2) found in 50% of Noonan syndrome patients [22] (**Figure 4B**). Recent studies revealed *PTPN11* loss combined with NF1 inactivation to promote malignancy of NF1-deficient cells [23].

As shown, biallelic loss of *Cdkn2a* is the most common and probably the earliest genomic event occurring in MPNSTs. We did not identify *Cdkn2a* loss in two of the 13 analyzed tumors, thus being probably not mandatory for tumor progression.

Successive MPNST grafting leads to pathogenic variant acquisition in tumor cells

We sought to determine whether the low number of variants and absence of pathogenic variants in TSGs like *Suz12*, *Eed*, and *Trp53*, considered to be a genetic hallmark of MPNSTs, were either due to our model's particularity or rapid tumor growth, dictating early animal sacrifice within 2-3 weeks. Therefore, we performed four successive transplantations of TCs isolated from two *Cdkn2a*- and two *Cdkn2a*+ MPNSTs. Approximately 100,000 TOM+ TCs from each primary tumor were grafted subcutaneously in nude recipients. After 2-4-week monitoring, tumors were extracted and dissociated, with a fraction (10,000 cells) re-grafted using the same procedure. After four passages, tumors were processed for WES followed by CNV and SNV analyses, with results detailed in **Figure 5**.

First, we observed accelerated growth of all tumors across successive passages, suggesting the acquisition of additional genetic aberrations promoting TC proliferation. Genomic analyses confirmed a significant increase in variants in all transplanted tumors (**Figure 5A**). Specifically, in both *Cdkn2a*+ transplanted MPNSTs, a loss of this marker was identified. In tumors from mice grafted with *Cdkn2a*- TCs, variants were found in *Trp53*, *Eed*, *Pdgfra*, and *Pdgfrb*, including a pathogenic variant of *Trp53* (c.515G>A, p.Arg172His) [24].

To assess the impact of these genetic aberrations, comparative IHC profiling of primary and transplanted tumors was performed (**Figure 5B**). All samples showed a loss of glial markers SOX10 (**Figure 5B**) and S100B (data not shown), with the acquisition of GMT marker SOX9 in TCs (**Figure 5B**). The rare SOX10+ labeling corresponded to TOM- SCs from tumor-infiltrating nerves (**Figure 5Bd**). Phospho-histone H3 (PH3) labeling revealed *Cdkn2a* inactivation in TCs, with a high proportion of variants including the pathogenic *Trp53* variant, correlating with increased proliferative activity (**Figure 5B-C**).

We performed immunolabeling for the tri-methylation of lysine 27 of histone H3 (H3K27me3), an epigenetic mark indicating Polycomb Repressive Complex 2 (PRC2) activity. PRC2 catalyzed this epigenetic mark, commonly associated with gene repression [25]. H3K27me3 proved often lost in MPNSTs from NF1 patients [19]. H3K27me3 labeling was significantly reduced in TOM+ TCs from *Cdkn2a*- primary and transplanted tumors versus sparse TOM- SCs from the same samples (arrowheads in **Figure 5Bm-o**) and *Cdkn2a*+ primary tumors (**Figure 5Bl**).

In summary, successive grafting of MPNST cells resulted in rapidly growing tumors with progressive accumulation of genetic aberrations, including pathogenic variants in *Trp53*. In most tumors, the initial event involved a biallelic *Cdkn2a* loss, followed by an increased number of variants and emerging pathogenic variants in other TSGs. This increased genetic burden correlated with accelerated proliferation and reduced PRC2 activity in TCs.

Sox9 knockdown impedes MPNST growth ***in vivo***

To investigate the role of genes whose activation in TCs correlates with GMT, we selected six candidates (*Sox9*, *Pdgfra*, *Pdgfrb*, *Mif*, *Cxcl12*, and *Crabp1*) based on their strong and sustained activation during GMT in both mouse and patient tumors (**Figure 3B**, **Figure S4C**). Some of them, including *Sox9*, are known to be involved in the development of MPNSTs and other cancers [26, 27]. Each candidate gene was inactivated in primary mouse MPNST cells with a *Sox10*-/*Sox9*+/*Cdkn2a*- signature using shRNA-mediated gene knockdown (KD) (see M&M). The efficacy of stable KD, measured by RT-qPCR, varied among the target genes (**Table S3**). For each candidate, at least two clones were used for transplantation into athymic nude mice. Approximately 10,000 TCs transduced with

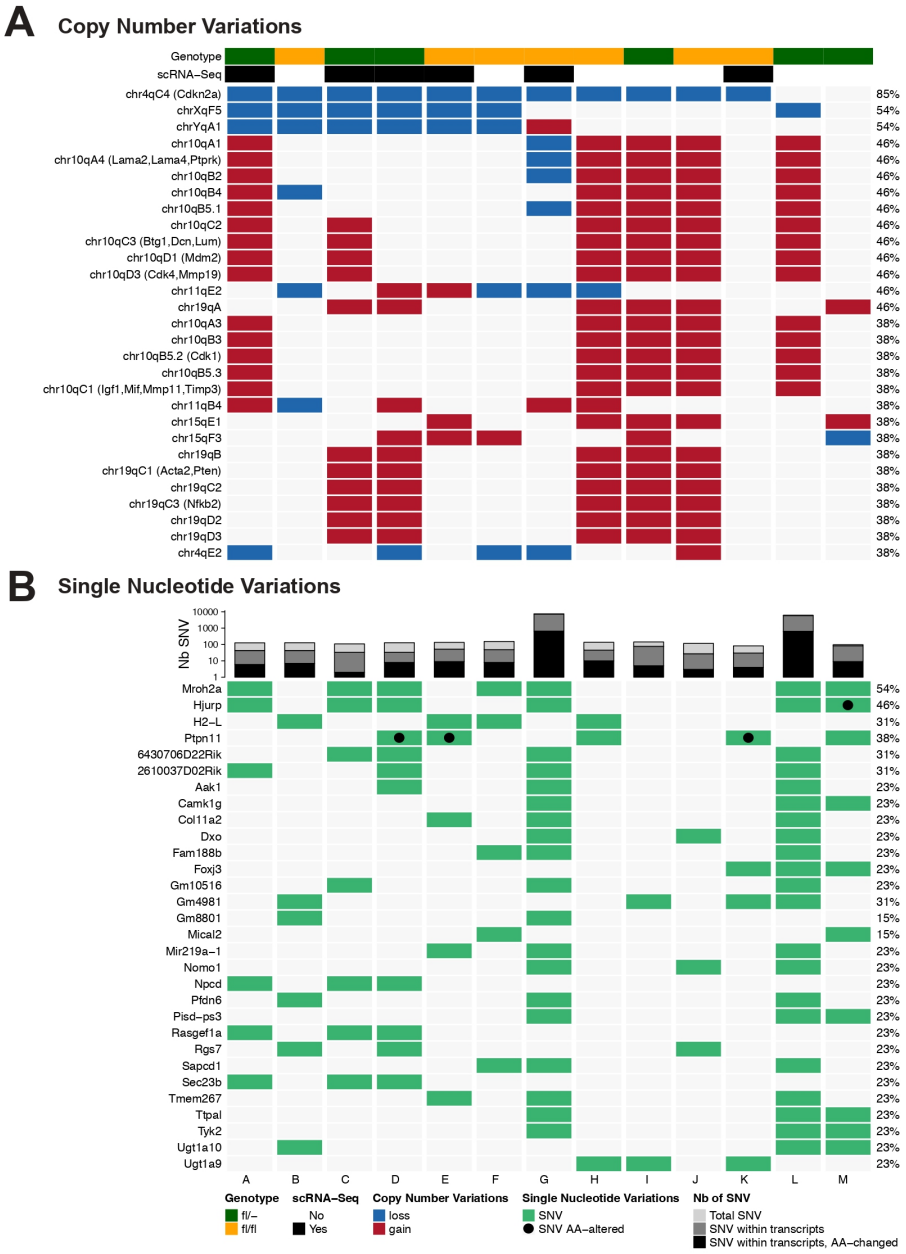


Figure 4. Genomic analysis of tumor cells isolated from mutant mouse MPNSTs. (A) Oncoplot showing CNV of tumor cells. Each column corresponds to an independent primary tumor. Top annotation shows *Nf1* genotype of mice and if the tumor has been characterized by scRNA-Seq. The rows below correspond to chromosomes with either gain or loss of genetic material. Note that 11/13 tumors show loss of parts of chromosome 4 carrying the *Cdkn2a* allele. (B) Oncoplot showing SNV of tumor cells. Columns are matched with panel (A). The top line represents the SNV number for each tumor. A selected list of genes carrying SNVs in different samples is shown on the rows below. SNVs with described pathogenic variants are indicated by a black dot. Note the presence of a pathogenic variant of Ptpn11 in 3/13 tumors. Frequencies of genetic events (CNV and SNV) are shown on the right. Color code for legends is shown below.

target shRNA or an empty vector (control; sh-ctrl) were injected into the left and right flanks, respectively, of six nude mice, and tumor growth was monitored over 2 weeks. All grafted mice were then sacrificed, and both the shRNA-silenced and control tumors were dissected,

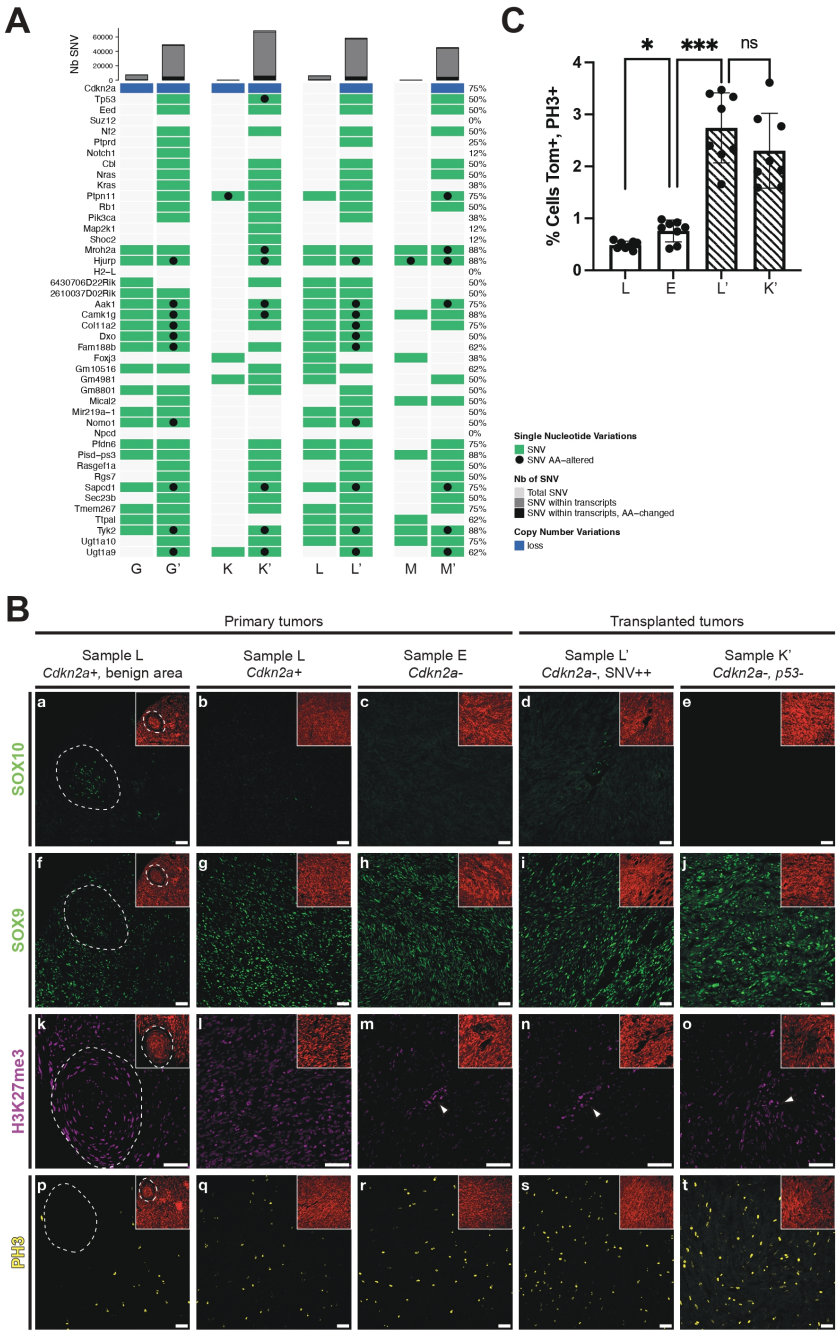


Figure 5. Genomic and IHC characterization of tumor cells from sequentially transplanted MPNSTs. (A) Oncoplot showing WES results of tumor cells. Each pair of columns corresponds to the primary (left) and its associated transplanted (right) tumor. The top line represents the SNV number for each tumor. The row below corresponds to CNV result at *Cdkn2a* locus. Remaining rows correspond to a selected list of genes carrying SNVs, with described pathogenic variants indicated by a black dot. Frequencies of genetic events and color code for legends are shown on the right. (B) Sections through primary and transplanted MPNSTs, immunolabeled for SOX10 (a-e), SOX9 (f-j), H3K27me3 (k-o), and PH3 (p-t). Insets in (a-t) show TOM labelling of the corresponding lesions. Scale bar = 50µm. Each column depicts sections from the same tumor sample. Dashed circle delineates benign areas (nerves) as internal control. Note increased H3K27me3 staining (m-o) in TOM- area pointed by arrowheads (m-o). (C) Quantification of TOM+, PH3+ cycling tumor cells in each tumor. Non-parametric t-test p-value: ns: non-significant, *: p-value <0.05, **: p-value <0.01, ***: p-value <0.001.

weighed, and measured. Despite heterogeneity in tumor mass within each group, significant growth differences occurred when *Crabp1*, *Mif*, and *Sox9* were targeted, highlighting their role in tumor growth (**Figure 6A**).

Sox9 is particularly interesting due to its pivotal role in regulating genes governing EMT. A 60% reduction in *Sox9* mRNA levels prior to grafting was shown to significantly impede tumor growth (**Figure 6A-B**). IHC characterization of two sh-ctrl and two sh-*Sox9* tumors using markers for TOM (TCs), SOX9, and PH3 (mitosis) revealed substantially lower SOX9 protein levels and reduced mitotic cells within the sh-*Sox9* tumors versus controls (**Figure 6C**). Additionally, TC density was notably decreased, aligning with an overall reduction in tumor volume, underscoring the crucial role of *Sox9* in MPNST growth. SC markers SOX10 and S100B were not reactivated.

These findings reinforce *Sox9*'s relevance in MPNST progression and its potential as a therapeutic target for preventing GMT.

Drugs targeting the RAS pathway efficiently inhibit ***Sox9*** expression in tumor cells from mouse MPNSTs

As *Sox9* inhibition in mouse MPNST cells reduced tumor growth after transplantation in nude mice, we hypothesized that compounds inhibiting *Sox9* expression would impede MPNST development. To identify such molecules, we conducted high-content screening of a biologically active compound library using a modified version of our primary MPNST culture system (see M&M). The cultured TCs (*Sox10*-/*Sox9*+/*Cdkn2a*-) were derived from primary mouse MPNSTs. In initial experiments, amplified TCs were transfected with siRNA targeting *Sox9* to measure inhibition efficacy, showing a $\geq$50% reduction in SOX9 protein levels (**Figure 7A-B**). Drug screening was conducted with the European Drug Discovery Consortium (EuLeadFactory), using a library of 3500 biologically active annotated compounds, following the schematic design presented in **Figure 7C**. Overall, 228 compounds leading to a $>$20% reduction in nuclear SOX9 signal intensity versus a (dimethyl sulfoxide) DMSO control were identified. These molecules were tested in dose-response curves, ranging from 30nM to 30µM. Using a nuclear SOX9 signal intensity reduction cut-off $<$50% of the DMSO control, we identified 12 candidates, all being FDA-approved drugs (**Table S4**). Measuring their impact on TC proliferation was challenging due to the short (24-hour) incubation time and induced morphology changes, rendering automatic cell counting inaccurate. Nevertheless, all 12 compounds are known inhibitors of distinct elements of the Ras signaling pathway, including the MAPK pathway, for which NF1 is a negative regulator (**Figure 7D**). Despite the presence of numerous inhibitors of other signaling pathways (Wnt, HH, TGF, Notch) involved in the terminal phase of malignant transformation [4], none of them was identified as potent inhibitor of *Sox9* expression, reinforcing our findings' relevance.

These results clearly validate our culture system for identifying drugs targeting MPNSTs. As a proof-of-concept, we used the GMT marker *Sox9*, since its inhibition delays tumor growth in vivo. *Sox9* expression in TCs is likely regulated by the RAS signaling pathway, thus being possibly a promising therapeutic target for combating MPNSTs.

Discussion

A ***Nf1***-KO mouse model faithfully recapitulating sporadic and NF1-derived MPNST pathogenesis

We herein characterized a NF1 mouse model that accurately recapitulates the development of sporadic and NF1-derived MPNSTs. In this model, MPNST development on wild-type or *Nf1*+/- backgrounds, assumedly representative of sporadic and NF1-derived tumors, suggested common pathogenic mechanisms. Future therapeutic agents are expected to be equally effective for both tumor types.

In NF1 patients, nodular pNFs are at risk of malignant progression through a dyNF intermediate step [28]. However, this pathway has not been firmly demonstrated, nor has the stepwise progression of dyNFs into MPNSTs deciphered. Moreover, nodular and diffuse pNFs were not identified and characterized in non-NF1 patients. In our model, both nodular and diffuse subcutaneous pNFs occurred. Recently, scRNAseq analyses performed on nodular and

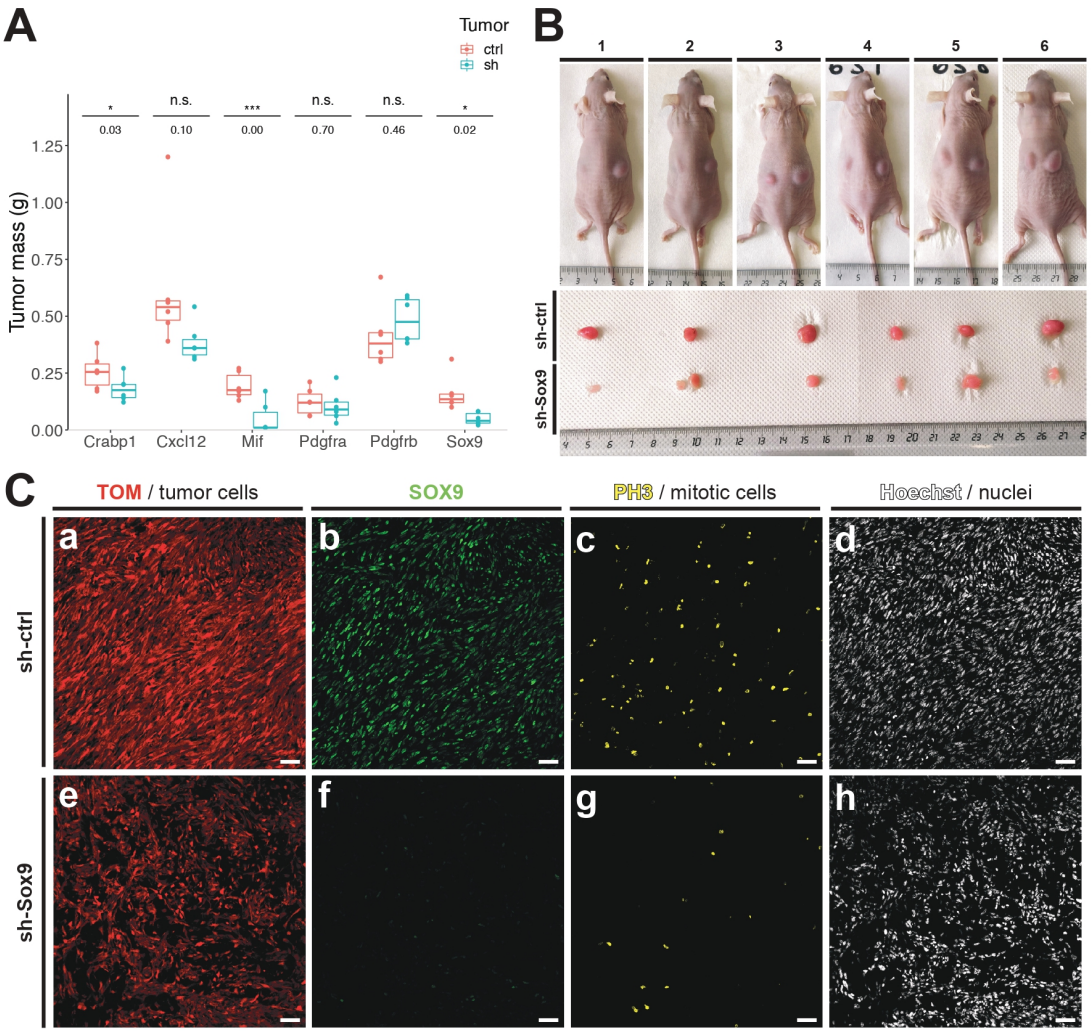


Figure 6. *Sox9* knockdown impedes MPNST growth *in vivo*. (A) Graph representing tumor mass comparison between tumors transfected with control shRNA (ctrl, red) and shRNA directed against the indicated candidate genes (sh, blue). Each dot corresponds to one transplanted tumor. Non-parametric t-test p-value: ns: non-significant, *: p-value <0.05, **: p-value <0.01, ***: p-value <0.001. (B) Macroscopic view of nude mice 2 weeks after subcutaneous injection of primary tumor cells, transfected with control shRNA (right flank) or anti-*Sox9* shRNA (left flank), along with dissected tumors on bottom part. (C) IHC of tumors developed in nude mice transplanted as indicated above, and labeled for TOM (tumor cells) (a,e), SOX9 (b,f), PH3 (mitotic cells) (c,g), and Hoechst (nuclei) (d,h). Scale bar = 50µm.

diffuse pNFs revealed the presence of CAFs (Fb population transiently present in dyNFs) and of tumor SC expressing dyNF markers (*Moxd1*, *Uchl1*) only in nodular NFs (unpublished results). Thus, nodular NFs are likely at the origin of dyNFs. As nodular NFs in NF1 patients are small and often asymptomatic, they could remain undetectable in non-NF1 individuals and, in some situations, become malignant. By grafting dyNF-derived TCs in nude mice, we demonstrated that they are at the origin of MPNSTs. In our mouse model, some nodular subcutaneous pNFs progressed into dyNFs, which transformed into MPNSTs, completely illustrating human MPNST pathogenesis.

Despite the numerous previous NF1 mouse models leading to pNF development, their frequent spontaneous progression into MPNSTs, as in our model, is unprecedented. This highlights that BC cells are likely at the origin

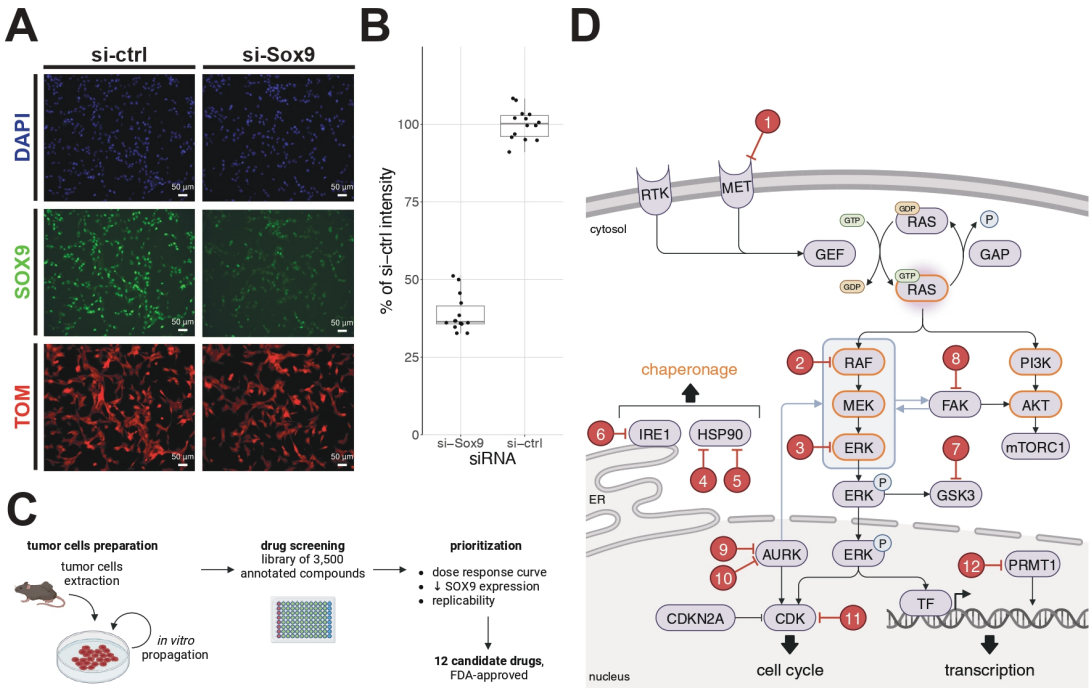


Figure 7. *Sox9* expression in tumor cells could be driven by the Ras signaling pathway.. (A) Tumor cells culture from primary MPNST, transfected with control siRNA (si-ctrl) or siRNA against *Sox9* transcript (si-*Sox9*), immunolabeled for DAPI (nuclei), SOX9 and TOM (tumor cells). Note decreased SOX9 labeling upon exposure to anti-*Sox9* siRNA. Scale bar = 50µm. (B) Quantification of SOX9 staining intensity in TOM+ cells, transfected with control siRNA (si-ctrl) or siRNA against *Sox9* transcript (si-*Sox9*). (C) Experimental design of in vitro screening of 3,500 annotated compounds, leading to the identification of 12 FDA-approved candidate drugs. (D) Representation of molecular targets for 12 candidate drugs in relation with Ras signaling pathway. IRE1 and HSP90 are involved in the protein chaperonage with orange bold border. Correspondence between drug number and identity is attached in **Table S4**. RTK: receptor tyrosine kinase, TF: transcription factor, GEF: guanine nucleotide exchange factor, GAP: GTPase-activating protein. Figure made with Biorender.com.

of MPNSTs. Recent studies identified rare glial stem-like cells (GSCs) in peripheral nerves, characterized by spherogenic capacity and nestin expression, as potential progenitors of MPNSTs [29]. Consistently, we revealed BC-derived glia from adult subcutaneous nerves to comprise spherogenic GSCs, supporting that these cells might give rise to MPNSTs. Further characterizing these cells will be decisive in determining the cancer stem cell origin of MPNSTs.

Glial-to-mesenchymal transition as a step to malignancy and potential therapeutic target

Although substantial progress in the cellular and molecular MPNST characterization was made over the last decade, the mechanisms underlying malignant initiation and progression remain poorly understood. This is largely due to their asymptomatic nature, being undetectable at their early malignancy phase, and the lack of animal models faithfully recapitulating this process. In our model, NFs frequently progressed to dyNFs, which evolved into MPNSTs, with transcriptional signatures sharing commonalities with MPNSTs from NF1-patients. This genetic tool appears thus appropriate to decipher the process of malignant transformation in both species. Using a scRNAseq approach, we built a comprehensive PNST transcriptomic atlas of more than 60,000 cells, including TCs and their microenvironment, such as pNFs, dyNFs, and MPNSTs. This is a major asset for the scientific community interested in MPNST pathogenesis and sarcoma development. Its analysis already revealed pNF-to-dyNF evolution to involve progressive extinction

of SC markers, yet with preserved expression of SC lineage marker *Sox10*. Contrarily, the dyNF-to-MPNST transition was characterized by *Sox10* expression loss and, in parallel, robust activation of mesenchymal markers. These include *Sox9* and several other transcription factors being involved in EMT and regulated by *Sox9* upon development and cancer formation [30]. In analyzed tumors, TCs never co-expressed *Sox10* and *Sox9*, suggesting a likely epigenetic mechanism responsible for simultaneous extinction of *Sox10* and activation of *Sox9* and other mesenchymal markers. We observed an increased TC proliferative activity, with morphological transition from thin bipolar glial cells to large multipolar mesenchymal cells, without genetic alterations. This argues in favor of an abrupt GMT, being possibly pivotal in malignant progression. Combined ATACseq/RNAseq of TCs during the GMT is particularly useful to investigate the underlying mechanisms. In all TCs analyzed, *Cdkn2a* loss only occurred when *Sox9* was activated, suggesting this latter event to precede genetic alterations. In dyNFs, we found GMT to occur without genetic alterations in *Cdkn2a* and other TSGs. This is not due to any of our model's particularities, as *Cdkn2a* was rapidly altered/lost in TCs following primary tumor transplantation in nude mice. Accordingly, genetic alterations are not mandatory for GMT, whereas GMT might promote genetic instability, starting with *Cdkn2a* loss followed by various additional alterations. Blocking GMT might thus prevent malignant transformation and reverse TCs to their initial glial fate.

Multi-targeted inhibition of the RAS pathway as a possible therapeutic approach to block MPNSTs

Numerous therapeutic options were explored to treat MPNSTs, with mostly negative or inconclusive results. Therefore, we developed a TC culture system derived from primary tumors that could be maintained over extended periods. This system preserves tumorigenic properties, including *Sox9* expression, rapid proliferation, and MPNST development after transplantation into nude mice. Using this system for drug screening, we employed a library of 3,500 annotated compounds, identifying 12 drugs that reduced *Sox9* expression. All were inhibitors directly or indirectly targeting the RAS pathway. This finding highly suggests a very strong link between the RAS pathway and *Sox9* expression and possibly GMT. Interestingly, the identified inhibitors target various nodes within the RAS pathway, including growth factor receptors and intracellular regulators. Therefore, rather than targeting a single branch of the RAS pathway, using a combination of compounds targeting multiple branches might constitute a way to completely silence the pathway, preventing *Sox9* expression and GMT, and possibly delaying or avoiding malignant transformation.

Materials and Methods

Animals

Mice used in this study were housed in a temperature- and humidity-controlled vivarium on a 12 h dark-light cycle with free access to food and water. We bred *Prss56*$^{Cre/Cre}$, *R26*$^{tdTom/tdTom}$, *Nf1*$^{fl/fl}$ mice with the *Nf1*$^{fl/fl}$ or *Nf1*+/- mice to obtain *Prss56*$^{Cre/+}$, *R26*$^{tdTom/+}$, *Nf1*$^{fl/fl}$ or *Prss56*$^{Cre/+}$, *R26*$^{tdTom/+}$, *Nf1*$^{fl/-}$ mutants, respectively. Genotyping strategies were described earlier [31]. All animal manipulations were approved by a French Ethical Committee (APAFIS #19581) and were performed according to French and European Union regulations.

Human samples

Tumor specimens were kindly provided by NF1 patients undergoing surgical resections at Henri Mondor hospital, Créteil, following written informed consent. Tumor samples were excised from 3 ANNUBP and 1 MPNST, placed in ice-cold RPMI and transported immediately to our laboratory for further processing. Information regarding patient datasets is provided in **Table S4**.

Histopathology and immunohistochemistry

Tumors were dissected, immersion fixed in 4% paraformaldehyde overnight at 4°C and processed for paraffin or cryo embedding. Paraffin sections were cut at 5µm and stained with hematoxyline-eosine (H&E) for histopathological

evaluation. Cryosections were cut at 14µm and immunolabeled with the following primary antibodies: rat anti-tdTomato (1:1000, Kerafast), rabbit anti-RFP (1:1000; Rockland), rabbit anti-S100B (1:500, Dako), rabbit anti-SOX10 (1:200, LSBio), rabbit anti-IBA1 (1:400, Wako), rabbit anti-CD3 (1/100, Abcam), rabbit anti-PH3 (1:200, Merck), rabbit anti-NGFR (a1:500, Merck), rat anti-L1CAM (1:100, Merck), goat anti-CHL1 (1:100, R&D Systems), goat anti-PDGFRA (1:100, R&D Systems), goat anti-PDGFRB (1:100, R&D Systems), rabbit anti-SOX9 (1:500, Abcam), goat anti-CD34 (1:50, R&D Systems), rabbit anti-SPP1 (1:100, Abcepta), rabbit anti-CRABP1 (1:100, Merck), rabbit anti-H3K27me3 (1:1000, Cell Signaling). Fluorophore-conjugated secondary antibodies were from Jackson Immuno Research. Briefly, sections were blocked/permeabilized for 1h at room temperature with immunoblock (IB) solution (4% bovine serum albumin (BSA)/ 0.3% Triton X-100 in PBS), then incubated overnight with the primary antibodies in IB at 4°C, washed and incubated with the secondary antibodies for 2h at room temperature. Cell nuclei were counterstained with Hoechst (Life Technologies). Images were acquired on Zeiss LSM 900 confocal microscope, then processed and assembled using ImageJ and Photoshop.

Sample processing and single-cell RNA sequencing

Mice were euthanised via isoflurane inhalation and perfused with PBS to remove circulating immune cells. Tumors were dissected and sampled for (immuno)histological and genomic analyses. Remaining tissue was minced into a slurry and digested in RPMI-1640 medium (Gibco, #42401-018) containing 0,25 mg/ml of Liberase HD (Roche, #05401089001) for 30 min at 37°C. Enzymatic digestion was halted by adding 10% FBS in RPMI (Gibcon #10270106) and tumor cell suspensions were collected by centrifugation (1200 rpmi for 5 min). After mechanical dissociation and filtering (30µm MACS SmartStratiners, Miltenyi Biotec, #130-098-458) single-cell suspensions were labelled with DAPI (Invitrogen, #D1306) and subjected to FACS to recover viable cells. Human tumors were processed immediately after surgical resection using the same procedure as described above. 20,000 of viable murine or human cells were loaded into one channel of the Chromium system using the V3 and V3.1 single-cell reagent kits (10X Genomics). Following capture and lysis, cDNAs were synthesized, then amplified by PCR for 12 cycles as per the manufacturer's protocol (10X Genomics). The amplified cDNAs were used to generate Illumina sequencing libraries that were each sequenced on one flow cell Nextseq500 Illumina.

Bioinformatic analysis

BCL files were processed using 10x Genomics Cell Ranger software (version 2 or 3). For mouse PNSTs, reads were mapped onto a custom mouse transcriptome based on GRCm38 transcriptome, in which we added tdTomato sequence. For original human PNSTs, mapping was made on GRCh37 transcriptome. Raw count matrices were used for subsequent analyses. Publicly available scRNA-Seq datasets from human PNSTs were downloaded as filtered count matrices.

Count matrices were processed using RStudio and R Markdown to generate fully traceable notebooks, including the figures included in this article. To ensure version stability, a Singularity container containing R (version 3.6.3) and all packages required for the analyses was developed and used to compile notebooks (**Table S5**). HTML reports, including the parameter values of all functions, and code to make the analyses and the figures, are available in a Github repository (see the Code and data availability Section).

From the count matrices, datasets were analyzed individually using Seurat V3 package [32]. Cells having less than 26 UMI or less than 600 genes were filtered out. Doublets were removed using the scDblFinder tool [33] and scds in hybrid mode [34]. Then, cells having more than 50 % of UMI related to mitochondrial genes or more than 30% related to ribosomal genes were filtered out. The UMI count matrix for remaining cells was log-normalized using the Seurat::NormalizeData function. Cell cycle annotation was made

using Seurat::AddCellCycle. Single cells were annotated for cell type using the Seurat::AddModuleScore function and cell type-specific marker gene sets listed in **Table S6** when annotating all cells, or in **Table S7** for the fibroblast dataset.

The same annotation procedure was applied to the human datasets, with an additional analysis of copy number variations using the infercnv package to assess the annotation of tumor cells, using macrophages as reference.

Individual datasets were merged to generate combined datasets. For the human PNST atlas, the intersection of Ensembl gene identifiers between original and published datasets was used to homogenize count matrices. For each individual dataset, cells were extracted based on cell type annotation, and then merged to generate the tumor cells and fibroblasts-specific datasets. Seurat::PCA and harmony [35], or FastMNN [36] were used to remove bath-effect. Cell projections were made using Seurat::RunUMAP, Seurat::RunTSNE, or diffusion map (DM) from the destiny package [37]. Batch-effect correction and projection methods are summarized in the table below.

Dataset	Cell type	Correction	Projection
Pan-PNST atlas	all	PCA - harmony	tSNE
Pan-PNST atlas	tumor cells	FastMNN	DM - UMAP
Pan-PNST atlas	fibroblasts	FastMNN	UMAP
MPNST atlas	all	PCA - harmony	tSNE
Primary/Transplanted	tumor cells	FastMNN	UMAP
Human PNST atlas	all	PCA - harmony	tSNE
Human PNST atlas	tumor cells	FastMNN	UMAP

Trajectory inference was made using two distinct methods: slingshot [13] and TInGa [38], using cells from the pNF datasets as trajectory root. For the functional analysis, the tumor cells were divided into three subsets using k-means function from the stats base package, on the batch-effect corrected space. Differential expression was conducted using Seurat::FindMarkers. Heatmaps were made using the ComplexHeatmap package [39]. Functional enrichment analyses were conducted using the msigdbr package (`https://CRAN.R-project.org/package=msigdbr`) for the gene sets database, and clusterProfiler package for the analyses [40].

Whole exome sequencing

Raw sequences were demultiplexed to FASTQ, taking care of UMIs, with bcl2fastq v2.20.0.422 using default parameters. Raw reads were trimmed for bad quality bases, adapter content and polymere tracks using fastp v0.23.2. General reads QC were assessed using fastqc v0.12.1, sample contamination with FastQ_Screen v0.15.3 and results were aggregated with multiQC v1.13. Mapping against the murine genome UCSC mm10 was performed with bwa-mem2 v.2.2.1 with default parameters, except for "-A 2 -E 1". Mapped reads QC was performed using samtools stats/flagstat/idxstats v1.18, and coverage was assessed using mosdepth v0.2.6. Deduplication of reads thanks to UMIs was performed using umi_tools v1.1.2. PCR and optical duplicates marking were performed using Picard tools v2.23.5. Somatic variants calling was performed using the GATK mutect2 v4.2.6.1 with recommended guidelines [41] . Variants annotation and filtering were performed using snpEff/snpSift v4.3.1t [42]. Copy number analysis of tumor/normal pairs was performed using EaCoN v0.2.8 using the ASCAT2 segmentation method, with default parameters [43].

Tumor cell culture

Cells dissociated from mouse MPNST were frozen in 90% FBS/10% DMSO and stored at -150°C. Once needed, vials were thawed by gentle agitation in a water bath at 37°C, DMSO was removed and cells were resuspended and cultured in DMEM, 10% FBS and 1% P/S in T75 flask (Thermofisher). As soon as the primary cells reach confluence, they are sub cultured in a new T75 flask by 1/10 dilutions, twice a week.

ShRNA-mediated Sox9 knockdown

Lentivirus encoding shRNA targeting *Sox9* (TRCN0000321420), *Cxcl12* (TRCN0000184347), *Pdgfra* (TRCN0000321928), *Pdgfrb* (TRCN0000321931), *Crabp1* (TRCN0000011959) and *Mif* (TRCN0000067344) were obtain from Sigma.

Viral infection was carried out as described by the manufacturer (MISSION Lentiviral Transduction Particles) at MOI 5. Briefly, murine primary MPNST cells were seeded at 5x103 in 96 well plates and transduced 24 hours later when wells became 50% confluent. Incubation with the virus was carried out overnight in 110µL DMEM containing 10%FBS, 1% P/S and 8µg/mL hexadimethrine bromide. Transduced cells were selected in medium containing 3µg/mL Puromycin for 72h, then cultured in a presence of 1,5 µg/mL Puromycin during clones' expansion. Non-targeting lentivirus (SHC201V and SHC201V, Sigma) were used as control.

Tumor grafting

Approximately 10,000 of *Tom*+ tumor cells from each primary tumor were subcutaneously grafted into nude recipients under isoflurane anesthesia. After monitoring for 2 weeks, tumors were extracted, dissociated and a fraction (10,000 cells) re-grafted using the same procedure.

SiRNA mediated ***Sox9*** knockdown

For siRNA knockdown targeting *Sox9* (ON-TARGETplus Mouse *Sox9* siRNA, L-059108-01-0005, Horizon Discovery) in MPNST cultured cells, we used lipofectamine RNAiMAX (Life Technologies) following the manufacturer's instructions. Non targeting siRNA (ON-TARGETplus Non-targeting Pool, D-001810-10-05) was used as control.

RNA preparation and real-time quantitative RT-PCR

Total RNAs were isolated using the RNAqueous-micro kit (Ambion). After DNAse treatment, concentration of RNA was determined using Nanodrop and QuBit technology. CDNA was synthesized using the Maxima First Strand cDNA Synthesis Kit and quantitative real-time PCR (qRT-PCR) was performed using the Maxima SYBER green/ROX qPCR Master Mix X2 (both from Thermofisher) and amplified using the Steponeplus real time PCR (Applied Biosystems). The relative abundance values of each amplification product were normalized to the internal controls (cyclophilin, tubulin and actin). Dedicated PCR primers are reported below.

Actin	F	TGTTACCAACTGGGACGACA
	R	GGGGTGTTGAAGGTCTCAAA
Tubulin	F	CGGGTCTCCAGGGCTTCT
	R	GGGCTGGGTAAATGGAGA
Cyclophilin	F	CCATCGTGTCATCAAGGACTT
	R	TTGCCATCCAGCCAGGAGGTC
Sox9	F	AGGAAGTCGGTGAAGAACGG
	R	GGACCCTGAGATTGCCCAGA
Cxcl12	F	GCCAACGTCAAGCATCTGAA
	R	TGGGCTGTTGTGCTTACTTG
PdgfrA	F	ACTTCGTGGACAGTTTTGGC
	R	ATCACCAACAGCACCAACAC
PdgfrB	F	ATCGCGCCACTTAATAAC
	R	TCCCTCATGATGTCTCGAGC
Mif	F	GTGAACACCAATGTTCCCCG
	R	TGCACTGCGATGTACTGTGC
Crabp1	F	GGACGCAAATGCAGGAGTTT
	R	AGCTCTCGGGTCCAGTAAGT

High-content drug screening

Compound Annotated Set library

The Annotated Set library is composed of the following individual compound libraries: CLOUD (CeMM Library of Unique Drugs, Enamine), Prestwick Chemical Library®(Prestwick Chemical), Tocriscreen Plus (Tocris) and SGC Chemical Probes. Compounds were stored in Echo 384-LDV plates (LP-0200) as 10mM stock solutions in DMSO.

Compound screening

MPNST cells were seeded at a concentration of 800 cells per well onto 384-well PhenoPlate microplates (6057302, Revvity) and incubated for 24h at 37℃, 5% CO2 prior to compound treatment. Compounds were dispensed onto cells using an Echo acoustic dispenser (Beckman Coulter Life Sciences) at a final concentration of 10µM for the primary screen and incubated for 12h. Dose response experiments were performed at 8 concentrations: 30nM, 100nM, 300nM, 1µM, 3µM, 6µM, 10µM (repeat of the primary screen concentration), and 30µM. Treatment was followed by fixation by addition of equal volume of 8% paraformaldehyde (Sigma) and 20minute incubation, blocking and permeabilization (0.3% Triton X-100, 4% BSA in PBS), incubation with anti-SOX9 primary antibody (1:1000 in blocking buffer, ab185230,

Abcam) and further incubation with secondary antibody (1:1000, Goat anti-Rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor™488, A11008, ThermoFisher Scientific) and DAPI (4',6-diamidino-2-phenylindole, 1:1000, ThermoFisher Scientific) in blocking buffer. All incubations of the staining protocol were carried out at room temperature for an hour unless otherwise specified. Wells were washed using PBS in between staining solutions. Plates were imaged using an IN Cell Analyzer 2200 (GE Healthcare Life Sciences). 6 fields of view were acquired per well, using the 20x objective (0.45 NA). Images were imported into an OMERO Plus database and analyzed using CellProfiler. Data was normalized to DMSO control wells (16 per plate) within the same plate. Two independent experiments were performed for both the primary screen (single technical replicate) and dose response confirmation test (two technical replicates). Compounds were considered potential hits if reducing SOX9 intensity under 80% of the DMSO control in one of the two replicates of the primary screen with more than 50 cells analyzed, or in both replicates with an average $\geq$100 cells analyzed per independent experiment. Hit re-confirmation was determined by similar reduction of SOX9 intensity at screening concentration (10µM) in the dose response experiment.

Code and data availability

scRNA-Seq data are available as FASTQ files and raw count matrices and Whole Exome Sequencing data as FASTQ files in ArrayExpress repository, under the accession number E-MTAB-14222. Detailed methodological steps, exhaustive parameter settings, along with intermediate figures, for the whole count matrices to figure process are available in HTML reports, shared through a Github repository (`https://github.com/audrey-onfroy/Radomska_et_al`).

Declaration of interests

The authors declare no competing interests.

Authors contribution

Conception and design: K.J. Radomska, F. Coulpier, P. Topilko; Development of methodology: K.J. Radomska, F. Coulpier, A. Onfroy, P. Topilko; Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): K.J. Radomska, F. Coulpier, L. Lecerf, T. El Jalkh, L. Sesma Sanz, N. Ortonne, P. Topilko; Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): A. Onfroy, K.J. Radomska, F. Coulpier, A. Beaude, B. Job, L. Sesma Sanz, P. Topilko: Writing and review of the manuscript: K.J. Radomska, A. Onfroy, F. Coulpier, P. Charnay, P. Topilko; Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): P. Wolkenstein, P. Topilko; Study supervision: P. Topilko

Acknowledgments

The P. Topilko group was financed by the Institut National de la Recherche Médicale (INSERM) and University Paris Est Créteil (UPEC). It has received financial support from: Fondation Maladies Rares, Cancéropole Ile de France, Association des Patients CAP NF, Aviesan-Inserm (Preneoplasia, C21049DS), and EuLeadFActory (Phenotypic screening of compounds preventing malignant transformation of neurofibromas in Neurofibromatosis type 1, ELFSC34_09). K.J. Radomska was financed by ARC, no. PDF20150603033 and CAP NF foundation.

We are grateful to the IMRB Imaging, genomic, and animal facilities. We thank the staff from FACS and animal facilities for technical assistance, and C. Colnot and L. Fertitta for critical manuscript reading.

References

1. Bandarigoda N. Somatilaka, Ali Sadek, Renee M. McKay, and Lu Q. Le. Malignant peripheral nerve sheath tumor: models, biology, and translation. *Oncogene*, 41(17):2405–2421, 4 2022.
2. Eline Beert, Hilde Brems, Bruno Daniëls, Ivo De Wever, Frank Van Calenbergh, Joseph

Schoenaers, Maria Debiec-Rychter, Olivier Gevaert, Thomas De Raedt, Annick Van Den Bruel, Thomy de Ravel, Karen Cichowski, Lan Kluwe, Victor Mautner, Raf Sciot, and Eric Legius. Atypical neurofibromas in neurofibromatosis type 1 are premalignant tumors. *Genes, Chromosomes and Cancer*, 50(12):1021–1032, 12 2011.

3. Markku M. Miettinen, Cristina R. Antonescu, Christopher D.M. Fletcher, Aerang Kim, Alexander J. Lazar, Martha M. Quezado, Karlyne M. Reilly, Anat Stemmer-Rachamimov, Douglas R. Stewart, David Viskochil, Brigitte Widemann, and Arie Perry. Histopathologic evaluation of atypical neurofibromatous tumors and their transformation into malignant peripheral nerve sheath tumor in patients with neurofibromatosis 1—a consensus overview. *Human Pathology*, 67:1–10, 9 2017.
4. Suganth Suppiah, Sheila Mansouri, Yasin Mamatjan, Jeffrey C. Liu, Minu M. Bhunia, Vikas Patil, Prisni Rath, Bharati Mehani, Pardeep Heir, Severa Bunda, German L. Velez-Reyes, Olivia Singh, Nazanin Ijad, Neda Pirouzmand, Tatyana Dalcourt, Ying Meng, Shirin Karimi, Qingxia Wei, Farshad Nassiri, Trevor J. Pugh, Gary D. Bader, Kenneth D. Aldape, David A. Largaespada, and Gelareh Zadeh. Multiplatform molecular profiling uncovers two subgroups of malignant peripheral nerve sheath tumors with distinct therapeutic vulnerabilities. *Nature Communications*, 14(1), 12 2023.
5. Aurélie Gresset, Fanny Coulpier, Gaspard Gerschenfeld, Alexandre Jourdon, Graziella Matesic, Laurence Richard, Jean Michel Vallat, Patrick Charnay, and Piotr Topilko. Boundary Caps Give Rise to Neurogenic Stem Cells and Terminal Glia in the Skin. *Stem Cell Reports*, 5(2):278–290, 2015.
6. Katarzyna J. Radomska, Fanny Coulpier, Aurelie Gresset, Alain Schmitt, Amal Debbiche, Sophie Lemoine, Pierre Wolkenstein, Jean Michel Vallat, Patrick Charnay, and Piotr Topilko. Cellular origin, tumor progression, and pathogenic mechanisms of cutaneous neurofibromas revealed by mice with Nf1 knockout in boundary cap cells. *Cancer Discovery*, 9(1):130–147, 1 2019.
7. Fanny Coulpier, Pernelle Pulh, Layna Oubrou, Julie Naudet, Laura Fertitta, Jean-Marc Gregoire, Arnaud Bocquet, Anne-Marie Schmitt, Pierre Wolkenstein, Katarzyna J. Radomska, and Piotr Topilko. Topical delivery of mitogen-activated protein kinase inhibitor binimetinib prevents the development of cutaneous neurofibromas in neurofibromatosis type 1 mutant mice. *Translational Research*, 261:16–27, 11 2023.
8. Meritxell Carrió, Bernat Gel, Ernest Terribas, Adriana Carolina Zucchiatti, Teresa Moliné, Inma Rosas, Álex Teulé, Santiago Ramón y Cajal, Juan Carlos López-Gutiérrez, Ignacio Blanco, Elisabeth Castellanos, Conxi Lázaro, Anat Stemmer-Rachamimov, Cleofé Romagosa, and Eduard Serra. Analysis of intratumor heterogeneity in Neurofibromatosis type 1 plexiform neurofibromas and neurofibromas with atypical features: Correlating histological and genomic findings. *Human Mutation*, 39(8):1112–1125, 8 2018.
9. Nicolas Ortonne, Pierre Wolkenstein, Jaishri O. Blakeley, Bruce Korf, Scott R. Plotkin, Vincent M. Riccardi, Douglas C. Miller, Susan Huson, Juha Peltonen, Andrew Rosenberg, Steven L. Carroll, Sharad K. Verma, Victor Mautner, Meena Upadhyaya, and Anat Stemmer-Rachamimov. Cutaneous neurofibromas. *Neurology*, 91(2_Supplement_1), 7 2018.
10. Matthew J. Carr, Jeremy S. Toma, Adam P.W. Johnston, Patrick E. Steadman, Scott A. Yuzwa, Neemat Mahmud, Paul W. Frankland, David R. Kaplan, and Freda D. Miller. Mesenchymal Precursor Cells in Adult Nerves Contribute to Mammalian Tissue Repair and Regeneration. *Cell Stem Cell*, 24(2):240–256, 2 2019.
11. Chencheng Han, Tongyan Liu, and Rong Yin. Biomarkers for cancer-associated fibroblasts. *Biomarker Research*, 8(1):64, 12 2020.
12. Laura González-González and Javier Alonso. Periostin: A Matricellular Protein With Multiple Functions in Cancer Development and Progression. *Frontiers in Oncology*, 8, 6 2018.
13. Kelly Street, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1):477, 12 2018.
14. Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, 5 2019.
15. Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 1(6):417–425, 12 2015.
16. Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna

Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 5 2000.

17. Zhao Huang, Zhe Zhang, Chengwei Zhou, Lin Liu, and Canhua Huang. Epithelial–mesenchymal transition: The history, regulatory mechanism, and cancer therapeutic opportunities. *MedComm*, 3(2), 6 2022.
18. Thomas Brabletz, Raghu Kalluri, M. Angela Nieto, and Robert A. Weinberg. EMT in cancer. *Nature Reviews Cancer*, 18(2):128–134, 2 2018.
19. Isidro Cortes-Ciriano, Christopher D. Steele, Katherine Piculell, Alyaa Al-Ibraheemi, Vanessa Eulo, Marilyn M. Bui, Aikaterini Chatzipli, Brendan C. Dickson, Dana C. Borcherding, Andrew Feber, Alon Galor, Jesse Hart, Kevin B. Jones, Justin T. Jordan, Raymond H. Kim, Daniel Lindsay, Colin Miller, Yoshihiro Nishida, Paula Z. Proszek, Jonathan Serrano, R. Taylor Sundby, Jeffrey J. Szymanski, Nicole J. Ullrich, David Viskochil, Xia Wang, Matija Snuderl, Peter J. Park, Adrienne M. Flanagan, Angela C. Hirbe, Nischalan Pillay, and David T. Miller. Genomic Patterns of Malignant Peripheral Nerve Sheath Tumor (MPNST) Evolution Correlate with Clinical Outcome and Are Detectable in Cell-Free DNA. *Cancer discovery*, 13(3):654–671, 3 2023.
20. John Richard McPherson, Choon-Kiat Ong, Cedric Chuan-Young Ng, Vikneswari Rajasegaran, Hong-Lee Heng, Willie Shun-Shing Yu, Benita Kiat-Tee Tan, Preetha Madhukumar, Melissa Ching-Ching Teo, Joanne Ngeow, Aye-Aye Thike, Steven George Rozen, Puay-Hoon Tan, Ann Siew-Gek Lee, Bin-Tean Teh, and Yoon-Sim Yap. Whole-exome sequencing of breast cancer, malignant peripheral nerve sheath tumor and neurofibroma from a patient with neurofibromatosis type 1. *Cancer Medicine*, 4(12):1871–1878, 12 2015.
21. Charlotte Philpott, Hannah Tovell, Ian M. Frayling, David N. Cooper, and Meena Upadhyaya. The NF1 somatic mutational landscape in sporadic human cancers. *Human Genomics*, 11(1):13, 12 2017.
22. Jeevana Praharsha Athota, Meenakshi Bhat, Sheela Nampoothiri, Kalpana Gowrishankar, Sanjeeva Ghanti Narayanachar, Vinuth Puttamallesh, Mohammed Oomer Farooque, and Swathi Shetty. Molecular and clinical studies in 107 Noonan syndrome affected individuals with PTPN11 mutations. *BMC Medical Genetics*, 21(1):50, 12 2020.
23. Ritsuko Harigai, Ryo Sato, Chikako Hirose, Toshiki Takenouchi, Kenjiro Kosaki, Takanori Hirose, Hideyuki Saya, and Yoshimi Arima. Mutation of PTPN11 (Encoding SHP-2) Promotes MEK Activation and Malignant Progression in Neurofibromin-Deficient Cells in a Manner Sensitive to BRAP Mutation. *Cancers*, 14(10):2377, 5 2022.
24. Geng Liu, Timothy J. McDonnell, Roberto Montes de Oca Luna, Mini Kapoor, Betsy Mims, Adel K. El-Naggar, and Guillermina Lozano. High metastatic potential in mice inheriting a targeted p53 missense mutation. *Proceedings of the National Academy of Sciences*, 97(8):4174–4179, 4 2000.
25. Raphaël Margueron and Danny Reinberg. The Polycomb complex PRC2 and its mark in life. *Nature*, 469(7330):343–349, 1 2011.
26. Shyra J. Miller, Walter J. Jessen, Japan Mehta, Atira Hardiman, Emily Sites, Sergio Kaiser, Anil G. Jegga, Hua Li, Meena Upadhyaya, Marco Giovannini, David Muir, Margaret R. Wallace, Eva Lopez, Eduard Serra, G. Petur Nielsen, Conxi Lazaro, Anat Stemmer-Rachamimov, Grier Page, Bruce J. Aronow, and Nancy Ratner. Integrative genomic analyses of neurofibromatosis tumours identify SOX9 as a biomarker and survival gene. *EMBO Molecular Medicine*, 1(4):236–248, 2009.
27. Munmun Panda, Surya Kant Tripathi, and Bijesh K. Biswal. SOX9: An emerging driving factor from cancer progression to drug resistance. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1875(2):188517, 4 2021.
28. James A. Walker and Meena Upadhyaya. Emerging therapeutic targets for neurofibromatosis type 1. *Expert Opinion on Therapeutic Targets*, 22(5):419–437, 5 2018.
29. Lai Man Natalie Wu, Feng Zhang, Rohit Rao, Mike Adam, Kai Pollard, Sara Szabo, Xuezhao Liu, Katie A. Belcher, Zaili Luo, Sean Ogurek, Colleen Reilly, Xin Zhou, Li Zhang, Joshua Rubin, Long-Sheng Chang, Mei Xin, Jiyang Yu, Mario Suva, Christine A. Pratilas, Steven Potter, and Q. Richard Lu. Single-cell multiomics identifies clinically relevant mesenchymal stem-like cells and key regulators for mpnst malignancy. *Sci. Adv*, 8, 2022.
30. Alice Jo, Sahitya Denduluri, Bosi Zhang, Zhongliang Wang, Liangjun Yin, Zhengjian Yan, Richard Kang, Lewis L. Shi, James Mok, Michael J. Lee, and Rex C. Haydon. The versatile functions of Sox9 in development, stem cells, and human diseases. *Genes and Diseases*, 1(2):149–161, 12 2014.
31. Yuan Zhu, Mario I. Romero, Pritam Ghosh, Zhengyi Ye, Patrick Charnay, Elizabeth J. Rushing, Jamey D. Marth, and Luis F. Parada. Ablation of NF1 function

in neurons induces abnormal development of cerebral cortex and reactive gliosis in the brain. *Genes & Development*, 15(7):859–876, 4 2001.

32. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902, 6 2019.
33. Pierre-Luc Germain, Aaron Lun, Carlos Garcia Meixide, Will Macnair, and Mark D. Robinson. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Research*, 10:979, 5 2022.
34. Abha S. Bais and Dennis Kostka. scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics*, 36(4):1150–1158, 2 2020.
35. Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 12 2019.
36. Laleh Haghverdi, Aaron T.L. Lun, Michael D. Morgan, and John C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 5 2018.
37. Philipp Angerer, Laleh Haghverdi, Maren Büttner, Fabian J. Theis, Carsten Marr, and Florian Buettner. destiny : diffusion maps for large-scale single-cell data in R. *Bioinformatics*, 32(8):1241–1243, 4 2016.
38. Helena Todorov, Robrecht Cannoodt, Wouter Saelens, and Yvan Saeys. TinGa: fast and flexible trajectory inference with Growing Neural Gas. *Bioinformatics*, 36(Supplement_1):i66–i74, 7 2020.
39. Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, 9 2016.
40. Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, 5 2012.
41. Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J. Daly, Ben Neale, Daniel G. MacArthur, and Eric Banks. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRXiv*, 2018.
42. Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, 6(2):80–92, 4 2012.
43. Peter Van Loo, Silje H. Nordgard, Ole Christian Lingjærde, Hege G. Russnes, Inga H. Rye, Wei Sun, Victor J. Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume, Charles M. Perou, Anne-Lise Børresen-Dale, and Vessela N. Kristensen. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915, 9 2010.

Supplementary Materials

Table S1. Mouse features, tumor histological characterization, and NGS identifiers annotation for all *Nf1* mutant mice

Table S2. Information regarding NF1 patient scRNA-Seq datasets

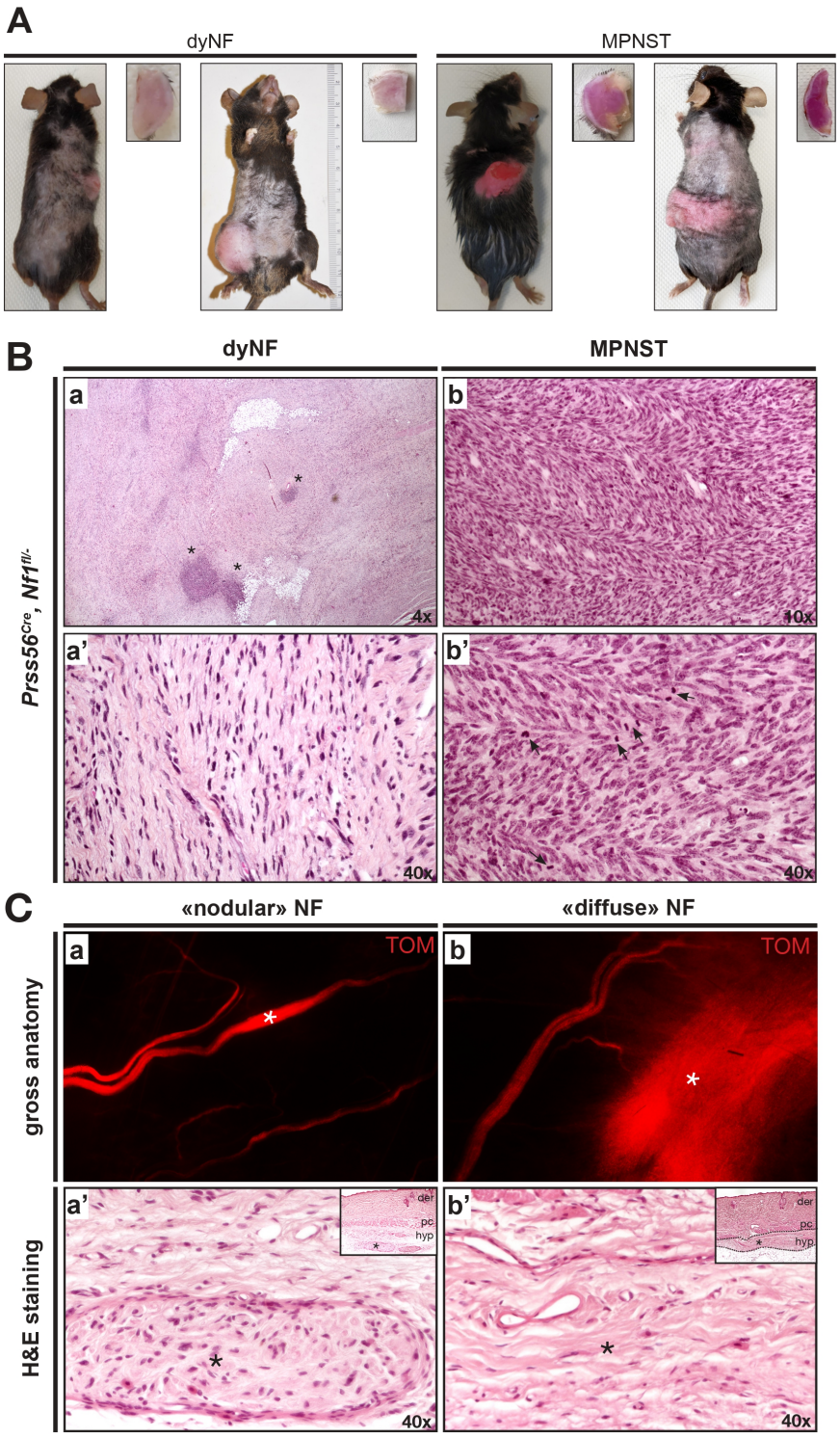
Table S3. Efficacy of stable inhibition induced by shRNA treatment and measured by qRT-PCR (n=4 for each treatment)

Table S4. Annotation of top 12 selected compounds after drug screening

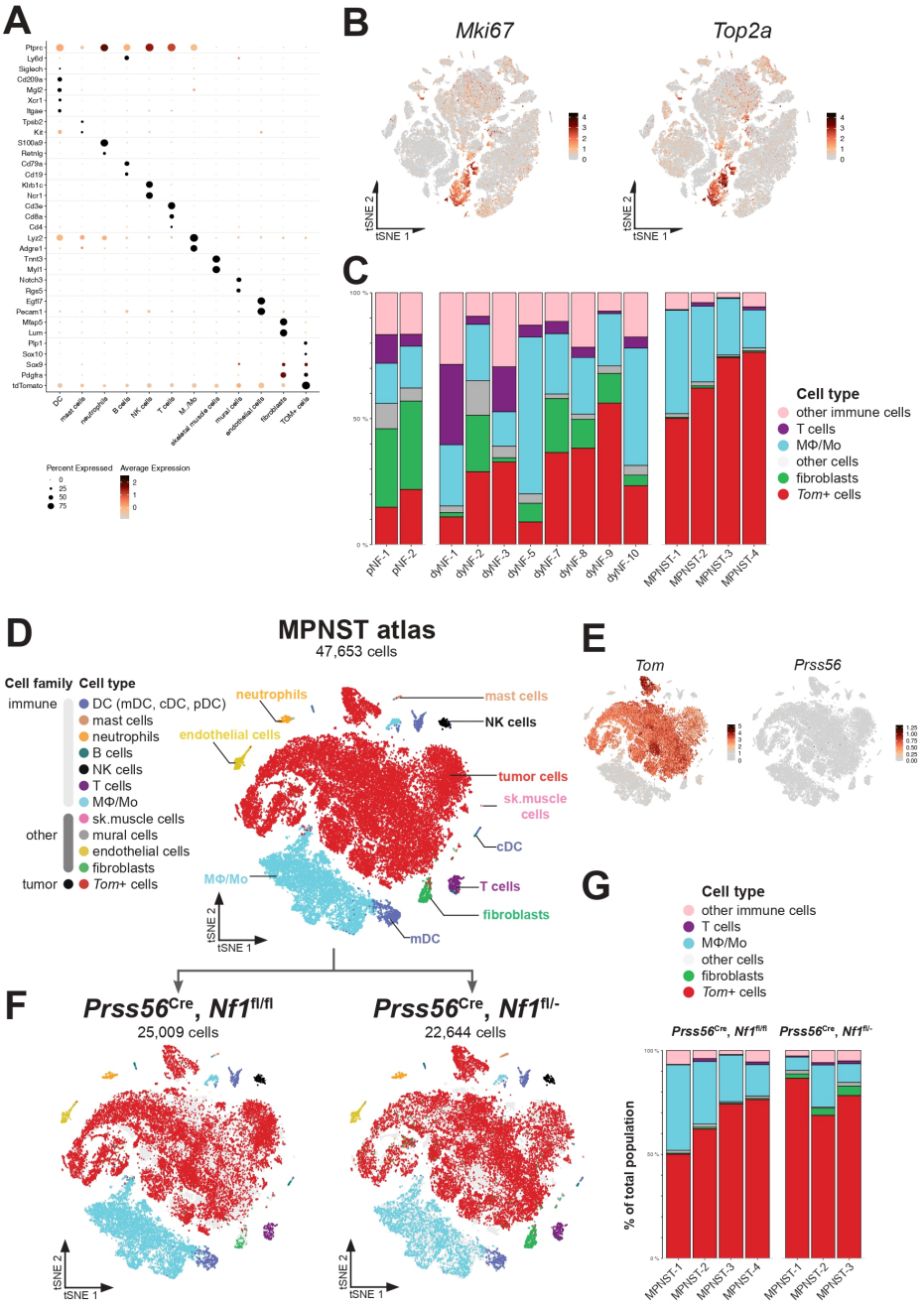
Table S5. Version of R packages used for scRNA-Seq data analysis

Table S6. Gene markers used for cell type annotation in all scRNA-Seq datasets

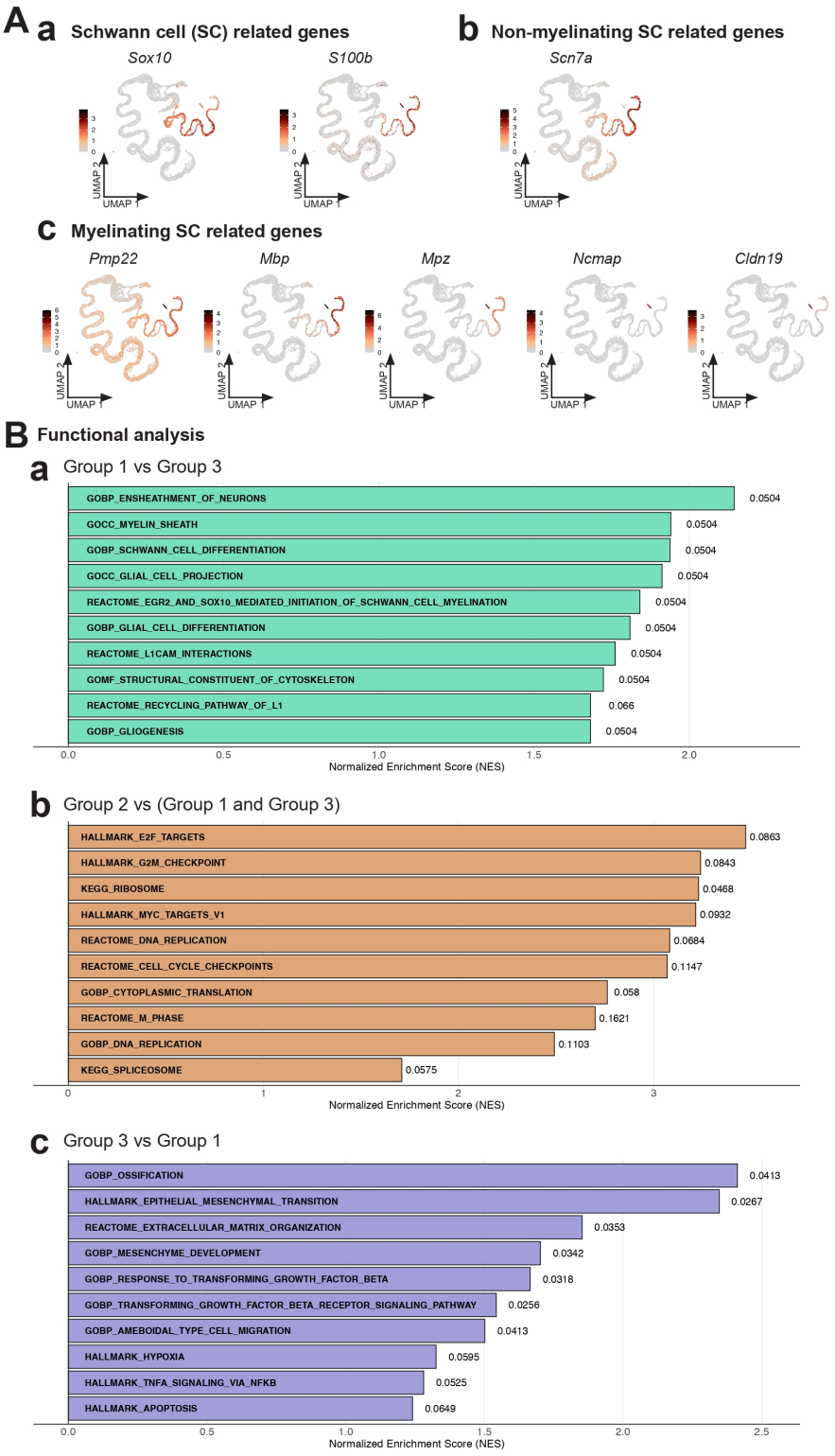
Table S7. Gene markers used for cell type annotation in the fibroblast dataset



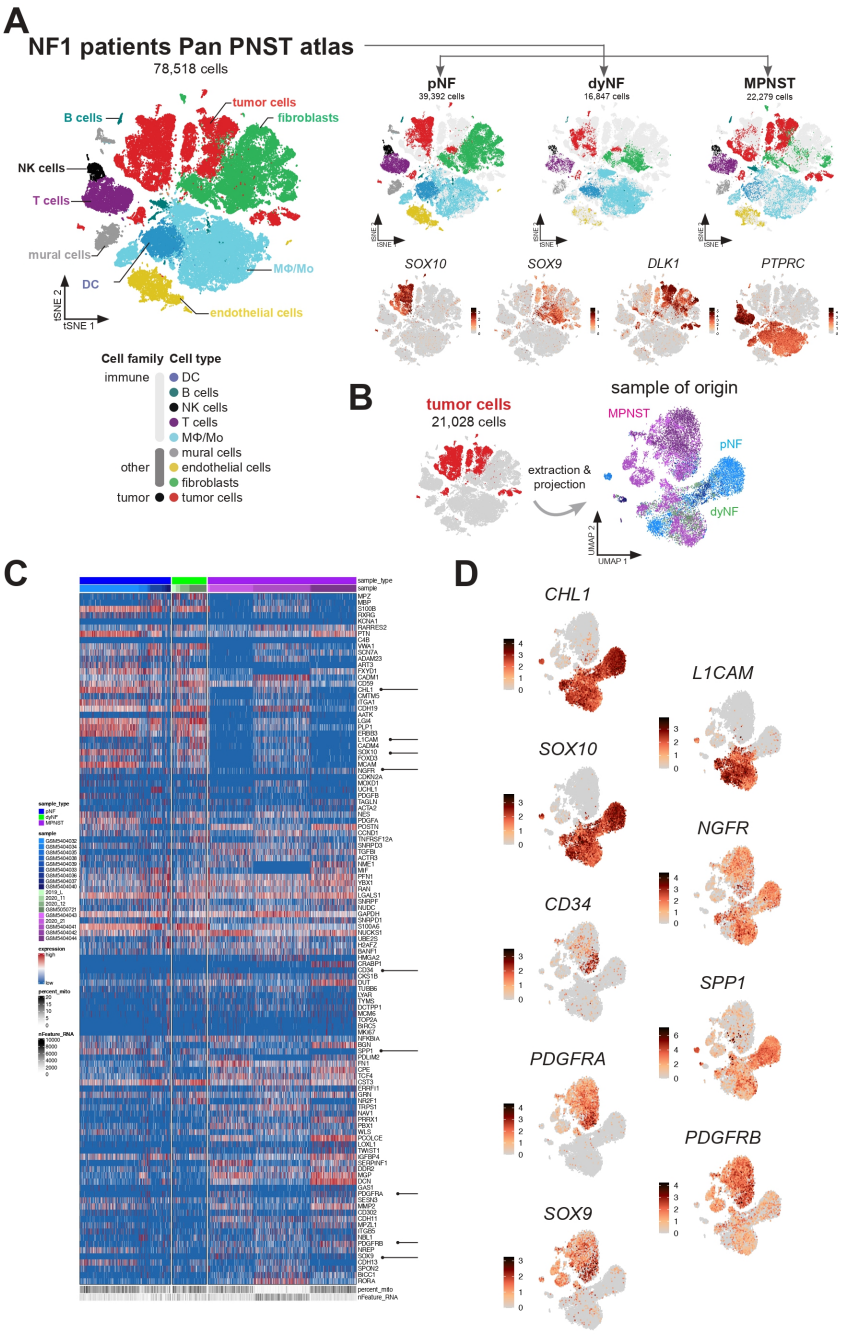
Supplementary Figure 1. Spectrum of peripheral nerve sheath tumors in *Nf1* mutant mouse model. (A) Gross anatomy of dyNF and MPNST-bearing *Prss56*^{Cre}, *Nf1*$^{fl/-}$ males. Dissected tumors are shown on the right. (B) H&E staining of dyNF (a, a') and MPNST (b, b') from *Prss56*^{Cre}, *Nf1*$^{fl/-}$ mice at two magnifications. Asterisks in (a) indicate clusters of mononuclear cells. Arrows in (b') indicate mitotic cells. (C) Comparison of "nodular" (a, a') and "diffuse" (b, b') pNFs (asterisks) in *Prss56*^{Cre}, Nf1fl /fl mice. Endogenous Tomato (TOM) fluorescence is shown in (a) and (b), and H&E staining of the corresponding lesions in (a') and (b'). Insets represent low magnification images showing tissue architecture and localization of respective pNFs. der: dermis, hypo: hypodermis, pc: panniculus carnosus muscle.



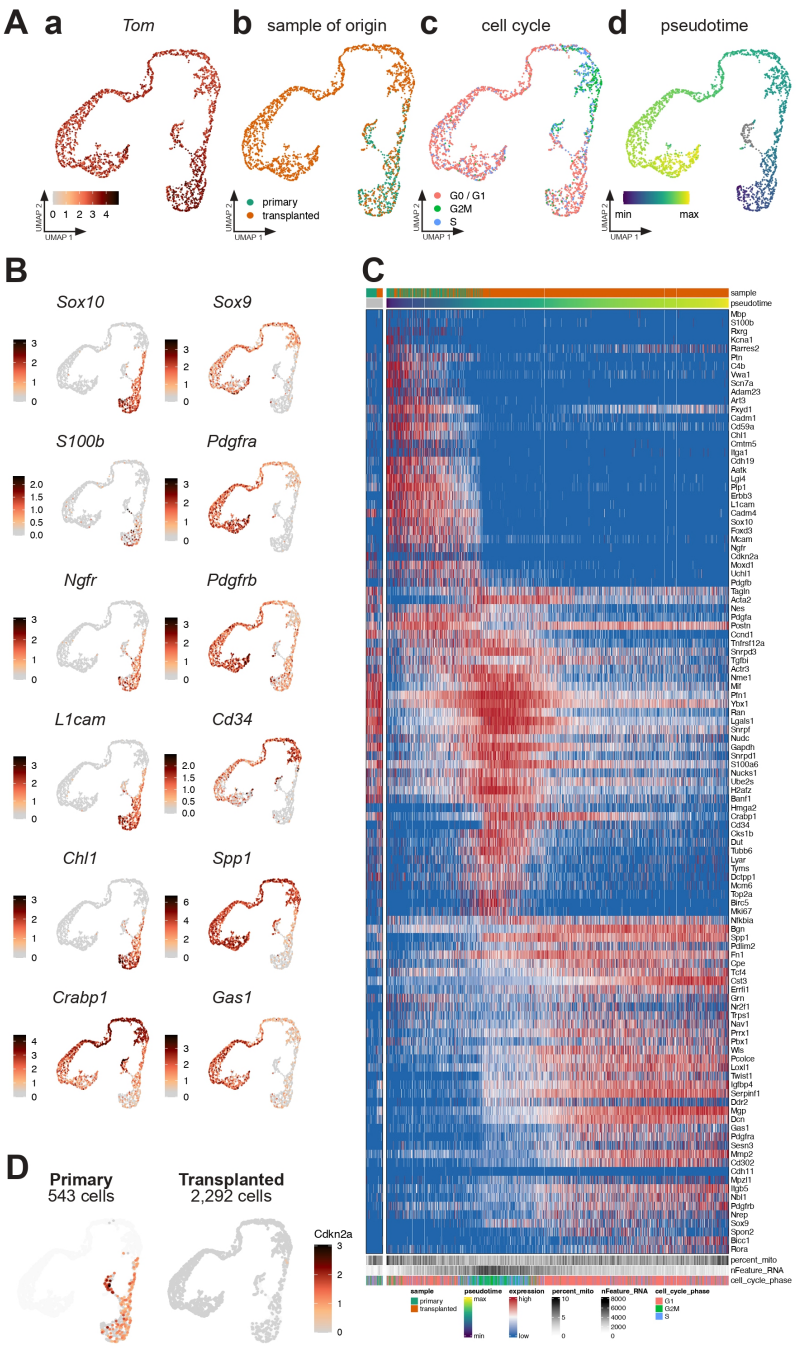
Supplementary Figure 2. Characterization of cellular composition of PNSTs from *Nf1* mutant mice. (A) Dotplot of average expression of markers of the different cell types composing the PNSTs. Dot size represents the percentage of cells in the tumor and color scale represents the average expression level across all cells within population. (B) tSNE plot of Pan MPNST atlas overlaid with the expression of cell cycle-related genes Mki67 and Top2a. (C) Barplot depicting the cellular composition of individual datasets including in the Pan PNST atlas. Note that datasets dyNF_4 and dyNF_6 are not shown because they originate from FACS sorting of TOM+ cells. (D) MPNST transcriptomic atlas assembled from *Prss56*^{Cre}, *Nf1*$^{fl/fl}$ and *Prss56*^{Cre}, *Nf1*$^{fl/-}$ primary MPNST datasets. Results are presented as tSNE plot to visualize cellular heterogeneity. Expression score-based annotation resolved 12 distinct cell populations. (E) tSNE plot of MPNST atlas overlaid with expression of Tom and Prss56. Cells are color-coded according to gene expression level. (F) tSNE plot of MPNST atlas split by genotype. Cells from complete atlas are shown as light gray background. (G) Barplot depicting the cellular composition of individual MPNSTs including in the MPNST atlas.



Supplementary Figure 3. Functional characterization of tumor cells extracted from the murine Pan-PNST atlas. (A) UMAP plot of tumor cells overlaid with the expression Schwann cell (SC) (a), non-myelinating SC (b) and myelinating SC (c) related transcripts. (B) Functional enrichment analyses of each group shown on **Figure 3Ae**. Group 1 was compared to group 3 (a) and conversely (c). Group 2 was compared to both groups 1 and 3 and depicts enrichment in cell-cycle related genes (b).



Supplementary Figure 4. Single-cell phenotyping of human PNSTs. (A) Pan PNST atlas assembled from 4 original and 14 published NF1 patient datasets (accession number GSE165826 and GSE179043, **Table S2**). (Left panel) Data are presented as tSNE plot to visualize cellular heterogeneity. Expression score-based annotation resolved nine distinct cell populations. (Right panel) tSNE plot of NF1 patient PNSTs split by tumor type. Cells from complete atlas are shown as light gray background. On bottom, cells are color-coded according to gene expression level. (B) Tumor cells are extracted from NF1 patient PNSTs atlas, projected using UMAP and colored based on tumor type. (C) Heatmap showing transcriptomic changes in tumor cells during pNF-to-MPNST progression, for the same genes identified in *Prss56*^{Cre}, *Nf1*$^{fl/fl}$ primary tumors. Cells are ordered by tumor types and samples of origin (top annotation). Bottom annotation show quality metrics: proportion of expression levels related to mitochondrial genes (percent.mito) and total number of genes detected by cell (nFeature_RNA). Note that fewer genes are detected in human scRNA-Seq compared to mice. (D) UMAP plot of NF1 patient tumor cells overlaid with the expression of genes validated at proteomic level in *Prss56*^{Cre}, *Nf1*$^{fl/fl}$ primary tumors.



Supplementary Figure 5. Dysplastic NF allografts recapitulate GMT program. (A) Tumor cells from Nude mice transplanted with a primary dyNF (*Prss56*^{Cre}, *Nf1*$^{fl/fl}$ mouse) were projected using UMAP. Cells were colored based on Tom expression (a), sample of origin (b), cell cycle phase annotation (c) and pseudotime inferred with slingshot (d). Gray cells in (d) did not belong to the linear trajectory branch inferred with slingshot and have no pseudotime value. (B) UMAP plot of tumor cells overlaid with the expression of genes validated at protein level in *Prss56*^{Cre}, *Nf1*$^{fl/fl}$ primary tumors. (C) Heatmap showing transcriptomic changes between primary and transplanted tumors. Cells were ordered by increasing pseudotime values. Top annotation shows pseudotime and sample of origin. Cells without pseudotime value were arbitrary placed on the right, as separated group. Bottom annotation shows cell cycle annotation and quality metrics: proportion of expression levels related to mitochondrial genes (percent.mito) and total number of genes detected by cell (nFeature_RNA). Note that cells without pseudotime values, in the left column, are characterized by high expression of mitochondrial genes, associated with low number of genes detected. (D) UMAP plot split by sample overlaid with the expression of *Cdkn2a*. Cells from complete dataset are shown as light gray background.

3.3 Complementary results

3.3.1 Histological assessment of tumor type and single-cell transcriptomics

The tumor types were assigned based either on histological evaluation or scRNA-Seq data. In scRNA-Seq, we identified two subtypes of tumor cells, either glial, characterized by *Sox10* expression, or mesenchymal, with *Sox9* expression (Figure 70A). The proportion of such subtypes was also used to discriminate dyNFs from MPNSTs. In this section, we evaluated the consistency between the histological assessment of tumor types and the transcriptomic phenotypes of tumor cells.

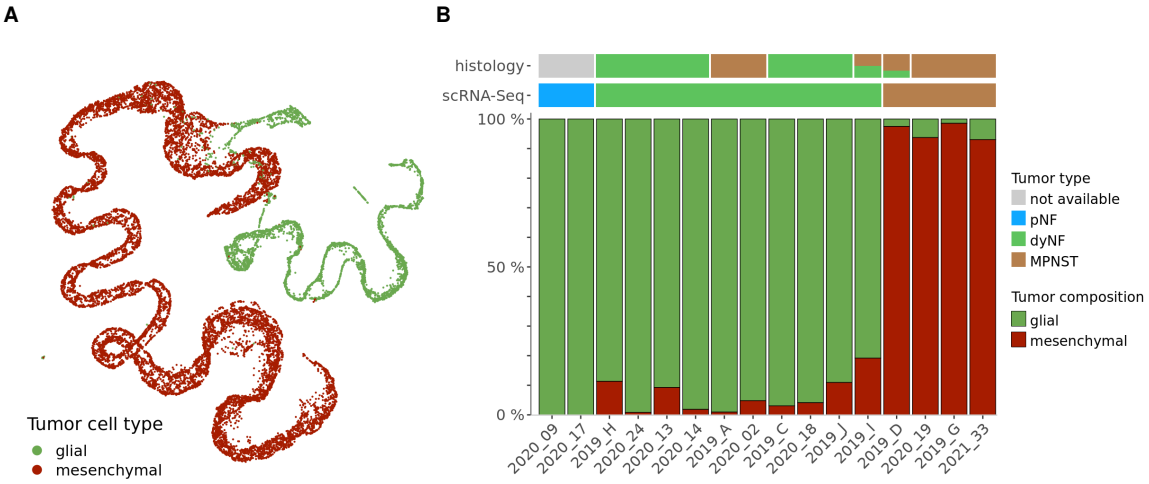


Figure 70 – ***Sox10+* and *Sox9+* tumor cells balance is associated with tumor type**. **A.** Tumor cells projection (Article 2, Figure 4A), colored according to the expression of glial (*Sox10+*) or mesenchymal (*Sox9+*) related genes. **B.** Barplot depicting the composition of each individual dataset in tumor cell type. The top annotation corresponds to tumor type, determined by scRNA-Seq or histological criteria. These figures are extracted from the compiled notebook **figures.html** from the Article 2.

We have matching histological and scRNA-Seq data for *Nf1*-KO^{wt} mice: 9 dyNF[10] and 4 MPNST. Their comparison showed that the diagnosis is coherent for all MPNST but 7/10 dyNF (Figure 70B). Discrepancies come from the spatial heterogeneity of transforming tumors, the number of mitotic figures and the infiltration by immune cells (Article 2, Table S1). The pNF nature of the two first datasets is known from the sample preparation. Unlike dyNF, they are deprived of mesenchymal cells (Figure 70B) and characterized by nerve-associated fibroblasts (Article 2, Figure 2E-G).

These results show that the nature of tumor cells, either glial or mesenchymal, can be used for the differential diagnosis of dyNF and MPNST. However, histological characterization remain necessary to distinguish subtypes of dyNF (cellular NFs, atypical

[10] 10 dyNF datasets are presented in the Article 2 but 2 correspond to the same primary tumor.

NFs or ANNUBP) and MPNST (low grade and high grade). Indeed, we did not observe differences between low grade and high grade MPNST using scRNA-Seq data. This distinction relies on the number of mitotic figures on the tissue sections.

3.3.2 Nodular pNFs are likely to be the origin of dyNF

As in NF1 patients, we identified two types of pNFs in our mouse model: nodular and diffuse. Clinical observations suggest that some nodular, but not diffuse pNFs progress to malignancy (Schaefer & Fletcher, 2015), but evidence is still missing. To further explore this issue, we generated two datasets, where nodular and diffuse pNFs have been distinguished from 8 *Nf1*-KO$^{fl/fl}$ mice of the same litter.

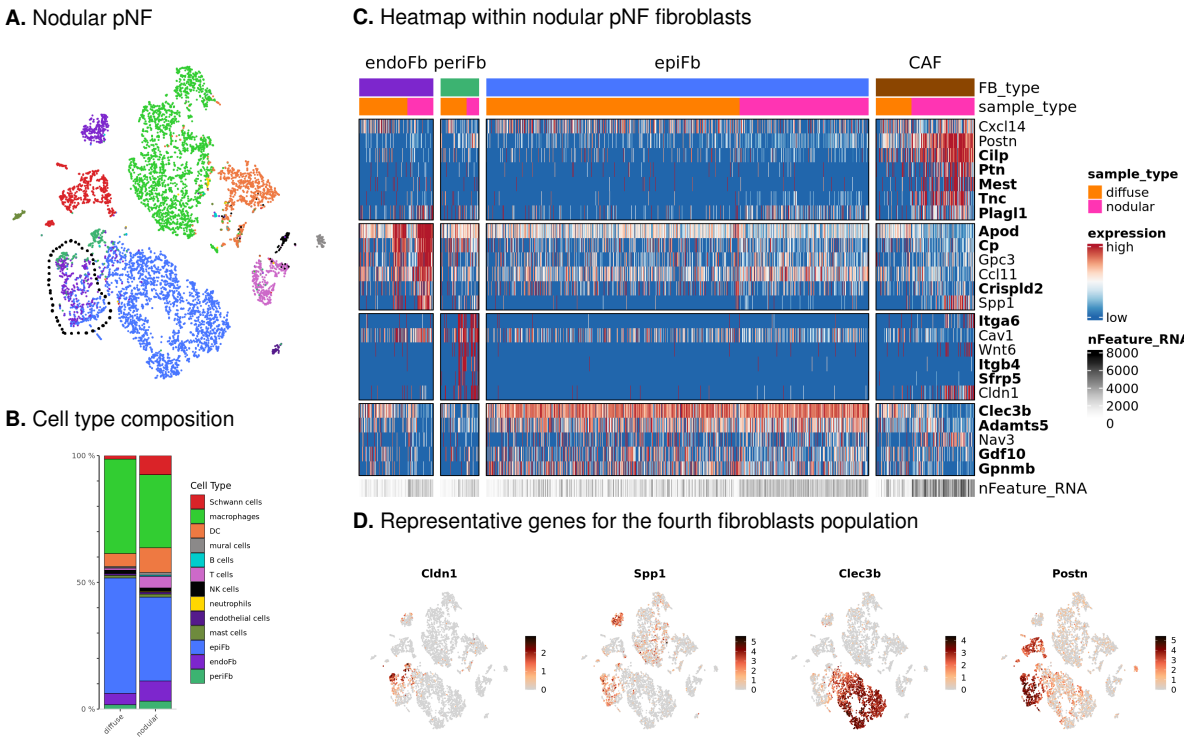


Figure 71 – **Nodular pNFs contain cancer-associated fibroblasts**. **A.** tSNE representation of 7,666 cells from the nodular pNF dataset, colored according to cell types. Cells were annotated using the nerve marker gene sets (Article 1). The fourth population of fibroblasts has been delineated manually. The color key is shared with panel B. **B.** Barplot showing cell populations representativeness in the diffuse (left bar) and the nodular (right bar) pNF datasets. **C.** Heatmap showing expression of selected genes in fibroblasts, in the diffuse and nodular pNFs dataset (top annotation). Fibroblasts from the two datasets have been annotated annotated for the four populations as in the Article 2. Gene expression is log-normalized, then scaled across cells. **D.** Representation of the expression levels for a marker gene of each fibroblast population, in the nodular pNF dataset: *Cldn1* (perineurial fibroblasts), *Spp1* (endoneurial fibroblasts), *Clec3b* (epineurial fibroblasts) and *Postn* (CAFs). These figures are extracted from the compiled notebooks **2024_02.html** (A-B) and **2_figures** (C-D) from the pNF project.

The diffuse pNF dataset contains few tumor SCs, the three nerve-associated fibroblasts (NAFs) and consistent macrophages and DCs populations as in the subcutaneous nerves atlas (Article 1). On the opposite, the nodular pNF contains more SCs and a fourth population of fibroblasts was visible on the tSNE (Figure 71A-B).

The **fourth population of fibroblasts** shares the annotation with the 3 NAFs (Figure 71A). To further assess if it depicts a similar transcriptomic profile with cancer-associated fibroblasts (CAFs) observed in dyNF datasets, we annotated all fibroblasts from both diffuse and nodular pNF datasets using the same markers as in the Article 2. Indeed, CAFs were annotated in both datasets, due to *Postn* expression. However, only fibroblasts from the nodular pNF dataset depict the expression of the same genes identified in dyNF datasets (Figure 71C). Unlike dyNFs fibroblasts, such CAFs retain a reminiscent NAFs identity (Figure 71C-D). Our results suggest that CAFs originate from the 3 NAFs, that converge to a homogeneous transcriptomic profile. In dyNFs, the absence of this reminiscent identity precluded this analysis (Article 2, Figure 2G).

We then investigated the **tumor population** by merging all SCs from diffuse and nodular pNFs with all tumor cells from our mouse models (Figure 72A). The contribution of diffuse and nodular pNFs to this representation suggest that nodular pNFs share the features with dyNFs (Figure 72B). Indeed, only the cells from nodular pNFs express *Moxd1* and *Uchl1*, features that we only observed in dyNF tumor cells.

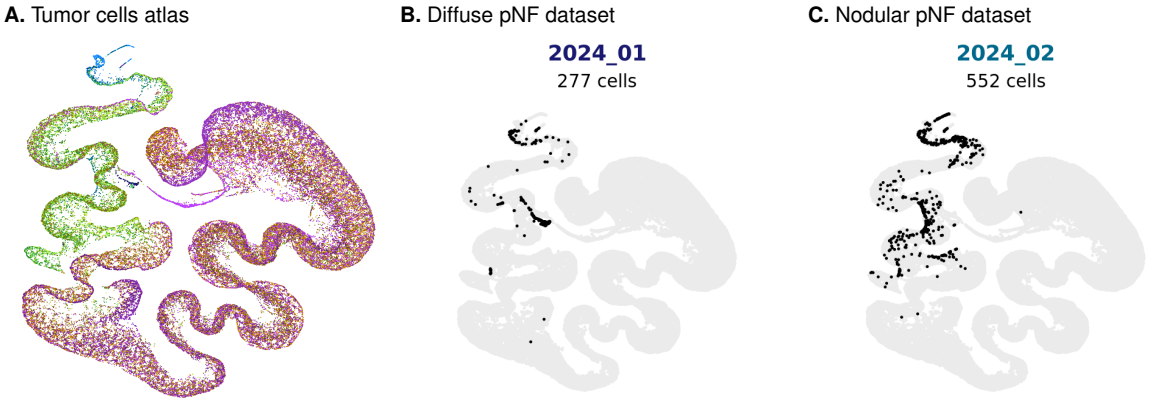


Figure 72 – **Nodular pNFs tumor cells share identity with dyNF tumor cells**. Representation of 73,619 tumor cells using the same methodology as in the Article 2. **A.** Cells are colored according to the sample of origin, with blue shades corresponding to pNFs, green to dyNFs and brown-purple to MPNSTs. **B-C.** Individual contributions of the diffuse (B) and nodular (C) pNF datasets to the projection. Cells from the complete dataset are shown as light gray background. Note the dyNF location of some cells from the nodular pNF dataset. Theses figures are extracted from the compiled notebook **1_tumor_cells_dataset** from the pNF project.

Since nodular pNFs have a more pronounced profile compared to diffuse pNFs, they do not constitute distinct stages but distinct entities. Our results suggest that nodular pNFs, but not diffuse pNFs, precedes dyNF (Figure 73). Their immune landscapes remain to be characterized.

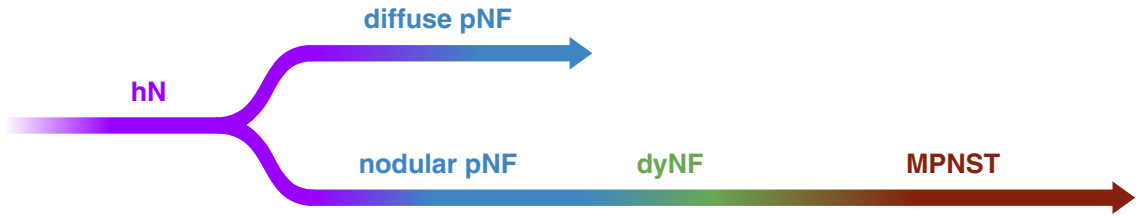


Figure 73 – Model of pNF progression. Hyperplastic nerves can progress into diffuse or nodular pNFs. Only nodular pNFs can evolve into dyNF then MPNST.

3.3.3 Discrepancies between RNA velocity and trajectory inference

In the Article 2, we were able to reconstruct the pNF-to-dyNF-to-MPNST progression using a projection combining diffusion map and UMAP and by trajectory inference from pNF datasets. Since RNA velocity is alternatively used to identify order between cells, we would like to investigate it on our data.

We analyzed the datasets 2020_02 (dyNF) and 2021_33 (MPNST) individually, as they constitute the two datasets with the highest number of tumor cells for their respective tumor type. We further mapped their inferred velocities on the same projection (Figure 74). These two datasets share a glial population (Figure 74C-D). Strikingly, velocities are different in these cells, with arrows going in the north-east direction for dataset 2020_02 and in the south-east direction for 2021_33 (Figure 74E-F). This observation suggests that RNA velocity is not robust across datasets, despite cells share very similar transcriptomic profile. In the mesenchymal population, velocities are coherent with the pseudotime and are oriented from proliferative to resting cells.

To conclude, estimating RNA velocity using scvelo is not consistent between two individual datasets and does not align with the trajectories inferred by slingshot or TInGa. However, since the trajectory inference also did not work at the individual dataset level, due to the low number of cells (data not shown). Therefore, we could evaluate RNA velocity on merged count matrices.

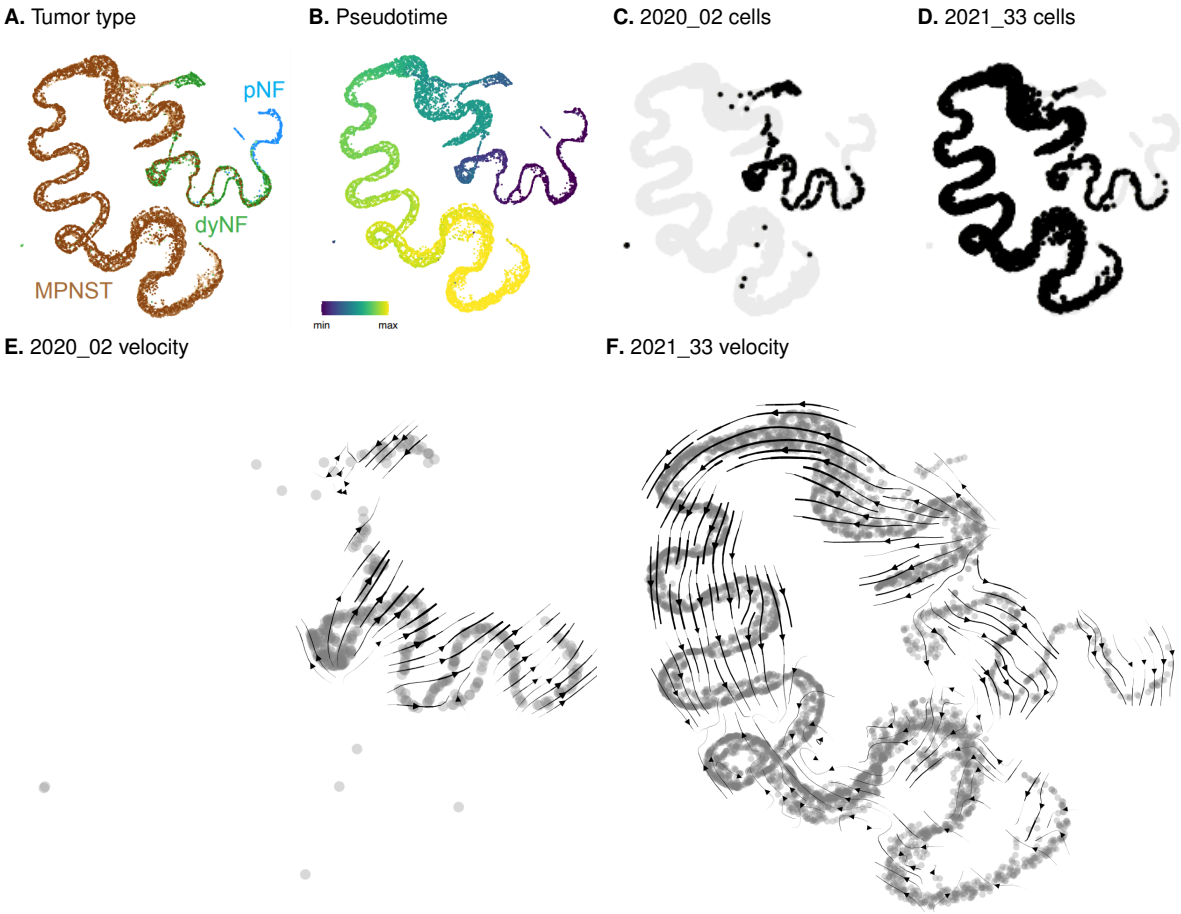


Figure 74 – **Comparison between trajectory inference and RNA velocity**. *Nf1*-KO^{wt} mice tumor cells are projected using FastMNN, diffusion map and UMAP, as explained in the Article 2. Panels A to D are issued from the Figure 3 of the Article 2. Cells are colored according to the tumor type of origin (**A**) or pseudotime, inferred with slingshot (**B**). Cells originating from datasets 2020_02 (**C**) or 2021_33 (**D**) are depicted as black dots, while other cells are represented as a light gray background. RNA velocity is represented on cells from 2020_02 (**E**) or 2021_33 (**F**). The projection is the same MNN-DM-UMAP presented in Figure 3 from the Article 2. These two figures are extracted from the compiled notebooks **scvelo_2020_02.html** and **scvelo_2021_33.html** from the RNA velocity project.

3.3.4 CNV informs about the intra-tumoral clonality

In NF1 patients, MPNSTs are subject to relapse after treatments, which are associated with the expansion of treatment-resistant clone. A clone corresponds to the expansion of specific cell, characterized by a unique genetic profile. We would like to evaluate the clonality of MPNSTs, once they are formed. While bulk genomic profiles can not inform about such clonality, inferring CNV from scRNA-Seq may provide preliminary results. To illustrate this, we chose the *Nf1*-KO^{wt} dataset containing the highest number of tumor cells: 2021_33.

The tSNE made from the gene expression matrix suggests at least 2 tumor subpopulations (Figure 75A). Additionally, inferred CNV on tumor cells showed a non homogeneous profile across tumor cells, indicating **several clones** (Figure 75B). Clones are clustered together, according to a clustering made from the CNV table. Based on the profile of each cluster, we built a phylogenetic tree that could reflect the parenthood between cells (Figure 75B, left). Notably, *Sox10*+ tumor cells does not constitute the older branch of the tree. It can reflect that the anomalies (+5q, +14p, -14q) occurred only after this population gave rise to the *Sox9*+ population.

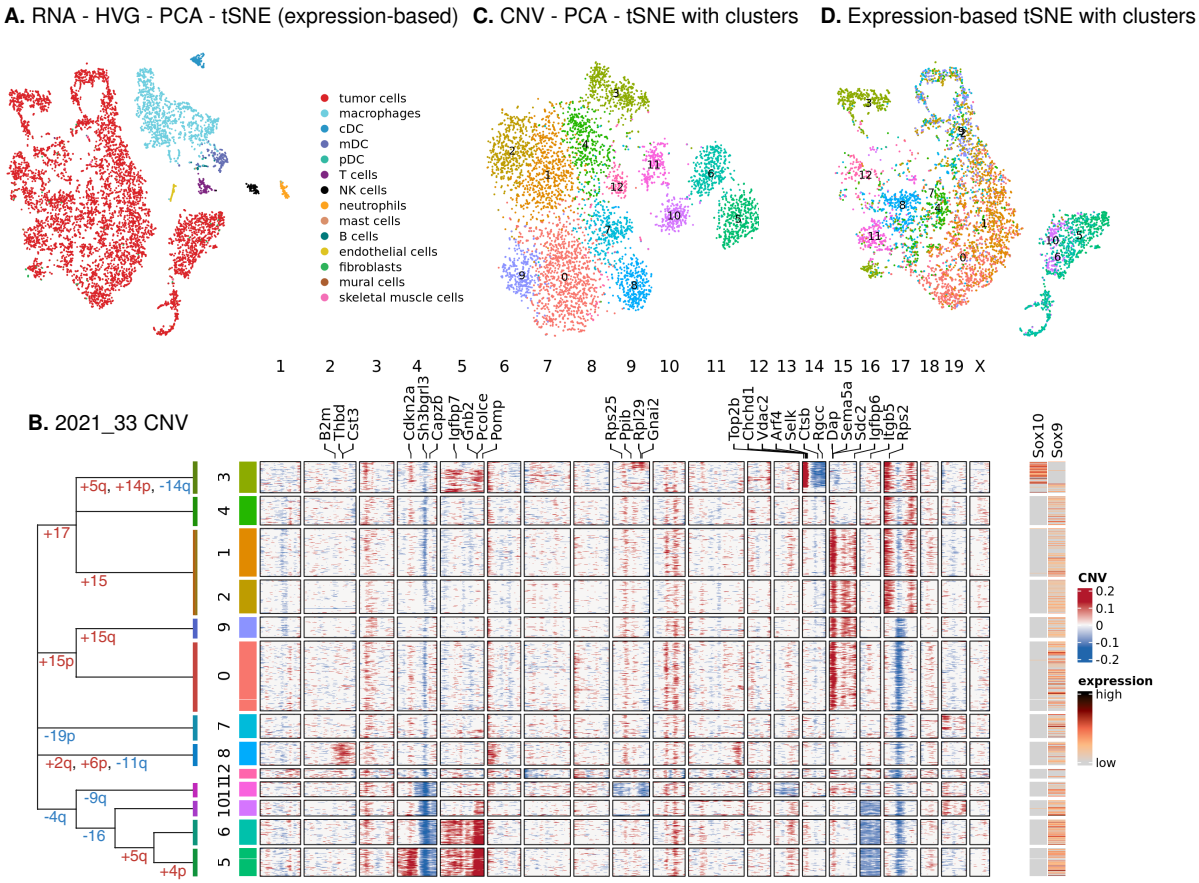


Figure 75 – **Distinct tumor subpopulations are associated with specific CNV in 2021_33**. This *Nf1*-KO^{wt} dataset corresponds to MPNST_4 in the Article 2. CNV were inferred using mDC as reference. **A.** Expression-based tSNE representation of 8,909 cells from 2021_33 dataset, colored according to cell types. The projection is made by identifying 3000 HVGs in the normalized gene expression matrix (RNA), running PCA and building tSNE. **B.** CNV values are represented on a heatmap, with gains in red and losses in blue. The columns correspond to genes, ordered according to chromosomes. Except *Cdkn2a*, the annotated genes are up-regulated or down-regulated in the cells depicted gain or loss for the associated chromosomic region, respectively. The rows correspond to cells, ordered by clusters. On the left, a possible phylogenetic tree, annotated for the genomic events, makes connection between clusters. On the right, *Sox10* and *Sox9* expression levels are indicated. Clusters were identified on the PCA made on the CNV table, for tumor cells only, and projected on the CNV-based tSNE (**C**) or on the expression-based tSNE (**D**). These figures are extracted from the compiled notebook **explore_CNV_2021_33.html** from the flox_flox_2021_33 project. The tree was made manually using BioRender.com.

Another representation of the CNV consists in building a tSNE from the CNV table. As expected, each clone/cluster is well delineated on this **CNV-based tSNE** (Figure 75C). To investigate whether the clones are associated with distinct transcriptomic profiles, we annotated the CNV clusters on the expression-based tSNE. Consistently, the two main clones correspond to the two subpopulations on the expression-based tSNE (Figure 75C-D).

With few inferred anomalies, clusters 7 and 12 are spread on this tSNE. On the opposite, although they have a distinct CNV profile, clusters 2 and 9 are overlapping. Indeed, they share the same proliferative activity (data not shown), which is correlated with lower expression of other genes. This explains their overlapping on the expression-based tSNE. Finally, clusters 5, 6 and 10 are distinct subclones, but depict a similar transcriptomic profile. This profile is marked by high expression levels of *Pcolce*, *Meg3* and *Wnt5a*. Although this population was specific to this dataset, expression of these three genes have already been described in patients (Yao *et al.*, 2023b; Wu *et al.*, 2022a; Thomson *et al.*, 2021).

We have paired scRNA-Seq and Whole Exome Sequencing (WES) data from this sample. Notably, the bulk CNV profile determined by **WES was consistent** with the results from infercnv (Figure 81 in Appendix). The coherence with infercnv and WES results was validated on a *Nf1*-KO^{het} MPNST (Figure 82 in Appendix). Overall, WES is able to detect the main CNV identified by infercnv, when they correspond to more than 30% of cells. At a single-cell level, we are informed about the co-existence of CNV and thus can define the genetic profile of clones. However, when applying infercnv using non immune cells (endothelial) as reference, results were discordant with WES (data not shown). In WES data, the reference corresponds to the spleen profile, which mainly consists in immune cells. Maybe using vessels as reference in WES would give similar results that infercnv using endothelial cells as reference.

These results suggest that tumor cells in MPNST can correspond to distinct genetic clones. To investigate the cellular origin of MPNST, we are currently setting up an experiment of retrograde barcoding. In culture, tumor cells from an early murine dyNF will be transduced by viral particles containing a unique genetic barcode. The transduced cells will be transplanted in a Nude recipient, which is supposed to develop MPNST after few weeks. The barcodes will be sequenced along with the transcriptome of tumor cells. In the subsequent scRNA-Seq data, we will quantify the diversity of barcodes. This analysis will assess the clonal origin of MPNSTs. It will inform whether the distinct clones we just defined are issued from distinct cells at the origin or if they result from a diverging progression after malignant transformation.

4 Article 3

Hidradenitis suppurativa (HS) is a chronic, recurrent and disabling inflammatory skin condition affecting 1% of the population. The disease is characterized by painful nodules, draining tunnels and scarring, located in hair-bearing skin folds. These features are thought to arise from the occlusion of hair follicles, followed by a rupture. Although lesions are characterized by both keratinization and inflammation, it is unknown whether one event precedes the other or if both mechanisms occur simultaneously.

4.1 Overview of the study

The team of Professor Sophie Hüe, in collaboration with our laboratory, generated the first single cell transcriptomics data from hair follicles (HFs) of HS patients. The aim of this study was to characterize the transcriptomic landscape of HS patients, compared to healthy donors. My contribution was to explore the scRNA-Seq data using the framework established for previous studies (Section 1), design the figures (except Figure 6) and write the computational sections of the manuscript.

Our scRNA-Seq atlas of HFs recapitulates the transcriptomic diversity obtained by Wu *et al.* (2022b) (Figure 76). HFs are composed of matrix cells, constituting the hair shaft ; non-matrix cells, supporting the structure ; and immune cells. By investigating the proportions of each cell population across datasets, we identified an imbalance among the non-matrix cells. While some patient datasets are enriched in cells from the inner bulge layer (IBL), other contain more middle outer root sheath (mORS) cells and are associated with an infiltration of T cells.

We were able to reconstruct the trajectory going from HF stem cells (HF-SCs) to IBL and mORS, a differentiation branching reported in the literature. Gene set enrichment analysis further revealed that IBL cells are responsible for keratinization while mORS cells are involved in inflammation (Figure 76). This observation holds for both healthy donors and HS patients, and was validated in an independent public dataset. Our results suggest a bias in the differentiation of HF-SCs, either toward IBL, responsible for a keratinizing profile in patients, or toward mORS, related to an inflammatory profile.

From this analysis, three marker genes were identified to quantify IBL, mORS and T cells in an independent cohort of 49 patients. Noteworthy, a quantification of their expression levels highlighted three groups of patients, which correlated with distinct clinical features (Figure 76). The clinical features include the gender, smoking status, body mass index, and the type of skin lesions.

Our study identified three groups of HS patients. The underlying group of each patient can be identified by their clinical phenotypes or by PCR using the set of three original markers. These results are promising for the development of personalized treatments in the context of HS, by targeting either the inflammation (mORS) or the keratinization (IBL). The markers of IBL and mORS are currently under patenting.

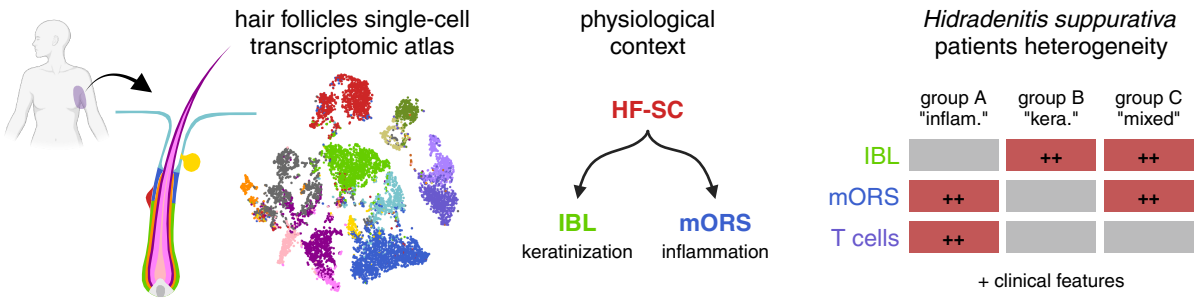


Figure 76 – Graphical abstract of the Article 3. A single-cell transcriptomic atlas of hair follicles highlighted two groups of patients, with distinct therapeutic implications. mORS: middle outer root sheath, IBL: inner bulge layer, HF-SCs: hair follicle stem cells.

Note: The manuscript is submitted. The Github link to make code available is broken and points to a private repository. It will be made public once the manuscript has been accepted.

The layout of the manuscript has been improved using a customized version of the Oxford University Press free layout, under the LaTeX Project Public License.

MANUSCRIPT SUBMITTED

The molecular signature of hair follicles classifies *Hidradenitis Suppurativa* patients into inflammatory, keratinizing, and mixed endotypes

Audrey Onfroy[1], Francette Jean-Louis[2], Philippe Lecorvoisier[3], Fanny Coulpier[1], Raphaele Arrouasse[3], Caroline Boucle[3], Salwa Abid[2], Émilie Sbidian[4], Christina Bergqvist[4], Pierre Wolkenstein[4], Véronique Godot[2], Yves Lévy[2], Piotr Topilko[1], Étienne Audureau[5] and Sophie Hüe[2,6*]

[1]INSERM U955-Team 9, Institut Mondor de Recherche Biomédicale (IMRB), Université Paris-Est Créteil (UPEC), Créteil, France, [2]INSERM U955 - Team 16, Institut Mondor de Recherche Biomédicale (IMRB), Université Paris-Est Créteil (UPEC), Créteil, France; Vaccine Research Institute (VRI), Créteil, France, [3]INSERM U955, Clinical Investigations Center 1430, AP-HP, DMU Saphire, Henri Mondor University Hospital, Creteil, France, [4]EpiDermE EA 7379, Université Paris Est Créteil, Créteil, F-94010, France; Department of Dermatology, Assistance Publique-Hôpitaux de Paris, Henri Mondor Hospital, Créteil, F-94010, France, [5]APHP, Hopital Henri-Mondor, Public Health Department, Creteil, France and [6]APHP, Hopital Henri-Mondor, Biologic Immunology-Hematology Department, DMU biologie, Creteil, France
*Corresponding author: sophie.hue@aphp.fr

Abstract

Hidradenitis suppurativa (HS) is a severe skin disorder affecting 1% of the global population, with a complex and poorly understood pathogenesis involving aberrant keratinization and autoinflammation. It remains unclear whether autoinflammatory events precede or follow hyperkeratotic changes in hair follicle (HF) epithelia. Using single-cell RNA sequencing, we compared the transcriptomic signatures of HF cells from HS patients and healthy donors. We identified three major cell populations: HF matrix cells, non-matrix cells, and immune cells. Unsupervised clustering revealed significant heterogeneity in non-matrix cells among HS patients, including HF stem cells (HF-SC) and two populations from the outer root sheath (ORS): the middle ORS (mORS) and the inner bulge layer (IBL). Trajectory analysis showed these two populations represent distinct differentiation branches from HF-SCs. Gene set enrichment analysis indicated IBL cells are enriched for keratinization genes, while mORS cells are enriched for interferon response genes. Analyzing an independent cohort of 49 HS patients, we stratified them into three groups based on clinical characteristics, each displaying specific enrichment in HF cells related to either keratinization (IBL) or inflammation (mORS). These findings suggest that the identified HF cell populations and their markers could serve as novel diagnostic biomarkers and therapeutic targets for HS.

Introduction

Hidradenitis Suppurativa (HS) is a chronic, severe, inflammatory skin disorder with a prevalence of around 1% and a profound impact on patients' quality of life [1]. Characteristic lesions such as inflammatory nodules, abscesses, and sinus tracts develop in the axillae, inguinal, and gluteal areas, typically during or after puberty. HS is a complex disease with contributory of genetic, epigenetic hormonal,

mechanical, microbial and lifestyle factors such as obesity and smoking. Multiple immunological pathways are described as overactivated, encompassing innate immune cells such as macrophages, immune soluble mediators and, in particular the complement and the Th1/Th17 pathways [2]. The treatment options are limited, leading to inadequate management of the disease for many patients. This is mainly due to the poor characterization of the events causing HS.

HS is recognized as an autoinflammatory keratinization disease because the pathomechanisms of HS are intimately associated with aberrant keratinization and autoinflammation [3]. In unaffected skin from predilection sites of patients with HS, the infundibular outer root sheath (ORS) shows hyperplasia and hyperkeratosis likely driven by an intrinsic dysregulation of the hair follicle stem cell (HF-SC) compartment. The ORS cells exhibit replication stress stimulating type I interferon responses [4]. Combined with factors such as obesity and smoking, this leads to follicle occlusion and stasis of keratin and bacteria within the dilating hair follicle. Subsequent cyst formation and rupture induce acute inflammation, characterized by an immune infiltrate of neutrophils, macrophages, dendritic cells, and T and B cells and increased expression of proinflammatory cytokines including IL1β, IL17, and TNFα [5]. Treatment targeting inflammation, such as adalimumab, achieved clinical response in only 40% to 60% of patients, and newer biological therapies have not significantly exceeded this response rate. A major barrier to identifying novel targets and achieving successful clinical translation is the lack of understanding of the interplay between keratinization and inflammation.

In this study, we generated a comprehensive single-cell RNA Sequencing (scRNA-Seq) atlas of hair follicle (HF) cells from HS patients by comparing the cell composition and transcriptomic signatures of these cells with those from healthy donors. Our analysis revealed that HF-SCs differentiate into two distinct cell types, each characterized by expression of specific markers and involved in either skin keratinization or the inflammatory response. By analyzing a cohort of 49 HS patients, we were able to stratify them into three groups based on their clinical characteristics. Each group demonstrated a distinct enrichment of an HF population associated with either keratinization or inflammation. These findings suggest that the identified HF cell populations and their markers could serve as novel diagnostic biomarkers and therapeutic targets for HS.

Materials and Methods

Ethic of human samples

The present study was conducted on a cohort comprising men and women with HS, as both sexes are affected by this condition. HS patients were recruited from the FoI-HYDRA study and enrolled from the dermatology department of the Henri Mondor Teaching hospital between June 2022 and December 2023. The Clinical Investigation Center performed skin biopsies.

For scRNA-Seq analysis, hair follicle samples were from discarded plastic surgery specimens from two healthy individuals and five HS patients (**Table S1**). This study was conducted in accordance with the declaration of Helsinki and was approved by the French institutional committee (reference N° 20.12.11.69413). The main inclusion criteria were 1) history of HS according to the diagnostic provided by the experimented investigator and 2) aged 18 years or older.

Human HF separation

Human scalp samples were digested in 1 U/mL Dispase (Stemcell Technologies) overnight at +4°C. Individual hair follicles were then pulled out one by one using forceps. To obtain single-cell suspension for scRNA-Seq, the isolated hair follicles were digested in TrypLE Express 1X (Gibco) for 30 minutes at room temperature, washed in DMEM + 10% FBS, and filtered through a 100 µm cell strainer.

Single-cell RNA sequencing

HF cells were sorted (BD InfluxTM Cell Sorter) using a nuclear labelling DRAQ7TM Far-Red Fluorescent Live-Cell Impermeant DNA Dye

(Abcam, ref: ab109202). Cells were collected in 0.04% BSA-RPMI solution.

For each individual, around 25,000 cells were loaded into one channel of the Chromium system using the V3.1 single-cell reagent kit (10X Genomics) to generate a single-cell gel bead emulsion. Following capture and lysis, cDNAs were synthesized, and then amplified by PCR for 12 cycles as per the manufacturer's protocol (10X Genomics). The amplified cDNAs were used to generate Illumina sequencing libraries that were sequenced on one flow cell Nextseq500 Illumina.

Bioinformatic analysis

Sequencing files were processed using 10x Genomics Cell Ranger 3.1.0. Reads were mapped on the GRCh37 (hg19) transcriptome. Raw count matrices were processed up to the figures presented here using RStudio, with R Markdown to generate fully traceable notebooks. To ensure version stability, a Singularity container containing R version 3.6.3 and all packages of interest was developed and used to compile notebooks. Version of packages is accessible in **Table S2**. HTML reports, including codes and figures, are available in a Github repository as an intuitive website (see the Code and Data Availability section).Individual datasets, associated with a single patient sample, were analyzed individually from the raw count matrix using Seurat V3 package [6]. Cells having less than 500 genes were filtered out. Doublet cells were removed using scDblFinder tool [7] and scds in hybrid mode [8]. Then, cells having more than 20% of UMI related to mitochondrial genes or more than 50% related to ribosomal genes were filtered out. The UMI count matrix for remaining cells was normalized using LogNormalize method implemented in Seurat::NormalizeData function. Single cells were annotated for cell type using a modified version of Seurat::AddModuleScore and cell type-specific marker gene sets in **Table S3**. According to previous publications [9–11] associated with scRNA-Seq from skin, they encompass CD4+ T cells, CD8+ T cells, Langerhans cells, macrophages, B cells, cuticle, cortex, medulla, IRS (inner root sheath), proliferative cells, IBL (inner bulge layer), ORS (outer root sheath), IFE (interfollicular epidermis), HF-SC (hair follicle stem cell), sebocytes and melanocytes. To smoothen eventual mis-annotation, single cell level annotation was grouped by cluster. Clustering was generated as follows. First, 3000 highly variable features (HVFs) were identified using Seurat::FindVariableFeatures. Then, we generated a principal component analysis (PCA) with 100 components, from those 3000 HVFs, using Seurat::RunPCA. Distances between cells were computed using top 20 components of the PCA, with Seurat::FindNeighbors. Finally, a clustering was generated by Seurat::FindClusters with resolution 2. The cluster annotation was defined as the most represented cell type by cluster.

To build the main "atlas" dataset (**Figure 1**), melanocytes clusters were removed in individual datasets prior to combination. Dataset combination was made using the base::merge function. For the populations-specific datasets associated with immune cells, HF-SCs (1) or IBL and mORS (2) and matrix cells, clusters of interest were selected in the atlas dataset, then merged. For the HFSCs + IBL + mORS dataset, datasets (1) and (2) were merged. For all 6 combined datasets, the same procedure. was applied. Genes expressed in less than 5 cells were removed. Merged count matrices were normalized using LogNormalize method and a 100 components PCA was made using the scaled expression of 2000 HVFs. Sample-specific effect was removed using harmony::RunHarmony on the PCA [12]. On the PCA, n dimensions were selected such that they summarize 60%, for the atlas, or 35%, for populations-specific datasets, of the variability between cells. This number of dimensions and the harmonized PCA were used as input to generate both clustering and t-distributed stochastic neighbor embedding (tSNE). The tSNE was built using Seurat::RunTSNE function.

For the dataset containing HF-SCs, IBL and ORS, a diffusion map was used to generate a projection representative of transcriptomic changes between cells, using the implementation from destiny package [13]. Trajectory inference was made using two distinct methods: slingshot [14] and TInGa [15], using cells in HF-SC

population as trajectory root, and the harmonized PCA as input. Slingshot output 4 lineages: two stopping in HF-SC population, one HF-SC to IBL and one HF-SC to mORS. TInGa can infer branching trajectory, thus was used for visualization purpose. Default parameters, except max_nodes set to 10, were used. Differential expression was conducted using Seurat::FindMarkers. Heatmaps were made using ComplexHeatmap package [16]. Functional enrichment analyses were conducted using msigdbr package [17] for the gene sets database, and clusterProfiler package for the analyses [18]. Gene sets score were computed using the Seurat::AddModuleScore function. When comparing the expression levels of scores between two populations, t-test implemented in the stats::t.test function was used.

Similarly, to validate results from IBL and mORS in our data, we reanalyzed the data from Wu and colleagues [9]. Briefly, FASTQ files were downloaded at accession number OEP002321 on biosino.org portal. Read mapping, individual samples analysis from raw count matrices, construction of the combined IBL + mORS dataset and its analysis were conducted using the exact same code as for our data. To visualize the whole atlas dataset (all samples, all cell types) coherently with the original publication, melanocytes were not removed.

RNA extraction and Real time Quantitative PCR

RNA extraction was performed according to the manufacturer's protocol (RNeasy Micro Kit, QIAGEN Inc.). RNA was converted to cDNA with QuantiTect reverse transcription Kit (QIAGEN). Quantitative PCRs were performed using the Brilliant II SYBR GREEN QPCR Master Mix kit (Agilent Technologies) for the expression of three genes specific for T cells, mORS and IBL. OAZ1 mRNA was used as a control for sample normalization. Expression of the gene OAZ1 was used as a reference and the relative levels of each gene were calculated using the $2\Delta\Delta CT$ method.

Clustering analysis of patients

Clustering analyses consisted of using the self-organizing maps (SOMs) algorithm developed by Kohonen. In a nutshell, the SOMs algorithm assigns each individual to a specific area on the map based on their characteristics, placing similar individuals in proximity and distinct ones in remote locations, thus allowing to draw visual comparisons of unique or overlapping patient characteristics and disease subtypes. The SOMs were constructed by applying the approach developed within the Numero package framework [19] for the R statistical platform 1) building the SOMs with statistical verification of the robustness of the contrasts observed by permutation tests and 2) determining suitable groupings based on the direct visualization of data patterns and key characteristics of the dataset.

For illustrative purposes, Gabriel's biplots were plotted to project the patients along the principal components axes from mixed principal component analysis according to their individual characteristics, colouring patients according to their final diagnosis or cluster and thus allowing for direct visual assessment of the discriminative ability of each subgrouping.

For descriptive statistics, categorical variables are expressed as proportions (%), quantitative variables as means (±standard deviation [SD]) or medians (interquartile range [IQR]), as appropriate. Biological parameters were log-transformed due to the non-normality of their distribution. We compared groups from clustering by means of one-way ANOVA or Kruskal-Wallis tests for continuous variables and chi-square tests or Fisher's exact tests for categorical variables. A p-value < 0.05 was considered significant. Descriptive statistics and between-clusters comparisons were realized using Stata software 15.1 (StataCorp, Tx, USA), and R 3.4.3 (R Foundation, Vienna, Austria; pca2d, Numero packages) for clustering analyses and visualizations.

Study approval

This monocentric, prospective study was conducted in accordance with the declaration of Helsinki and was approved by the appropriate ethics committee (CPP North-West IV: 2021-A02352-39, 13/01/2022). All patients gave written informed consent before study enrolment.

Code and data availability

Single cell RNA sequencing data are available as FASTQ files, raw count matrices and individual Seurat object in Gene Expression Omnibus repository, under the accession number GSE273109. Detailed methodological steps, exhaustive parameter settings, along with intermediate figures, for the raw count matrices to figure process are available in HTML reports, shared through a Github repository (`https://github.com/audrey-onfroy/scrnaseq_hair_follicles_hidradenitis_suppurativa/index.html`).

Results

Single-Cell RNA Sequencing of hair follicle cells reveals differences between HS patients and healthy donors

We obtained discarded human skin samples from plastic surgery and isolated hair follicles (HFs) from them by dispase digestion followed by individually pulling the HFs out of dermis following the protocol described by Wu and colleagues [9]. Skin samples were collected from 5 HS patients in perilesional lesions and from 2 healthy donors (HD) (see **Table S1** for patients' information). We performed scRNA-Seq profiling of HF cells using 10x Chromium platform (see Materials and Methods) and obtained 12,111 cells passing quality control steps. Cells were annotated for cell types based on established panel of HF markers (**Table S3**) [9–11]. They were assigned to three categories of cells: i) immune cells characterized by expression of PTPRC ii) matrix characterized by expression of transcriptional repressor MSX2 and iii) non-matrix characterized by expression of KRT14 (**Figure 1A**).

Immune cell category included five cellular populations: i) *CD4*+ and ii) *CD8*+ T cells expressing respectively *CD3* and *CD4* or *CD8A* markers, iii) Langerhans cells co-expressing *AIF1* and *CD207*, iv) macrophages co-expressing *AIF1* with *TREM2* and *MSR1* and v) B cells expressing *CD79* (**Figure 1B-C**).

The matrix category can be further divided based on specific transcriptomic signatures into five populations as follow: i) cuticle (*KRT32*, *KRT35*), ii) cortex (*KRT31*), iii) medulla (*ALDH1A3*, *BAMBI*), iv) inner root sheath (IRS) (*KRT71*, *KRT73*) and v) proliferative cells, expressing high levels of *TOP2A* and *MCM5* genes (**Figure 1B-C**).

The non-matrix category also regroups five populations: i) hair follicle stem cells (HF-SCs) expressing high levels of *CXCL14*, *KRT15*, *COL17A1*, *DIO2* and *TCEAL2*, ii) middle outer root sheath cells (mORS) expressing *KRT15* and *COL17A1* but lacking *DIO2* and *TCEAL2* expression, (iii) IBL cells expressing *KRT16* and *KRT6C* [9], (iv) interfollicular epidermis (IFE) expressing *SPINK5* and *LY6D* and (v) sebocytes expressing CLMP and PPARG (**Figure 1B-C**).

This comprehensive molecular annotation of the three main categories identified 15 transcriptionally distinct populations in HS and healthy donors' hair follicles, suggesting that the cellular composition of hair follicles remains similar (**Figure 1D**). We then investigated whether there were differences in cell proportions among the three categories between individuals. The findings revealed a higher proportion of immune cells accompanied by a reduction in the proportion of matrix cells in HS patients (**Figure 1E**).

To provide more specific details on the proportions of each cell type within the three categories, the cells were clustered independently of their annotation (**Figure 1F**). Subsequently, the abundance of cells within each cluster was compared across the different samples. The unsupervised clustering of abundances allowed for the differentiation between HS and HD samples (**Figure 1G**). Additionally, it identified two distinct categories of HS patients (**Figure 1G**). The most notable discrepancies between HS patients were observed in the non-matrix cell populations. The first group (HS3 and HS5) was enriched in the IBL population (cluster 1), while the second group (HS1, HS2 and HS4) showed an enrichment in the mORS population (cluster 0) and T cells (cluster 2).

In conclusion, while the overall cellular composition of HFs from patients with HS remains unchanged compared with healthy donors, there is a notable increase in the proportion of immune

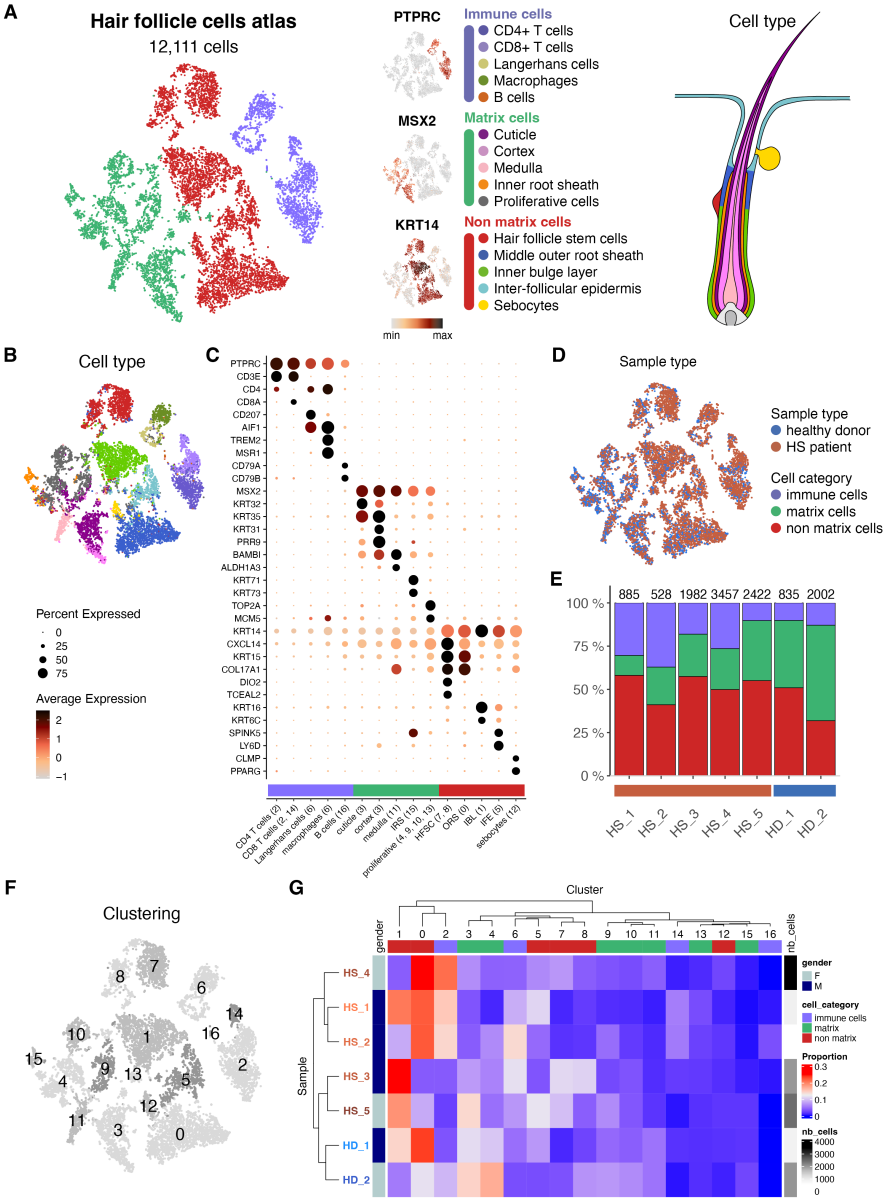


Figure 1. Transcriptomic atlas of hair follicles cells from healthy donors (HD) and Hidradenitis suppurativa (HS) patients. (A) tSNE plot showing 12,111 hair follicle cells from 2 HD and 5 HS patient samples. Cells are colored based on three cell categories, characterized by *PTPRC* (immune cells), *MSX2* (matrix cells) or *KRT14* (non-matrix cells) expression. Each category corresponds to 5 cell types, represented on a hair follicle draw. Created with Biorender.com. (B) tSNE plot of hair follicle cells, colored based on cell type annotation. Legend is shared with (A). (C) Dot plot assessing cell type annotation. Dot size represents the percentage of cells with positive expression for each marker. Dot color represents the scaled expression level of markers across all cells within each cell type. Number in brackets in X-axis corresponds to cluster number in (F). (D) tSNE plot of hair follicle cells, colored based on sample type. All cell types are shared across all samples. (E) Barplot depicting the cell category composition of individual samples. Total number of cells within each sample is indicated on top. Legend is shared with (D). (F) tSNE plot of hair follicles cells, colored based on clusters. (G) Heatmap showing the proportions of cells within each cluster (column), for each sample (row). Each sample is annotated by gender (F, female or M, male) and by the number of cells it contains. Clusters are annotated by the cell category they represent.

cells in HS samples. A more detailed analysis showed that the non-matrix populations (IBL and mORS) and T lymphocytes can distinguish two different types of HS patients.

The molecular characteristics of immune cells from patients with HS depicted activation compared to those from healthy donors

As HS is an inflammatory disease, we further interrogated the potential heterogeneity among immune cells between HS patients and HD. We extracted all immune cells from the complete dataset and projected them on a tSNE (**Figure 2A**). All immune populations were shared across all samples, suggesting similar cellular composition. However, samples differ from the abundance of immune cell types. We discriminated two groups of HS patients: one enriched in macrophages (HS3 and HS5) and the second enriched in T cells (HS1, HS2 and HS4) (**Figure 2B**). To investigate the differences in macrophages and T cells between HD and all HS patients, we performed differential expression analysis.

Macrophages showed an increase in Major Histocompatibility Class (MHC) class I and MHC class II expression in all HS patients compared to HD. Moreover, complement pathway is enriched in all HS patients (**Figure 2C**). Because IL1β is suspected to be the dominant pathway in HS disease [20], we investigated *IL1B* expression within macrophages. As expected, *IL1B* levels are significantly elevated in HS patients ($p = 0.03$) although there is some variability in its expression due to the heterogeneity across patients (**Figures 2D; S1A**). Additionally, the only FDA-approved treatment for HS is adalimumab, targeting TNF pathway [21]. As with *IL1B*, we observed a significant increase in TNF expression among HS patients ($P < 0.001$) (**Figure 2E; S1B**). Note that *IL6* expression was not observed in the macrophages of HS patients (**Figure S1D**). Overall, these results indicate a pro-inflammatory phenotype of macrophages located around the hair follicle in the perilesional areas of HS patients.

We then focused on *CD4*+ T cells as the Th17 axis is a central feature of inflammation in this disease and has been previously characterized in HS tissues [2]. *IFNG* and *IL17* expressing *CD4*+ T cells were observed in HS patients with considerable inter-individual variability (**Figure S1E, F**). In contrast, almost no *IFNG* or *IL17*-expressing *CD4*+ T cells were present in healthy donors. Interestingly, CD4+ T cells have a cytotoxic phenotype with unusual high expression of granzyme A (*GZMA*) and *KLRB1* (**Figure 2F, G, S1C**). The presence of cytotoxic CD4+ T cells was validated by immunofluorescence on skin biopsies from HS patients with GZMA staining (**Figure 2H**).

Subsequently, an analysis of cytotoxic *CD8*+ T cells was conducted, which revealed a greater prevalence of *IFNG*, *GZMB* and *PRF1*-expressing *CD8*+ T cells in HS patients compared to healthy donors (**Figure S1E, G-H**).

Given that the generation of cytotoxic *CD4*+ and *CD8*+T cells requires type I interferon and *IL12* signaling [22], and considering our previous demonstration of an interferon signature in HS [4], the presence of these cytotoxic *CD4*+ and *CD8*+ T cells further emphasizes the crucial role of interferon in HS pathogenesis.

HF-SCs from HS patients exhibit molecular features of inflammation and activation

We have previously shown a disturbance in the homeostasis of HF-SCs in HS disease [4]. To gain further insight into the alterations in transcriptomic signature within the HF-SCs, we extracted these cells and performed a clustering analysis. Investigation of the differentially expressed genes between subclusters identified nine subclusters among the HF-SCs (**Figure 3A**).

Unsupervised clustering based on the abundance of HF-SCs in these nine subclusters revealed two distinct groups of individuals (**Figure 3B**). The first group (HD1, HS1, and HS2) was enriched in subcluster 2 (SC2), in contrast to the second group (HS3, HS4, HS5, and HD2). Differential expression analysis indicated that SC2 was characterized by the expression of angiopoietin-like 7 (*ANGPTL7*) (**Figure 3B**), which is known to be expressed during the telogen phase of the HF cycle [23]. These two groups also correlated with the overall proportion of HF-SCs in each sample. Specifically, the first group (HS1, HS2, and HD1) had approximately

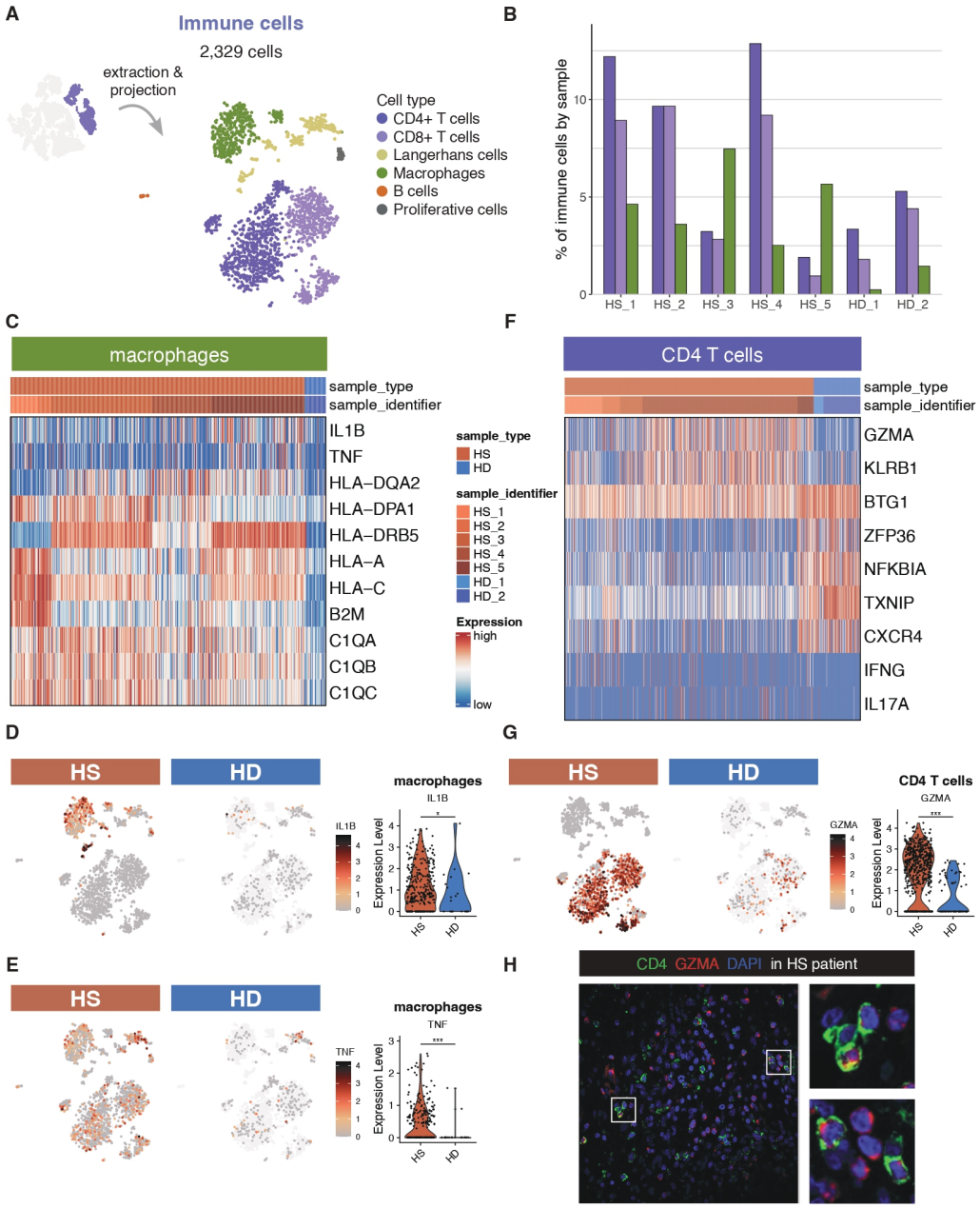


Figure 2. Immune cells in HS patients are activated compared to healthy donors. (A) tSNE plot showing 2,329 immune cells extracted from the main atlas (**Figure 1**). Cells are colored based on cell type annotation. (B) Barplot depicting the proportion of T cells and macrophages among all cells in each sample. Legend is shared with (A). (C) Heatmap showing differentially expressed genes in macrophages, between HS and HD samples. (D-E) Feature plots and violin plots showing *IL1B* (D) and *TNF* (E) expression levels, split by sample type. On the tSNE, cells from complete immune cells dataset are shown as light background. The violin plots consider only macrophages. (F) Heatmap showing differentially expressed genes in *CD4*+ T cells, between HS and HD samples. Legend is shared with (C). (G) Feature plot and violin plot showing *GZMA* expression level, split by sample types. On the tSNE, cells from complete immune cells dataset are shown as light background. The violin plots consider only CD4+ T cells. (H) Section through hair follicles from a HS patient, immunolabeled from CD4, GZMA and DAPI. High magnification areas are denoted by square and represented on the right. Statistical t-test p-value, ns: > 0.05, *: < 0.05 and > 0.01, **: < 0.01 and > 0.001 and ***: < 0.001. For additional features in immune cells, refer to **Figure S1**.

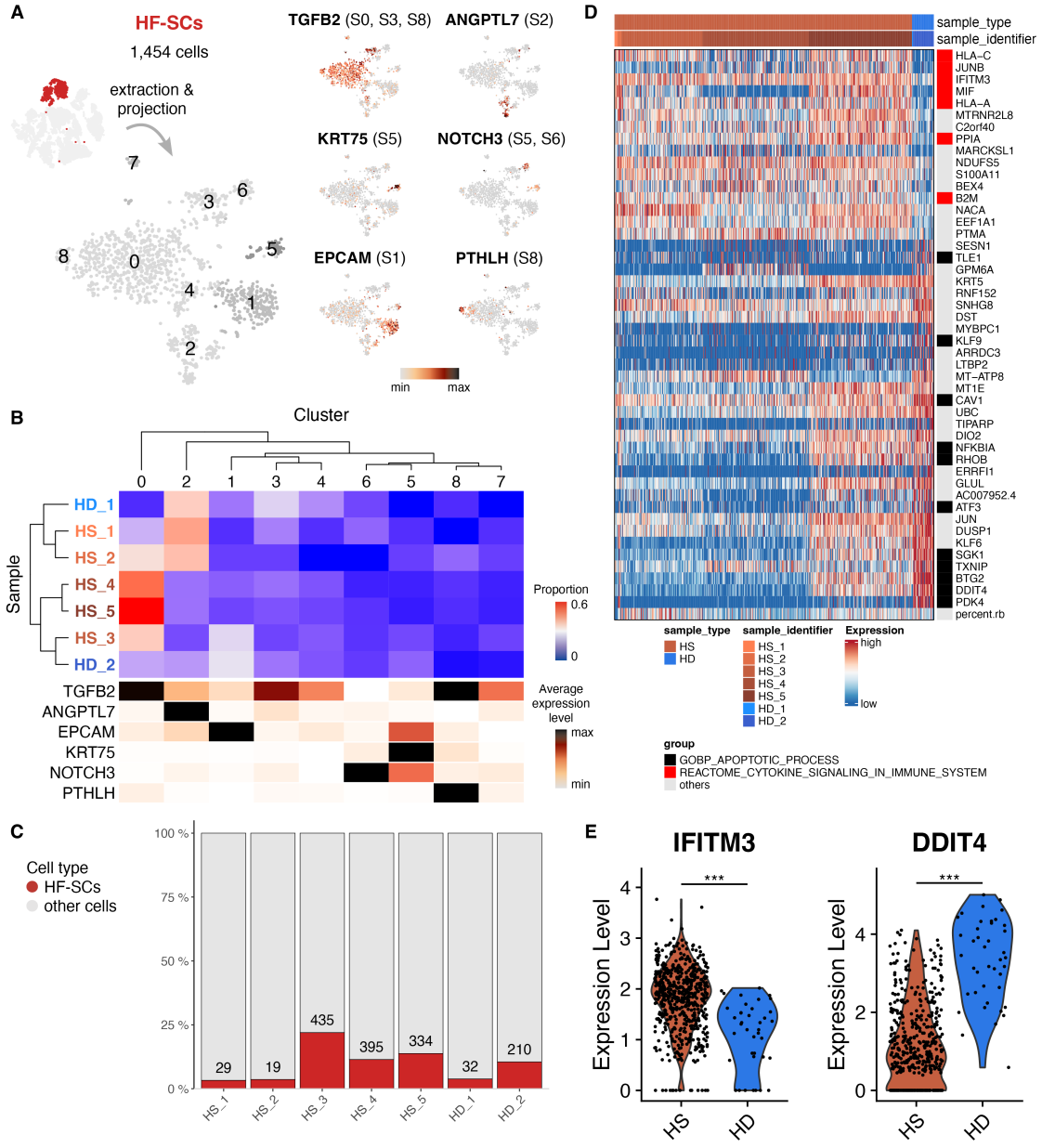


Figure 3. Heterogeneity of hair follicle stem cells (HF-SCs) shows enrichment in inflammation markers in HS patients. (A) tSNE plot showing 1,454 HF-SCs extracted from the main atlas (Figure 1). Cells are colored based on a clustering. (Right panel) Feature plots showing expression levels of representative markers for each subcluster, indicated in brackets. (B) Heatmap showing the proportions of cells within each subcluster (column), for each sample (row). Bottom annotation depicts the average expression levels for each marker within the subcluster. (C) Barplot depicting the proportion of HF-SCs, among all cells, in each sample. Total number of HF-SCs within each sample is indicated. (D) Heatmap showing differentially expressed genes in clusters S0 and S8 (TGFB2+), between HS and HD samples. Gene names are color-coded according to two gene sets from the molecular signature database (MSigDB). (E) Violin plots showing IFITM3 and DDIT4 expression levels, among HF-SCs, split by sample type. Statistical t-test p-value, ns: > 0.05, *: < 0.05 and > 0.01, **: < 0.01 and > 0.001 and ***: < 0.001.

3% HF-SCs, while the second group (HS3, HS4, HS5, and HD2) had more than 10% HF-SCs (Figure 3C). Overall, these results suggest that the observed dichotomy in the seven samples is

linked to the stage of the HF cycle. The first group is associated with the telogen (resting phase), while the second is associated with the anagen phase, during which HF-SCs proliferate.

Nevertheless, an enrichment in subcluster 0 (SC0), which expresses the highest level of *TGFB2*, was observed in all HS patients (**Figure 3B**). Differential expression analysis within this SC0 subpopulation identified fundamental transcriptomic changes between HD and HS patients (**Figure 3D**). SC0 from HS patients have an inflammatory phenotype, as evidenced by an enrichment in HLA class I molecule, *MIF* and *IFITM3* (**Figure 3D-E**). In addition, SC0 was enriched for ribosome genes in HS patients suggesting that it represented HF-SCs poised for activation. In contrast, SC0 from healthy donors exhibited an enrichment in genes involved in DNA repair, such as *DDIT4*, as well as the apoptosis pathway (**Figure 3D-E**). These pathways are crucial for preventing excessive stem cell proliferation and removing damaged cells, thereby maintaining the integrity of the hair follicle [24].

Our results indicate HF-SCs from HS patients are highly enriched in inflammatory and activation markers and are less susceptible to apoptosis. As we previously demonstrated a loss of HF-SC homeostasis associated with replication stress [4], these results highlight the significant role of HF-SCs in the pathogenesis of HS.

IBL and mORS exhibit distinct molecular characteristics and function

We previously identified two groups of HS patients based on respective enrichment in IBL (cluster 1) and mORS (cluster 0) (**Figure 1G**). We thus extracted IBL and mORS from the complete dataset (**Figure 4A**) and investigated differences between both populations, in all HS and HD samples. Notably, both populations displayed fundamental transcriptomic changes (**Figure 4B**). To further identify signaling pathways underlying differentially expressed genes, we used Gene Set Enrichment Analysis (GSEA). Noteworthy, GSEA revealed significant enrichment for genes involved in keratinization in IBL, whereas mORS showed enrichment for interferon response genes (**Figure 4C**). By assigning expression scores for both gene sets to all cells, we showed that these enrichments reflect the physiological characteristics of IBL and mORS, rather than being specific to HS disease (**Figure S2A**).

To extend our findings, we re-analyzed 6 previously published scRNA-Seq datasets from HFs isolated from healthy individuals [9]. Using the same methodology applied to our data, we identified 11 transcriptionally distinct populations shared across all six samples (**Figure S2B-D**). We then extracted IBL and mORS from the complete dataset (**Figure S2E,F**). Enrichment analysis confirmed our results. Specifically, IBL and mORS exhibit similar profiles in terms of keratinization and inflammation (**Figure S2G,H**).

To determine whether IBL and mORS represent successive or distinct differentiation branches from HF-SCs, we generated a diffusion map. This method preserves continuous transcriptomic changes in cell populations, allowing us to visualize potential differentiation branches. **Figure 4D** shows that IBL and mORS form two separate branches from HF-SCs, suggesting distinct differentiation processes rather than a sequential progression from one to the other. The representation does not highlight differences between HD and HS patients, suggesting that the global structure of the three populations is conserved in HS (**Figure 4E**). To further evaluate the differentiation pathways, we used trajectory inference starting from the HF-SC population. The Slingshot tool, which is suited for inferring linear branches, identified two distinct differentiation pathways: HF-SCs to IBL and HF-SCs to mORS (data not shown). Additionally, the TInGa tool confirmed these findings by illustrating the two lineages in a single branching trajectory (**Figure 4F**).

HF-SCs express a set of highly enriched genes, including transcription factors (*SOX9*, *LHX2*, *RUNX1*), which maintain these cells in a quiescent, undifferentiated state (**Figure S3A-D**). Differentiated progeny from HF-SCs maintain expression of numerous stem cell markers [25]. Notably, we observed that mORS maintain *SOX9*, *RUNX1* and *COL17A1* as stem cell markers whereas IBL maintain *SOX9*, *LHX2* and *FGF18* (**Figure S3B-F**).

Overall, these results from scRNA-Seq of HFs suggest HF-SCs differentiate in two cell types with distinct characteristics: (i) IBL, which has a keratinizing phenotype, and (ii) mORS, which has an inflammatory phenotype.

IBL and mORS depicted inflammatory molecular signature in HS patients

First, we investigated differences in IBL population between HS patients and HD. Our analysis revealed that IBL from HS patients exhibit an inflammatory phenotype with overexpression of *MIF*, *S100A9* and display an active differentiation process with overexpression of galectin-7 (*LGALS7*), a lectin that is almost exclusively expressed in stratified epithelia [26] and ARF5, a small GTPase that is involved in splicing complex, ribosome synthesis and function, intracellular material transport [27]. In contrast, IBL from HD showed up-regulation of *DUSP1* and *DDIT4*, which play a crucial role in cell cycle control (**Figure 5A**).

For mORS, we observed an identical profile with an increase in the inflammatory profile (*CCL2*, *S100A9*) and a type 1 interferon signature (*IFI27* and *IFITM3*) and a decrease in genes involved in the cell cycle control (*DUSP1*, *KLF6*) and in genes involved in cell adhesion (*CTGF* and *CLDN1*) for HS patients (**Figure 5A**).

By splitting the representation between HS and HD, we noticed a subpopulation of mORS was enriched in HS samples (**Figure 5B-C**). This population, characterized by *IL1R2* and *WNT3* expression, was mostly present in the mORS-enriched HS patients and absent in the IBL-enriched HS patients (**Figure 5C**). Their paucity in HD patients was confirmed in an already published dataset (**Figure S2F**). Transcriptomic profile of this *IL1R2*+*WNT3*+ mORS subpopulation indicates that these cells are inflammatory (*IFI27*, *IL17* and *CXCL14*), responsive to IL1 (*IL1R2*) and proliferative (*WNT3*, *WNT4*) (**Figure 5D**).

Finally, the matrix cells were subjected to analysis (**Figure S4A**). The samples exhibited a notable variation in the abundance of matrix cell populations. Our observations revealed that HS3 and HS5 patients displayed an enrichment in cortex cells (**Figure S4B**). GSEA within this cortex population between HS patients and HD identified transcriptomic changes associated with keratin filaments and the control of the spatial distribution of intermediate filaments (**Figure S4C**). Consistent with these findings, *KRT32* which belongs to the family of keratins, a type of intermediate filament proteins that form intricate filament systems in epithelial cells was significantly increased in HS patients compared to HD (**Figure S4D**).

In conclusion, transcriptomic profiling of hair follicles identified two distinct profiles of HS patients. The first profile is marked by an abundance of IBL, macrophages, and cortex. The second profile is characterized by an abundance of mORS, which includes a highly inflammatory subpopulation, along with cytotoxic T cell infiltration. All non-matrix populations (HF-SCs, IBL and mORS) in HS patients exhibit an inflammatory phenotype and show reduced susceptibility to apoptosis.

Distinct physiopathology processes observed in HS patients

We sought to determine whether such a stratification of patients into two groups overlap with clinical phenotype. To achieve this, we isolated 5 fresh hair follicles from skin biopsies issued from the Fol-Hydra cohort of 49 HS patients (**Table 1**). After extracting RNA from the hair follicle cells, we conducted qPCR analysis on specific genes for IBL, mORS and T cells. Given the considerable clinical heterogeneity among HS patients, we began our analysis by characterizing the patients based on their clinical features. Using the self-organizing maps (SOMs) methodology, we performed clustering analysis, which displayed polarized distributions of patient characteristics across the maps. This was evidenced by the contrasting colourings, ranging from blue (lowest values) to red (highest values) (**Figure 6A**). From these results, 3 clusters of HS patients with homogenous features were built (**Figure 6B**). Cluster boundaries are shown as solid black lines across the maps, with detailed clinical characteristics presented in **Table 1**. The first group (cluster A, n=14) is characterized by the presence of young, non-smoking women with a high body mass index

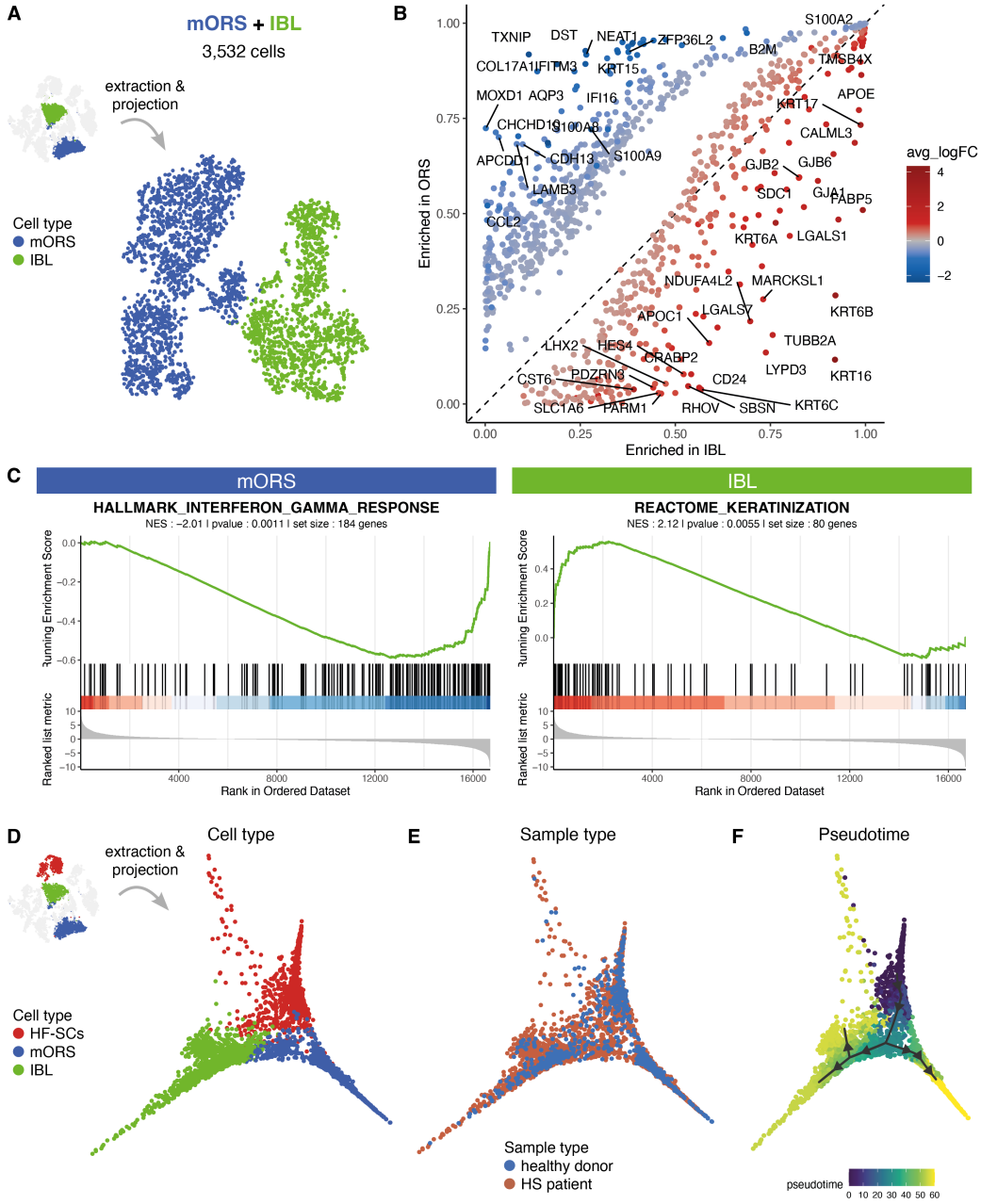


Figure 4. Inner bulge layer cells (IBL) and middle outer root sheath cells (mORS) represent two branches of differentiation from HF-SCs, with distinct transcriptomic activities. (A) tSNE plot showing 3,532 IBL and mORS cells from the main atlas (Figure 1). Cells are colored based on cell type annotation. (B) Representation of differentially expression genes between mORS and IBL from all samples. X- and Y-axis respectively correspond to the proportion of cells within IBL and mORS with positive expression of the gene. Dashed line corresponds to genes expressed in same proportion of cells in IBL and mORS. One dot, corresponding to one gene, is colored according to average log fold change (avg_logFC) between IBL and mORS. (C) GSEA plots made from average fold change of all genes, between IBL and mORS. (D) Diffusion map of merged HF-SCs, IBL and mORS cells from the main atlas. Cells are colored based on cell type annotation. (E) Diffusion map of HF-SCs, IBL and mORS colored based on sample type, showing overlapping between HD and HS patients cells. (F) Diffusion map of HF-SCs, IBL and mORS colored based on pseudotime, resulting from trajectory inference using TInGa, starting in HF-SCs.

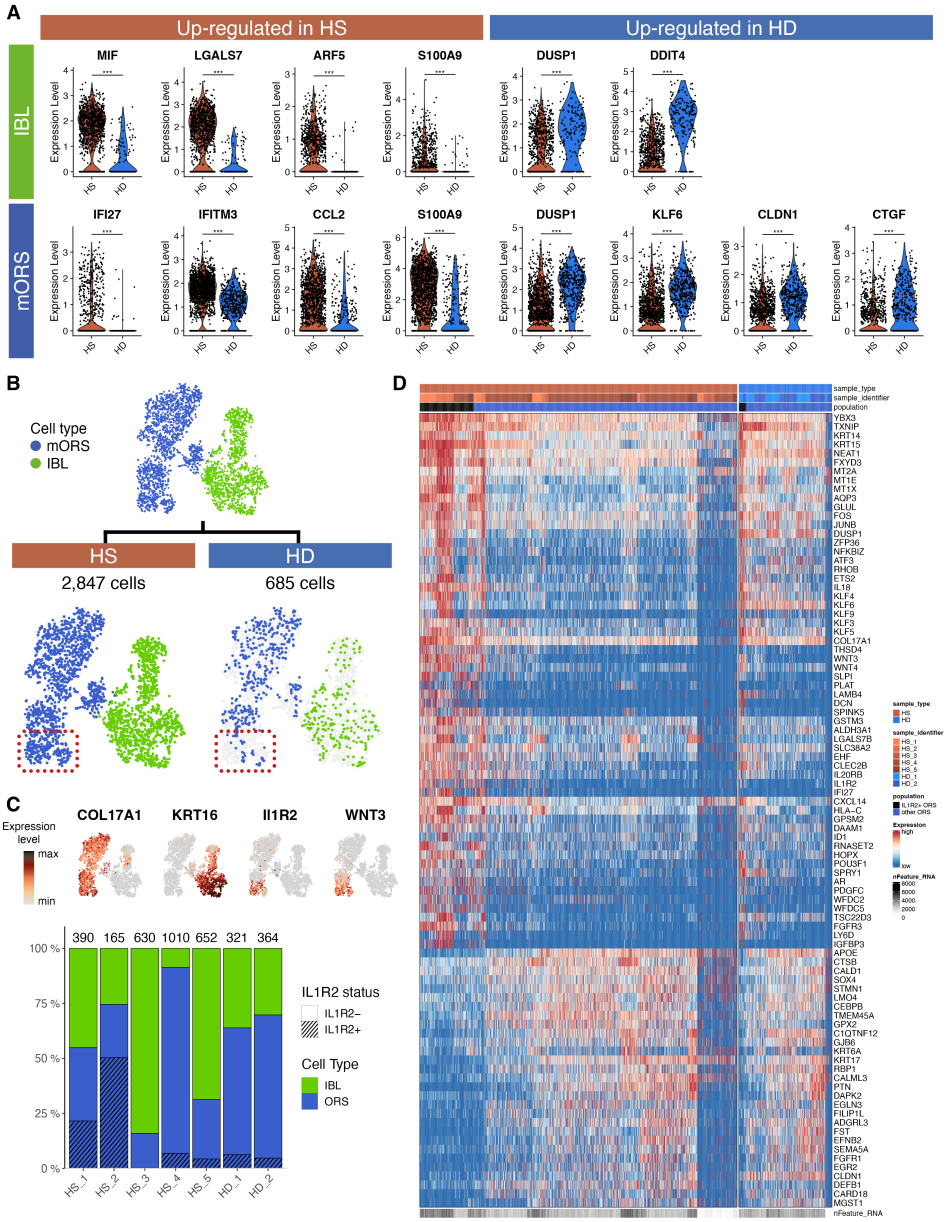


Figure 5. IBL and mORS depicts enrichment in inflammatory signature in HS patients compared to healthy donors. (A) Violin plots showing the expression level of differentially expressed genes between HS and HD, split by sample type, only within mORS (top) or within IBL (bottom). Statistical t-test p-value, ns: > 0.05, *: < 0.05 and > 0.01, **: < 0.01 and > 0.001 and ***: < 0.001. (B) tSNE plot of the IBL and mORS dataset, split by sample type. Cells are colored based on cell type annotation. Cells from the complete dataset are shown as light gray background. A dashed red box highlights a subpopulation in mORS, enriched in HS patients compared to HD. (C) (Top panel) tSNE plot of the IBL and mORS dataset, where cells are colored based on expression levels of selected markers for IBL, mORS and the HS-enriched mORS subpopulation. (Bottom panel) Barplot depicting the representativness of IBL and mORS in individual samples. The HS-enriched mORS subpopulation is represented as hatched bars. Total number of IBL and mORS within each sample is indicated on top. (D) Heatmap showing differentially expressed genes between the *IL1R2+* population and other mORS, from HD and HS patient cells. The top annotation corresponds to sample type, sample identifier and population, where the HS-enriched mORS subpopulation is indicated in black. The bottom annotation corresponds to the number of genes in each cell. Low levels indicate lower cell quality.

(BMI) and an early onset of disease. 11 out of 14 patients had an LC1 phenotype and 12 had inflammatory skin nodules (**Figure 6A, B**). Their hair follicles were mainly enriched in T cells and mORS (**Figure 6C**). The second group (cluster B, n=16) is also characterized by the presence of women with a predominant LC1 phenotype, but they are older, smokers and have a normal BMI and a later onset of disease. Most of them report a HS familial history and their nodules are devoid of inflammatory characteristics (**Figure 6A, B**). In this group, the HF cells are enriched in IBL with no T cells infiltrate (**Figure 6C**). The final group (Cluster C, n=19) was predominantly male (57.9%), smokers, with a later age at diagnosis and little family history (1 out of 19). Six out of ten patients exhibiting an LC3 phenotype fall within this cluster C. They differed from the others because of patchy scarring, fistulas and continuous involvement. These patients exhibited the highest HS-PGA (*Hidradenitis Suppurativa* Physician Global Assessment) scores (**Figure 6A, B**). Their hair follicle cells are enriched in IBL and mORS, but there is no T cell infiltrate (**Figure 6C**).

These results demonstrated a correlation between the subtypes identified by transcriptomic analysis and distinct clinical features. In conclusion, these data suggest that the clinical features arise from three distinct pathophysiological processes: (i) a process in which inflammation is predominant (cluster A), (ii) a process in which keratinization occurs without inflammation (cluster B), and (iii) a process in which the two mechanisms are intertwined (cluster C).

Discussion

Decoding the primary events that trigger follicular hyperplasia in HS remains extremely challenging due to the absence of appropriate HS animal models. In contrast to published studies on the cells involved in HS lesions, our research focused on the cells that comprise the hair follicle, as we have previously demonstrated their involvement in the early stages of the disease. Using scRNA-Seq analysis of hair follicle cells, we provided several critical insights into the pathogenesis of HS. We established that two populations derived from HF-SCs play two major roles, one in keratinization process and the other associated with inflammation. All non-matrix populations (HF-SCs, IBL and mORS) in HS patients exhibit an inflammatory phenotype and show reduced susceptibility to apoptosis. We revealed three distinct endotypes of HS patients based on hair follicle cell composition: i) inflammatory endotype: young non-smoking women with a high BMI and inflammatory nodules. Their hair follicle cells exhibited an enrichment in mORS cells and T cells, indicating a predominance of inflammation; ii) keratinizing endotype: smoking women with a normal BMI and no inflammatory nodules. Their hair follicle cells exhibited an enrichment in inner bulge layer (IBL) cells, suggesting that keratinization was predominant; iii) mixed endotype: smoking men with fistula and continuous involvement. Their hair follicle cells showed an enrichment in both IBL and mORS cells, without T-cell infiltration, indicating a combination of keratinization and inflammation. The data presented herein provide the first evidence that two hair follicle-derived cell types with distinct molecular characteristics can stratify HS patients either related to inflammation, to keratinization or to both mechanisms. These cells could serve as potential markers for personalized medicine in HS patients.

Immune dysregulation surrounding the hair follicle following follicular occlusion is hypothesized to initiate the aberrant inflammatory process characteristic of HS. Here, we found an infiltration of cytotoxic CD4+ T cells (CD4+ CTLs) which, though commonly expanded in cancer and autoimmune conditions, also appear in healthy supercentenarians and acute infections like SARS-CoV-2 [28]. These cells are likely restricted due to their potential to cause immunopathology. In influenza A infections, CD4+ CTLs develop from CD4 effectors in infected tissues, requiring local antigen presentation and type I interferon, which induces interleukin 15, without needing CD28 co-stimulation [29]. Therefore, type I interferon produced by ORS [4] and the presence of an infiltration of B cells as antigen presenting cells observed in HS lesions [30] cells may contribute to the development

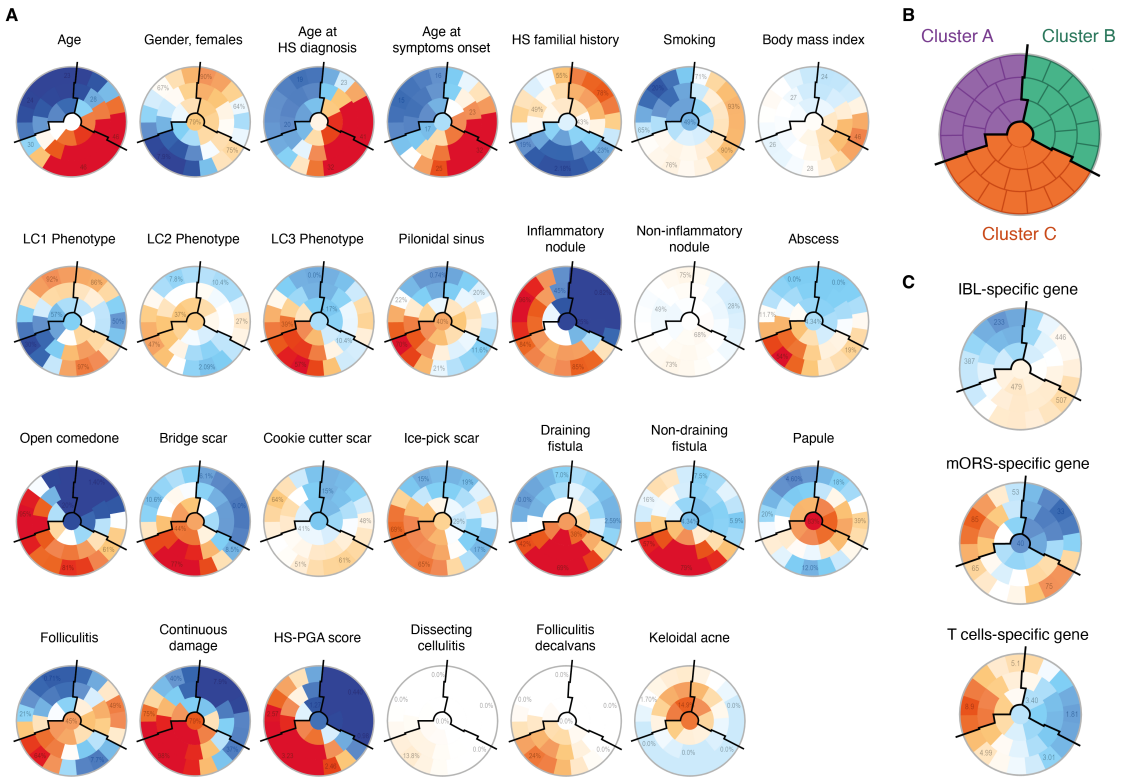


Figure 6. Clustering analysis of HS patients based on the Self-Organizing Maps (SOMs) methodology. (A) Unsupervised analysis by SOM placed all HS patients identified as globally similar for age, gender age at HS diagnosis, HS familial history and cigarette smoking within 1 of 40 small groupings (districts) throughout the maps. The more individuals resemble in terms of phenotyping, the closer they are placed on the map. Each individual map shows the mean values per district for each characteristic, blue colors indicating the lowest average values and red colors the highest, with detailed numbers written for a selection of representative districts in each SOM. (B) Close districts were combined to provide 3 suitable clusters of HS patients, labelled A to C. The clusters boundaries are delimited by solid black lines. (C) Representation of gene expression levels on the SOMs. These values were not taken into account while running the algorithm.

of CD4+ CTLs. Future experiments should investigate the functional role of CD4+ CTLs, particularly their involvement in the disruption of hair follicle homeostasis.

Immune cells have emerged as critical regulators in development, tissue regeneration, and repair [24]. Resident macrophages play diverse roles in HF-SC regulation. Interestingly, dermal TREM2+ macrophages produce oncostatin M, which regulates HF-SC quiescence [31]. Furthermore, TNFα, secreted mainly by activated macrophages following injury has been demonstrated to stimulate new hair growth even beyond the plucked area [32]. On the other hand, HF-SCs must resist immune attack due to their critical role in long-term tissue renewal. Consequently, HF-SCs evade detection and destruction by both innate and adaptive immune cells through the systemic downregulation of the antigen presentation machinery, including MHC class I and TAP proteins [33]. We observed that hair follicle cells in the perilesional zone were enriched with macrophages displaying the specific pro-inflammatory signature of HS characterized by increased expression of *IL1B*, *TNF*α and complement [20]. Additionally, a subcluster of HF-SC was identified which expressed high level of *TGFB2* and MHC class-I genes. We postulate the secretion of *IL1B* and *TNF*α by macrophages may induce the MHC class I expression [34] observed in this subcluster. This would result in a vicious

Characteristic	N	Cluster A N=14[1]	Cluster B N=16[1]	Cluster C N=19[1]	p-value[2]
Clustering features					
Age	49	23.5 [20.3;25.8]	31.0 [23.0;40.8]	38.0 [29.0;45.0]	**<0.001**
Gender, females	49	9 (64.3%)	12 (75.0%)	8 (42.1%)	0.13
Age at HS diagnosis	48	19.0 [17.3;21.8]	27.5 [22.0;36.0]	26.0 [20.3;32.8]	**0.001**
Age at symptoms onset	49	14.0 [14.0;15.8]	21.0 [17.0;23.8]	25.0 [17.5;26.5]	**<0.001**
HS familial history	38	7 (53.8%)	5 (55.6%)	1 (6.3%)	**0.006**
Smoking	35	3 (37.5%)	9 (81.8%)	12 (75.0%)	0.12
Body mass index	49	25.5 [24.3;31.3]	23.5 [21.0;30.5]	27.0 [25.0;32.5]	0.22
LC1 Phenotype	49	11 (78.6%)	11 (68.8%)	13 (68.4%)	0.85
LC2 Phenotype	49	3 (21.4%)	3 (18.8%)	4 (21.1%)	>0.99
LC3 Phenotype	49	2 (14.3%)	2 (12.5%)	6 (31.6%)	0.41
Pilonidal sinus	49	2 (14.3%)	3 (18.8%)	7 (36.8%)	0.29
Inflammatory nodule	49	12 (85.7%)	1 (6.3%)	15 (78.9%)	**<0.001**
Non-inflammatory nodule	49	8 (57.1%)	8 (50.0%)	12 (63.2%)	0.74
Abscess	49	1 (7.1%)	1 (6.3%)	5 (26.3%)	0.23
Open comedone	49	8 (57.1%)	4 (25.0%)	13 (68.4%)	**0.033**
Bridge scar	49	3 (21.4%)	0 (0.0%)	10 (52.6%)	**<0.001**
Cookie cutter scar	49	7 (50.0%)	5 (31.3%)	9 (47.4%)	0.51
Ice-pick scar	49	6 (42.9%)	3 (18.8%)	10 (52.6%)	0.11
Draining fistula	49	0 (0.0%)	1 (6.3%)	11 (57.9%)	**<0.001**
Non-draining fistula	49	2 (14.3%)	1 (6.3%)	12 (63.2%)	**<0.001**
Papule	49	2 (14.3%)	5 (31.3%)	6 (31.6%)	0.51
Folliculitis	49	1 (7.1%)	5 (31.3%)	7 (36.8%)	0.14
Continuous damage	49	8 (57.1%)	4 (25.0%)	15 (78.9%)	**0.006**
HS-PGA score	49	2.00 [2.00;3.00]	1.00 [0.00;1.00]	3.00 [2.00;3.00]	**<0.001**
Dissecting cellulitis	49	0 (0.0%)	0 (0.0%)	1 (5.3%)	>0.99
Folliculitis decalvans	49	0 (0.0%)	0 (0.0%)	2 (10.5%)	0.32
Keloidal acne	49	1 (7.1%)	0 (0.0%)	0 (0.0%)	0.29
Illustrative features					
KRT16	49	261 [91.4;392]	416 [317;464]	398 [321;612]	**0.027**
GPX2	49	61.1 [46.3;67.7]	30.5 [23.8;43.8]	66.2 [48.0;78.4]	**0.015**
CD3e	49	4.24 [2.35;7.0]	2.06 [1.04;2.50]	2.52 [1.60;5.2]	**0.016**

[1]Median [25%;75%]; n (%)

[2]Kruskal-Wallis rank sum test; Pearson's Chi-squared test; Fisher's exact test

Table 1. Clinical characteristic of the 49 HS patients from the FoI-HYDRA cohort.

circle whereby immune attack contributes to the disruption of HF homeostasis and HF-SCs quiescence.

The results presented here suggest that HF-SCs give rise to two distinct cell types, IBL and mORS. Our findings indicate that human IBL cells are enriched with stem cell markers, including *SOX9*, *LHX2*, and *FGF18*, which are important for maintaining quiescence [35]. Additionally, we observed an increased expression of keratin 6 and keratin 16 genes, with the K16-K6 keratin pair enhancing cellular integrity through improved cell-cell and cell-matrix interactions [36]. We postulate that the human IBL cells are most analogous to the inner bulge cells in mice. Murine CD34-K6+ inner bulge cells express high levels of inhibitory signals, such as FGF18 and bone morphogenetic protein 6 (BMP6), to maintain HF-SCs quiescence. Despite being quiescent, these cells actively contribute to the niche environment [24]. Furthermore, cadherin-mediated adherens junctions between K6+ inner bulge cells and bulge HF-SCs form a tightly sealed barrier, preventing

bacteria from entering the skin through the hair orifice. In HS patients, IBL cells display active differentiation characterized by keratinization, proliferation, and an inflammatory phenotype. We hypothesize that excessive IBL proliferation leads to follicular occlusion and plugging, resulting in cyst formation in HS patients with IBL enrichment. Consequently, we identified these patients as exhibiting a keratinization profile.

The mORS population are highly enriched for hallmarks of the immune response characterized by interferon gamma response. This observation is true for both healthy donors and HS patients and validated in a public dataset. Lay K and colleagues demonstrated that when HF-SCs identified damage, they initiated the process of multiplication and triggered an immune response. This prompts the stem cells to multiply and repair the barrier while regulating the inflammatory response [37]. On the other hand, skin stem cells that drive regeneration can themselves instruct an immune response by coordinating the recruitment of Treg cells and neutrophils during wound healing. Treg cells create a tolerogenic environment to support epithelial regeneration, while neutrophils induce inflammation to prevent infection [38]. We previously demonstrated that mORS cells from HS patients can secrete cytokines and chemokines, thereby recruiting immune cells to the site of inflammation [39]. In line with this, we observed that HS patients with enriched ORS showed T infiltration. Moreover, scRNA-Seq analysis revealed the presence of a highly proliferative HS-specific mORS subpopulation with an inflammatory profile. We propose that inflammatory mORS may modify the immune microenvironment. Overall, we anticipate that follicular occlusion in these cases is related to an inflammatory process, with these patients being identified as having an inflammatory profile.

Subsequently, a study was conducted on 49 patients with HS to determine whether HF cells composition is correlated with three different endotypes. The first phenotype identified a correlation between young non-smoking women with a high BMI, inflammatory nodules with the mild HSPGA score and a prevalence of mORS and T cells in HF cells. This suggests that inflammation may be a predominant factor in this demographic. This group may correspond to the regular phenotype defined by Van Der Zee [40] or LC1 defined by Canoui-Poitrine [41] or inflammatory pattern described by Martorell [42]. The second group comprised women smokers with a normal BMI and no inflammatory nodules, whose HF cells were found to be enriched in IBL, indicating that keratinization was the predominant process. This population may belong to the follicular pattern described by Martorell [42]. The last group in which both keratinization and inflammation are involved is made up of most of the smoking men with fistulae and continuous involvement. It should be noted, however, that some smoking women with a very high BMI also belong to this group. As a result, it is challenging to relate this group to the classic phenotype described. Of the six patients with a mixed phenotype, four belong to this group. Consequently, these patients will require treatment strategies that target both the IBL and mORS cell populations. Overall, our results suggest that follicular hyperplasia in HS may arise from independent mechanisms: inflammation, keratinization, or both. This heterogeneity poses a challenge for treatment and suggests that a tailored therapeutic approach may improve patient management.

Our study is not without any limitations. One such limitation is that we were unable to carry out functional studies due to the significant challenges associated with maintaining either IBL or ORS cells population in culture following cell sorting. The reason why some patients are enriched in mORS and others in IBL remains unanswered and will require further investigation.

The substantial clinical heterogeneity observed among the various clusters within the cohort makes it unlikely that inflammation precedes keratinization, or vice versa. However, it cannot be excluded that patients in group C represent the most severe cases from groups A or B, having progressed to a mixed phenotype. It would therefore be interesting to follow patients over time to determine whether they can evolve from one phenotype to another. Furthermore, it would be highly beneficial to study patients' response to treatment as a function of their pathophysiological profile.

In conclusion, our results suggest that distinct pathophysiological mechanisms contribute to the development of HS. The composition of hair follicle cells provides a promising method for selecting or tailoring therapeutic regimens, allowing for personalized treatment strategies based on whether a patient is identified as belonging to the inflammatory, keratinizing, or mixed categories.

Authors contribution

Conceptualization (AO, YL, PT, EA, SH); Data Acquisition (FJL, PLC, FC, RA, CBo, SA, ES, CBe); Analysis (AO, FJL, FC, EA); Interpretation (AO, FJL, EA, SH); Visualization (AO, EA, SH); Writing original draft (AO, SH); Writing and review (AO, FJL, PLC, ES, PW, VG, PT, EA, SH); Project Administration (PW, VG, YL, PT, EA, SH); Funding Acquisition (SH).

Acknowledgments

This work was supported by the Institut National de la Santé et de la Recherche Médicale (INSERM), the University Paris Est Créteil (UPEC). It has received financial support from French ”Agence Nationale de la Recherche” (ANR) under project ANR-20-CE17-0019 and ”Société Française de Dermatologie et de Pathologie Sexuellement Transmissible”.

All authors have disclosed any financial or personal relationships with organizations that could influence the results. We are grateful to the IMRB genomic facility. We thank the staff from FACS for technical assistance.

References

1. Łukasz Matusiak. Profound consequences of hidradenitis suppurativa: a review. *British Journal of Dermatology*, 183(6), 12 2020.
2. Christos C. Zouboulis, Farida Benhadou, Angel S. Byrd, Nisha S. Chandran, Evangelos J. Giamarellos-Bourboulis, Gabriella Fabbrocini, John W. Frew, Hideki Fujita, Marcos A. González-López, Philippe Guillem, Wayne P. F. Gulliver, Iltefat Hamzavi, Yildiz Hayran, Barbara Hórvath, Sophie Hüe, Robert E. Hunger, John R. Ingram, Gregor B.E. Jemec, Qiang Ju, Alexa B. Kimball, Joslyn S. Kirby, Maria P. Konstantinou, Michelle A. Lowes, Amanda S. MacLeod, Antonio Martorell, Angelo V. Marzano, Łukasz Matusiak, Aude Nassif, Elena Nikiphorou, Georgios Nikolakis, André Nogueira da Costa, Martin M. Okun, Lauren A.V. Orenstein, José Carlos Pascual, Ralf Paus, Benjamin Perin, Errol P. Prens, Till A. Röhn, Andrea Szegedi, Jacek C. Szepietowski, Thrasyvoulos Tzellos, Baoxi Wang, and Hessel H. van der Zee. What causes hidradenitis suppurativa ?-15 years after. *Experimental Dermatology*, 29(12):1154–1170, 12 2020.
3. Toshifumi Nomura. Hidradenitis Suppurativa as a Potential Subtype of Autoinflammatory Keratinization Disease. *Frontiers in Immunology*, 11, 5 2020.
4. Cindy Orvain, Yea-Lih Lin, Francette Jean-Louis, Hakim Hocini, Barbara Hersant, Yamina Bennasser, Nicolas Ortonne, Claire Hotz, Pierre Wolkenstein, Michele Boniotto, Pascaline Tisserand, Cécile Lefebvre, Jean-Daniel Lelièvre, Monsef Benkirane, Philippe Pasero, Yves Lévy, and Sophie Hüe. Hair follicle stem cell replication stress drives IFI16/STING-dependent inflammation in hidradenitis suppurativa. *Journal of Clinical Investigation*, 130(7):3777–3790, 6 2020.
5. John W. Frew. Hidradenitis suppurativa is an autoinflammatory keratinization disease: A review of the clinical, histologic, and molecular evidence. *JAAD International*, 1(1):62–72, 7 2020.
6. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902, 6 2019.
7. Pierre-Luc Germain, Aaron Lun, Carlos Garcia Meixide, Will Macnair, and Mark D. Robinson. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Research*, 10:979, 5 2022.
8. Abha S. Bais and Dennis Kostka. scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics*, 36(4):1150–1158, 2 2020.
9. Sijie Wu, Yao Yu, Caiyue Liu, Xia Zhang, Peiying Zhu, You Peng, Xinyu Yan, Yin Li, Peng Hua, Qingfeng Li, Sijia Wang, and Liang Zhang. Single-cell transcriptomics reveals lineage trajectory of human scalp hair follicle and informs mechanisms of hair graying. *Cell Discovery*, 8(1):49, 5 2022.
10. Rie Takahashi, Adrienne Grzenda, Thomas F. Allison, Jeffrey Rawnsley, Samuel J. Balin, Shan Sabri, Kathrin Plath, and William E. Lowry. Defining Transcriptional Signatures of Human Hair Follicle Cell States. *Journal of Investigative Dermatology*, 140(4):764–773, 4 2020.

11. Simon Joost, Karl Annusver, Tina Jacob, Xiaoyan Sun, Tim Dalessandri, Unnikrishnan Sivan, Inês Sequeira, Rickard Sandberg, and Maria Kasper. The Molecular Anatomy of Mouse Skin during Hair Growth and Rest. *Cell Stem Cell*, 26(3):441–457, 3 2020.
12. Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 12 2019.
13. Philipp Angerer, Laleh Haghverdi, Maren Büttner, Fabian J. Theis, Carsten Marr, and Florian Buettner. destiny : diffusion maps for large-scale single-cell data in R. *Bioinformatics*, 32(8):1241–1243, 4 2016.
14. Kelly Street, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1):477, 12 2018.
15. Helena Todorov, Robrecht Cannoodt, Wouter Saelens, and Yvan Saeys. TinGa: fast and flexible trajectory inference with Growing Neural Gas. *Bioinformatics*, 36(Supplement_1):i66–i74, 7 2020.
16. Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, 9 2016.
17. Igor Dolgalev. *msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format*, 2022. R package version 7.5.1.9001.
18. Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, 5 2012.
19. Song Gao, Stefan Mutter, Aaron Casey, and Ville-Petteri Mäkinen. Numero: a statistical framework to define multivariable subgroups in complex population-based datasets. *International Journal of Epidemiology*, 48(2):369–374, 4 2019.
20. Ellen Witte-Händel, Kerstin Wolk, Athanasia Tsaousi, Marie Luise Irmer, Rotraut Mößner, Orr Shomroni, Thomas Lingner, Katrin Witte, Desiree Kunkel, Gabriela Salinas, Stefan Jodl, Nicole Schmidt, Wolfram Sterry, Hans-Dieter Volk, Evangelos J. Giamarellos-Bourboulis, Anna Pokrywka, Wolf-Dietrich Döcke, Sylke Schneider-Burrus, and Robert Sabat. The IL-1 Pathway Is Hyperactive in Hidradenitis Suppurativa and Contributes to Skin Infiltration and Destruction. *Journal of Investigative Dermatology*, 139(6):1294–1305, 6 2019.
21. Margaret M. Lowe, Haley B. Naik, Sean Clancy, Mariela Pauli, Kathleen M. Smith, Yingtao Bi, Robert Dunstan, Johann E. Gudjonsson, Maia Paul, Hobart Harris, Esther Kim, Uk Sok Shin, Richard Ahn, Wilson Liao, Scott L. Hansen, and Michael D. Rosenblum. Immunopathogenesis of hidradenitis suppurativa and response to anti–TNF-α therapy. *JCI Insight*, 5(19), 10 2020.
22. Laiqing Hua, Shuyu Yao, Duy Pham, Li Jiang, Jeffrey Wright, Deepali Sawant, Alexander L. Dent, Thomas J. Braciale, Mark H. Kaplan, and Jie Sun. Cytokine-Dependent Induction of CD4+ T cells with Cytotoxic Potential during Influenza Virus Infection. *Journal of Virology*, 87(21):11884–11893, 11 2013.
23. Shiri Gur-Cohen, Hanseul Yang, Sanjeethan C. Baksh, Yuxuan Miao, John Levorse, Raghu P. Kataru, Xiaolei Liu, June de la Cruz-Racelis, Babak J. Mehrara, and Elaine Fuchs. Stem cell–driven lymphatic remodeling coordinates tissue regeneration. *Science*, 366(6470):1218–1225, 12 2019.
24. Ya-Chieh Hsu and Elaine Fuchs. Building and Maintaining the Skin. *Cold Spring Harbor Perspectives in Biology*, 14(7):a040840, 7 2022.
25. Elaine Fuchs. The Tortoise and the Hair: Slow-Cycling Cells in the Stem Cell Race. *Cell*, 137(5):811–819, 5 2009.
26. Gaëlle Gendronneau, Sadaf Sanii, Tien Dang, Frédérique Deshayes, Delphine Delacour, Evelyne Pichard, Tamara Advedissian, Sukhvinder S. Sidhu, Mireille Viguier, Thierry Magnaldo, and Francoise Poirier. Overexpression of Galectin-7 in Mouse Epidermis Leads to Loss of Cell Junctions and Defective Skin Repair. *PLOS ONE*, 10(3):e0119031, 3 2015.
27. Qian Li, Fang Li, Xinqiu Song, Ning Lu, Xintao Jing, Hua Wen, Peihan Ma, Hua Zhang, Wenzhu Yao, Xiaofei Wang, and Mingxin Zhang. Pan-cancer analysis of ARFs family and ARF5 promoted the progression of hepatocellular carcinoma. *Heliyon*, 10(7):e29099, 4 2024.
28. Teresa Preglej and Wilfried Ellmeier. CD4+ Cytotoxic T cells – Phenotype, Function and Transcriptional Networks Controlling Their Differentiation Pathways. *Immunology Letters*, 247:27–42, 7 2022.
29. Priyadharshini Devarajan, Allen M. Vong, Catherine H. Castonguay, Noah J. Silverstein, Olivia Kugler-Umana, Bianca L. Bautista, Karen A. Kelly, Jeremy Luban, and Susan L. Swain. Cytotoxic CD4 development requires CD4 effectors to concurrently recognize local antigen and encounter type I IFN-induced IL-15. *Cell Reports*,

42(10):113182, 10 2023.

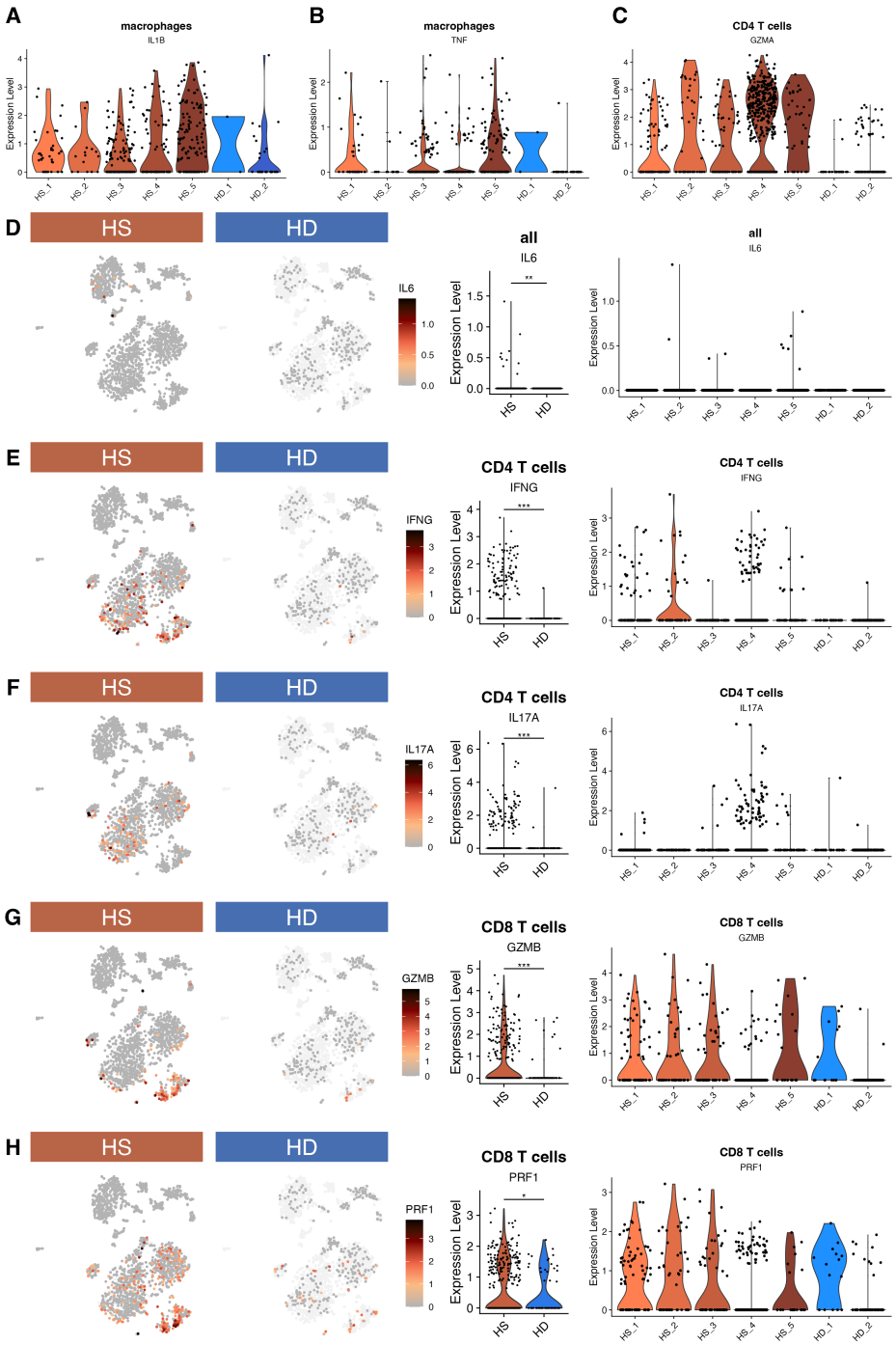
30. Johann E. Gudjonsson, Lam C. Tsoi, Feiyang Ma, Allison C. Billi, K.R. van Straalen, A.R.J.V. Vossen, H.H. van der Zee, Paul W. Harms, Rachael Wasikowski, Christine M. Yee, Syed M. Rizvi, Xianying Xing, Enze Xing, Olesya Plazyo, Chang Zeng, Matthew T. Patrick, Margaret M. Lowe, Richard E. Burney, Jeffrey H. Kozlow, Jill R. Cherry-Bukowiec, Yanyun Jiang, Joseph Kirma, Stephan Weidinger, Kelly C. Cushing, Michael D. Rosenblum, Celine Berthier, Amanda S. MacLeod, John J. Voorhees, Fei Wen, J. Michelle Kahlenberg, Emanual Maverakis, Robert L. Modlin, and Errol P. Prens. Contribution of plasma cells and B cells to hidradenitis suppurativa pathogenesis. *JCI Insight*, 5(19), 10 2020.
31. Etienne C.E. Wang, Zhenpeng Dai, Anthony W. Ferrante, Charles G. Drake, and Angela M. Christiano. A Subset of TREM2+ Dermal Macrophages Secretes Oncostatin M to Maintain Hair Follicle Stem Cell Quiescence and Inhibit Hair Growth. *Cell Stem Cell*, 24(4):654–669, 4 2019.
32. Chih-Chiang Chen, Lei Wang, Maksim V. Plikus, Ting Xin Jiang, Philip J. Murray, Raul Ramos, Christian F. Guerrero-Juarez, Michael W. Hughes, Oscar K. Lee, Songtao Shi, Randall B. Widelitz, Arthur D. Lander, and Cheng Ming Chuong. Organ-Level Quorum Sensing Directs Regeneration in Hair Stem Cell Populations. *Cell*, 161(2):277–290, 4 2015.
33. Judith Agudo, Eun Sook Park, Samuel A. Rose, Eziwoma Alibo, Robert Sweeney, Maxime Dhainaut, Koichi S. Kobayashi, Ravi Sachidanandam, Alessia Baccarini, Miriam Merad, and Brian D. Brown. Quiescent Tissue Stem Cells Evade Immune Surveillance. *Immunity*, 48(2):271–285, 2 2018.
34. Ralf Paus, Natsuho Ito, Masahiro Takigawa, and Taisuke Ito. The Hair Follicle and Immune Privilege. *Journal of Investigative Dermatology Symposium Proceedings*, 8(2):188–194, 10 2003.
35. Ya-Chieh Hsu, H. Amalia Pasolli, and Elaine Fuchs. Dynamics between Stem Cells, Niche, and Progeny in the Hair Follicle. *Cell*, 144(1):92–105, 1 2011.
36. Fengrong Wang, Song Chen, Hans B. Liu, Carole A. Parent, and Pierre A. Coulombe. Keratin 6 regulates collective keratinocyte migration by altering cell–cell and cell–matrix adhesion. *Journal of Cell Biology*, 217(12):4314–4330, 12 2018.
37. Kenneth Lay, Shaopeng Yuan, Shiri Gur-Cohen, Yuxuan Miao, Tianxiao Han, Shruti Naik, H. Amalia Pasolli, Samantha B Larsen, and Elaine Fuchs. Stem cells repurpose proliferation to contain a breach in their niche barrier. *eLife*, 7, 12 2018.
38. Jingyun Luan, Cynthia Truong, Aleksandra Vuchkovska, Weijie Guo, Jennifer Good, Bijun Liu, Audrey Gang, Nicole Infarinato, Katherine Stewart, Lisa Polak, Hilda Amalia Pasolli, Emma Andretta, Alexander Y. Rudensky, Elaine Fuchs, and Yuxuan Miao. CD80 on skin stem cells promotes local expansion of regulatory T cells upon injury to orchestrate repair within an inflammatory environment. *Immunity*, 57(5):1071–1086, 5 2024.
39. Claire Hotz, Michele Boniotto, Aurélie Guguin, Mathieu Surenaud, Francette Jean-Louis, Pascaline Tisserand, Nicolas Ortonne, Barbara Hersant, Romain Bosc, Florence Poli, Henri Bonnabau, Rodolphe Thiébaut, Véronique Godot, Pierre Wolkenstein, Hakim Hocini, Yves Lévy, and Sophie Hüe. Intrinsic Defect in Keratinocyte Function Leads to Inflammation in Hidradenitis Suppurativa. *Journal of Investigative Dermatology*, 136(9):1768–1780, 9 2016.
40. Hessel H. van der Zee and Gregor B.E. Jemec. New insights into the diagnosis of hidradenitis suppurativa: Clinical presentations and phenotypes. *Journal of the American Academy of Dermatology*, 73(5):S23–S26, 11 2015.
41. Florence Canoui-Poitrine, Aurélie Le Thuaut, Jean E. Revuz, Cédric Viallette, Germaine Gabison, Florence Poli, Florence Pouget, Pierre Wolkenstein, and Sylvie Bastuji-Garin. Identification of Three Hidradenitis Suppurativa Phenotypes: Latent Class Analysis of a Cross-Sectional Study. *Journal of Investigative Dermatology*, 133(6):1506–1511, 6 2013.
42. A. Martorell, A. Jfri, S.B.L. Koster, P. Gomez-Palencia, M. Solera, A. Alfaro-Rubio, L. Hueso, and V. Sanz-Motilva. Defining hidradenitis suppurativa phenotypes based on the elementary lesion pattern: results of a prospective study. *Journal of the European Academy of Dermatology and Venereology*, 34(6):1309–1318, 6 2020.

Supplementary tables and figures

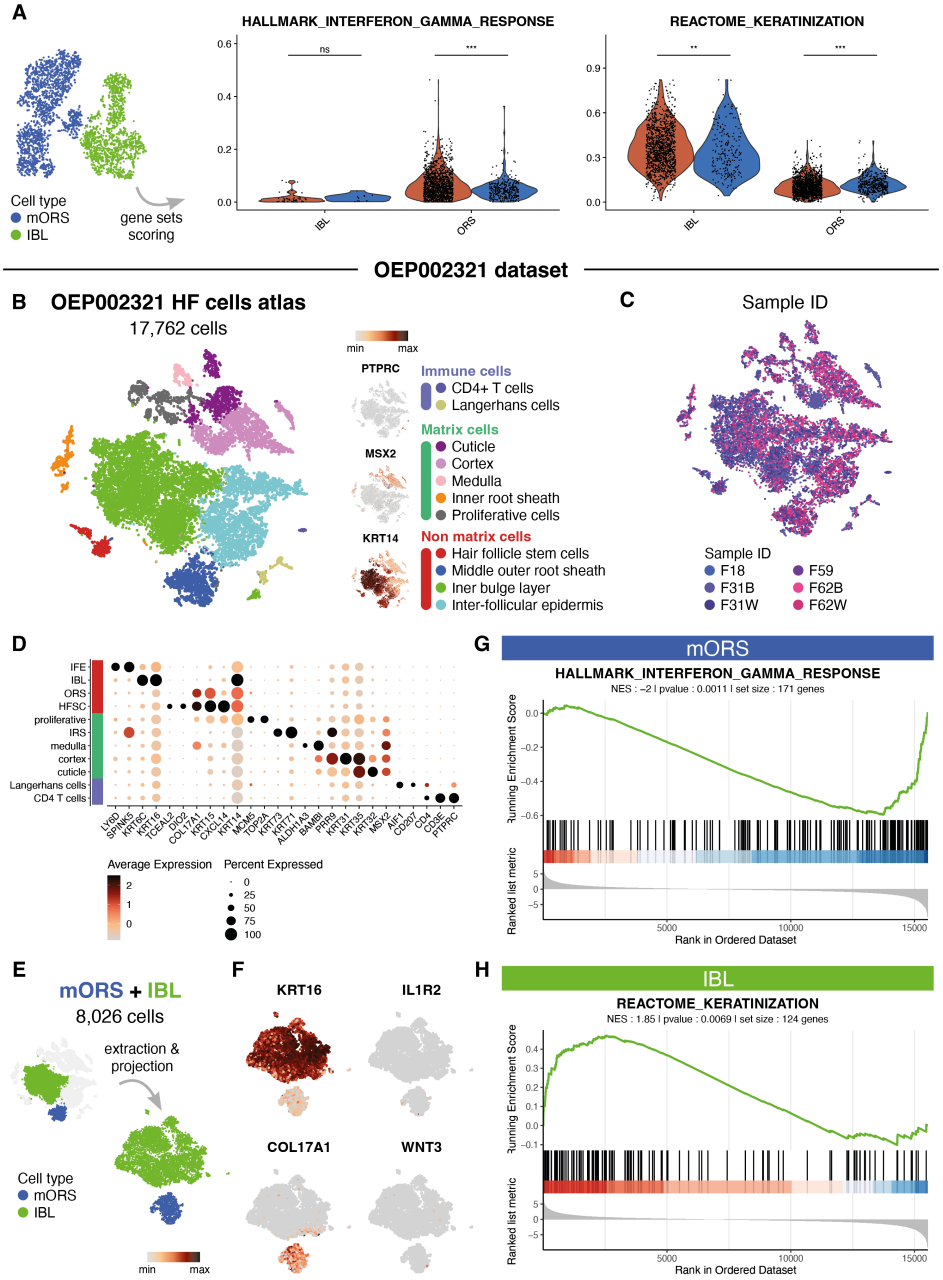
Supplementary Table 1. Clinical characteristics of patients (n=7) for scRNA-Seq analysis.

Supplementary Table 2. Version of R packages for the analysis of scRNA-Seq data.

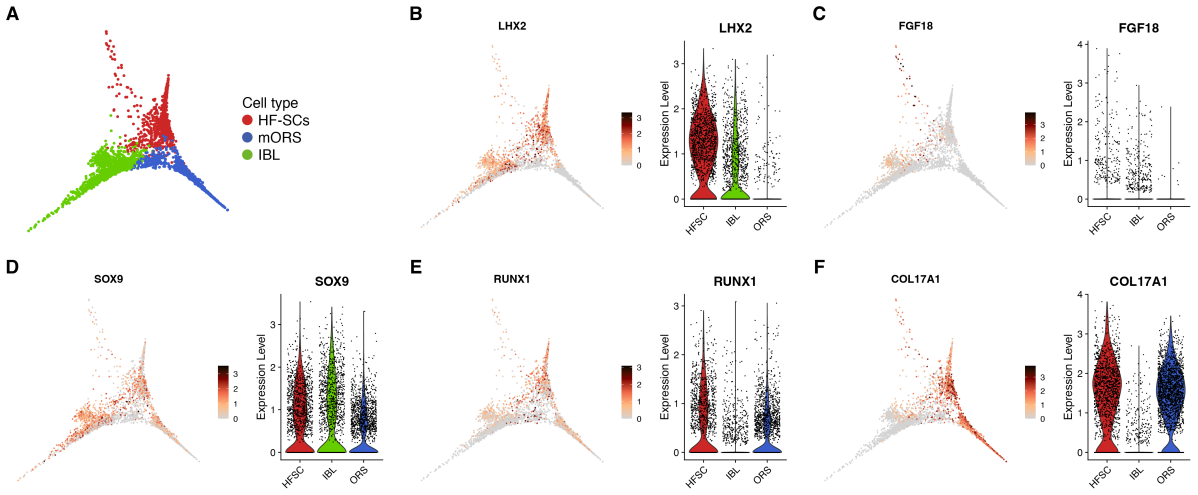
Supplementary Table 3. Gene markers for single-cell level cell type annotation of scRNA-Seq data.



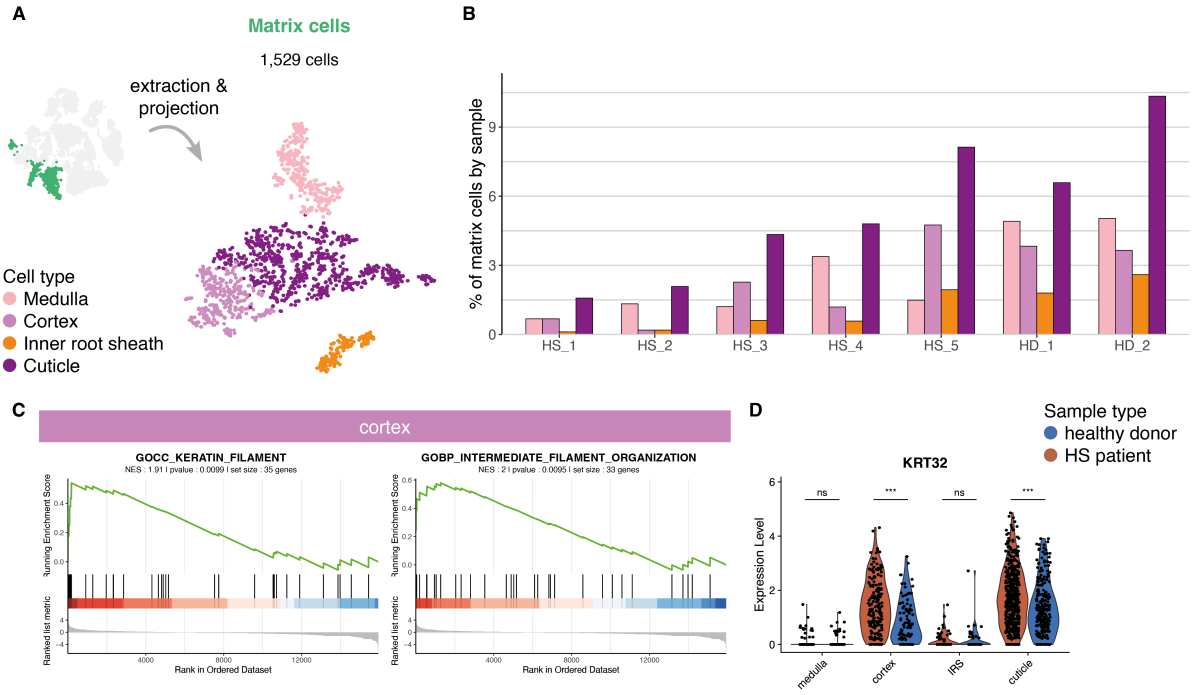
Supplementary Figure 1. Hair follicle immune cells lack IL6 expression in both HD and HS patients. (A-C) Violin plots showing three genes from the main heatmaps (**Figure 2B,D**), split by sample identifier. Each panel only considers the population indicated in the subtitle. (D-H) Feature plots and violin plots showing gene expression levels, split by sample type or sample identifier. On the tSNE, cells from complete immune cells dataset are shown as light background. The violin plots consider the T cells population indicated in the subtitle. Statistical t-test p-value, ns: > 0.05, *: < 0.05 and > 0.01, **: < 0.01 and > 0.001 and ***: < 0.001.



Supplementary Figure 2. IBL and mORS depict distinct biological activities. (A) Violin plots showing single cell level scoring for the expression levels of all genes from two gene sets, split by IBL and mORS, and by sample type. Each population has a specific signature, which shows significant statistical variations between sample type. Statistical t-test p-value, ns: > 0.05, *: < 0.05 and > 0.01, **: < 0.01 and > 0.001 and ***: < 0.001. (B-C) tSNE plot of 17,762 cells from the OEP002321 dataset, colored based on cell type annotation (B) or sample identifier (C). (D) Dot plot assessing cell type annotation in the OEP002321 dataset. Dot size represents the percentage of cells with positive expression for each marker. Dot color represents the scaled expression level of markers across all cells within each cell type. (E-F) tSNE plot showing 8,026 IBL and mORS cells from the OEP002321 dataset, colored based on (E) cell type annotation or (F) expression level of representative markers is depicted. (G-H) GSEA plots made from average fold change of all genes, between IBL and mORS, in the OEP002321 dataset.



Supplementary Figure 3. Expression of selected markers on the hair follicle atlas. (A) Diffusion map of HF-SCs, IBL and mORS, colored based on cell type annotation. (B-F) Feature plots and violin plots showing expression levels of selected genes. Violin plots are split by cell types.



Supplementary Figure 4. Inner bulge layer cells (IBL) and middle outer root sheath cells (mORS) represent two branches of differentiation from HF-SCs, with distinct transcriptomic activities. (A) tSNE plot showing 1,529 matrix cells extracted from the main atlas (**Figure 1**). Cells are colored based on cell type annotation. (B) Barplot depicting the proportion of each matrix cell population among all cells in individual samples. Legend is shared with (A). (C) GSEA plots made from average fold change of all genes, in the cortex, between HS patients and HD. (D) Violin plot showing *KRT32* expression level, split by matrix population and by sample type. Note the enrichment in the cortex and cuticle in HS patients compared to HD. Statistical t-test p-value, ns: > 0.05, *: < 0.05 and > 0.01, **: < 0.01 and > 0.001 and ***: < 0.001.

4.3 Outlook

As for the Article 1 and Article 2, a Github repository is attached to the manuscript, with a very similar layout as shown in Section 2.3.1. It contains all the R Markdown notebooks, and their compiled HTML versions, used to process the raw count matrices and obtain the final figures.

From a computational point of view, this story validated the analysis framework (Section 1.4). This framework, which consists of a directory tree, a methodology for performing the analyses, and a Singularity container for compiling the notebooks, was conceived in a generic way and could be applied to a new project. The R Markdown notebooks can be adapted to a new experimental design with few changes.

The most time consuming part of this project was the **cell type annotation**. The methodology consisted of clustering cells and performing differential expression analysis between clusters. The differentially expressed genes were then compared to the literature. In particular, two publications were considered (Joost *et al.*, 2020; Wu *et al.*, 2022b). In Joost *et al.* (2020), scRNA-Seq experiments were performed on mouse skin. This sample is broader and the dataset is more diverse compared to human hair follicles,. Thus, the representativeness of the cell populations was largely different from our data. However, the publication presents robust marker genes that were helpful in annotating our cells. This publication was also a source of inspiration for the design of our figures. Although the experiments of Wu *et al.* (2022b) were performed on human hair follicles, the marker genes were not robust enough to be applied to our study. However, reanalysis of their data validated our set of marker genes for cell type annotation (Article 3, Figure S2).

This version of the manuscript has been shared to all the authors for comments. With the last author, we will consider the suggestions from our colleagues to update the text and the figures, before submission. The sequencing data (FASTQ files and count matrices) further need to be uploaded on a public repository.

V Discussion

1 Summary of the contributions

The purpose of my PhD thesis was to decipher the cellular and molecular mechanisms driving the initiation of plexiform neurofibromas and the malignant transformation, through reproducible analysis of scRNA-Seq data from our mouse model of NF1 disease. This work was part of a team work, including a computational contribution from a Master student, who left the laboratory before my arrival.

1.1 Computational contributions

For the processing of the sequencing files to count matrix, the pre-existing Python workflow was converted into a Nextflow workflow. It enables more readability and modularity. Attaching a Singularity container to all processes further allowed the portability of the project using the same software environment. The reproducibility of this processing step was ensured by running it twice, using the same input data, on two distinct high performance computing clusters. This processing step is documented and the documentation is used by a colleague to apply the pipeline on her original data.

Starting from the count matrix, the pre-existing pipeline focusing on the processing of the individual dataset was substantially improved. The improvement includes (i) a better traceability, readability and shareability using computational notebooks, (ii) the documentation of all custom and wrapper functions into a package, called aquarius, and versioned through Git, (iii) the management of all other packages of interest and their dependencies in a Singularity container, and (iv) the establishment of a standardized directory tree to manage the input data and the analyses. All these aspects ensure the reproducibility of the computational treatment of our single-cell transcriptomic data.

In order to address distinct biological questions, the notebook to process individual datasets was enriched by many others. The most accomplished include the study of the heterogeneity of a specific cell population through many datasets, differential gene expression, functional enrichment and trajectory inference. Other notebooks are still under development. They focus on the biological pathway and transcription factor activity inferences, RNA velocity, and inference of CNV on scRNA-Seq data.

To avoid any repetition of the same computational steps and standardize the methodology to answer our questions, a notebook tree was set up. This notebook structure is compatible with the directory tree and is applicable to distinct projects. Additionally, the aquarius package was enriched with new functions, mainly wrappers

of existing tools. Among original functions, our tool for cell type annotation is discussed later (Section 2.2.1). Finally, the compiled notebooks associated with scientific publications are embedded in a main web page and shared through the Code Availability section of manuscripts, thereby complying with the open science and reproducible research precepts.

The resulting framework enables an analysis of scRNA-Seq data in a ready-to-publish way, and to be more efficient while exploring data. Although it was set up in the context of NF1, it was successfully applied to another study (Article 3).

1.2 Biological contributions

In the Article 1, we showed that both *Nf1*-KO^{wt} and *Nf1*-KO^{het} mice develop thicker nerves compared to control mice. They are at the origin of pNFs. The first events preceding pNFs formation are the expansion of pro-fibrotic *Nf1*$^{-/-}$ Schwann cells, the progression of endoneurial fibroblasts toward an inflammatory profile, and the infiltration of endoneurial macrophages. In the context of skin trauma or haploinsufficiency of *Nf1* in the micro-environment in three month-old mice, a stronger phenotype was observed. These mice developed pNFs faster than mutant mice from other conditions.

We made a cellular and molecular characterization of mice subcutaneous nerves, in various conditions, through single-cell transcriptomics. The resulting atlas can serve as a resource for the community. We identified the first events preceding the formation of pNFs, which can be further investigated for early diagnosis or to develop preventive therapeutic options.

In the Article 2, we showed that both *Nf1*-KO^{wt} and *Nf1*-KO^{het} mice spontaneously develop MPNST from benign tumors. As in NF1 patients, the malignant transformation is characterized by a glial-to-mesenchymal transition and by genetic events, including the loss of copies of *Cdkn2a*. The over-activation of Ras signaling pathways is responsible for the sustained expression of *Sox9* in MPNSTs. Consequently, we suggested to target Ras signaling to limit the progression of MPNSTs.

Our mouse model is only altered for *Nf1* and is the first to recapitulate the malignant transformation of neurofibromas with an incidence of 70%. It is a powerful tool to characterize the early events driving malignant transformation, and identify and develop new therapeutic targets.

In the following sections, I first discuss elements relative to the computational aspects (Section 2) before focusing on the biological ones (Section 3).

2 Discussion on the computational part

This chapter starts with remarks about the analysis of scRNA-Seq data (Section 2.1). Then, the perspectives about the methods are discussed (Section 2.2). Although our framework enables the reproducibility of the analyses, it is not yet reusable by others (Section 2.3). The chapter ends with an outlook about making research reproducible in computational biology (Section 2.4).

2.1 Hindsight regarding the analysis of scRNA-Seq data

2.1.1 Biological relevance

Each scRNA-Seq dataset captures a snapshot of a biological structure at a specific time, yet it is prone to selection biases during the cell dissociation process. Depending on the technology used, labeling transcripts by their cell of origin can result in low coverage of individual transcriptomic landscapes. It thus limits the interpretation of cell abundance in the data. Additionally, the absence of gene counts may indicate either lack of expression or missing data. Several solutions address these challenges. Spatial transcriptomics offers insights into cellular diversity within tissue sections and cell-cell interactions (Moses & Pachter, 2022). The Smart-Seq2 protocol enhances the gene detection per cell compared to droplet-based methods (Picelli *et al.*, 2013). When datasets exhibit high drop-out rates, computational methods like MAGIC can impute missing gene expression values effectively (Dijk *et al.*, 2018).

2.1.2 Variability of the discrete cell clustering

In the Article 3, a clustering was used to compare relative abundances across datasets. However, the clustering of single cells involves partitioning a distance matrix between cells, which is sensitive to input parameters, such as the resolution. In scRNA-Seq datasets, cellular abundances may be biased, but comparisons between datasets obtained with consistent procedures are feasible. To bypass the clustering, differential abundance can be evaluating using k-nearest neighbor graphs (Dann *et al.*, 2022).

Similarly, differential expression analysis often relies on cluster-based comparisons, impacting the list of differentially expressed genes and subsequent analyses. Different clustering resolutions may overlook rare populations with specific profiles. Recent preprints have introduced cluster-free methods for differential expression (Missarova *et al.*, 2023; Ahlmann-Eltze & Huber, 2024), which could be valuable for studying population heterogeneity without manual adjustment.

2.1.3 Trajectory inference as a continuous clustering

Trajectory inference is not a temporal analysis. Similarly to the discrete cell clustering, trajectory inference evaluates relations between cells, but defines a **continuous clustering**, called pseudotime. In the Article 2 and the Article 3, they were used to define an order between cell populations, which is already known from the literature. Discovering new biological successive events would require additional validations.

2.1.4 Limits of the gene sets enrichment analyses

Enrichment analyses between cell populations in scRNA-Seq suffer from the high drop-out rates. Typically, gene set enrichment analysis (GSEA) is applied in bulk RNA-Seq using specific values from differentially expressed (DE) genes. It has been demonstrated that GSEA can be effectively applied in scRNA-Seq by using the fold changes of all genes as input (Holland *et al.*, 2020). During my thesis, I compared GSEA results using fold changes of **all genes versus DE genes** only. Using DE genes often resulted in few enriched gene sets, leading to numerous false negatives. These false negatives may further result from the drop-out events that limit the number of significantly DE genes. Conversely, using all genes yielded many gene sets with a higher incidence of false positives. For instance, non-specific genes, such as *Jun*, *Egr1*, and *Dusp**, are shared by multiple unrelated gene sets. These genes were thus not considered when evaluating the consistency of enriched gene sets.

The issue with **non-specific genes** underscores a fundamental limitation of gene set analyses (Cantini *et al.*, 2017). In current approaches, all genes within a gene set are treated equally. They do not take into account that some genes encode hub proteins involved in multiple pathways, while others are more specialized. The PROGENy database addresses this by assigning genes related to signaling pathways with scores and p-values, enabling their prioritization (Schubert *et al.*, 2018). In the MSigDB, the Hallmark collection consists in robust and non-redundant gene sets. To enhance the MSigDB, one improvement would be to annotate genes that are shared across numerous unrelated gene sets, considering them with reduced importance during enrichment analysis. The InfoSigMap has been designed to visually evaluate the redundancy in the gene sets from collections, such as MSigDB (Cantini *et al.*, 2017). But, no straightforward way enables to treat differentially the redundant genes while running enrichment analyses.

Despite these challenges, it is generally preferable to use all genes rather than only DE genes when investigating gene set enrichment in scRNA-Seq data.

2.2 Perspectives about the methods

2.2.1 Benefits and future directions of the cell type annotation method

One of the most time-consuming steps in analyzing scRNA-Seq data is the cell type annotation, often tailored to the specific study context. Despite numerous methods and databases offering cell type-specific signatures, few studies compare these methods to help users select the most suitable one for their projects (Huang *et al.*, 2021). Many authors manually annotate cell clusters based on the expression of curated markers, while others develop new annotation methods. Our approach, similar to recent proposals (Mikolajewicz *et al.*, 2022), utilizes the Seurat::AddModuleScore function with a custom set of markers.

Marker sets were identified through differential expression analyses across datasets, manually refined to minimize annotation issues. This exploratory approach, though effective, is labor-intensive, repetitive, and challenging to automate or trace. Alternatively, **automatic methods** for identifying cell type-specific markers from scRNA-Seq data (Aevermann *et al.*, 2021; Cortal *et al.*, 2021; Dumitrascu *et al.*, 2021) hold promises for improving traceability and efficiency in marker identification.

In our method, each cell receives scores for positive and negative marker sets, facilitating cell type annotation. This scoring approach, leveraging negative markers, is particularly useful for distinguishing closely related cell types like mesenchymal tumor cells and fibroblasts. Unlike cluster-based methods, our approach is cluster-free. It allows the annotation of rare cells that are not assigned to specific clusters and facilitates seamless integration across datasets without re-running the cell annotation tool. However, to correct potential single-cell mis-annotations, a smoothing over clusters may be required. This lacks robustness when the clustering parameters change but ensures clear cell type representation in atlas projections.

Currently, our method annotates all the cells in a dataset. Enabling a "**not assigned**" annotation could enhance the method, aligning with approaches that include this feature (Huang *et al.*, 2021).

A future direction could involve developing an **annotation tree structure**, organizing cell types hierarchically based on nested lists of specific markers. This approach could better manage annotation competition and be adaptable to new cell types without modifying the existing marker sets extensively. For example, distinguishing various T cell subpopulations could be achieved by nesting markers hierarchically, starting broadly with markers for all T cells (*Cd3*) and refining annotations progressively with specific

markers for cytotoxic, regulatory, and helper T cells as needed. Such hierarchical annotation trees offer flexibility in addressing diverse biological questions.

2.2.2 Methodological improvement

The harmony package effectively removed **batch-effects** in datasets containing distinct cell populations, using PCA based on log-normalized, scaled expression levels of highly variable genes (HVGs). It performed well on scRNA-Seq data from hair follicles and PNSTs from NF1 mice, yielded good results on NF1 patient data but poorer outcomes for two single-nuclei datasets from NF1 mice (data not shown).

In the original publication of harmony, authors recommended using the SCTransform method for identifying HVGs, normalizing and scaling gene expression levels (Hafemeister & Satija, 2019; Korsunsky *et al.*, 2019). However, SCTransform was not adopted due to its restriction to HVGs. Applying SCTransform on the whole count matrix is not recommended, and is further highly time and memory consuming. Of note, the normalization of the whole count matrix is necessary for differential expression and gene expression visualization.

A future direction could involve evaluating the substitution of log-normalization with SCTransform for batch-effect correction across our datasets. Additionally, enhancing the notebook for datasets integration could involve retaining log-normalization solely for analysis purposes and SCTransform for visualization purposes. This approach aims to optimize both batch-effect correction and downstream analyses efficiently.

Removing the batch-effect on the (normalized) count matrix rather than on the PCA may be another alternative.

We alternatively used Seurat::AddModuleScore or the AUCell package to attribute **gene sets scores** to each individual cell (Street *et al.*, 2018; Aibar *et al.*, 2017) and visually evaluated their distribution on the cell projection (UMAP or tSNE). Many tools enable such a scoring (Zhang *et al.*, 2020b). A recent method enables to attribute a statistical significance to the individual scores, which would be useful to compare cells (Najm *et al.*, 2024). Uniquely, this method can identify shifted gene set scores. This is valuable to automatically investigate many gene sets, without individually assessing the distribution of their activity score on the cell projection.

2.2.3 The ideal tool is not always the one being used

scRNA-Seq data offer vast analytical possibilities with numerous available methods and benchmarking studies comparing them. The ideal tool for any analysis should be

highly robust and provide biologically meaningful results. However, the tool chosen is often dictated by factors like documentation, ease of installation and use, maintenance, and community support, rather than solely by its pure analytical performance.

2.3 Missing documentation for re-usability

2.3.1 Limitations of the aquarius package

The aquarius package was developed at the beginning of my thesis to automate the use of documented functions and manage their evolution through versioning. While some functions originated from a Master student internship, they were documented later. Currently, the package serves as a methodological toolkit for performing custom analyses and avoid redundant code sections in the notebook. It is not yet suitable for broad use due to two main limitations: documentation and dependencies.

A website was built to ensure clear definitions of the package and comprehensive documentation of its code and function parameters. However, there is a **lack of documentation** on practical application scenarios, which could be addressed through notebook templates providing concrete examples (see Section 2.3.2).

Despite containing numerous functions, many were ultimately not used because they lacked generalization or did not significantly add to existing functions and methods. To improve and lighten the package, the functions should be made more flexible in handling input data and customizable parameter values. Additionally, functions that serve as mere wrappers of existing ones, with predefined parameter values, should be annotated as deprecated and removed for clarity.

Due to diverse functionalities, aquarius relies on a considerable **number of dependencies**, complicating its installation. To mitigate this, the package could be modularized into sub-packages with specific purposes, or differentiate between essential and non-essential dependencies that are only required for specific functions.

Lastly, certain functions in aquarius require **optimization**, as they are excessively time or memory-consuming even on standard datasets. For instance, the function for ordering features along pseudotime (aquarius::find_features_order) is particularly slow due to the stats::loess function, which is inefficient for large datasets. Similarly, aquarius::run_foldchange requires excessive memory, indicating a need to replace it with optimized alternatives from well-established packages like edgeR or limma.

2.3.2 Re-usability of the analysis framework

The framework aims to establish an efficient methodology for reproducible data analysis. Currently, the pipeline from the sequencing files to the count matrix is documented and reused on the IFB Core Cluster. However, the notebook tree, which starts from the count matrix, lacks documentation and is not sufficiently generalized for easy reuse. In the lab, colleagues simplified the notebook tree for their projects by sometimes skipping steps like compiling with the Singularity container or maintaining an index page to link notebooks. They relied on the compiled notebooks from mature projects as templates, such as those from the Article 1 and the Article 2, but suffered from the lack of documentation to independently perform advanced analyses.

A future direction could involve creating a public Github repository with standardized R Markdown notebook templates, independent of specific projects. A documentation would clarify the purpose of each notebook and outline customizable parameter values with concrete examples from mature projects. An annotation of each notebook, by keywords, should enable easy findability and comparison with alternative similar notebooks (Cohen-Boulakia *et al.*, 2017).

Establishing a repository for single-cell transcriptomic analysis, akin to initiatives like nf-core[1], would greatly benefit the community. Such a repository would enhance modularity across analysis tools and methods, improve visibility, and save more time for the analysis than for the set up.

2.4 Making research reproducible

2.4.1 Chronology of the five pillars

This paragraph is an outlook about the chronological setting up of the five pillars of reproducibility in computational biology: code, versioning, environment, documentation, and data (Ziemann *et al.*, 2023).

Writing **code** is a daily task. To ensure traceability, R Markdown notebooks were utilized during my thesis, compiled once their structure was finalized. The repeatability was ensured by compiling notebooks multiple times to validate consistent results. The reproducibility was assessed for some notebooks, by running them on a distinct machine and obtaining exactly the same figures. Indeed, due to the numerous consecutive steps while analyzing scRNA-Seq data, ensuring the stability of the final representation

[1] nf-core website: `https://nf-co.re/` for Nextflow users and Github page: `https://github.com/nf-core`

of the data assesses the stability of the computational analysis. The **code documentation** involved explaining the purposes of custom function and analysis approaches. This was done in the meantime as developing code.

The **versioning** followed the code development, with stable functions integrated into a package. Significant updates were tagged and versioned, with notebook footers indicating the version used. The notebooks are not versioned, but the compilation date is included in the compiled document.

Post-analysis, identifying relevant packages established the **environment**. Functions like aquarius::repro_dependency_tree facilitated dependency management and the construction of Singularity container. For an easier environment definition and containerization, we could took advantage of R projects and the renv package[2].

The final pillar, **data**, involves storing raw data locally and count matrices linked via symbolic links to each project. Metadata were managed within experimental notebooks and in the first step of the notebook tree. When preparing the scientific contributions, raw data and count matrices were gathered and uploaded on a platform (ArrayExpress[3]) by a colleague. Since we analyzed NF1-related data, it should be useful for the NF1 community to link them in the **NF Data Portal**[4].

To enhance data traceability and access, a **database** was developed in collaboration with the University IT team, aimed at real-time biological material evaluation. Deployed on a server in Summer 2024, the database and its interface will undergo usability beta-testing and be utilized for data correlation studies and metadata semi-automatic assignation.

All the tools and methods exist to make the computational analysis of scRNA-Seq data reproducible. The documentation should not be forgotten as a continuous involvement. The most time challenging part concerned the environment control. Once ensured, it enables the reproducibility of the analyses and to progress faster.

Making research reproducible leads to publications with **fragmented resources**, due to various format and size files. Connecting and standardizing resources via **FAIR Digital Objects** alongside scientific publications could further unify research outputs (Niehues *et al.*, 2024).

[2]renv website: `https://rstudio.github.io/renv/articles/renv.html`
[3]ArrayExpress web page: `https://www.ebi.ac.uk/biostudies/arrayexpress`
[4]NF Data Portal website: `https://nf.synapse.org/`

2.4.2 The reproducibility as a methodological step in computational biology

In computational biology, the reproducibility can be defined as obtaining the same results, using the same data, methods and environment on a different machine. However, the results of sequencing data analysis can be compared to the temperature forecasts (Figure 77). Sequencing data represent a fixed observation within a biological context, and various analysis steps can yield widely differing results from similar data. Ensuring reproducibility locks in a specific analysis method (blue line) and could lead to results that deviate significantly from an "average" unknown outcome (red and black lines). This raises the question of whether reproducibility is always desirable.

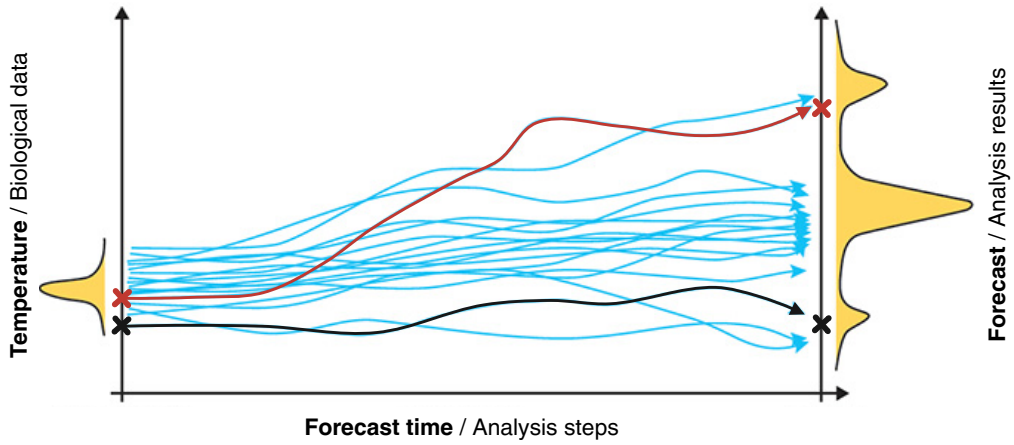


Figure 77 – **Fixing the analyses can lead to non robust results**. The figure represents the evolution of the temperature forecast across the time, written in bold. It can be compared to the processing of biological data to obtain results. Making the input data slightly varying will give a range of possible results (red and black crosses). Each blue line represents a way to forecast temperature or to analyze data. As forecasting temperature in a near future is robust, making few analysis steps of data output stable results. However, a lot of analysis steps output a result which may completely change if the input (data or parameters) slightly differ (red and black lines). A reproducible analysis consists in fixing one line, without exploring the others. A similar figure was presented at the 52nd ORAP forum[5].

Since methodological decisions impact analyses, it is crucial to evaluate the robustness of the analyses after establishing their reproducibility. Thus, making research reproducible is a methodological step toward ensuring robust analyses and results across different biological data or teams. Evaluating robustness involves varying the experimental context or methodology and assessing the results. A result is robust if it remains consistent under different contexts.

[5] 52nd ORAP forum: `http://orap.irisa.fr/52ieme-forum-reproductibilite/`

3 Discussion on the biological part

After a summary of the main features of our mouse model (Section 3.1), the cellular origin of the tumors are discussed (Section 3.2). Then, we discuss the two main triggers of the tumor progression: *Nf1* haploinsufficiency and inflammation (Section 3.3). The three next sections concern the tumor initiation (Section 3.4), tumor progression (Section 3.5), and malignant transformation (Section 3.6). The discussion ends with therapeutic options along the tumor development process (Section 3.7).

3.1 The *Nf1*-KO mouse model

In the peripheral nervous system, the expression of *Prss56* is restricted to a specific population, called boundary cap cells (BCCs). BCCs are transiently present during the embryonic development in mammals. In the *Nf1*-KO mouse model, the loss of exons 31 and 32 in *Nf1* is restricted to *Prss56*+ BCCs, making the inactivation of NF1 a controlled event. This suggests that the various symptoms developed by the model are issued from this population and its derivatives.

3.1.1 The *Nf1*-KO mouse model recapitulates several features of NF1 disease

The *Nf1*-KO mouse model recapitulates several symptoms of the NF1 disease, including various types of peripheral nerve sheath tumors (PNSTs) and an accumulation of melanocytes in the dermis, which can be compared to café-au-lait macules on the skin of NF1 patients (Radomska *et al.*, 2019). Unpublished work at the Institute has shown that mice also develop ocular lesions and bone abnormalities such as spinal deformities and pseudarthrosis (C. Colnot lab).

Among PNSTs, both *Nf1*-KO male and female develop small cutaneous neurofibromas (cNFs), paraspinal and sub-cutaneous plexiform neurofibromas (pNFs) (Radomska *et al.*, 2019). Male mice housed together are naturally aggressive. In the context of bite-induced skin trauma, cNFs are bigger and pNFs can transform into dysplastic neurofibromas (dyNFs) and MPNSTs (Article 2 and Figure 78).

3.1.2 Limitation of PNSTs studies to male mice

NF1 disease affects both men and women, but with distinct phenotypes. NF1 female patients are predisposed to breast cancer, potentially due to interactions between the NF1 protein and hormones (Madanikia *et al.*, 2012). Pregnancy further increases the

development of cutaneous neurofibromas (Sbidian *et al.*, 2016). However, hormonal changes during puberty do not affect the growth of pNFs (Dagalakis *et al.*, 2013).

Our MPNST mouse model is limited to male mice. Since female mice are not aggressive, we could perform skin lesions to assess MPNST development in them. Preliminary results show that they develop at least dyNFs (datasets dyNF_3 (2020_13) and dyNF_4 (2020_14) in the Article 2). Although tumor cell transcriptomic profiles are similar between males and females, the micro-environment needs further exploration. More datasets are needed to ensure the robustness of this analysis.

Skin lesions could also be used to better control the malignant transformation. Such a protocol is used in the context of cNFs development to better control their phase of progression (Article 4).

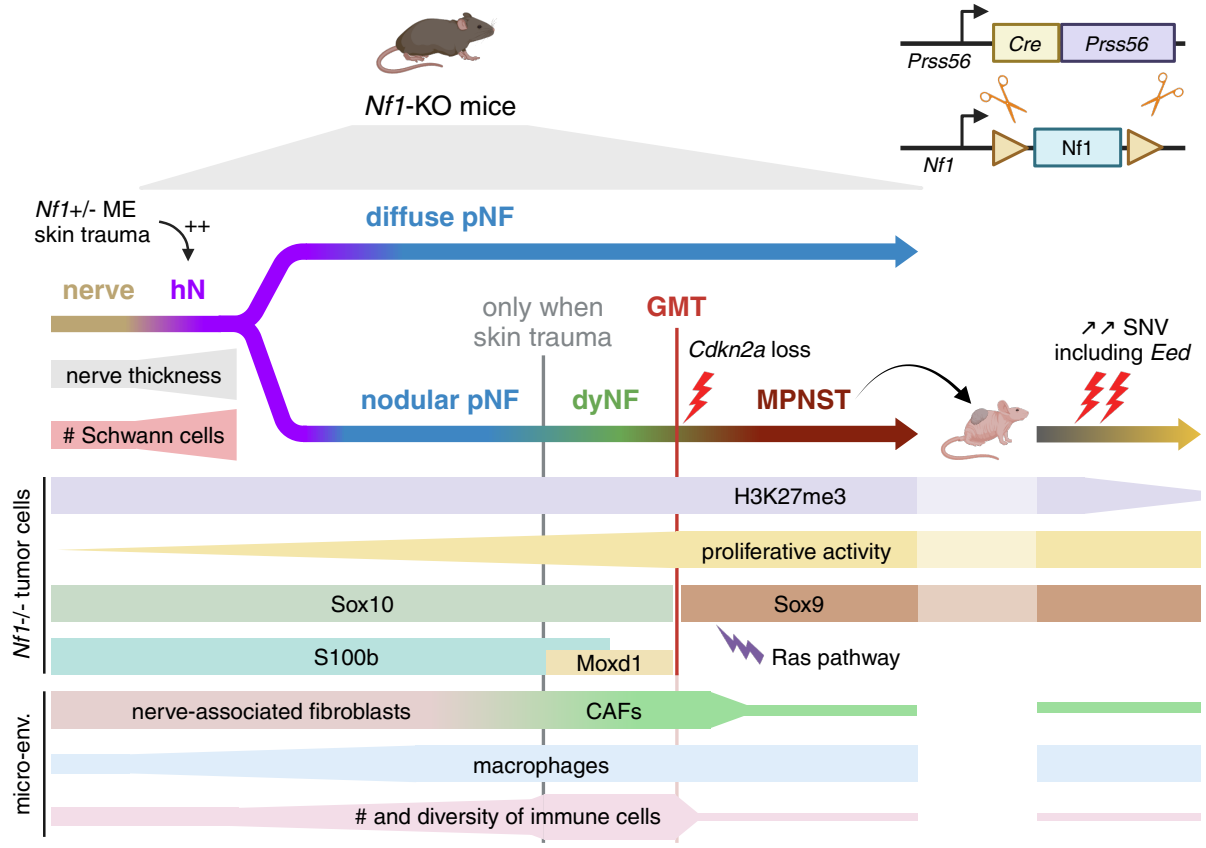


Figure 78 – *Nf1*-KO mouse model recapitulates the development of PNSTs. The figure summarizes key elements from the Article 1, Article 2 and complementary results. The construction of the mouse model is indicated in the top right corner. The mouse model recapitulates the development of hyperplastic nerves (hN), various subtypes of plexiform neurofibromas (pNF), and, in the context of chronic skin trauma when male mice bite each others, dysplastic neurofibromas (dyNF) and malignant peripheral nerve sheath tumors (MPNST). The main features are indicated along the development of peripheral nerve sheath tumors (PNSTs). The arrow with "++" indicates that *Nf1* haploinsufficiency or skin trauma accelerates the development of hN. #: number, CAFs: cancer-associated fibroblasts.

3.2 Fate of boundary cap cells in the peripheral nervous system

3.2.1 Distinct fate between BC-derived and NC-derived cells

Our mouse model is the first to recapitulate cNFs, pNFs, and their malignant transformation with high penetrance (70%). This indicates that *Prss56*+ boundary cap (BC) cells and their derivatives are the origin of these tumors, while other Schwann cell (SC) lineages are less likely to form cNFs and MPNSTs. Indeed, only the *Hoxb7*^{Cre};*Nf1*$^{fl/fl\ or\ fl/-}$ mouse model recapitulates the development of cNFs (Chen *et al.*, 2019) and *Hoxb7*+ cells include the BC cells (Li *et al.*, 2020). Furthermore, all other mouse models developing MPNSTs are targeted for the inactivation of *Cdkn2a*, *Pten*, *Trp53* or PRC2-encoding genes (Somatilaka *et al.*, 2022). This suggests that BC-derived but not other neural crest (NC)-derived cells have derivatives that can give rise to cNFs and MPNSTs.

The dissociation of tissue is crucial for producing scRNA-Seq data. In subcutaneous nerves, myelin removal was necessary for high-quality cell suspension. But, it resulted in few SCs in the final data, thus limiting the analysis of SC heterogeneity. Single nuclei RNA sequencing (snRNA-Seq) can be considered to bypass myelin removal, retrieving more SCs in nerves (Yim *et al.*, 2022). To characterize the heterogeneity and differences between BC-derived SCs (*Sox10*+*Tom*+) and other NC-derived SCs (*Sox10*+*Tom*-), we could generate snRNA-Seq datasets from nerves of the *Prss56*^{Cre};*Rosa26*^{Tom} traced model.

3.2.2 Predisposition of BC-derived cells to form neurofibromas

In other mouse models, additional mutations are necessary to promote MPNST development (Somatilaka *et al.*, 2022). In our model, some of these mutations occur spontaneously, either in male mice subjected to skin trauma, or in immunocompromised environment after allografting. The housing conditions in other mouse models are not clearly specified in the associated publications, but ethical considerations in the Unites States suggest they are housed individually (personal communication).

Despite this lack of information, this suggests that BC-derived cells are specifically implicated in MPNST formation, unlike other NC-derived cells. The susceptibility of these cells to acquire genetic aberrations may be linked to their epigenetic profile. Comparing the epigenetic profiles of pNF tumor cells from various mouse models could shed light on the reasons of the predisposition of BC-derived cells to form MPNSTs. Such insights could be pivotal in evaluating the malignant potential of pNFs in patients and exploring preventive epigenetic therapeutic options (Lu *et al.*, 2020).

A preliminary study in 3 mice suggests that ***Prss56*^{Cre};*Rosa26*^{Tom};*Nf1*$^{fl/+}$** mice develop MPNSTs. This suggests that only the loss of one allele of *Nf1* is sufficient for BC-derived cells to have an advantage over other cells, to form tumors. We are currently investigating if the second allele of *Nf1* has been affected or not. To further investigate if only BC-derived cells give rise to MPNSTs, we propose developing a ***Prss56*^{Cre};*Rosa26*^{Tom};*Nf1*$^{+/-}$** mouse model. All *Prss56* derivatives would express the *Tomato* reporter gene but all cells would be heterozygous for *Nf1*. This model would clarify if MPNSTs exclusively originate from BC cells based on their color. Indeed, only the tumors originating from BC cells will be red (*Tom*+).

Single nucleotide polymorphisms (SNPs) allow for individual identification. Additionally, it is possible to retrieve the tissue of origin of circulating cell-free DNA using specific epigenetic signatures, offering implications for early tumor detection (Oberhofer *et al.*, 2022). From these two observations, we are wondering whether specific epigenetic footprints could potentially trace the embryonic origin of cells, or not. In the context of MPNSTs, this approach would involve establishing the epigenetic profiles of all subtypes of SCs, in relation with their embryonic origin. Subsequently, correlating these profiles with those found in MPNSTs could inform about their embryonic origin.

3.3 Triggers of tumor development and malignant progression

3.3.1 *Nf1* haploinsufficiency in the micro-environment

Our lab previously showed that *Nf1*-KO^{het} mice develop symptoms (cNF and pNF) earlier than *Nf1*-KO^{wt} (Radomska *et al.*, 2019). We further showed that *Nf1* haploinsufficiency correlates with a more severe phenotype in hN compared to wild-type background (Article 1). This is consistent with the literature (White & Rhodes, 2024).

In the Article 2, we showed *Nf1* haploinsufficiency does not accelerate MPNST development, as there were no differences in incidence, age, or tumor count. This contrasts with literature suggesting *Nf1* haploinsufficiency limits the malignant transformation (White & Rhodes, 2024). However, such conclusions were reached after studies on MPNST mouse models developed in the United States, maybe housed alone, and engineered to carry additional genetic alterations than *Nf1*. In our model, MPNSTs are observed only in male mice housed together. We cannot rule out that skin trauma effects outweigh the differences induced by the *Nf1* haploinsufficiency.

3.3.2 Skin trauma and inflammation

The skin trauma is necessary for malignant progression in our mouse model, regardless of *Nf1* status in the TME (Article 2). It may be involved in the pNF-to-dyNF transition, as dyNF and MPNST are observed only in house-grouped male mice. This aligns with the increased **immune infiltration** reported in dyNF from our mouse model (Article 2) and in NF1 patients (Article 5). Since genetic aberrations were identified in MPNSTs, we are wondering whether they results from the immune infiltration in dyNFs or another mechanism than inflammation, triggered by the skin trauma.

Before the occurrence of pNFs, we showed that skin trauma increases phenotype severity in subcutaneous nerves of 3-month-old mice (Article 1). We believe the superficial skin trauma induced by bites promotes a **local inflammation** that affects subcutaneous nerves but not nerve roots. Indeed, we did not report more paraspinal pNF in house-grouped males compared to mice housed alone.

We observed more macrophages in *Nf1*-KO house-grouped male mice, compared to *Nf1*-KO mice kept alone or control mice. This further emphasizes the impact of skin trauma on an inflammation in mutant nerves. The expansion and activation of endoneurial macrophages (*Cx3cr1*+) but not epineurial macrophages was described after sciatic nerve injury (Ydens *et al.*, 2020). The roles of CX3CR1 include pro-inflammatory, anti-apoptotic, and pro-angiogenic activities. But, contradictory implications have been identified in cancer (Szukiewicz, 2024). In the context of NF1, making the **CX3CR1-CX3CL1 axis** a potential therapeutic target remains to be evaluated.

After NF1 loss in SCs, **inflammation** has been proposed as a **second trigger** driving pNF formation, notably through STAT3 (Fletcher *et al.*, 2020). The inactivation of STAT3 reduces the number of macrophages and pNF growth in the *Dhh*^{Cre};*Nf1*$^{fl/fl}$ mouse model (Fletcher *et al.*, 2019). Similarly, the *Mpz*^{Cre};*Nf1*$^{fl/fl}$ model does not develop pNF unless nerve crush injury occurs (Ribeiro *et al.*, 2013). The comparison between pNFs and unrepaired injured nerves is addressed in Section 3.4.

In our model, the mechanic trauma is uncontrolled as it is induced by bites. Since tumor size is not correlated with the dyNF or MPNST stage, the tumor type is known only after sacrifice. Consequently, the dyNF stage, and particularly early dyNF, is very difficult to identify. For a better control of tumor progression, we could manually perform skin lesions on mice housed alone, as done to control development of cNFs Article 4. To identify the mechanisms initiating nerve inflammation after skin injury, we could test various topical delivery of pro-inflammatory drugs.

The relevance of skin trauma in NF1 patients needs investigation. Indeed, pNFs are thought to be congenital, as over 50% of pediatric pNFs are detected before 3 years old (Prada *et al.*, 2012b). However, 17% of isolated pNFs in adults occur after trauma (Ho *et al.*, 2022). A **systemic inflammation** might directly drive pNF progression, rather than a mechanical skin trauma. Indeed, systemic inflammation resulting from external stimuli, food, or psychological stress can predispose to cancer (Zitvogel *et al.*, 2017; Dai *et al.*, 2020). We showed that *Nf1* haploinsufficiency in the TME induces an inflammatory profile in endoneurial fibroblasts (Article 1), suggesting *NF1*$^{+/-}$ status could also trigger an inflammation in NF1 patients. Coherently, it has been shown that NF1 patients are characterized by a systemic inflammation (Nobeyama *et al.*, 2022).

3.4 From hN to pNF: the hypothesis of unrepaired wound

3.4.1 Repair phenotype in hyperplastic nerves and pNFs

In a non-pathological context, repair SCs arise from mSCs and nmSCs after nerve injury. They are characterized by immature markers (**dedifferentiation**) and an inflammatory profile (**activation**) (Jessen & Mirsky, 2016). Nerve injury response also involves changes in macrophages, from infiltration, pro-inflammatory, anti-inflammatory profiles to exclusion.

Given the activated profile of tumor SCs, immune cell infiltration, and abundance of pro-inflammatory macrophage in pNFs (Choi *et al.*, 2017), they are compared to injured nerves that fail to repair (Parrinello & Lloyd, 2009; Gomez-Sanchez *et al.*, 2013; Fletcher *et al.*, 2019). Our scRNA-Seq data from *Nf1*-KO mice nerves shows similar features (Article 1). Our results further indicate that a similar population to repair-associated fibroblasts, transiently present after nerve injury (Carr *et al.*, 2018), is found in nodular pNFs and constitute the identity of all fibroblasts in dyNFs (next Section). Taken together, these observations supports the comparison of neurofibromas with wounds that fail to heal. As a consequence, the molecular events driving the physiological nerve repair process can be considered as therapeutic targets.

3.4.2 From nerve-associated fibroblasts to cancer-associated fibroblasts

In pNFs, three classical nerve-associated fibroblasts are present: epineurial, perineurial, and endoneurial fibroblasts (Kershner *et al.* (2022) and Article 2). Conversely, dyNFs exclusively contain cancer-associated fibroblasts (CAFs) characterized by the expression of *Postn*, *Cilp*, *Ptn*, *Mest*, *Tnc*, and *Plagl1* (Article 2 and Section 3.3.2 of Results). These CAFs strongly resemble the "differentiating fibroblasts" identified in

transcriptomic studies following injury of the sciatic nerves (Carr *et al.*, 2018). This suggests that the CAFs observed in nodular pNFs and dyNFs arise through a similar process as described for differentiating fibroblasts. Authors demonstrated that differentiating fibroblasts originate from the endoneurial population. In the context of pNF-to-dyNF progression, we suggest they could originate from the three nerve populations. The transcriptomic similarities between CAFs and differentiating fibroblasts strengthen the comparison between neurofibromas and unwounded injured nerves.

In cancer, various subtypes of CAFs have been described. Exploring whether differentiating fibroblasts are implicated in tumor development, we noted that the expression of *Postn* is a common feature shared between differentiating fibroblasts and CAFs, but not the other genes (Sahai *et al.*, 2020; Lavie *et al.*, 2022). Notably, CAFs in neurofibromas exhibit high expression of ECM-related genes (Article 2, Figure 2G), suggesting they could be categorized as "matrix CAFs" (Lavie *et al.*, 2022). Despite their diverse roles and interactions with tumor cells in pNFs and dyNFs, the implications of CAFs in tumor progression and their exclusion from MPNSTs warrant further investigation.

3.4.3 Repair phenotype of tumor SCs

Repair SCs are characterized by Ras signaling activation, leading to demyelination if necessary, myelinophagy, and macrophage recruitment (see Introduction, Section 1.4.2). Since NF1 negatively regulates RAS activity, a repair phenotype in tumor SCs may begin once a Ras pathway is being activated. Then, the lack of NF1 prevents the down-regulation of Ras signaling, hindering the ability of SCs to re-differentiate in mature SCs and become quiescent.

In NF1 disease, mutant SCs in neurofibromas are sometimes compared to dedifferentiated SCs (Iribar *et al.*, 2016; Kershner *et al.*, 2022). Indeed, unlike repair SCs, SC precursors (SCPs) and immature SCs cannot promote the axonal regeneration after nerve injury in rats (Endo *et al.*, 2022). Furthermore, our data showed that mutant SCs over-express cytokines like *Ccl2* and other inflammation-related genes (*Phlda1*, *Jun*), compared to wild-type SCs (Article 1). This profile is specific to repair SCs (Endo *et al.*, 2022; Zhao *et al.*, 2022). Additionally, *Cdh19* and *Sox2*, used as SCP markers (Iribar *et al.*, 2016), are also expressed by repair SCs in the dataset from (Zhao *et al.*, 2022).

Conversely, repair SCs are characterized by *Bdnf*, *Gdnf*, and *Artn* expression, which we did not identify in our data. Thus, tumor cells in pNFs can be compared to repair SCs, although they are not fully reparative. This may explain a failure to resolve the wound, resulting in pNF.

3.4.4 Late events of pNF initiation

Studying the events driving pNF initiation is not feasible in NF1 patients, due to their asymptomatic nature. Our study focused on a characterization of hyperplastic nerves (hN) in 3-month-old *Nf1*-KO mice (Article 1). We identified three main features of this phenotype: the pro-fibrotic activity of tumor SCs, the inflammatory profile of endoneurial fibroblasts, and the expansion of endoneurial macrophages.

All hNs do not progress to pNFs. This progression may involve a second wave of events. Such a phase likely includes a disruption of nerve bundles, characterized by the loss of tight junctions in perineurial fibroblasts, the subsequent infiltration of mast cells and T cells (Fletcher *et al.*, 2019) and the emigration of tumor cells from the nerve fascicles (Article 1, Figure S1). Coherently, lower expression of tight functions (*Cldn1* and *Tjp1*) and mast cells infiltration were observed in our pNF datasets, issued from mice older than 7 months old, compared to hN from 3 months old mice. Studying hNs in 4 to 6 months old mice may help to order the late events preceding pNFs.

3.5 From pNF to dyNF: a trigger in the nerve micro-environment

Although pNFs can transform in dyNFs and then in MPNSTs, the malignant transformation of cNFs has never been observed. This can be related to distinct epigenetic profiles (Grit *et al.*, 2021) that may be visible at a transcriptomic level. Combining the scRNA-Seq of cNFs (Article 4) and pNFs (Article 2 and ongoing) will reveal distinct transcriptomic profiles in tumor cells.

Since nerve micro-environmental factors are essential for pNF formation, they may also influence progression to dyNF and malignant transformation (Liao *et al.*, 2016). Indeed, while cNFs develop in the nerve endings in the skin, pNFs are found at nerve roots and along the peripheral nerves. Their malignant transformation could be related to differences in the micro-environment, particularly immune cells (Article 5).

There is a significant clinical variability among neurofibromas, lacking a standardized classification. However, among pNFs, nodular and diffuse subtypes are recognized. If they constitute successive stages, the diffuse subtype should be physically the second step. Due to a more severe phenotype in nodular pNFs compared to diffuse pNFs, our preliminary findings suggest that they represent distinct entities, with nodular neurofibromas preferentially giving rise to dyNFs (Figure 78). This is consistent with literature (Schaefer & Fletcher, 2015; Cortes-Ciriano *et al.*, 2023).

In our scRNA-Seq data, we observed a decrease in the number of perineurial fibroblasts expressing *Tjp1* (10% less cells) and *Cldn1* (30% less cells), in diffuse

pNFs compared to nodular pNFs. These genes encode tight junctions and enable the perineurial fibroblasts to maintain the structure of nerve fascicles (data not shown). In diffuse pNFs, this may lead to a decompaction of nerve fascicles, facilitating tumor cells to escape from the nerve. Considering the importance of nerve micro-environment for pNF development, this escape could explain why diffuse pNFs are less prone to malignant transformation.

Histological sections of nodular pNFs should be done to investigate their differences with dyNFs. Additionally, to evaluate the propensity of nodular pNFs tumor cells, but not diffuse, to give rise to dyNF, one could transplant tumor cells from both types of pNFs into the sciatic nerves of recipient mice and monitor tumor development.

In our mouse model, even in the context of chronic skin trauma, only a subset of pNFs transform into MPNSTs. Currently, it is not possible to predict the transformation of pNFs. Supposing that young *Nf1*-KO mice contain more pNFs susceptible to progress, compared to old mice, differential expression analyses in tumor cells and the micro-environment in pNFs between young and old mice could reveal crucial markers. These markers would aid in understanding the mechanisms driving pNF-to-dyNF progression and predicting which pNFs are at risk of malignant transformation in patients.

Unlike pNFs, we showed that dyNF tumor cells express *Moxd1*, *Uchl1*, and *Pdgfb*, with some losing *S100b* (Article 2). Their validation at a proteic level if further needed to evaluate if they could help in identifying pNFs that will progress toward malignancy. Finally, there were too few tumor cells in the pNF datasets to robustly identify genes that are lost in dyNFs. Additional pNFs datasets are required to conduct such an analysis.

3.6 From dyNF to MPNST: the glial-to-mesenchymal transition

From E13, *Prss56* is no longer expressed or *de novo* reactivated. In our model, it ensures that all *Nf1* mutant cells originate exclusively from boundary cap cells. The constitutive and definitive expression of the *Tomato* reporter gene is indispensable for tracking all mutant cells. By analyzing *Tom+* cells during tumor progression, we identified the glial-to-mesenchymal transition (GMT) as the primary mechanism underlying malignant transformation. Indeed, almost all MPNST tumor cells exhibit a mesenchymal profile. We further demonstrated that allografted dyNFs progress into MPNSTs in nude recipients, confirming they are the precursors of MPNSTs (Beert *et al.* (2011) and Article 2).

3.6.1 Chronology of genetic events and GMT

The invalidating translocation or genetic loss of copies of *Cdkn2a* is the initial genetic event driving the progression of pNFs to dyNFs in NF1 patients (Beert *et al.*, 2012; Magallón-Lorenz *et al.*, 2021; Cortes-Ciriano *et al.*, 2023). Subsequently, a GMT occurs, along with two proposed ways of progression: (A) maintenance of immune infiltration and PRC2 activity, coupled with susceptibility to mitotic errors and chromosomal instability, or (B) loss of PRC2 methyltransferase function followed by gain of copies in the chromosome 8 and reduced immune cell infiltration (Cortes-Ciriano *et al.*, 2023).

Our mouse model predominantly recapitulates the way (A). The sustained macrophage infiltration in MPNSTs and whole exome sequencing showing *Cdkn2a* copy loss (11/13 samples) but no alteration in PRC2-related genes further support this way. To support the way (B), regarding the synteny of chromosome 8 in human and mouse, we should observe gains at the very beginning of chromosomes 3 and 4 and the small arm of chromosome 8, which is not the case. This way (B) may not always be a distinct branch but could follow (A). Coherently, a lower staining of H3K27me3 and genetic variants in *Eed* were observed after the successive allografting of primary mouse MPNSTs (Article 2, Figure 4).

In the two MPNSTs lacking the loss of copies of *Cdkn2a* locus, the GMT of tumor cells was confirmed. The loss of *Cdkn2a* occurs after transplantation, suggesting that it was not altered before (Article 2, Figure 5). Consequently, we suggest that the loss of *Cdkn2a* can occur after the GMT. Given this possibility, we supposed that the consequence of the genetic loss of *Cdkn2a* is to promote the over-proliferation of tumor cells. The loss of PRC2 activity, through loss of functions in *Eed* or *Suz12*, causing the loss of H3K27me3 mark, can be responsible for the over-activation of genes. This may impact the activity of tumor cells and change their aggressiveness and metastatic potential.

A crucial next step involves the whole genome sequencing of dyNFs to identify CNVs, SNVs, or translocations. However, sustained *Cdkn2a* expression in tumor cells of dyNFs from our mouse model (Article 2) challenges the typical *Cdkn2a* loss observed in dyNFs compared to pNFs in patients (Beert *et al.*, 2012).

In our scRNA-Seq data, a few *Sox10*+ tumor cells persisted in MPNSTs, all of which were *Cdkn2a*-. Differential expression analysis between *Sox10*+*Cdkn2a*+ (dyNF) and *Sox10*+*Cdkn2a*- (MPNST) tumor cells could unveil early events driving GMT. However, validating genetic alterations causing *Cdkn2a* loss in these cells will be challenging due to their rarity. Fluorescence in situ hybridization (FISH) could clarify the event.

3.6.2 The hypothesis of cancer stem cells

Since all tumor cells in MPNSTs are *Tom*+, it emphasizes that *Nf1*$^{-/-}$ BC cells are at their origin. However, whether MPNST tumor cells arises from a rare cancer stem cell (CSC) within dyNFs or if numerous dyNF tumor cells have the potential to acquire a mesenchymal identity remains unknown.

CSCs are characterized by rare mitoses. By asymmetrical division, they generate highly proliferative cells responsible for tumor growth and aggressiveness (Blanpain, 2013; Kreso & Dick, 2014; Batlle & Clevers, 2017). Indeed, treatments that target only dividing cells may not affect the CSC populations at the origin of malignant tumors, which may explain further relapse (Kreso & Dick, 2014). In MPNSTs, proposed markers for such a CSC population include CD133, OCT4 and NES, as well as decreased expression of NCAM and CD90 (Spyra *et al.*, 2011; Sun *et al.*, 2021).

In our mouse model, except for *Nes*, all these genes are either expressed similarly in all tumor cells from pNFs to MPNSTs, or not expressed at all. Nestin (NES) is broadly expressed at the transcriptomic level. Our trajectory analysis shows *Nes* expression peaks just before the malignant transformation, indicating its potential role in the malignant transformation, yet proteomic investigations are necessary for validation.

MPNSTs can be polyclonal (see Results, Section 3.3.4). But, we cannot exclude this clonal heterogeneity to be posterior to the malignant transformation. To explore the propensity of tumor cells from dyNFs to generate mesenchymal tumor cells, other members of the lab are developing an *in vitro* barcoding experiment followed by transplantation into recipient mice. This approach will assess the clonal origins of MPNSTs after few weeks of growth, by investigating the representativness of barcodes.

3.6.3 A population-level transition or a cellular switch ?

The malignant transformation is characterized by the GMT, an identity change in the tumor cell population. However, we did not demonstrate that the GMT represents also a cellular transition.

In support of a **population-level transition**, each dyNF scRNA-Seq dataset consistently contained at least one *Sox9*+ *Sox10*- tumor cell. Such cells were overlapping with the mesenchymal tumor cells of MPNSTs, supporting their mesenchymal identity. These rare cells may be the cellular origin of MPNSTs. Further supporting a population transition is the abrupt transcriptomic changes between glial and mesenchymal tumor

cells (Article 2, Figure 3B). Particularly, no co-expression of *Sox9* and *Sox10* was observed at the mRNA level. Furthermore, the linkage between glial and mesenchymal cells in the diffusion map UMAP correlates with the cell cycle. In our data compared to others, tumor cells particularly exhibit a strong cell cycle identity characterized by elevated expression of cell cycle-related genes and down-regulation of genes responsible for cell identity. Despite attempts to mitigate cell cycle bias to smooth the transition between glial and mesenchymal cells on the projection, these efforts were unsuccessful and disrupted the projection of tumor cells.

Arguments in favor of the **cellular switch** hypothesis include the near absence of *Sox10*$^{+}$ cells in MPNSTs. No death-related signaling pathways were identified in our scRNA-Seq data, suggesting these cells do not undergo apoptosis. Moreover, we did not observe a significant hyper-proliferation of *Sox9*$^{+}$ cells, implying they quickly outcompete the glial tumor cells. Finally, the rare *Sox9*+ tumor cells observed in dyNFs may correspond to *Sox10*+ cells that have already transformed, underlying their pre-malignant nature. The low number of *Sox9*+ cells in dyNFs and the absence of dead *Sox10*+ in MPNSTs suggest a massive cellular-level transition of tumor cells. To support this hypothesis, preliminary results showed that mouse dyNF *Sox10*+ tumor cells acquire mesenchymal features after few days in cell culture.

If the GMT is a cellular switch, the mechanisms responsible for this dramatic identify change remains to be elucidated. Given that it can occur without major genetic alterations, the trigger may reside as an epigenetic predisposition in tumor cells, awaiting for a external signal from the micro-environment.

3.6.4 Relevance in the sporadic context

Nearly half of MPNSTs occur in sporadic contexts (Somatilaka *et al.*, 2022). Half of them further depict alterations in *Nf1* (Cortes-Ciriano *et al.*, 2023).

Our *Nf1*-KO^{wt} mouse model can be considered as suited for studying these sporadic MPNSTs. In this model, MPNSTs arise from dyNFs. Although nodular pNFs and one ANNUBP (dyNF) have been reported sporadically, their malignant transformation has never been observed (Ho *et al.*, 2022; Tsuchie *et al.*, 2023). This suggests than sporadic MPNSTs can arise *de novo*. Conversely, sporadic MPNSTs could arise from very small neurofibromas that are not detectable due to their slow growth and absence of symptoms. Acquisition of additional mutations in cell cycle-related genes, including *CDKN2A*, would make them rapidly growing and symptomatic. This is consistent with the fact that the *Nf1* wild-type background predisposes to malignant transformation

compared to the *Nf1* haploinsufficient background (White & Rhodes, 2024).

MPNSTs are malignant tumors arising in the peripheral nervous system. Although the Schwann cell origin of NF1-related MPNSTs is commonly accepted, sporadic MPNSTs may originate from a distinct cell population. Considering the mesenchymal nature of tumor cells, the nerve-associated fibroblasts could give rise to MPNSTs.

3.7 Therapeutic hopes

The aim of therapeutic strategies is to prevent, delay, or cure the tumors. In this section, three targets are suggested: the ECM, the TME or tumor cells.

3.7.1 Target the fibrosis via periostin

Few genes consistently exhibit a sustained expression across all peripheral nerve sheath tumors (PNSTs): pNFs, dyNFs, and MPNSTs. Notably, periostin, encoded by *Postn*, is one of these genes that further distinguishes tumor cells from wild-type SCs (Kershner *et al.* (2022) and Article 1). Periostin is a secreted protein with physiological expression in various tissues, particularly in collagen-rich regions (Dorafshan *et al.*, 2022). It interacts notably with integrins on cell membranes and its downstream signaling pathways include Ras signaling, extracellular matrix (ECM) remodeling, cell adhesion, migration, and epithelial-to-mesenchymal transition (Dorafshan *et al.*, 2022).

Abnormally high expression of periostin is implicated in fibrosis, wound healing, and tumorigenesis (Dorafshan *et al.*, 2022). Under the hypothesis that neurofibromas are unrepaired wound (Section 3.4) and the fibrosis can impede nerve repair, the fibrosis may be a mechanism leading to the formation of PNSTs. In our mouse model, the expression of cytokines like *Pdgfa* and *Pdgfb* can activate *Postn* in glial tumor cells. In mesenchymal tumor cells, *Sox9* and *Twist1* may contribute to its expression (Dorafshan *et al.*, 2022). Given its diverse roles in cancer, *Postn* emerges as a potential therapeutic target.

Despite its prevalent expression in tumor cells across many cancers, *Postn* is absent or expressed at low levels in most cancer cell lines (Dorafshan *et al.*, 2022). This suggests that *Postn* expression in tumors is context-dependent, influenced by signals from the tumor micro-environment. Preliminary transcriptomic analysis of 3D sphere cultures of tumor cells from our mouse model revealed sustained *Postn* expression in glial cells, indicating an intrinsic regulatory mechanism, whereas its expression was

absent in mesenchymal tumor cells. This observation complicates the evaluation of *Postn*-targeted therapies in culture settings.

3.7.2 A therapeutic window through the immune infiltrate of dyNF

In MPNSTs, the immune landscape predominantly contains macrophages, with T cells notably absent (Wu *et al.*, 2022a; White & Rhodes, 2024). Recent studies propose strategies to re-activate T cell infiltration and then use immune checkpoint blockade for effective tumor cell elimination (Somatilaka *et al.*, 2024).

Conversely, dyNFs exhibit enrichment in immune cells, suggesting an active cross-talk between tumor cells and the micro-environment (TME). Investigating this interaction could unveil novel therapeutic avenues to potentially delay or prevent malignant transformation. Ongoing research by another team utilizes our transcriptomic atlas of *Nf1*-KO^{wt} mouse tumors for this purpose. Spatial transcriptomics may further provide insights into the specific TME cell types interacting physically with tumor cells, thereby aiding in the prioritization of therapeutic strategies.

3.7.3 Ras signaling and MPNST progression

3.7.3.1 Ras signaling as therapeutic target

Through a screening of 3,500 compounds, we showed that Ras signaling pathway is related to the sustained expression of SOX9 in mesenchymal tumor cells (Article 2). Currently, combined treatments implicating Ras pathways successfully delayed MPNST growth (Kohlmeyer *et al.* (2023) and presentations at the 2024 Global NF Conference). Further targeting Ras signaling in dyNF may delay or prevent the GMT, and thus, the malignant transformation.

3.7.3.2 Ras activity and scRNA-Seq data

We were wondering whether the Ras signaling pathways were over-activated or not, in mesenchymal tumor cells compared the glial ones, using our scRNA-Seq data. Using GSEA, we did not identify consistent enrichment of Ras-related gene sets. Several factors could account for this observation. Firstly, the activity of Ras pathways may be comparable between dyNF and MPNST cells despite resulting in distinct downstream effects. Secondly, scRNA-Seq may not accurately capture the pathway activity, particularly since Ras signaling involves phosphorylation cascades rather than changes in mRNA levels. To delve deeper, we explored the activity of

transcription factors (TFs) involved in the PI3K and MAPK pathways, two branches of Ras signaling. We used the CollecTRI database and the univariate linear model implemented in the decoupleR package (Müller-Dott *et al.*, 2023; Badia-i-Mompel *et al.*, 2022). Our findings suggest varied TF activity patterns, some TFs being activated exclusively in glial cells, mesenchymal cells, both, or neither. Furthermore, the inferred activity was sometimes contradictory to the TF expression. The reliability of this type of analysis is therefore questionable.

To more accurately characterize the biological activity of cells, single-cell proteomic and post-translational modifications analysis would be ideal, though these methods are still evolving (Marx, 2019). Alternatively, kinome profiling evaluates the phosphorylation status of specific proteins, providing a more direct assessment of pathway activity compared to transcriptomic data (Sikkema *et al.*, 2009).

In MPNSTs from patients, the successive activation of the Ras, Wnt-β catenin, and Shh pathways has been reported through bulk RNA-Sequencing analysis (Suppiah *et al.*, 2023). Bulk RNA-Sequencing provide a better transcriptome coverage compared to 10X Genomics scRNA-Seq protocol. It enables a more robust analysis of signaling pathways activity. To assess if our mouse model recapitulates the activation of Wnt and Shh pathways, we could propagate primary mouse MPNSTs into nude recipients, followed by such an analysis.

VI Appendix

1 List of genes and proteins

This section summarizes the list of genes and proteins cited in this thesis, excluding the manuscripts. By convention, they are identified through their Ensembl ID, gene symbol and gene name, for *Homo sapiens*.

ENSG00000213088	**ACKR1/DARC**	Atypical Chemokine Receptor 1
ENSG00000107796	**ACTA2**	Actin Alpha 2, Smooth Muscle
ENSG00000137845	**ADAM10**	ADAM Metallopeptidase Domain 10
ENSG00000174837	**ADGRE1**	Adhesion G Protein-Coupled Receptor E1
ENSG00000106624	**AEBP1**	AE Binding Protein 1
ENSG00000139154	**AEBP2**	AE Binding Protein 2
ENSG00000188157	**AGRN**	AGRN
ENSG00000204472	**AIF1**	Allograft Inflammatory Factor 1
ENSG00000142208	**AKT**	AKT Serine/Threonine Kinase 1
ENSG00000165092	**ALDH1A1**	Aldehyde Dehydrogenase 1 Family Member A1
ENSG00000134982	**APC**	APC Regulator Of WNT Signaling Pathway
ENSG00000117407	**ARTN**	Artemin
ENSG00000139352	**ASCL1**	Achaete-Scute Family BHLH Transcription Factor 1
ENSG00000087586	**AURKA**	Aurora Kinase A
ENSG00000103126	**AXIN**	Axin 1
ENSG00000176697	**BDNF**	Brain Derived Neurotrophic Factor
ENSG00000164434	**BFABP/FABP7**	Fatty Acid Binding Protein 7
ENSG00000182492	**BGN**	Biglycan
ENSG00000089685	**BIRC5**	Baculoviral IAP Repeat Containing 5
ENSG00000125378	**BMP4**	Bone Morphogenetic Protein 4
ENSG00000153162	**BMP6**	Bone Morphogenetic Protein 6
ENSG00000108691	**CCL2/MCP1**	C-C Motif Chemokine Ligand 2
ENSG00000110092	**CCND1**	Cyclin D1
ENSG00000007062	**CD133/PROM1**	Prominin 1
ENSG00000116031	**CD207**	CD207 Molecule
ENSG00000090659	**CD209**	CD209 Molecule
ENSG00000174059	**CD34**	CD34 Molecule
ENSG00000198851	**CD3E**	CD3 Epsilon Subunit Of T-Cell Receptor Complex
ENSG00000010610	**CD4**	CD4 Molecule
ENSG00000026508	**CD44**	CD44 Molecule
ENSG00000081237	**CD45/PTPRC**	Protein Tyrosine Phosphatase Receptor Type C
ENSG00000109956	**CD57**	Beta-1,3-Glucuronyltransferase 1
ENSG00000019582	**CD74**	CD74 Molecule
ENSG00000007312	**CD79B**	CD79b Molecule
ENSG00000153563	**CD8A**	CD8 Subunit Alpha
ENSG00000154096	**CD90/THY1**	Thy-1 Cell Surface Antigen

ENSG00000039068	**CDH1**	Cadherin 1
ENSG00000071991	**CDH19**	Cadherin 19
ENSG00000170558	**CDH2**	Cadherin 2
ENSG00000113361	**CDH6**	Cadherin 6
ENSG00000081138	**CDH7**	Cadherin 7
ENSG00000135446	**CDK4**	Cyclin Dependent Kinase 4
ENSG00000105810	**CDK6**	Cyclin Dependent Kinase 6
ENSG00000147889	**CDKN2A**	Cyclin Dependent Kinase Inhibitor 2A
ENSG00000147883	**CDKN2B**	Cyclin Dependent Kinase Inhibitor 2B
ENSG00000172757	**CFL**	Cofilin 1
ENSMUSG00000040809	**CHIL3**	(*Mus musculus* only) Chitinase Like 3
ENSG00000138615	**CILP**	Cartilage Intermediate Layer Protein
ENSG00000163347	**CLDN1**	Claudin 1
ENSG00000184113	**CLDN5**	Claudin 5
ENSG00000132514	**CLEC10A**	C-Type Lectin Domain Containing 10A
ENSG00000163815	**CLEC3B**	C-Type Lectin Domain Family 3 Member B
ENSG00000168542	**COL3A1**	Collagen Type III Alpha 1 Chain
ENSG00000184371	**CSF1**	Colony Stimulating Factor 1
ENSG00000182578	**CSF1R**	Colony Stimulating Factor 1 Receptor
ENSG00000118523	**CTGF**	Cellular Communication Network Factor 2
ENSG00000163599	**CTLA4**	Cytotoxic T-Lymphocyte Associated Protein 4
ENSG00000036257	**CUL3**	Cullin 3
ENSG00000164619	**CV2**	BMP BindingENdothelial Regulator
ENSG00000006210	**CX3CL1**	C-X3-C Motif Chemokine Ligand 1
ENSG00000168329	**CX3CR1**	C-X3-C Motif Chemokine Receptor 1
ENSG00000107562	**CXCL12**	C-X-C Motif Chemokine Ligand 12
ENSG00000121966	**CXCR4**	C-X-C Motif Chemokine Receptor 4
ENSG00000011465	**DCN**	Decorin
ENSG00000080166	**DCT**	Dopachrome Tautomerase
ENSG00000175084	**DES**	Desmin
ENSG00000139549	**DHH**	Desert Hedgehog Signaling Molecule
ENSG00000185559	**DLK1**	Delta Like Non-Canonical Notch Ligand 1
ENSG00000198719	**DLL1**	Delta Like Canonical Notch Ligand 1
ENSG00000105880	**DLX5**	Distal-Less Homeobox 5
\	**E2F**	(family) E2F Transcription Factor
ENSG00000074266	**EED**	Embryonic Ectoderm Development
ENSG00000172889	**EGFL7**	EGF Like Domain Multiple 7
ENSG00000146648	**EGFR**	Epidermal Growth Factor Receptor
ENSG00000120738	**EGR1/KROX24**	Early Growth Response 1
ENSG00000122877	**EGR2/KROX20**	Early Growth Response 2
ENSG00000179388	**EGR3**	Early Growth Response 3
ENSG00000130427	**EPO**	Erythropoietin

ENSG00000141736	**ERBB2**	Erb-B2 Receptor Tyrosine Kinase 2
ENSG00000065361	**ERBB3**	Erb-B2 Receptor Tyrosine Kinase 3
ENSG00000102882	**ERK1/MAPK3**	Mitogen-Activated Protein Kinase 3
ENSG00000100030	**ERK2/MAPK1**	Mitogen-Activated Protein Kinase 1
ENSG00000091831	**ER/ESR1**	Estrogen Receptor 1
ENSG00000108799	**EZH1**	ENhancer Of Zeste 1 PRC2 Subunit
ENSG00000106462	**EZH2**	ENhancer Of Zeste 2 PRC2 Subunit
ENSG00000113194	**FAF2**	Fas Associated Factor Family Member 2
ENSG00000169398	**FAK/PTK2**	Protein Tyrosine Kinase 2
ENSG00000113578	**FGF1**	Fibroblast Growth Factor 1
ENSG00000138685	**FGF2**	Fibroblast Growth Factor 2
ENSG00000077782	**FGFR1**	Fibroblast Growth Factor Receptor 1
ENSG00000115414	**FN1**	Fibronectin 1
ENSG00000187140	**FOXD3**	Forkhead Box D3
ENSG00000111206	**FOXM1**	Forkhead Box M1
ENSG00000118689	**FOXO3**	Forkhead Box O3
ENSG00000104290	**FZD3**	Frizzled Class Receptor 3
ENSG00000172020	**GAP43**	Growth Associated Protein 43
ENSG00000266524	**GDF10**	Growth Differentiation Factor 10
ENSG00000143869	**GDF7**	Growth Differentiation Factor 7
ENSG00000168621	**GDNF**	Glial Cell Derived Neurotrophic Factor
ENSG00000131095	**GFAP**	Glial Fibrillary Acidic Protein
\	**GFP**	Green Fluorescent Protein
ENSG00000176402	**GJC3**	Gap Junction Protein Gamma 3
ENSMUSG00000030048	**GKN3**	(*Mus musculus* only) Gastrokine 3
ENSG00000177885	**GRB2**	Growth Factor Receptor Bound Protein 2
ENSG00000082701	**GSK3B**	Glycogen Synthase Kinase 3 Beta
ENSG00000145681	**HAPLN**	Hyaluronan And Proteoglycan Link Protein 1
ENSG00000019991	**HGF**	Hepatocyte Growth Factor
ENSG00000100644	**HIF1A**	Hypoxia Inducible Factor 1 Subunit Alpha
ENSG00000143437	**HIF1B**	Aryl Hydrocarbon Receptor Nuclear Translocator
ENSG00000275410	**HNF1B**	HNF1 Homeobox B
ENSG00000080824	**HSP90**	Heat Shock Protein 90 Alpha Family Class A Member 1
ENSG00000142798	**HSPG2**	Heparan Sulfate Proteoglycan 2
ENSG00000066044	**HUR/ELAVL1**	ELAV Like RNA Binding Protein 1
ENSG00000115738	**ID2**	Inhibitor Of DNA Binding 2
ENSG00000115457	**IGFBP2**	Insulin Like Growth Factor Binding Protein 2
ENSG00000163453	**IGFBP7**	Insulin Like Growth Factor Binding Protein 7
ENSG00000115008	**IL1A**	Interleukin 1 Alpha
ENSG00000125538	**IL1B**	Interleukin 1 Beta
ENSG00000136244	**IL6**	Interleukin 6
ENSG00000115232	**ITGA4**	Integrin Subunit Alpha 4

ENSG00000150093	**ITGB1**	Integrin Subunit Beta 1
ENSG00000008083	**JARID2**	Jumonji And AT-Rich Interaction Domain Containing 2
ENSG00000132510	**JMJD3/KDM6B**	Lysine Demethylase 6B
ENSG00000107643	**JNK/MAPK8**	Mitogen-Activated Protein Kinase 8
ENSG00000177456	**JUN**	Jun Proto-Oncogene, AP-1 Transcription Factor Subunit
ENSG00000121361	**KCNJ8**	Potassium Inwardly Rectifying Channel Subfamily J Member 8
ENSG00000157404	**KIT**	KIT Proto-Oncogene, Receptor Tyrosine Kinase
ENSG00000111796	**KLRB1**	Killer Cell Lectin Like Receptor B1
ENSG00000198910	**L1CAM**	L1 Cell Adhesion Molecule
ENSG00000149357	**LAMTOR1**	Late Endosomal/Lysosomal Adaptor, MAPK And MTOR Activator 1
ENSG00000131023	**LATS1**	Large Tumor Suppressor Kinase 1
ENSG00000150457	**LATS2**	Marge Tumor Suppressor Kinase 2
ENSG00000128342	**LIF**	LIF, Interleukin 6 Family Cytokine
ENSG00000106683	**LIMK1**	LIM Domain Kinase 1
ENSG00000182541	**LIMK2**	LIM Domain Kinase 2
ENSG00000118046	**LKB1**	Serine/Threonine Kinase 11
ENSG00000162761	**LMX1A**	LIM Homeobox Transcription Factor 1 Alpha
ENSG00000171236	**LRG1**	Leucine Rich Alpha-2-Glycoprotein 1
ENSG00000138095	**LRPPRC**	Leucine Rich Pentatricopeptide Repeat Containing
ENSG00000139329	**LUM**	Lumican
ENSG00000133800	**LYVE1**	Lymphatic VesselENdothelial Hyaluronan Receptor 1
ENSG00000099949	**LZTR1**	Leucine Zipper Like Transcription Regulator 1
ENSG00000204103	**MAFB**	MAF BZIP Transcription Factor B
ENSG00000197971	**MBP**	Myelin Basic Protein
ENSG00000110492	**MDK**	Midkine
ENSG00000135679	**MDM2**	MDM2 Proto-Oncogene
ENSG00000169032	**MEK1/MAP2K1**	Mitogen-Activated Protein Kinase Kinase 1
ENSG00000126934	**MEK2/MAP2K2**	Mitogen-Activated Protein Kinase Kinase 2
ENSG00000065559	**MEK4/MAP2K4**	Mitogen-Activated Protein Kinase Kinase 4
ENSG00000076984	**MEK7/MAP2K7**	Mitogen-Activated Protein Kinase Kinase 7
ENSG00000005102	**MEOX1**	Mesenchyme Homeobox 1
ENSG00000106484	**MEST**	Mesoderm Specific Transcript
ENSG00000105976	**MET/HGFR**	MET Proto-Oncogene, Receptor Tyrosine Kinase
ENSG00000197614	**MFAP5**	Microfibril Associated Protein 5
ENSG00000187098	**MITF**	Melanocyte Inducing Transcription Factor
ENSG00000076984	**MKK7/MAP2K7**	Mitogen-Activated Protein Kinase Kinase 7
ENSG00000138722	**MMRN1**	Multimerin 1
ENSG00000079931	**MOXD1**	Monooxygenase DBH Like 1
ENSG00000158887	**MPZ**	Myelin Protein Zero
ENSG00000260314	**MRC1**	Mannose Receptor C-Type 1
ENSG00000163132	**MSX1**	Msh Homeobox 1
ENSG00000198793	**MTOR**	Mechanistic Target Of Rapamycin Kinase

ENSG00000118513	**MYB**	MYB Proto-Oncogene, Transcription Factor
ENSG00000065534	**MYLK**	Myosin Light Chain Kinase
ENSG00000067798	**NAV3**	Neuron Navigator 3
ENSG00000149294	**NCAM**	Neural Cell Adhesion Molecule 1
ENSG00000189430	**NCR1**	Natural Cytotoxicity Triggering Receptor 1
ENSG00000132688	**NES**	Nestin
ENSG00000162992	**NEUROD1**	Neuronal Differentiation 1
ENSG00000178403	**NEUROG2**	Neurogenin 2
ENSG00000196712	**NF1**	Neurofibromin 1
ENSG00000134259	**NGF**	Nerve Growth Factor
ENSG00000064300	**NGFR**	Nerve Growth Factor Receptor
ENSG00000116962	**NID1**	Nidogen 1
ENSG00000087303	**NID2**	Nidogen 2
ENSG00000101198	**NKAIN4**	Sodium/Potassium Transporting ATPase Interacting 4
ENSG00000183691	**NOG**	Noggin
\	**NOTCH**	(family) Notch Receptor
ENSG00000157168	**NRG1**	Neuregulin 1
ENSG00000185652	**NT3**	Neurotrophin 3
ENSG00000174546	**NUPR1**	Nuclear Protein 1, Transcriptional Regulator
ENSG00000204531	**OCT4/POU5F1**	POU Class 5 Homeobox 1
ENSG00000185668	**OCT6/POU3F1**	POU Class 3 Homeobox
ENSG00000184221	**OLIG1**	Oligodendrocyte Transcription Factor 1
ENSG00000177468	**OLIG3**	Oligodendrocyte Transcription Factor 3
ENSG00000198561	**P120/CTNND1**	Catenin Delta 1
ENSG00000112062	**P38/MAPK14**	Mitogen-Activated Protein Kinase 14
ENSG00000149269	**PAK1**	P21 (RAC1) Activated Kinase 1
ENSG00000135903	**PAX3**	Paired Box 3
ENSG00000112511	**PCL1/PHF1**	PHD Finger Protein 1
ENSG00000143033	**PCL2/MTF2**	Metal Response Element Binding Transcription Factor 2
ENSG00000119403	**PCL3/PHF19**	PHD Finger Protein 19
ENSG00000188389	**PD1**	Programmed Cell Death 1
ENSG00000197461	**PDGFA**	Platelet Derived Growth Factor Subunit A
ENSG00000100311	**PDGFB**	Platelet Derived Growth Factor Subunit B
ENSG00000134853	**PDGFRA**	Platelet Derived Growth Factor Receptor Alpha
ENSG00000113721	**PDGFRB**	Platelet Derived Growth Factor Receptor Beta
ENSG00000120217	**PDL1**	CD274 Molecule
ENSG00000261371	**PECAM1**	Platelet AndENdothelial Cell Adhesion Molecule 1
ENSG00000139289	**PHLDA1**	Pleckstrin Homology Like Domain Family A Member 1
ENSG00000109132	**PHOX2B**	Paired Like Homeobox 2B
ENSG00000121879	**PI3K**	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha
ENSG00000145287	**PLAC8**	Placenta Associated 8
ENSG00000118495	**PLAGL1**	PLAG1 Like Zinc Finger 1

ENSG00000123560	**PLP(1)**	Proteolipid Protein
ENSG00000109099	**PMP22**	Peripheral Myelin Protein 22
ENSG00000152192	**POU4F1**	POU Class 4 Homeobox 1
ENSG00000116132	**PRRX1**	Paired Related Homeobox 1
ENSG00000237412	**PRSS56**	Serine Protease 56
ENSG00000185920	**PTCH**	Patched 1
ENSG00000171862	**PTEN**	Phosphatase And Tensin Homolo
ENSG00000105894	**PTN**	Pleiotrophin
ENSG00000136238	**RAC1**	Rac Family Small GTPase 1
ENSG00000132155	**RAF(1)**	Raf-1 Proto-Oncogene, Serine/Threonine Kinase
ENSG00000133703	**(K,N,H)RAS**	RAS Proto-Oncogene, GTPase
ENSG00000139687	**RB(1)**	RB Transcriptional Corepressor 1
ENSG00000162521	**RBBP4**	RB Binding Protein 4, Chromatin Remodeling Factor
ENSG00000102054	**RBBP7**	RB Binding Protein 7, Chromatin Remodeling Factor
ENSMUSG00000061100	**RETNLA**	(*Mus musculus* only) Resistin Like Alpha
ENSMUSG00000022651	**RETNLG**	(*Mus musculus* only) Resistin Like Gamma
ENSG00000143248	**RGS5**	Regulator Of G Protein Signaling 5
ENSG00000163914	**RHO**	Rhodopsin
ENSG00000143878	**RHOB**	Ras Homolog Family Member B
ENSG00000067900	**ROCK**	Rho Associated Coiled-Coil Containing Protein Kinase 1
ENSG00000143546	**S100A8**	S100 Calcium Binding Protein A8
ENSG00000160307	**S100B**	S100 Calcium Binding Protein B
ENSG00000108443	**S6K1**	Ribosomal Protein S6 Kinase B1
ENSG00000049130	**SCF**	Stem Cell Factor, KIT Ligand
ENSG00000136546	**SCN7A**	Sodium Voltage-Gated Channel Alpha Subunit 7
ENSG00000168497	**SDPR**	Caveolae Associated Protein 2
ENSG00000135622	**SEMA4F**	Semaphorin 4F
ENSG00000132386	**SERPINF1**	Serpin Family F Member 1
ENSG00000160691	**SHC**	SHC Adaptor Protein 1
ENSG00000164690	**SHH**	Sonic Hedgehog Signaling Molecul
ENSG00000179295	**SHP2/PTPN11**	Protein Tyrosine Phosphatase Non-Receptor Type 11
ENSMUSG00000051504	**SIGLECH**	(*Mus musculus* only) Sialic Acid Binding Ig-like Lectin H
ENSMUSG00000030237	**SLCO1A4**	(*Mus musculus* only) Solute Carrier Organic Anion Transporter Family, member 1a4
ENSG00000108055	**SMC3**	Structural Maintenance Of Chromosomes 3
ENSG00000128602	**SMO**	Smoothened, Frizzled Class Receptor
ENSG00000124216	**SNAI1**	Snail Family Transcriptional Repressor 1
ENSG00000019549	**SNAI2**	Snail Family Transcriptional Repressor 2
ENSG00000115904	**SOS1**	SOS Ras/Rac Guanine Nucleotide Exchange Factor 1
ENSG00000181449	**SOX2**	SRY-Box Transcription Factor 2
ENSG00000005513	**SOX8**	SRY-Box Transcription Factor 8
ENSG00000125398	**SOX9**	SRY-Box Transcription Factor 9
ENSG00000100146	**SOX10**	SRY-Box Transcription Factor 10

ENSG00000118785	**SPP1**	Osteopontin, Secreted Phosphoprotein 1
ENSG00000166068	**SPRED1**	Sprouty Related EVH1 Domain Containing 1
ENSG00000184895	**SRY**	Sex Determining Region Y
ENSG00000168610	**STAT3**	Signal Transducer And Activator Of Transcription 3
ENSG00000178691	**SUZ12**	SUZ12 PRC2 Subunit
ENSG00000141510	**T(R)P53**	Tumor Protein P53
ENSG00000018408	**TAZ**	WW Domain Containing Transcription Regulator 1
ENSG00000071564	**TCF3**	Transcription Factor 3
ENSG00000196628	**TCF4**	Transcription Factor 4
ENSG00000108064	**TFAM**	Transcription Factor A, Mitochondrial
ENSG00000137203	**TFAP2A**	Transcription Factor AP-2 Alpha
ENSG00000105329	**TGFB**	Transforming Growth Factor Beta 1
ENSG00000035862	**TIMP2**	TIMP Metallopeptidase Inhibitor 2
ENSG00000104067	**TJP1**	Tight Junction Protein 1
ENSG00000041982	**TNC**	Tenascin C
ENSG00000232810	**TNF**	Tumor Necrosis Factor
\	**(td)Tom(ato)**	(*Mus musculus* only) tandem Tomato
ENSG00000197253	**TPSB2**	Tryptase Beta 2
ENSG00000095970	**TREM2**	Triggering Receptor Expressed On Myeloid Cells 2
ENSG00000258947	**TUBB3**	Tubulin Beta 3 Class III
ENSG00000122691	**TWIST1**	Twist Family BHLH Transcription Factor 1
ENSG00000233608	**TWIST2**	Twist Family BHLH Transcription Factor 2
ENSG00000154277	**UCHL1**	Ubiquitin C-Terminal Hydrolase L1
ENSG00000112715	**VEGF(A)**	Vascular Endothelial Growth Factor A
ENSG00000102755	**VEGFR1**	Vascular Endothelial Growth Factor Receptor 1
ENSG00000128052	**VEGFR2**	Kinase Insert Domain Receptor
ENSG00000026025	**VIM**	Vimentin
ENSG00000156076	**WIF1**	WNT Inhibitory Factor 1
ENSG00000125084	**WNT1**	Wnt Family Member 1
ENSG00000108379	**WNT3**	Wnt Family Member 3
ENSG00000173578	**XCR1**	X-C Motif Chemokine Receptor 1
ENSG00000137693	**YAP**	Yes1 Associated Transcriptional Regulator
\	**YFP**	Yellow Fluorescent Protein
ENSG00000148516	**ZEB1**	Zinc Finger E-Box Binding Homeobox 1
ENSG00000169554	**ZEB2**	Zinc Finger E-Box Binding Homeobox 2
ENSG00000156925	**ZIC3**	Zic Family Member 3
ENSG00000174963	**ZIC4**	Zic Family Member 4

2 R packages in the R 3.6.3 Singularity container

The table below lists the R packages installed in the R 3.6.3 Singularity container. The packages are in the order of dependencies, meaning that when installing the n^{th} package, its dependencies are among the $n-1$ previously installed packages.

order	package name	version
1	acepack	1.4.1
2	ADGofTest	0.3
3	backports	1.2.1
4	base64enc	0.1-3
5	BH	1.72.0-3
6	bit	4.0.4
7	bitops	1.0-6
8	brew	1.0-6
9	brio	1.1.2
10	carData	3.0-4
11	caTools	1.18.0
12	commonmark	1.7
13	corrplot	0.84
14	cpp11	0.4.0
15	curl	4.3
16	emoa	0.5-0.1
17	fansi	0.4.1
18	farver	2.0.3
19	fastcluster	1.2.3
20	fastmap	1.1.0
21	fastmatch	1.1-0
22	findpython	1.0.7
23	FNN	1.1.3
24	formatR	1.7
25	futile.options	1.0.1
26	GenomeInfoDbData	1.2.2
27	gitcreds	0.1.1
28	gower	0.2.1
29	gsl	2.1-6
30	HiddenMarkov	1.8-13
31	highr	0.8
32	ica	1.0-2
33	ini	0.3.1
34	jpeg	0.1-8.1

order	package name	version
35	labeling	0.3
36	lambda.r	1.2.4
37	lazyeval	0.2.2
38	listenv	0.8.0
39	lsei	1.2-0
40	magrittr	2.0.1
41	matrixStats	0.56.0
42	mco	1.0-15.1
43	misc3d	0.8-4
44	nloptr	1.2.2.1
45	numDeriv	2016.8-1.1
46	plogr	0.2.0
47	png	0.1-7
48	polyclip	1.10-0
49	praise	1.0.0
50	prettyunits	1.1.1
51	R6	2.4.1
52	RANN	2.6.1
53	rappdirs	0.3.3
54	RColorBrewer	1.1-2
55	RcppParallel	5.1.4
56	RcppProgress	0.4.2
57	rematch	1.0.1
58	Rhdf5lib	1.8.0
59	rjson	0.2.20
60	rprojroot	2.0.2
61	rstudioapi	0.13
62	sourcetools	0.1.7
63	SQUAREM	2020.2
64	sys	3.3
65	systemfonts	1.0.4
66	tsne	0.1-3
67	tzdb	0.1.2
68	utf8	1.1.4

order	package name	version
69	viridisLite	0.3.0
70	whisker	0.4
71	wk	0.5.0
72	yaml	2.2.1
73	zip	2.2.0
74	zlibbioc	1.32.0
75	codetools	0.2-16
76	askpass	1.1
77	assertthat	0.2.1
78	BiocManager	1.30.10
79	checkmate	2.0.0
80	clipr	0.7.0
81	digest	0.6.25
82	downloader	0.4
83	futile.logger	1.4.3
84	globals	0.14.0
85	here	1.0.1
86	iterators	1.0.12
87	mime	0.9
88	openssl	1.4.1
89	pkgconfig	2.0.3
90	ps	1.6.0
91	R.methodsS3	1.8.1
92	Rhtslib	1.18.1
93	rlang	1.0.2
94	rvcheck	0.1.8
95	snow	0.4-3
96	svglite	2.0.0
97	beeswarm	0.4.0
98	boot	1.3-25
99	cachem	1.0.6
100	DEoptimR	1.0-9
101	ellipsis	0.3.2
102	fdrtool	1.2.15
103	foreach	1.5.0
104	Formula	1.2-3
105	gridGraphics	0.5-0
106	gtable	0.3.0
107	htmltools	0.5.2
108	jquerylib	0.1.4
109	KernSmooth	2.23-17

order	package name	version
110	lattice	0.20-41
111	latticeExtra	0.6-29
112	locfit	1.5-9.4
113	memoise	2.0.0
114	nlme	3.1-147
115	nnet	7.3-14
116	parallelly	1.28.1
117	pbapply	1.4-2
118	polynom	1.4-0
119	processx	3.5.2
120	proxy	0.4-24
121	pspline	1.0-18
122	purrr	0.3.4
123	randomForest	4.6-14
124	scatterplot3d	0.3-41
125	shape	1.4.6
126	stabledist	0.7-1
127	statmod	1.4.34
128	stringi	1.4.6
129	vipor	0.4.5
130	withr	2.5.0
131	xfun	0.35
132	xopen	1.0.0
133	xtable	1.8-4
134	zoo	1.8-10
135	cluster	2.1.0
136	rpart	4.1-15
137	abind	1.4-5
138	BBmisc	1.11
139	BiocGenerics	0.32.0
140	BiocParallel	1.20.1
141	bit64	4.0.5
142	callr	3.7.0
143	clue	0.3-60
144	coda	0.19-4
145	colorspace	1.4-1
146	crayon	1.3.4
147	data.table	1.14.2
148	data.tree	1.0.0
149	DBI	1.1.1
150	DiceKriging	1.5.6
151	diffobj	0.3.5

order	package name	version
152	doParallel	1.0.15
153	evaluate	0.18
154	fontawesome	0.2.2
155	foreign	0.8-76
156	fs	1.5.2
157	future	1.22.1
158	future.apply	1.5.0
159	generics	0.1.0
160	GlobalOptions	0.1.2
161	glue	1.4.2
162	graph	1.64.0
163	gridExtra	2.3
164	gtools	3.8.2
165	hexbin	1.28.1
166	jsonlite	1.7.2
167	kernlab	0.9-29
168	lifecycle	1.0.1
169	limma	3.42.2
170	lmtest	0.9-38
171	markdown	1.1
172	MASS	7.3-54
173	Matrix	1.3-2
174	MatrixModels	0.4-1
175	modeltools	0.2-23
176	munsell	0.5.0
177	mvtnorm	1.1-1
178	npsurv	0.4-0
179	parallelMap	1.5.0
180	ParamHelpers	1.14
181	pcaPP	1.9-73
182	pdist	1.2
183	plot3D	1.3
184	R.oo	1.24.0
185	R.utils	2.11.0
186	Rcpp	1.0.9
187	RcppAnnoy	0.0.16
188	RcppArmadillo	0.9.870.2.0
189	RcppEigen	0.3.3.7.0
190	RcppHNSW	0.2.0
191	RcppTOML	0.1.7
192	RCurl	1.98-1.2
193	remotes	2.4.2
194	reticulate	1.26
195	rhdf5	2.30.1
196	rjags	4-13
197	RJSONIO	1.3-1.4
198	robustbase	0.93-9
199	RSpectra	0.16-0
200	rsvd	1.0.3
201	Rtsne	0.15
202	s2	1.0.7
203	S4Vectors	0.24.4
204	sandwich	3.0-1
205	sass	0.4.0
206	scales	1.1.1
207	segmented	1.5-0
208	sitmo	2.0.1
209	softImpute	1.4
210	sp	1.4-1
211	SparseM	1.78
212	stringr	1.4.0
213	survival	3.2-13
214	TH.data	1.1-0
215	timeDate	3043.102
216	tinytex	0.38
217	triebeard	0.3.0
218	tweenr	1.0.1
219	units	0.7-2
220	urltools	1.7.3
221	vcd	1.4-8
222	vctrs	0.3.8
223	webshot	0.5.2
224	xgboost	1.4.1.1
225	XML	3.99-0.3
226	xml2	1.3.2
227	xts	0.12.1
228	class	7.3-17
229	mgcv	1.8-31
230	ape	5.3
231	argparse	2.1.5
232	Biobase	2.46.0
233	BiocNeighbors	1.4.2
234	blob	1.2.1
235	bslib	0.3.1

order	package name	version
236	circlize	0.4.16
237	cli	3.0.1
238	copula	1.0-0
239	credentials	1.3.1
240	crosstalk	1.1.0.1
241	desc	1.4.1
242	dqrng	0.2.1
243	e1071	1.7-3
244	edgeR	3.28.1
245	fitdistrplus	1.0-14
246	GA	3.2
247	gdata	2.18.0
248	gert	1.4.1
249	GetoptLong	1.0.5
250	gplots	3.0.3
251	hms	1.1.1
252	htmlwidgets	1.5.4
253	httr	1.4.2
254	igraph	1.2.5
255	influenceR	0.1.0
256	IRanges	2.20.2
257	irlba	2.3.3
258	knitr	1.41
259	knn.covertree	1.0
260	laeken	0.5.2
261	later	1.3.0
262	lava	1.6.7
263	leiden	0.3.3
264	lhs	1.0.2
265	libcoin	1.0-9
266	lubridate	1.7.9
267	maptools	1.0-1
268	minqa	1.2.4
269	mixtools	1.2.0
270	ModelMetrics	1.2.2.2
271	multcomp	1.4-17
272	openxlsx	4.1.5
273	pcaMethods	1.78.0
274	pillar	1.6.3
275	pkgbuild	1.3.1
276	pkgload	1.2.2
277	plyr	1.8.6
278	princurve	2.1.4
279	pROC	1.16.2
280	prodlim	2019.11.13
281	progress	1.2.2
282	promises	1.1.0
283	proxyC	0.1.5
284	pryr	0.1.4
285	quantreg	5.55
286	ranger	0.12.1
287	reshape	0.8.8
288	reshape2	1.4.4
289	rje	1.10.15
290	rmarkdown	2.18
291	ROCR	1.0-7
292	roxygen2	7.1.2
293	RSQLite	2.2.0
294	rversions	2.1.1
295	selectr	0.4-2
296	sessioninfo	1.1.1
297	tibble	3.1.5
298	tidyselect	1.1.0
299	TTR	0.24.2
300	uwot	0.1.8
301	visNetwork	2.0.9
302	vroom	1.5.5
303	XVector	0.26.0
304	AnnotationDbi	1.48.0
305	Biostrings	2.54.0
306	carrier	0.1.0
307	cellranger	1.1.0
308	classInt	0.4-3
309	coin	1.4-2
310	ComplexHeatmap	2.13.1
311	DelayedArray	0.12.3
312	DO.db	2.9
313	dplyr	1.0.7
314	dynutils	1.0.5.9000
315	europepmc	0.3
316	forcats	0.5.0
317	GenomeInfoDb	1.22.1
318	GenomicRanges	1.38.0
319	gh	1.3.0

order	package name	version
320	GO.db	3.10.0
321	GOSemSim	2.13.1
322	graphlayouts	0.7.0
323	HDF5Array	1.14.4
324	htmlTable	1.13.3
325	httpuv	1.5.2
326	ipred	0.9-12
327	lmds	0.1.0
328	lme4	1.1-27.1
329	org.Mm.eg.db	3.10.0
330	pbkrtest	0.4-8.6
331	rcmdcheck	1.3.3
332	readr	2.0.2
333	readxl	1.3.1
334	rematch2	2.1.1
335	Rsamtools	2.2.3
336	rvest	0.3.5
337	sf	1.0-3
338	shiny	1.7.1
339	smoother	1.1
340	SummarizedExperiment	1.16.1
341	tidyr	1.1.4
342	usethis	2.0.1
343	waldo	0.3.1
344	annotate	1.64.0
345	babelgene	22.3
346	babelwhale	1.0.1
347	beachmat	2.2.1
348	BiocSingular	1.2.2
349	broom	0.7.9
350	dbplyr	1.4.4
351	DelayedMatrixStats	1.8.0
352	dyndimred	1.0.3
353	genefilter	1.68.0
354	geneplotter	1.64.0
355	GenomicAlignments	1.22.1
356	gridpattern	0.3.1
357	GSEABase	1.48.0
358	haven	2.3.1
359	kableExtra	1.3.4
360	msigdbr	7.5.1
361	recipes	0.1.17

order	package name	version
362	rio	0.5.16
363	rtracklayer	1.46.0
364	SingleCellExperiment	1.8.0
365	slingshot	1.4.0
366	testthat	3.1.0
367	tidygraph	1.1.2
368	zinbwave	1.8.0
369	AUCell	1.8.0
370	BiocFileCache	1.10.2
371	biomaRt	2.42.1
372	car	3.0-11
373	devtools	2.4.3
374	DropletUtils	1.6.1
375	gridtext	0.1.1
376	isoband	0.2.1
377	rstatix	0.5.0
378	scds	1.2.0
379	VIM	6.1.1
380	ggplot2	3.3.5
381	ggplotify	0.0.5
382	ggrepel	0.9.1
383	ggridges	0.5.2
384	ggsci	2.9
385	ggsignif	0.6.0
386	ggtext	0.1.0
387	ggthemes	4.2.4
388	ggvenn	0.1.9
389	infercnv	1.2.1
390	mlr	2.17.1
391	patchwork	1.1.2
392	plotly	4.9.2.1
393	qvalue	2.18.0
394	sctransform	0.2.1
395	smoof	1.6.0.2
396	UpSetR	1.4.0
397	viridis	0.5.1
398	caret	6.0-86
399	cowplot	1.0.0
400	dendextend	1.15.1
401	DiagrammeR	1.0.6.1
402	EnhancedVolcano	1.4.0
403	fgsea	1.12.0

order	package name	version
404	ggbeeswarm	0.6.0
405	ggdendro	0.1-20
406	ggforce	0.3.1
407	ggpattern	0.3.1
408	ggplot.multistats	1.0.0
409	ggpubr	0.3.0
410	ggraph	2.0.3
411	harmony	0.1.0
412	Hmisc	4.4-0
413	mlrMBO	1.1.4
414	randomForestSRC	2.12.1
415	scater	1.14.6
416	scran	1.14.6
417	Seurat	3.1.5
418	batchelor	1.2.4
419	clustree	0.4.3
420	DESeq2	1.26.0
421	destiny	3.0.1
422	DOSE	3.12.0
423	dynparam	1.0.0
424	dynwrap	1.2.1
425	enrichplot	1.6.1
426	gng	0.1.0
427	nichenetr	0.1.0
428	scBFA	1.0.0
429	scDblFinder	1.1.8
430	clusterProfiler	3.14.3
431	dynfeature	1.0.0.9000
432	dynmethods	1.0.5
433	dynplot	1.0.2.9000
434	dyneval	0.9.9
435	TInGa	0.0.0.9000
436	hdf5r	1.3.2
437	ggwordcloud	0.5.0

3 Composition of Conda environments

This section details the version of tools in the two Conda environments related to RNA velocity (Materials and Methods, Section 4.2.2).

tool	version
click	8.0.4
cython	3.0.0
h5py	3.7.0
matplotlib	3.7.2
numba	0.56.4
numpy	1.23.5
samtools	1.9
scikit-learn	1.3.0
scipy	1.10.1
velocyto	0.17.17

Table 15 – Conda environment to use velocyto. The versions written in red were forced while creating the environment. Including the dependencies installed automatically in the latest compatible version, this environment contains 117 tools.

tool	version	comment
ipywidgets	8.1.1	installed to suppress a warning
leidenalg	0.10.1	
matplotlib	3.5.0	3.8.0 leads to AttributeError and mplDeprecation 3.4.1 leads to "cannot import colormaps" error
numba	0.57.1	
numpy	1.21.1	> 1.21.1 leads to ValueError in scvelo
pandas	1.5.3	scvelo requires pandas < 2
python-igraph	0.10.8	
scanpy	1.9.5	
scipy	1.10.1	
scvelo	0.2.5	

Table 16 – Conda environment to use scvelo. The versions written in red were forced while creating the environment. Including the dependencies installed automatically in the latest compatible version, this environment contains 224 tools.

4 Supplementary materials

4.1 The Seurat object format

The Seurat package (Stuart *et al.*, 2019), a suite of tools for scRNA-Seq processing and analysis, was used to prepare data for downstream analysis. This package provides a complete structure for data storage (Figure 79).

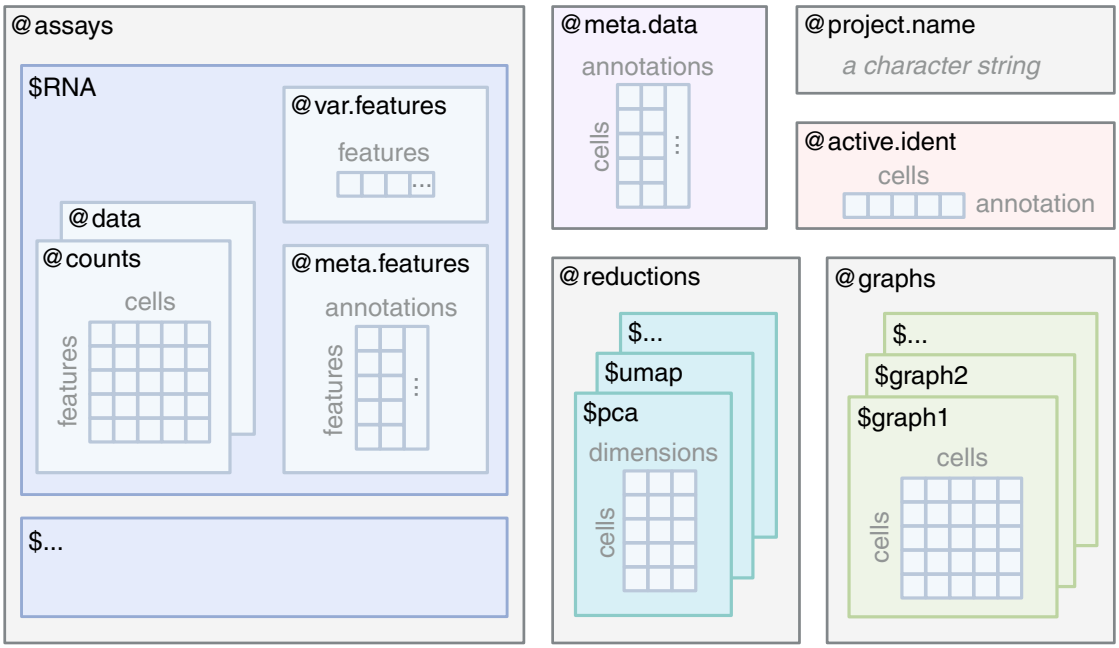


Figure 79 – Overview of the Seurat object structure, in Seurat version 3. A Seurat object (**sobj**) is a data structure combining several simple R data formats, such as matrices, data-frames or character strings, shown in gray. The three small dots highlight where new elements can be added to the object. All simple formats are organized in a nested structure, shown as rectangles. Words in black indicate how each component can be extracted from the object, e.g. **sobj@assays$RNA@meta.features**. Only elements treated in next sections are depicted in the figure. The word "slot" corresponds to elements accessible by the @ symbol.

Each Seurat object is associated with a project name, i.e. a simple character string. Initially, the active identifiers of cells (**active.ident**) corresponds to the project name but can be easily modified. I used this slot to perform differential expression analysis (Methods, Section 4.1.1). All cell annotations, such as cell type, cell cycle or quality metrics, are stored in the metadata slot.

The initial main content of a scRNA-Seq data object is the UMI count matrix, a cells-by-features matrix, stored in the **RNA** assay of a Seurat object. Features, such as genes, can be annotated in the **meta.features** data-frame (Methods, Section

3.1.2). A subset of features, such as HVGs, can be stored in **var.features** for further processing. All of these elements are related to the **RNA** assay, but the **assays** slot can contain multiple assays (next Section 4.2). Further processing, such as dimensionality reductions and clustering, are stored in the **reductions** and **graphs** slots (Methods, Section 3.1.3).

4.2 Additional assays

To answer specific questions, the Seurat object was enriched with new assays (Figure 80), including the CNV assay. Please refer to the Section 4.3 in Materials and Methods chapter for methodological details regarding the CNV analysis.

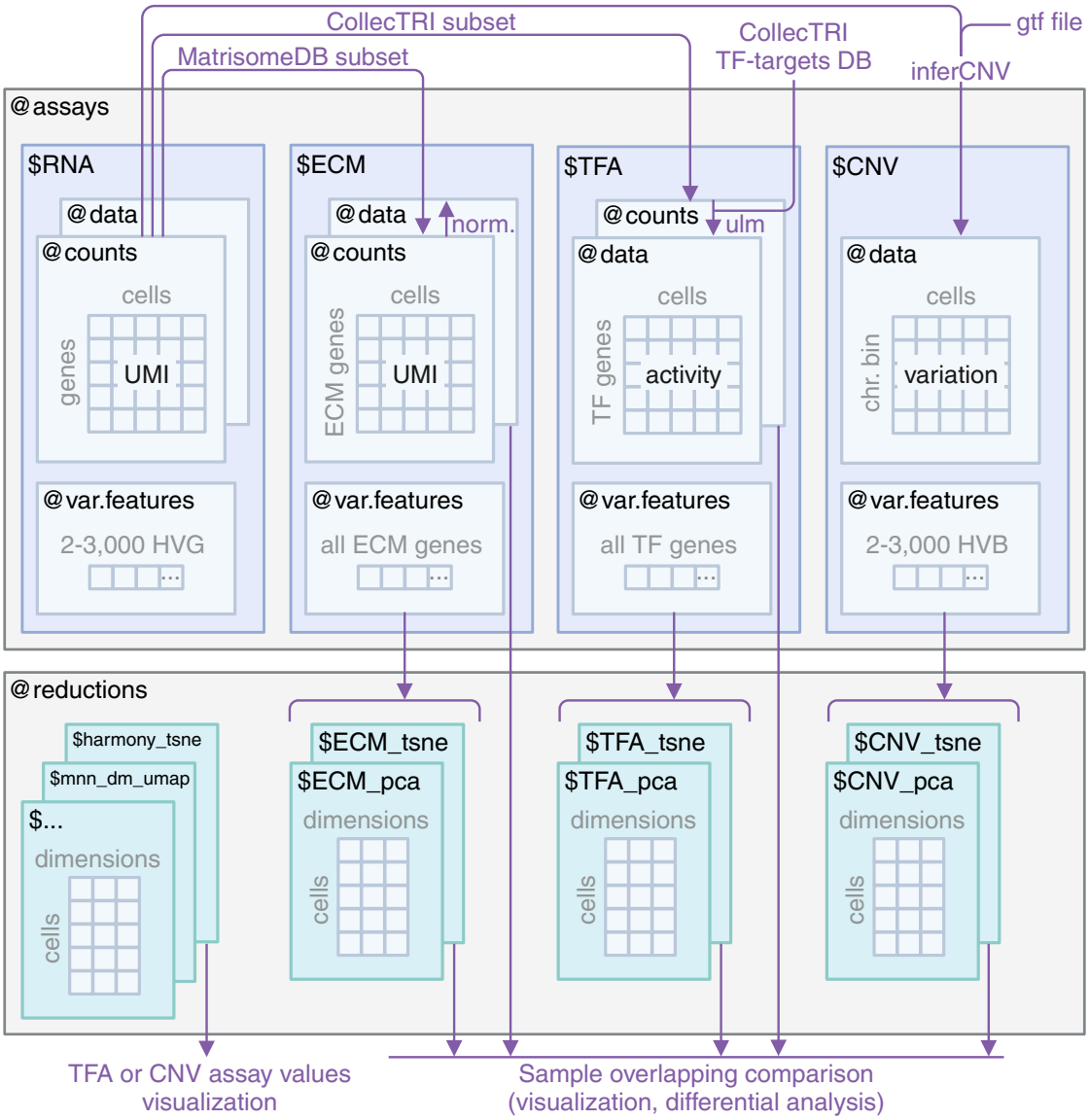


Figure 80 – Construction of additional assays and their usage. This figure completes the Figure 79 with additional assays and their associated reductions. HVG/B: highly variable genes/bins, chr.: chromosome, norm.: normalization, TF: transcription factors, ulm: univariate linear model.

4.3 Differential expression analyses

In this paragraph, I provide details regarding differential expression analysis using the Seurat::FindMarkers function. As shown in the code section below, the two populations of interest are passed to the function through ident.1 and ident.2 parameters. Before, the column of interest, in the metadata slot of Seurat object, is set as active identifiers of cells.

```
Seurat::Idents(sobj) = sobj$COLUMN_OF_INTEREST
mark = Seurat::FindMarkers(sobj,
                           assay = "RNA",
                           ident.1 = "population 1",
                           ident.2 = "population 2")
```

The function outputs a data-frame, here named "mark", consisting of 5 columns, rows corresponding to genes:

- **avg_logFC**: log fold-change of the average gene expression between population 1 and population 2 ;
- **pct.1**: proportion of cells in population 1 within strictly positive expression for the gene. I named this value the **representativness** of the gene. If the gene is perfectly representative of the population 1, pct.1 is equal to 1 ;
- **pct.2**: proportion of cells in population 2 within strictly positive expression for the gene. I named this value the **selectivity** of the gene. If the gene is selective of the population 1, pct.2 is equal to 0 ;
- **pval**: p-value of the Wilcoxon test ;
- **p_val_adj**: p-value to the Wilcoxon test, corrected by the Bonferroni method, meaning that the p-value is multiplied by the number of genes in the count matrix. This value was used to define gene as "differentially expressed" (DE), with a threshold set to 0.05.

The metrics pct.1 and pct.2 were particularly used to identify a suitable set of markers of the cell type annotation (Section 6, Poster 1).

The relation between performing differential expression between population 1 and 2, or the opposite, is summarized in Table 17.

metric	population 1 vs population 2	population 2 vs population 1
average log fold-change	avg_logFC	-avg_logFC
representativness	pct.1	1 - pct.1
selectivity	pct.2	1 - pct.2
p-value	pval	pval
adjusted p-value	p_val_adj	p_val_adj

Table 17 – Output of differential expression analysis, using Seurat::FindMarkers. The `mark` differential expression results is generated using the code section above, between population 1 and population 2. The same output data-frame can be used to compare population 2 vs population 1.

The DESeq2 package takes as input the UMI count matrix, before normalization. Due to error returned by downstream functions, a count of 1 was added to all values. The standard piece of code used to perform differential expression analysis using DESEq2 package was:

```
# Define DESeq2 input
countData = Matrix::as.matrix(sobj@assays$RNA@counts + 1) # genes x cells
colData = data.frame(ident = sobj$COLUMN_OF_INTEREST)     # cells x annotation

# Build DESeq2 dataset object
dds = DESeq2::DESeqDataSetFromMatrix(countData = countData,
                                     colData = colData,
                                     design = formula(~ ident))

# Perform differential expression
dds = DESeq2::DESeq(dds, test = "LRT", reduced = formula(~ 1),
                    useT = TRUE, minmu = 1e-6, minReplicatesForReplace = Inf,
                    parallel = TRUE, BPPARAM = BiocParallel::MulticoreParam(8))

# Extract results between population 1 and population 2
res = DESeq2::results(dds, contrast = c("ident", "population 1", "population 2")) %>%
  as.data.frame() %>%
  dplyr::filter(padj < 0.05) %>%
  dplyr::arrange(-log2FoldChange)
```

5 Supplementary results

This section completes the results from Section 3.3.4. Whole Exome Sequencing (WES) data were analyzed as described in the Article 2, using the spleen of individual mice as reference. The Figure 81 highlights the consistency between infercnv results from scRNA-Seq data, showed in Figure 75. The main events: small loss in chromosome 4, gains in chromosomes 15 et 17, are shared between the two analyses.

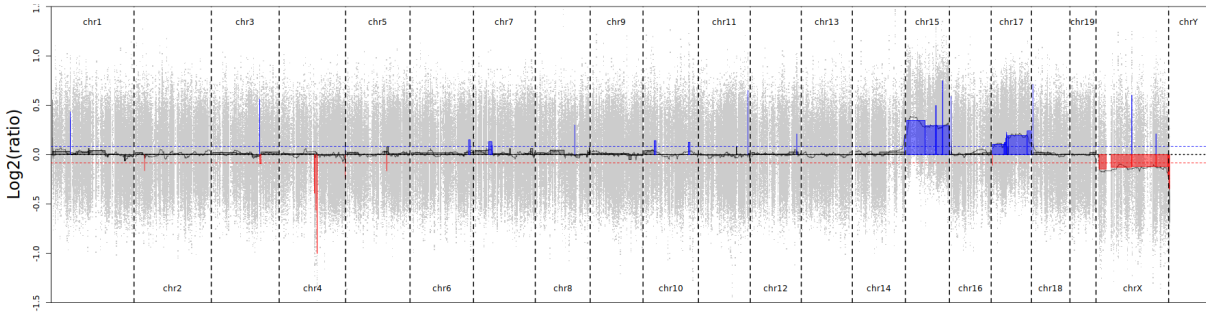


Figure 81 – Whole Exome Sequencing genomic profile of 2021_33. This *Nf1*-KO^{wt} dataset corresponds to MPNST_4 in Article 2. The Manhattan plot shows CNV detected from WES data, with gains in blue and losses in red (Courtesy of Bastien Job).

As a second example, we compared CNV results for tumor cells from the MPNST 2022_26, corresponding to a *Nf1*-KO^{het} mouse (Figure 82). This dataset is composed of tumor cells and macrophages (Figure 82A). Using infercnv on scRNA-Seq, we inferred CNV on tumor cells only, taking all immune cells from the micro-environment as reference. Most cells are characterized by gains in chromosome 19 (Figure 82B). The results suggest two main clones: one (clusters 1 and 3) depicting gains in chromosomes 11 and 15, and a second (other clusters) with loss in chromosome 2, contains subclones with gains in chromosomes 6, 10 and X. The main alterations detected by infercnv are consistent with the WES data (Figure 82C). Consistently, a projection of the infercnv table using tSNE, and annotation of clusters, showed that clusters 1 and 3 are isolated from the main population (Figure 82D). On the expression-based representation, despite the existence of genetic subclones, the main clone is characterized by an homogeneous transcriptomic profile. Strikingly, the second clone depicts a distinct transcriptomic profile (clusters 1 and 3 on the left) (Figure 82E).

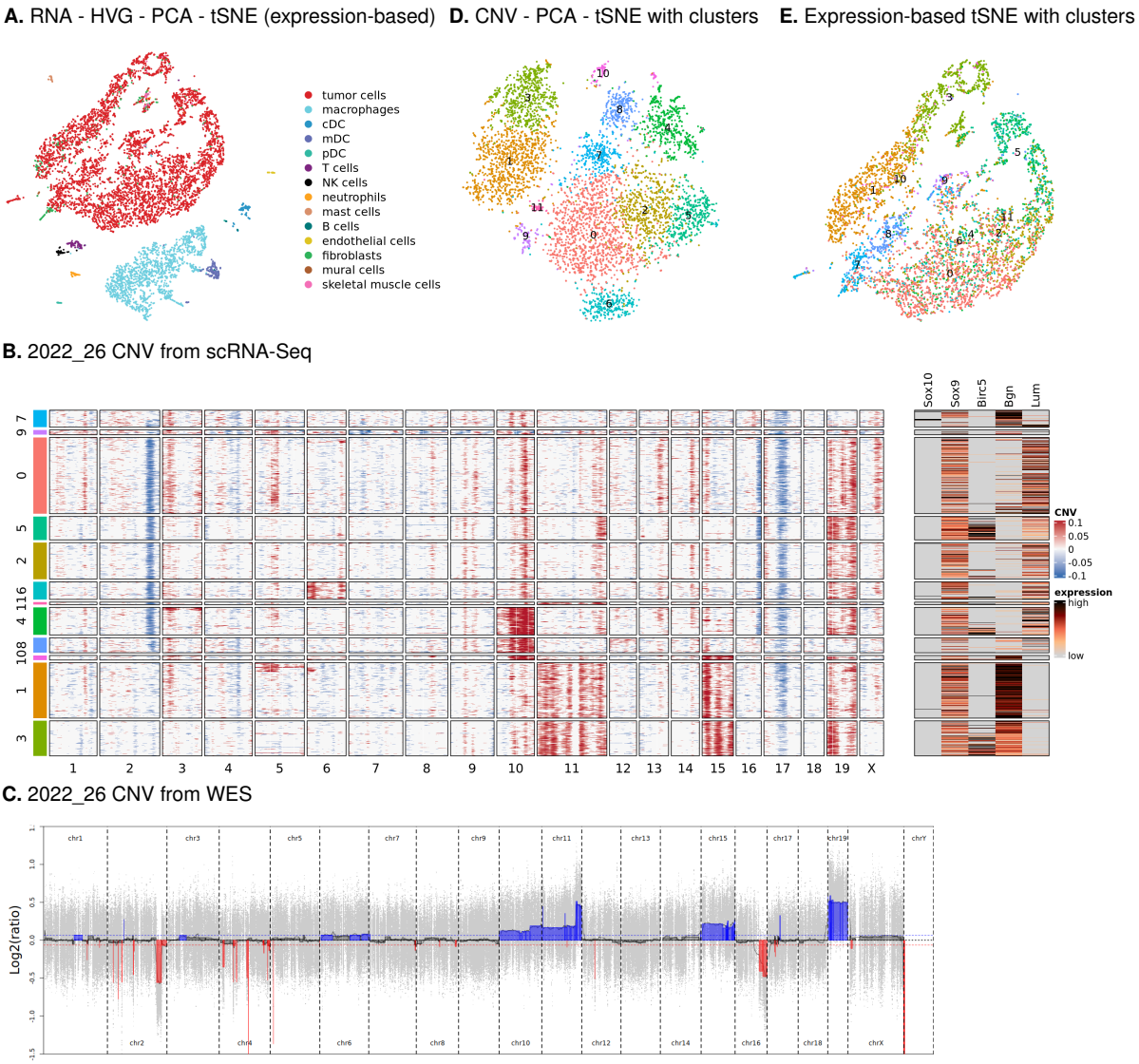


Figure 82 – Distinct tumor subpopulations are associated with specific CNV in 2022_26. This *Nf1*-KO^{het} dataset corresponds to MPNST_2 in the Article 2. CNV were inferred using immune cells as reference. **A.** Expression-based tSNE representation of 8,521 cells from 2022_26 dataset, colored according to cell types. The projection is made by identifying 3000 HVGs in the normalized gene expression matrix (RNA), running PCA and building tSNE. **B.** CNV results are represented on a heatmap, made with the ComplexHeatmap R package (Gu *et al.*, 2016). Each column corresponds to a gene. Each row corresponds to the CNV profile of a single cell, with gains in red and losses in blue. Cells are ordered by clusters. On the right, expression levels of genes are indicated. **C.** Manhattan plot showing CNV detected from WES data, with gains in blue and losses in red (Courtesy of Bastien Job). **E-D.** The clusters were identified on a PCA made from the infercnv table, containing only results for tumor cells. Tumor cells were projected on a CNV-based tSNE (**C**) or on the expression-based tSNE (**D**), and colored according to the CNV clusters. The colors of the tumor cells clusters are consistent across the panels. These figures (A,B,D,E) are extracted from the compiled notebook **explore_CNV_2022_26.html** from the flox_minus_2022_26 project.

6 Participation to scientific events

During my thesis, I attended to the following events with a poster or an oral communication:

- JOBIM 2022 (Open Days for Biology, Informatics and Mathematics): **Poster 1**, about cell type annotation
- MaDICS Symposium 2022: **Poster 1**, about cell type annotation
- Doctoral School Day 2022: **Poster 2**, about the Article 2, winner of the best CoDoPoDo poster award
- Global NF Conference 2023: **Poster 3**, about the impact of *Nf1* haploinsufficience on tumor development
- Doctoral School Day 2023: oral communication, winner of the best oral communication award
- Global NF Conference 2024: **Poster 4**, about the Article 2

Posters are attached in the next pages.

Deciphering mechanisms driving malignant transformation in Neurofibromatosis Type I from single-cell transcriptomics

Audrey ONFROY[1,2], Fanny COULPIER[1], Katarzyna RADOMSKA[1], Denis THIEFFRY[2] and Piotr TOPILKO[1]

[1] Institut Mondor de Recherche Biomédicale, Inserm U955 - Team 9, Créteil, France
[2] Institut de Biologie de l'École Normale Supérieure, UMR CNRS 8197, Inserm U1024, PSL Université, Paris, France

Contact : audrey.onfroy@inserm.fr

Poster 1

NF1 IS A GENETIC DISEASE

- Affects 1 / 3,000 births worldwide
- Bi-allelic inactivation of **Nf1 gene** in Schwann cells
- A lot of symptoms, including development of **neurofibromas** (NF) : benign nerve sheaths tumors

purpose of my PhD project

nerve sheath → plexiform NF → dysplastic NF → MPNST*

no treatment to date

*MPNST = malignant peripherical nerve sheaths tumor

Nf1-KO MOUSE MODEL

- Bi-allelic inactivation of **Nf1 gene** in Schwann cells
- Nf1 mutant cells are traced by the **Tomato** fluorescent reporter
- Nf1-KO mice share common features with NF1 patients, including development of pNFs that spontaneously evolve into MPNSTs through a dysplastic stage

What are the molecular events driving malignant transformation of pNF into MPNST, in tumor forming cell populations, in the mouse model and in NF1 patients ?

SINGLE-CELL TRANSCRIPTOMICS DATA

Single-cell RNA sequencing : for each biological sample, we have a genes x cells expression matrix

pNF	dyNF	MPNST
2	10	4
1	2 + 1 [2]	1

How to annotate cells for cell type ?

- cluster cells into groups with similar expression profiles
- manual or automatic annotation, using existing references

MARKERS IDENTIFICATION IN TUMOR FORMING POPULATIONS

Motivations

- **No existing reference** markers for tumor cells in NF1 disease to date
- In Nf1-KO mice, Nf1 mutant cells are traced by the **Tomato** fluorescent reporter
- All Nf1-KO mice datasets share similar **tumor forming cell populations**, for which we know few markers
- Distinction between tumor cells and fibroblasts in NF1 patients is difficult

→ **Refine population markers in the specific context of NF1**

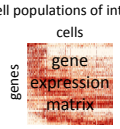
Methods

1- Input

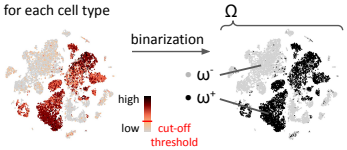
- cell types and known markers :

cell type	marker.s
A	α
B	β, δ
C	γ

- scRNA-Seq dataset containing cell populations of interest



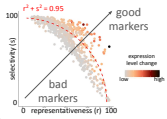
2- Identification of positive and negative populations for each cell type



Marker.s expression score (left) helps to define cells (right) related to the corresponding cell type.

3- Identification of positive markers candidates by **differential expression** between ω^+ and ω^-

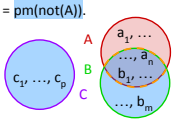
4- Markers selection, for each cell type
Goal : maximize **representativeness** (r) and **selectivity** (s)
Criteria : **max($r^2 + s^2$)**
Keep top 50 markers



representativeness (r) : % of cells in ω^+ having a positive number of counts for the marker
selectivity (s) :
% of cells in ω^- having no count for the marker

5- Output : **markers for each cell type**
The positive markers for cell type A (pm(A)) are $a_1, \ldots, a_n$. The negative markers are defined as the union of positive markers for all other cell types, i.e. nm(A) = pm(not(A)).

The intersection between pm(A) and pm(not(A)) is not necessarily empty : a gene can be a marker for several cell types.



AUTOMATIC ANNOTATION AT A SINGLE CELL LEVEL

Motivations

- Cluster-based annotation can lead to confusion between cell types
- **Expression scores** are used to predict cell cycle phase in Seurat's package [3]
- Using **negative markers** can strenghten cell type annotation

→ **Score positive and negative markers expression levels, to annotate single cells**

Methods

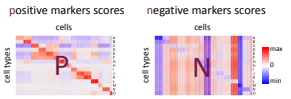
1- Input

- cell types and positive markers :

cell type	markers
A	$a_1, \ldots, a_n$
B	$b_1, \ldots, b_m$
C	$c_1, \ldots, c_p$
…	…
O	$o_1, \ldots, o_m$

- scRNA-Seq dataset to annotate

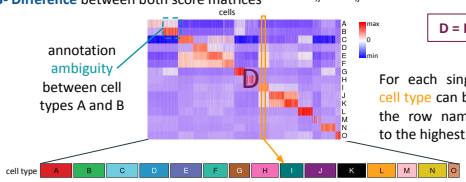
2- Markers expression scoring with Seurat



For cell type **i** and cell **j** :

- $P_{i,j} > 0$: **j** expresses positive markers for **i**
- $N_{i,j} > 0$: **j** expresses negative markers for **i**
- $P_{i,j}$ - | $N_{i,j}$ | > 0 : **j** may be from cell type **i**

3- Difference between both score matrices



D = P - | N |

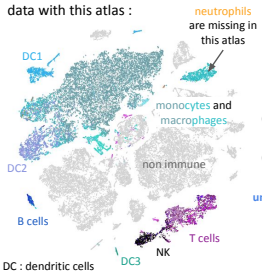
For each single cell (column), cell type can be annotated from the row names corresponding to the highest score.

RESULTS

- We used **4** Nf1-KO mice datasets to generate markers for each cell type
- We applied the annotation on **16** Nf1-KO mice datasets

Immune cell type annotation is consistent with the Tumor Immune Cell Atlas [4]

Annotation our of data with this atlas :

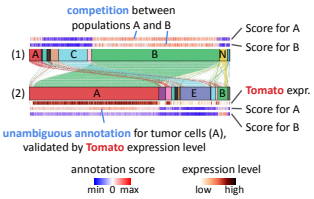


DC : dendritic cells

New identified markers strengthen the annotation

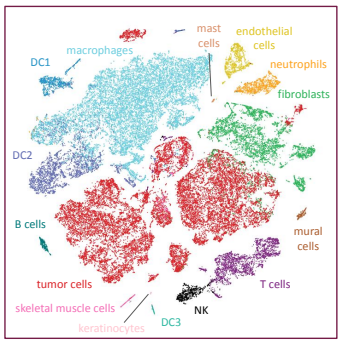
known markers : 1-3 markers / population
new markers : 50 markers / population

Annotation changes for 11,593 cells (23 %) considering known (1) or new (2) markers



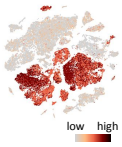
Cell type annotation at a single cell level, for 50,036 cells

Annotation pipeline runtime : 435 seconds (7.3 minutes)

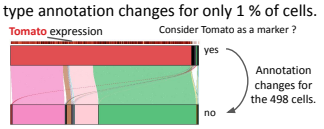


Cell type annotation is robust to Tomato expression consideration

Tomato expression level enables tumor cells identification



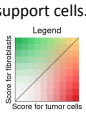
Without considering **Tomato** as a marker, cell type annotation changes for only 1 % of cells.

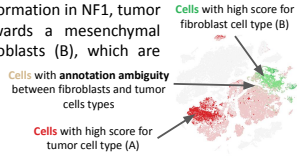


Most of those 498 cells whose annotation changes are tumor cells (A), which becomes mis-annotated as stromal cells (H, M, B).

Expression scores enable to keep track of overlapping signatures

During malignant transformation in NF1, tumor cells (A) progress towards a mesenchymal profile, related to fibroblasts (B), which are tissue support cells.





TAKE-HOME MESSAGES

Context

- NF1 scRNA-Seq datasets share similar cell populations
- In Nf1-KO mouse model, tumor cells express the **Tomato** reporter gene
- To date, there is no marker to identify tumor cells in NF1 scRNA-Seq datasets

Cell type annotation at a single cell level

- Identification of new markers leads to more **robust** cellular signatures
- Expression scoring enables to keep track of **overlapping signatures**
- **No clustering** is required : single cell annotation enables the identification of rare populations and the distinction of closely related populations

PROSPECTS

Add annotation quality metrics

- **ambiguity metric** : relative weight of the retained cell type compared all possible types
- **competition metric** : number of cell types with "high" score

Perform a sensitivity analysis of the annotation, regarding :

- the **number of positive markers** identified for each cell type : impact on runtime ?
- the number, or **proportion**, **of cells** taken into account to identify positive markers for one cell type

Apply this approach to other NF1 scRNA-Seq datasets, and build a NF1 atlas

Analyse novel types of data : scRNA-Seq long read (**isoforms**) and single-nuclei ATAC-Seq (**chromatin accessibility**)

Summer 2022

REFERENCES

[1] Radomska et al., Cellular Origin, Tumor Progression, and Pathogenic Mechanisms of Cutaneous Neurofibromas Revealed by Mice with Nf1 Knockout in Boundary Cap Cells, *Cancer Discovery* (2019)
[2] Sun et al., Stem-like cells drive NF1-associated MPNST functional heterogeneity and tumor progression, *Cell Stem Cell* (2021)
[3] Stuart et al., Comprehensive Integration of Single-Cell Data, *Cell* (2019)
[4] Nieto et al., A single-cell tumor immune atlas for precision oncology, *Genome Research* (2021)

Poster 2

Deciphering mechanisms driving malignant transformation in Neurofibromatosis Type I, from single-cell transcriptomics

Audrey Onfroy[1,2], Fanny Coulpier[1], Katarzyna Radomska[1], Denis Thieffry[2] and Piotr Topilko[1]
[1] Institut Mondor de Recherche Biomédicale, Inserm U955 - Team 9, Créteil, France
[2] Institut de Biologie de l'École Normale Supérieure, UMR CNRS 8197, Inserm U1024, PSL Université, Paris, France

Contact : audrey.onfroy@inserm.fr

NF1 is a genetic disease

- Affects 1 / 3,000 births worldwide
- Bi-allelic mutation of ***NF1*** in Schwann cell lineage
- Numerous symptoms, including development of benign nerve sheath tumors : **neurofibromas** (NFs), and their progression to malignant peripheral nerve sheath tumors (**MPNSTs**)

purpose of my PhD project

nerve sheath → plexiform NF → dysplastic NF → MPNST

no treatment to date

Nf1-KO mouse model [1]

- Bi-allelic mutation of ***Nf1*** in Schwann cell lineage
- *Nf1* mutant cells are traced by the **Tomato** reporter
- *Nf1*-KO mice share common features with NF1 patients, including development of pNFs and their progression into MPNSTs through a dysplastic stage

What are the molecular events driving malignant transformation of pNF into MPNST, in tumor forming cell populations, in the mouse model and in NF1 patients ?

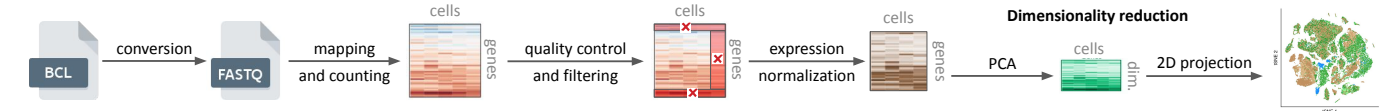
Available datasets

Single-cell transcriptomic technologies enable the study of NFs and MPNSTs transcriptional landscapes considering their cellular heterogeneity.

	pNF	dyNF	MPNST
Number of datasets	2	10	4

Single-cell RNA sequencing : for each biological sample, we have a genes x cells expression matrix

Methods



Results

Peripheral Nerve Sheath Tumors atlas from Nf1-KO mice

What is the identity of each cell population ?

Figure 1 - tSNE showing 60,000 cells from 16 datasets. One dot represents one single cell. Cells are colored by sample of origin. Batch-effect has been removed using harmony [2]. Some cell populations are specific from pNF (blue) or MPNST (brown).

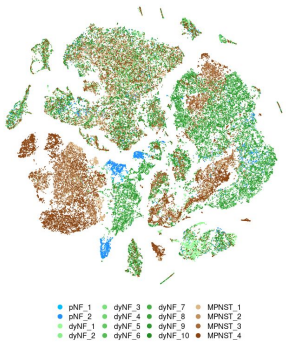
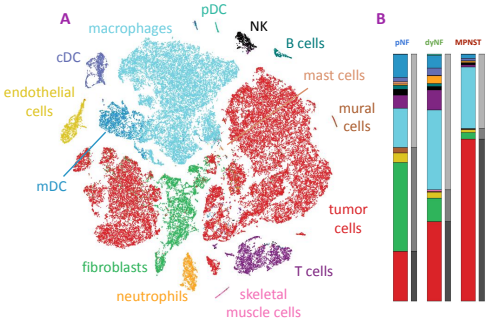


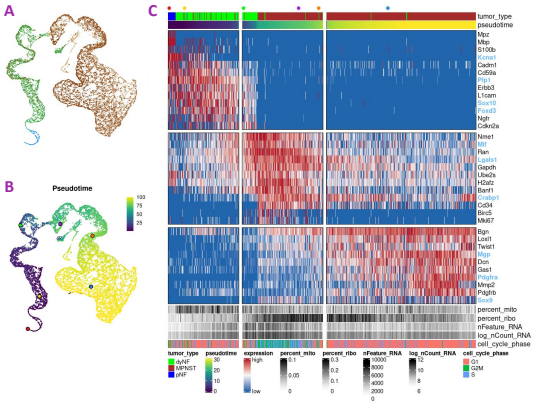
Figure 2 - **(A)** tSNE showing 60,000 cells from 16 datasets. We designed an automatic cell type annotation method. Annotation is validated using **Tomato** expression level, a tumor immune cell atlas [3], and known marker genes. **(B)** Barplot depicting tumor type specific composition. Tumor cell population (■) becomes predominant in MPNST, while stromal cells (■) decline, and immune cells (■) amount peaks in dysplastic tumors.



Tumor cell populations transcriptional landscape from pNF to MPNST

What is the transcriptomic profile of tumor cells ?

Figure 3 - **(A)** Representation of 20,000 tumor cells from 16 datasets. Cells are colored by sample of origin. Same legend as Fig. 1. **(B)** Pseudotime inference using Slingshot [4] reconstructs cell-to-cell transcriptomic changes from pNFs to MPNSTs. **(C)** Heatmap showing gene expression changes along pseudotime. Expression of genes in blue have been validated by immunolabeling. Tumor cells in pNF carry a glial signature (Sox10) whereas MPNST cells have a mesenchymal signature (Sox9).



Conclusions & Prospects

Single-cell transcriptomics allow tumor forming populations study at a single cell level

- Tumors are formed by about 10 cell populations, whose proportions and transcriptomic activities changes during malignant progression
- Tumor cell population becomes predominant in MPNST, while stromal cells decline, and immune cells amount peaks in dysplastic tumors
- Trajectory inference shows a switch in tumor cells identity, confirmed by immunolabeling

Prospects

- Study the tumor micro-environment, and their interactions with the tumor cell populations
- Functional validation of candidate genes implicated in malignant transformation
- Integrate other types of data : mutational landscape, single-nuclei ATAC-Seq and NF1 patients scRNA-Seq

References

[1] Radomska et al., Cellular Origin, Tumor Progression, and Pathogenic Mechanisms of Cutaneous Neurofibromas Revealed by Mice with Nf1 Knockout in Boundary Cap Cells, Cancer Discovery (2019)

[2] Stuart et al., Comprehensive Integration of Single-Cell Data, Cell (2019)

[3] Nieto et al., A single-cell tumor immune atlas for precision oncology, Genome Research (2021)

[4] Street et al., Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics, BMC Genomics (2018)

NF1 genetic background in the tumor micro-environment governs transcriptomic landscape of tumor cells in malignant peripheral nerve sheath tumors

Audrey Onfroy[1,2], Fanny Coulpier[1], Katarzyna Radomska[1], Denis Thieffry[2] and Piotr Topilko[1]

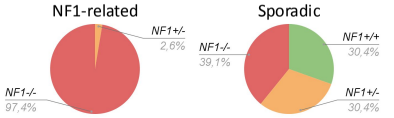
[1] Institut Mondor de Recherche Biomédicale, Inserm U955 - Team 9, Créteil, France [2] Institut de Biologie de l'École Normale Supérieure, UMR CNRS 8197, Inserm U1024, PSL Université, Paris, France Contact : audrey.onfroy@inserm.fr

1- Primary MPNSTs context

Context of MPNST development :

- Neurofibromatosis Type I (NF1)-related
- Sporadic
  - ≈ 40 % are *NF1*-affected

Primary MPNSTs are affected on *NF1* gene



Charts based on fig1C from [1], focusing on primary MPNSTs

2- *Nf1*-KO mouse models develop MPNSTs

- Bi-allelic inactivation of ***Nf1* gene** in Prss56 lineage : *Nf1*-KO mice
- *Nf1*-KO mice share common features with NF1 patients, including spontaneous transformation of ANNUBP into MPNSTs [3]

	$Prss56^{Cre},R26^{tdTom},Nf1^{flox/flox}$ [2]	$Prss56^{Cre},R26^{tdTom},Nf1^{flox/-}$ [3]
simple name	*Nf1*-KO^{wt}	*Nf1*-KO^{het}
ME *Nf1* status	*Nf1*+/+	*Nf1*+/-
model for	*NF1*-affected sporadic MPNSTs	NF1-related MPNSTs

Figure 2 - Two alternative mouse models regarding **genetic background**

3- Materials

What is the impact of *Nf1* background in the tumor micro-environment on tumor cells progression ?

Characterization of tumors developed by both *Nf1*-KO mouse models using :

- histology
- immunohistochemistry
- single-cell RNA sequencing (scRNA-Seq)
- exome sequencing

	pNF	ANNUBP	MPNST
Nf1-KO^{wt}	2	10	4
Nf1-KO^{het}	*coming soon*	2	3

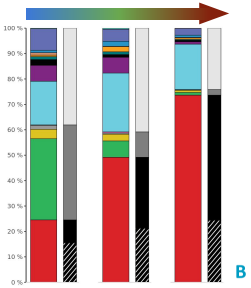
Figure 3 - Number of scRNA-Seq datasets. Each dataset corresponds to one tumor sample from one mouse. Tumor stratification defined from histology and Sox10/Sox9 expression balance in tumor cells.

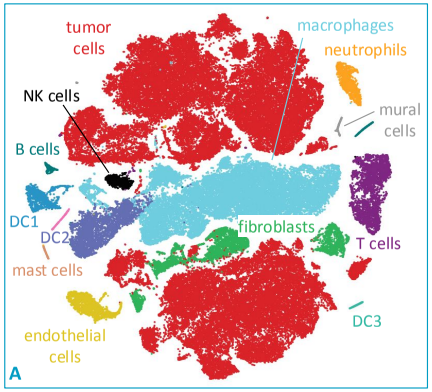
4- Single-cell atlas from *Nf1*-KO mouse tumor samples

Methods : data structure using Seurat [4], sample-effect correction using harmony [5]

Figure 4 - **(A)** tSNE showing 100,000 cells from 21 datasets. Cell type annotation is validated using **Tomato** expression level, a tumor immune cell atlas [6], and known marker genes. **(B)** Barplot depicting tumor type specific composition. Tumor cell population (■) becomes predominant in MPNST, while stromal cells (■) decline, and immune cells (■) amount peaks in ANNUBP.

Same diversity was observed between *Nf1*-KO^{wt} and *Nf1*-KO^{het} samples at the same progression stage. Less macrophages were present on Nf1$^{+/-}$ background (not shown).





5- Tumor population depicts same GMT

Methods : extraction of tumor cells from our atlas, sample-effect correction using FastMNN [7] and trajectory inference on using slingshot [8]

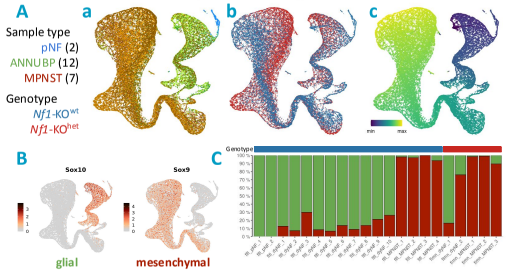


Figure 5 - **(A)** UMAP showing 45,000 tumor cells, colored according to 21 samples of origin **(a)**, genotype **(b)** or pseudotime **(c)**. Root has been manually set in a pNF dataset. **(B)** Feature plots showing Sox10 and Sox9 expression levels. **(C)** Barplot depicting glial (Sox10+) and mesenchymal (Sox9+) distribution trend in each sample.

Genes whose expression varies along pseudotime shown same trends in both genotypes : both genotypes depict same **glial-to-mesenchymal transition**.

6- Mesenchymal tumor cells slighly differ

Methods : fold change computation within mesenchymal tumor cells between *Nf1*-KO^{wt} and *Nf1*-KO^{het} samples, then enrichment analysis using clusterProfiler [9]

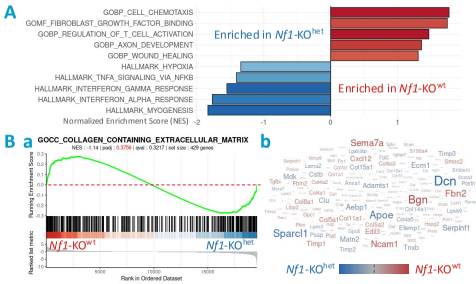


Figure 6 - GSEA in Sox9+ tumor cells between genotypes. **(A)** Enrichment plot showing 10 significantly enriched gene sets (p-value < 0.05). **(B) (a)** GSEA plot showing enrichment score for ECM-related genes. Similar figure was obtained in glial population. Note that the curve is symmetrical around the dotted line, reflecting two-sided enrichment. **(b)** Gene contribution to this enrichment. Color corresponds to genotype and size to contribution.

7- Take-Home Messages & Perspectives

- Mice from both genotypes develop ANNUBP that spontaneously transform into MPNSTs
- Sox10 and Sox9 expression levels can be used to stratify tumor between ANNUBP and MPNST
- Tumors share similar TME populations across genotypes
- Mesenchymal tumor cells slighly differ by their contribution to extra-cellular matrix (ECM)

⇒ What are the transcriptomic changes in TME responsible for differences between genotypes ?

References

[1] Cortes-Ciriano *et al.*, Cancer Discovery (2023)
[2] Radomska *et al.*, Cancer Discovery (2019)
[3] Radomska *et al.*, (in prep)
[4] Stuart *et al.*, Cell (2019)
[5] Korsunski *et al.*, Nature Methods (2019)
[6] Nieto *et al.*, Genome Research (2021)
[7] Haghverdi *et al.*, Nature Biotechnology (2018)
[8] Street *et al.*, BMC Genomics (2018)
[9] Wu *et al.*, Innovation (2021)

(T)ME : (tumor) micro-environment
pNF : plexiform neurofibroma
GMT : glial-to-mesenchymal transition
GSEA : gene set enrichment analysis

Poster 3

2023 NF Conference

Glial-to-mesenchymal transition drives malignant progression of plexiform neurofibromas in a mouse model of NF1

Katarzyna J. Radomska[1], **Audrey Onfroy**[1,4], Laure Lecerf[1], Bastien Job[2], Aurélien Beaude[1], Laura Sesma-Sanz[3], Tatiana El Jalkh[1], Denis Thieffry[4], Patrick Charnay[4], Pierre Wolkenstein[1,5], Nicolas Ortonne[1,6], Fanny Coulpier[1] and Piotr Topilko[1]

Contact: piotr.topilko@inserm.fr

Poster 4

1- Understand MPNST development

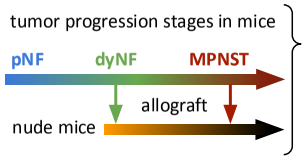
MPNST are aggressive soft tissue sarcomas with no therapy to date.

➢ The goal of the study was to gain knowledge in the cellular and molecular mechanisms governing malignant transformation using a novel NF1 mouse model.

Mouse model to study malignant transformation :

- Bi-allelic inactivation of ***Nf1*** in neural crest-derived boundary cap cells and their derivatives : *Nf1*-KO mice
- *Nf1*-KO mice develop pNFs that spontaneously undergo stepwise progression into MPNST [3]

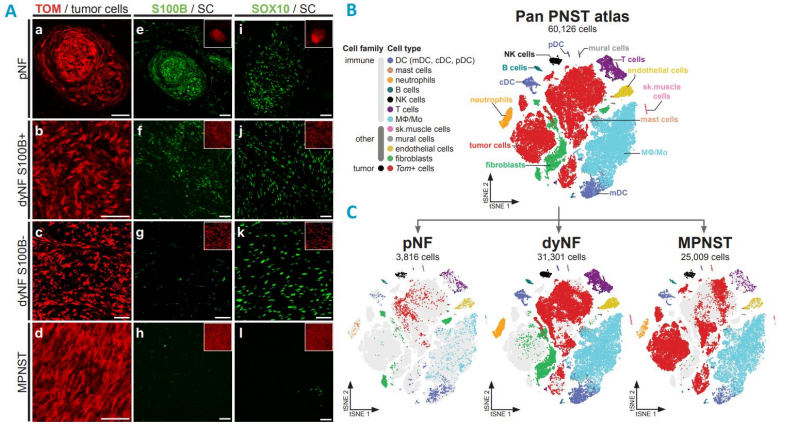
2- Materials



Characterization of tumors using :

- histological evaluation
- IHC characterization
- single-cell transcriptomic profiling
- whole exome sequencing

3- *Nf1*$^{-/-}$ tumor cells gradually loss their glial identity and proliferate during malignant progression



➢ Tumors progressively extinguish Schwann cells markers (S100B and SOX10) in *Nf1*-KO mice

➢ Tumor cells change their transcriptomic profile during progression

➢ Nerve fibroblasts signature evolves into cancer-associated fibroblasts between pNF and dyNF

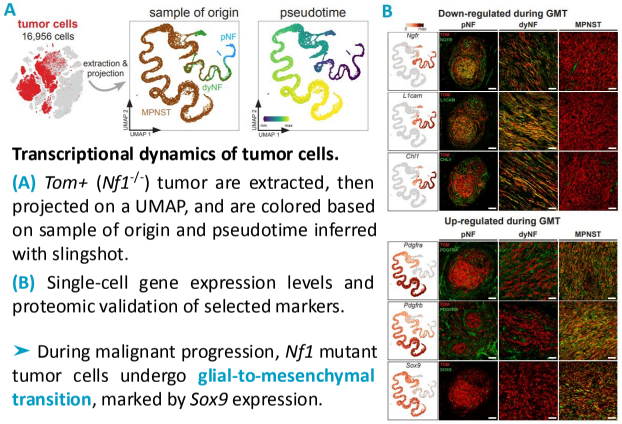
Spectrum of peripheral nerve sheath tumors (PNST) in *Nf1*-KO mice.

(A) Sections through pNF, dyNF and MPNST immunolabeled for TOM (tumor cells), S100ß and SOX10 (glial lineage). Insets show TOM staining of the corresponding lesions.

(B) Pan PNST transcriptomic atlas assembled from *Nf1*-KO derived pNF (n=2), dyNF (n=10) and MPNST (n=4) datasets. Expression score-based annotation resolved 12 distinct cell populations.

(C) Pan PNST atlas split by tumor type. Cells from complete atlas are shown as light grey background. Note distinct location of fibroblasts and tumor cells between tumor types, indicating distinct transcriptomic profiles.

4- Tumor cells undergo Glial-to-Mesenchymal Transition



Transcriptional dynamics of tumor cells.

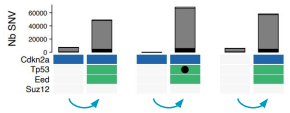
(A) *Tom+* (*Nf1*$^{-/-}$) tumor are extracted, then projected on a UMAP, and are colored based on sample of origin and pseudotime inferred with slingshot.

(B) Single-cell gene expression levels and proteomic validation of selected markers.

➢ During malignant progression, *Nf1* mutant tumor cells undergo **glial-to-mesenchymal transition**, marked by *Sox9* expression.

Evolution of the transcriptomic profile of tumor cells is conserved between NF1 patients and *Nf1*-KO mice.

5- Glial-to-Mesenchymal Transition preceeds *Cdkn2a* loss in tumor cells



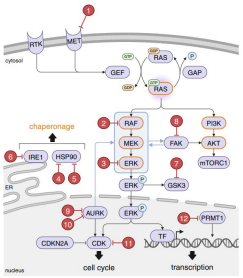
Genomic characterization of tumor cells from *Nf1*-KO mice and transplanted MPNSTs. Each pair of columns corresponds to the primary MPNST (left) and its associated transplanted (right) tumor. The top line represents the number of SNVs for each tumor. The rows below correspond to CNV loss (blue) and SNV (green), including pathogenic variant (dot).

➢ MPNST can be propagated in nude recipients and GMT preceeds *Cdkn2a* loss in tumor cells.

6- Ras signaling pathway is involved in *Sox9* expression

- Partial *Sox9* inhibition (shRNA and siRNA) delays tumor growth *in vivo*.
- Screening of a library of 3,500 compounds on tumor cells, using SOX9 as read out identified 12 FDA-approved drugs, all related to the inhition of Ras pathway.

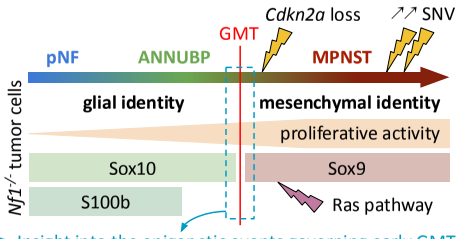
Representation of molecular targets for top 12 candidate drugs, in relation by Ras signaling pathway.



➢ *Sox9* and Ras pathway are therapeutic targets for treatment of MPNSTs

7- Summary and future directions

- *Nf1*-KO mice recapitulates malignant transformation



➢ Insight into the epigenetic events governing early GMT

[1]Mondor Institute for Biomedical Research, INSERM, Créteil, France
[2]Gustave Roussy Institute of Oncology, Villejuif, France
[3]National Phenotypic Screening Centre, School of Life Sciences, University of Dundee, Dundee, United Kingdom
[4]École Normale Supérieure, PSL Research University, CNRS, INSERM, IBENS, Paris, France
[5]Department of Dermatology, Reference Centre for Neurofibromatoses, Henri-Mondor hospital, AP-HP, Créteil, France
[6]Department of Pathology, Henri-Mondor hospital, AP-HP, Créteil, France

Funding : Association Neurofibromatoses et Recklinghausen, Foundation Maladies Rares, Canceropole IDF, ITMO Cancer

2024 NF Conference

Abreviations

dyNF : dysplastic neurofibroma, including ANNUBP
GMT : glial-to-mesenchymal transition
MPNST : malignant peripheral nerve sheath tumor
pNF : plexiform neurofibroma
PNST : peripheral nerve sheath tumor

References

[1] Beert *et al.*, Genes, Chromosomes & Cancer (2011)
[2] Radomska *et al.*, Cancer Discovery (2019)
[3] Radomska *et al.* (in preparation)
[4] Sun *et al.*, Cell Stem Cell (2021)
[5] Wu *et al.*, Science Advances (2022)
[6] Rhodes *et al.*, Human Molecular Genetics (2019)
[7] Magallón-Lorenz *et al.*, Human Genetics (2021)
[8] Suppiah *et al.*, Nature Communications (2023)

References

Ackerman, Sarah D. & Kelly R. Monk (2016). “The scales and tales of myelination: using zebrafish and mouse to study myelinating glia”. *Brain Research*, Vol.1641.
DOI: 10.1016/j.brainres.2015.10.011.

Aevermann, Brian *et al.* (2021). “A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell RNA sequencing”. *Genome Research*, Vol.31 (10).
DOI: 10.1101/gr.275569.121.

Ahlmann-Eltze, Constantin & Wolfgang Huber (2023). “Comparison of transformations for single-cell RNA-seq data”. *Nature Methods*, Vol.20 (5).
DOI: 10.1038/s41592-023-01814-1.

— (2024). “Analysis of multi-condition single-cell data with latent embedding multivariate regression”. *BioRXiv*.
DOI: 10.1101/2023.03.06.531268.

Aibar, Sara *et al.* (2017). “SCENIC: single-cell regulatory network inference and clustering”. *Nature Methods*, Vol.14 (11).
DOI: 10.1038/nmeth.4463.

Akshintala, Srivandana *et al.* (2020). “Longitudinal evaluation of peripheral nerve sheath tumors in neurofibromatosis type 1: growth analysis of plexiform neurofibromas and distinct nodular lesions”. *Neuro-Oncology*, Vol.22 (9).
DOI: 10.1093/neuonc/noaa053.

Amann, Lukas & Marco Prinz (2020). “The origin, fate and function of macrophages in the peripheral nervous system-an update”. *International Immunology*, Vol.32 (11).
DOI: 10.1093/INTIMM/DXAA030.

Amstutz, Peter *et al.* (2024). *Existing Workflow systems. Common Workflow Language wiki, GitHub*. https://s.apache.org/existing-workflow-systems. Accessed: 2024-02-21.
URL: https://s.apache.org/existing-workflow-systems.

Anastasaki, Corina *et al.* (2017). “Updated nomenclature for human and mouse neurofibromatosis type 1 genes”. *Neurology Genetics*, Vol.3 (4).
DOI: 10.1212/NXG.0000000000000169.

Angerer, Philipp *et al.* (2016). “destiny: diffusion maps for large-scale single-cell data in R”. *Bioinformatics*, Vol.32 (8).
DOI: 10.1093/bioinformatics/btv715.

Argelaguet, Ricard *et al.* (2021). “Computational principles and challenges in single-cell data integration”. *Nature Biotechnology*, Vol.39 (10).
DOI: 10.1038/s41587-021-00895-7.

Armingol, Erick *et al.* (2021). “Deciphering cell–cell interactions and communication from gene expression”. *Nature Reviews Genetics*, Vol.22 (2).
DOI: 10.1038/s41576-020-00292-x.

Arthur-Farraj, Peter J. *et al.* (2012). “c-Jun Reprograms Schwann Cells of Injured Nerves to Generate a Repair Cell Essential for Regeneration”. *Neuron*, Vol.75 (4).
DOI: 10.1016/j.neuron.2012.06.021.

Aslam, Bilal *et al.* (2017). “Proteomics: Technologies and Their Applications”. *Journal of Chromatographic Science*, Vol.55 (2).
DOI: 10.1093/chromsci/bmw167.

Badia-i-Mompel, Pau *et al.* (2022). “decoupleR: ensemble of computational methods to infer biological activities from omics data”. *Bioinformatics Advances*, Vol.2 (1).
DOI: 10.1093/bioadv/vbac016.

Báez-Flores, Juan *et al.* (2023). “The therapeutic potential of neurofibromin signaling pathways and binding partners”. *Communications Biology*, Vol.6 (1).
DOI: 10.1038/s42003-023-04815-0.

Bais, Abha S. & Dennis Kostka (2020). “scds: computational annotation of doublets in single-cell RNA sequencing data”. *Bioinformatics*, Vol.36 (4).
DOI: 10.1093/bioinformatics/btz698.

Baker, Monya (2016). “1,500 scientists lift the lid on reproducibility”. *Nature*, Vol.533 (7604).
DOI: 10.1038/533452a.

Batlle, Eduard & Hans Clevers (2017). “Cancer stem cells revisited”. *Nature Medicine*, Vol.23 (10).
DOI: 10.1038/nm.4409.

Becht, Etienne *et al.* (2019). “Dimensionality reduction for visualizing single-cell data using UMAP”. *Nature Biotechnology*, Vol.37 (1).
DOI: 10.1038/nbt.4314.

Beert, Eline *et al.* (2011). “Atypical neurofibromas in neurofibromatosis type 1 are premalignant tumors”. *Genes, Chromosomes and Cancer*, Vol.50 (12).
DOI: 10.1002/gcc.20921.

Beert, Eline *et al.* (2012). “Biallelic inactivation of NF1 in a sporadic plexiform neurofibroma”. *Genes Chromosomes and Cancer*, Vol.51 (9).
DOI: 10.1002/gcc.21969.

Benito, Cristina *et al.* (2017). “STAT3 controls the long-term survival and phenotype of repair schwann cells during nerve regeneration”. *Journal of Neuroscience*, Vol.37 (16).
DOI: 10.1523/JNEUROSCI.3481-16.2017.

Bergen, Volker *et al.* (2020). “Generalizing RNA velocity to transient cell states through dynamical modeling”. *Nature Biotechnology*, Vol.38 (12).
DOI: 10.1038/s41587-020-0591-3.

Bergoug, Mohammed *et al.* (2020). “Neurofibromin Structure, Functions and Regulation”. *Cells*, Vol.9 (11).
DOI: 10.3390/cells9112365.

Blanpain, Cédric (2013). “Tracing the cellular origin of cancer”. *Nature Cell Biology*, Vol.15 (2).
DOI: 10.1038/ncb2657.

Bonsang, Benjamin *et al.* (2022). “VEGF and VEGFR family members are expressed by neoplastic cells of NF1-associated tumors and may play an oncogenic role in malignant peripheral nerve sheath tumor growth through an autocrine loop”. *Annals of Diagnostic Pathology*, Vol.60.
DOI: 10.1016/j.anndiagpath.2022.151997.

Bottillo, Irene *et al.* (2008). “Germline and somatic NF1 mutations in sporadic and NF1-associated malignant peripheral nerve sheath tumours”. *The Journal of Pathology*, Vol.217 (5).
DOI: 10.1002/path.2494.

Bray, Nicolas L. *et al.* (2016). “Near-optimal probabilistic RNA-seq quantification”. *Nature Biotechnology*, Vol.34 (5).
DOI: 10.1038/nbt.3519.

Britsch, Stefan *et al.* (2001). “The transcription factor Sox10 is a key regulator of peripheral glial development”. *Genes and Development*, Vol.15 (1).
DOI: 10.1101/gad.186601.

Brockman, Qierra R. *et al.* (2022). “PRC2 loss drives MPNST metastasis and matrix remodeling”. *JCI Insight*, Vol.7 (20).
DOI: 10.1172/jci.insight.157502.

Brodeur, Abel *et al.* (2023). “Replication games: how to make reproducibility research more systematic Setting the agenda in research”. *Nature*, Vol.621.
DOI: 10.1038/d41586-023-02997-5.

Brosseau, Jean-Philippe *et al.* (2018). “NF1 heterozygosity fosters de novo tumorigenesis but impairs malignant transformation”. *Nature Communications*, Vol.9 (1).
DOI: 10.1038/s41467-018-07452-y.

Brüning, Ralf Schulze *et al.* (2022). “Comparative analysis of common alignment tools for single-cell RNA sequencing”. *GigaScience*, Vol.11.
DOI: 10.1093/gigascience/giac001.

Burks, Ciersten A. *et al.* (2019). “Ketotifen Modulates Mast Cell Chemotaxis to Kit-Ligand, but Does Not Impact Mast Cell Numbers, Degranulation, or Tumor Behavior in Neurofibromas of Nf1-Deficient Mice”. *Molecular Cancer Therapeutics*, Vol.18 (12).
DOI: 10.1158/1535-7163.MCT-19-0123.

Button, Katherine S. *et al.* (2013). “Power failure: why small sample size undermines the reliability of neuroscience”. *Nature Reviews Neuroscience*, Vol.14 (5).
DOI: 10.1038/nrn3475.

Calabrò, Sonia *et al.* (2023). “Impact of Gut Microbiota on the Peripheral Nervous System in Physiological, Regenerative and Pathological Conditions”. *International Journal of Molecular Sciences*, Vol.24 (9).
DOI: 10.3390/ijms24098061.

Cantini, Laura *et al.* (2017). “Classification of gene signatures for their information value and functional redundancy”. *npj Systems Biology and Applications*, Vol.4 (1).
DOI: 10.1038/s41540-017-0038-8.

Cao, Ru *et al.* (2002). “Role of Histone H3 Lysine 27 Methylation in Polycomb-Group Silencing”. *Science*, Vol.298 (5595).
DOI: 10.1126/science.1076997.

Carlson, Bruce M. (2014). “Formation of Germ Layers and Early Derivatives”. *Human Embryology and Developmental Biology*. Fifth Edition. Elsevier. Chap. 5.
DOI: 10.1016/B978-1-4557-2794-0.00005-X.

Carr, Matthew J. *et al.* (2018). “Mesenchymal Precursor Cells in Adult Nerves Contribute to Mammalian Tissue Repair and Regeneration”. *Cell Stem Cell*, Vol.24 (2).
DOI: 10.1016/j.stem.2018.10.024.

Carrió, Meritxell *et al.* (2018). “Analysis of intratumor heterogeneity in Neurofibromatosis type 1 plexiform neurofibromas and neurofibromas with atypical features: Correlating histological and genomic findings”. *Human Mutation*, Vol.39 (8).
DOI: 10.1002/humu.23552.

Carroll, Steven L. (2012). “Molecular mechanisms promoting the pathogenesis of Schwann cell neoplasms”. *Acta Neuropathologica*, Vol.123 (3).
DOI: 10.1007/s00401-011-0928-6.

Casadevall, Arturo *et al.* (2016). “A Framework for Improving the Quality of Research in the Biological Sciences”. *mBio*, Vol.7 (4).
DOI: 10.1128/mBio.01256-16.

Chandrasekaran, Prashant *et al.* (2022). “CXCL12 defines lung endothelial heterogeneity and promotes distal vascular growth”. *Development (Cambridge)*, Vol.149 (21).
DOI: 10.1242/dev.200909.

Chaney, Katherine E. *et al.* (2020). “Cdkn2a loss in a model of neurofibroma demonstrates stepwise tumor progression to atypical neurofibroma and mpnst”. *Cancer Research*, Vol.80 (21).
DOI: 10.1158/0008-5472.CAN-19-1429.

Chari, Tara & Lior Pachter (2023). “The specious art of single-cell genomics”. *PLoS Computational Biology*, Vol.19 (8).
DOI: 10.1371/journal.pcbi.1011288.

Chen, Bing *et al.* (2021). “Single Cell Transcriptome Data Analysis Defines the Heterogeneity of Peripheral Nerve Cells in Homeostasis and Regeneration”. *Frontiers in Cellular Neuroscience*, Vol.15.
DOI: 10.3389/fncel.2021.624826.

Chen, Zhiguo *et al.* (2014). “Cells of Origin in the Embryonic Nerve Roots for NF1-Associated Plexiform Neurofibroma”. *Cancer Cell*, Vol.26 (5).
DOI: 10.1016/j.ccell.2014.09.009.

Chen, Zhiguo *et al.* (2019). “Spatiotemporal Loss of NF1 in Schwann Cell Lineage Leads to Different Types of Cutaneous Neurofibroma Susceptible to Modification by the Hippo Pathway”. *Cancer Discovery*, Vol.9 (1).
DOI: 10.1158/2159-8290.CD-18-0151.

Chernoff, Karen A. & Julie V. Schaffer (2016). “Cutaneous and ocular manifestations of neurocutaneous syndromes”. *Clinics in Dermatology*, Vol.34 (2).
DOI: 10.1016/j.clindermatol.2015.11.003.

Choi, Kwangmin *et al.* (2017). “An inflammatory gene signature distinguishes neurofibroma Schwann cells and macrophages from cells in the normal peripheral nervous system”. *Scientific Reports*, Vol.7 (1).
DOI: 10.1038/srep43315.

Chowdhury, Hussain Ahmed *et al.* (2020). “(Differential) Co-Expression Analysis of Gene Expression: A Survey of Best Practices”. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol.17 (4).
DOI: 10.1109/TCBB.2019.2893170.

Cleven, Arjen H. G. *et al.* (2016). “Loss of H3K27 tri-methylation is a diagnostic marker for malignant peripheral nerve sheath tumors and an indicator for an inferior survival”. *Modern Pathology*, Vol.29 (6).
DOI: 10.1038/modpathol.2016.45.

Cohen-Boulakia, Sarah *et al.* (2017). “Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities”. *Future Generation Computer Systems*, Vol.75.
DOI: 10.1016/j.future.2017.01.012.

Coifman, R.R. *et al.* (2005). “Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps”. *Proceedings of the National Academy of Sciences*, Vol.102 (21).
DOI: 10.1073/pnas.0500334102.

Cokelaer, Thomas *et al.* (2023). “Reprohackathons: promoting reproducibility in bioinformatics through training”. *Bioinformatics*, Vol.39 (Supplement_1).
DOI: 10.1093/bioinformatics/btad227.

Cortal, Akira *et al.* (2021). “Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID”. *Nature Biotechnology*, Vol.39 (9).
DOI: 10.1038/s41587-021-00896-6.

Cortes-Ciriano, Isidro *et al.* (2023). “Genomic Patterns of Malignant Peripheral Nerve Sheath Tumor (MPNST) Evolution Correlate with Clinical Outcome and Are Detectable in Cell-Free DNA”. *Cancer discovery*, Vol.13 (3).
DOI: 10.1158/2159-8290.CD-22-0786.

Coulpier, Fanny *et al.* (2009). “Novel features of boundary cap cells revealed by the analysis of newly identified molecular markers”. *Glia*, Vol.57 (13).
DOI: 10.1002/glia.20862.

Court, Felipe A. *et al.* (2006). “Basal lamina: Schwann cells wrap to the rhythm of space-time”. *Current Opinion in Neurobiology*, Vol.16 (5).
DOI: 10.1016/j.conb.2006.08.005.

Courtès, Ludovic (2013). “Functional Package Management with Guix”. *ArXiv*.
DOI: 10.48550/arXiv.1305.4584.

Creus-Bachiller, Edgar *et al.* (2023). “Expanding a precision medicine platform for malignant peripheral nerve sheath tumors: New patient-derived orthotopic xenografts, cell lines and tumor entities”. *Molecular Oncology*, Vol.18 (4).
DOI: 10.1002/1878-0261.13534.

Dagalakis, Urania *et al.* (2013). “Puberty and Plexiform Neurofibroma Tumor Growth in Patients with Neurofibromatosis Type I”. *The Journal of Pediatrics*, Vol.164 (3).
DOI: 10.1016/j.jpeds.2013.10.081.

Dahl, Andrew A. (2021). *Ophthalmologic Manifestations of Neurofibromatosis Type 1 (NF-1)*.
URL: https://emedicine.medscape.com/article/1219222-overview.

Dai, Shirui *et al.* (2020). “Chronic Stress Promotes Cancer Development”. *Frontiers in Oncology*, Vol.10.
DOI: 10.3389/fonc.2020.01492.

Dann, Emma *et al.* (2022). “Differential abundance testing on single-cell data using k-nearest neighbor graphs”. *Nature Biotechnology*, Vol.40 (2).
DOI: 10.1038/s41587-021-01033-z.

De Paoli-Iseppi, Ricardo *et al.* (2021). “Isoform Age - Splice Isoform Profiling Using Long-Read Technologies”. *Frontiers in Molecular Biosciences*, Vol.8.
DOI: 10.3389/fmolb.2021.711733.

De Raedt, T. *et al.* (2003). “Elevated Risk for MPNST in NF1 Microdeletion Patients”. *The American Journal of Human Genetics*, Vol.72 (5).
DOI: 10.1086/374821.

“Decoding noncoding RNAs” (2022). *Nature Methods*, Vol.19 (10).
DOI: 10.1038/s41592-022-01654-5.

Dehner, Carina *et al.* (2021). “Chromosome 8 gain is associated with high-grade transformation in MPNST”. *JCI Insight*, Vol.6 (6).
DOI: 10.1172/jci.insight.146351.

Dijk, David van *et al.* (2018). “Recovering Gene Interactions from Single-Cell Data Using Data Diffusion”. *Cell*, Vol.174 (3).
DOI: 10.1016/j.cell.2018.05.061.

Ding, Jiarui *et al.* (2020). “Systematic comparison of single-cell and single-nucleus RNA-sequencing methods”. *Nature Biotechnology*, Vol.38 (6).
DOI: 10.1038/s41587-020-0465-8.

Dobin, Alexander *et al.* (2013). “STAR: ultrafast universal RNA-seq aligner”. *Bioinformatics*, Vol.29 (1).
DOI: 10.1093/bioinformatics/bts635.

Dorafshan, Shima *et al.* (2022). “Periostin: biology and function in cancer”. *Cancer Cell International*, Vol.22 (1).
DOI: 10.1186/s12935-022-02714-8.

Drouet, Alain *et al.* (2004). “Neurofibromatosis 1-associated neuropathies: a reappraisal”. *Brain*, Vol.127 (9).
DOI: 10.1093/brain/awh234.

Dumitrascu, Bianca *et al.* (2021). “Optimal marker gene selection for cell type discrimination in single cell analyses”. *Nature Communications*, Vol.12 (1).
DOI: 10.1038/s41467-021-21453-4.

Dunn, Gavin P. *et al.* (2013). “Role of resection of malignant peripheral nerve sheath tumors in patients with neurofibromatosis Type 1”. *Journal of Neurosurgery*, Vol.118 (1).
DOI: 10.3171/2012.9.JNS101610.

Eberwine, James *et al.* (1992). “Analysis of gene expression in single live neurons”. *Proceedings of the National Academy of Sciences*, Vol.89 (7).
DOI: 10.1073/pnas.89.7.3010.

Edwards, Marc A. & Siddhartha Roy (2017). “Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition”. *Environmental Engineering Science*, Vol.34 (1).
DOI: 10.1089/ees.2016.0223.

Eisenbarth, Stephanie C. (2019). “Dendritic cell subsets in T cell programming: location dictates function”. *Nature Reviews Immunology*, Vol.19 (2).
DOI: 10.1038/s41577-018-0088-1.

Elefteriou, Florent *et al.* (2009). “Skeletal abnormalities in neurofibromatosis type 1: Approaches to therapeutic options”. *American Journal of Medical Genetics Part A*, Vol.149A (10).
DOI: 10.1002/ajmg.a.33045.

Endo, Takeshi *et al.* (2022). “Mature but not developing Schwann cells promote axon regeneration after peripheral nerve injury”. *npj Regenerative Medicine*, Vol.7 (1).
DOI: 10.1038/s41536-022-00205-y.

Erickson, Alek G. *et al.* (2023). “The transcriptional portraits of the neural crest at the individual cell level”. *Seminars in Cell and Developmental Biology*.
DOI: 10.1016/j.semcdb.2022.02.017.

Evans, D. Gareth R. *et al.* (2011). “Mortality in neurofibromatosis 1: in North West England: an assessment of actuarial survival in a region of the UK since 1989”. *European Journal of Human Genetics*, Vol.19 (11).
DOI: 10.1038/ejhg.2011.113.

Evans, D.G. *et al.* (2010). “Birth incidence and prevalence of tumor-prone syndromes: Estimates from a UK family genetic register service”. *American Journal of Medical Genetics Part A*, Vol.152A (2).
DOI: 10.1002/ajmg.a.33139.

Ewels, Philip *et al.* (2016). “MultiQC: summarize analysis results for multiple tools and samples in a single report”. *Bioinformatics*, Vol.32 (19).
DOI: 10.1093/bioinformatics/btw354.

Farid, Mohamad *et al.* (2014). “Malignant Peripheral Nerve Sheath Tumors”. *The Oncologist*, Vol.19 (2).
DOI: 10.1634/theoncologist.2013-0328.

Ferner, Rosalie E. *et al.* (2006). “Guidelines for the diagnosis and management of individuals with neurofibromatosis”. *Journal of Medical Genetics*, Vol.44 (2).
DOI: 10.1136/jmg.2006.045906.

Fertitta, Laura *et al.* (2024). “Whole-Body Positron Emission Tomography with 18F-Fluorodeoxyglucose/Magnetic Resonance Imaging as a Screening Tool for the Detection of Malignant Transformation in Individuals with Neurofibromatosis Type 1”. *Journal of Investigative Dermatology*.
DOI: 10.1016/j.jid.2024.01.028.

Fisher, Michael J. *et al.* (2022). “Management of neurofibromatosis type 1-associated plexiform neurofibromas”. *Neuro-Oncology*, Vol.24 (11).
DOI: 10.1093/neuonc/noac146.

Fletcher, Jonathan S *et al.* (2020). “After Nf1 loss in Schwann cells, inflammation drives neurofibroma formation”. *Neuro-Oncology Advances*, Vol.2 (Supplement_1).
DOI: 10.1093/noajnl/vdz045.

Fletcher, Jonathan S. *et al.* (2019). “STAT3 inhibition reduces macrophage number and tumor growth in neurofibroma”. *Oncogene*, Vol.38 (15).
DOI: 10.1038/s41388-018-0600-x.

Furlan, Alessandro & Igor Adameyko (2018). “Schwann cell precursor: a neural crest cell in disguise?” *Developmental Biology*, Vol.444.
DOI: 10.1016/j.ydbio.2018.02.008.

Ge, Ling Ling *et al.* (2022). “Neurofibroma Development in Neurofibromatosis Type 1: Insights from Cellular Origin and Schwann Cell Lineage Development”. *Cancers*, Vol.14 (18).
DOI: 10.3390/cancers14184513.

Gelman, Andrew & Eric Loken (2019). “The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time”.
URL: https://api.semanticscholar.org/CorpusID:198164638.

Gerber, Daniel *et al.* (2021). “Transcriptional profiling of mouse peripheral nerves to the single-cell level to build a sciatic nerve atlas (Snat)”. *eLife*, Vol.10.
DOI: 10.7554/ELIFE.58591.

Germain, Pierre-Luc *et al.* (2022). “Doublet identification in single-cell sequencing data using scDblFinder”. *F1000Research*, Vol.10.
DOI: 10.12688/f1000research.73600.2.

Gitler, Aaron D. *et al.* (2003). “Nf1 has an essential role in endothelial cells”. *Nature Genetics*, Vol.33 (1).
DOI: 10.1038/ng1059.

Gladdy, Rebecca A. *et al.* (2010). “Do radiation-associated soft tissue sarcomas have the same prognosis as sporadic soft tissue sarcomas?” *Journal of Clinical Oncology*, Vol.28 (12).
DOI: 10.1200/JCO.2009.25.1728.

Glenn, Thomas D. & William S. Talbot (2013). “Signals regulating myelination in peripheral nerves and the Schwann cell response to injury”. *Current Opinion in Neurobiology*, Vol.23 (6).
DOI: 10.1016/j.conb.2013.06.010.

Gomez-Sanchez, Jose Antonio *et al.* (2013). “Epigenetic induction of the Ink4a/Arf locus prevents Schwann cell overproliferation during nerve regeneration and after tumorigenic challenge”. *Brain*, Vol.136 (7).
DOI: 10.1093/brain/awt130.

Gresset, Aurélie *et al.* (2015). “Boundary Caps Give Rise to Neurogenic Stem Cells and Terminal Glia in the Skin”. *Stem Cell Reports*, Vol.5 (2).
DOI: 10.1016/j.stemcr.2015.06.005.

Griffin, John W. & Wesley J. Thompson (2008). “Biology and pathology of nonmyelinating schwann cells”. *Glia*, Vol.56 (14).
DOI: 10.1002/glia.20778.

Grit, Jamie L. *et al.* (2021). “Distinctive epigenomic alterations in NF1-deficient cutaneous and plexiform neurofibromas drive differential MKK/p38 signaling”. *Epigenetics and Chromatin*, Vol.14 (1).
DOI: 10.1186/s13072-020-00380-6.

Grover, Shabnam Bhandari *et al.* (2021). “Imaging diagnosis of plexiform neurofibroma- unravelling the confounding features: A report of two cases”. *Radiology Case Reports*, Vol.16 (9).
DOI: 10.1016/j.radcr.2021.06.025.

Gu, Zuguang *et al.* (2016). “Complex heatmaps reveal patterns and correlations in multidimensional genomic data”. *Bioinformatics*, Vol.32 (18).
DOI: 10.1093/bioinformatics/btw313.

Hafemeister, Christoph & Rahul Satija (2019). "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". *Genome Biology*, Vol.20 (1).
DOI: 10.1186/s13059-019-1874-1.

Hagel, Christian *et al.* (2006). "Histopathology and clinical outcome of NF1-associated vs. sporadic malignant peripheral nerve sheath tumors". *Journal of Neuro-Oncology*, Vol.82 (2).
DOI: 10.1007/s11060-006-9266-2.

Haghverdi, Laleh *et al.* (2018). "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors". *Nature Biotechnology*, Vol.36 (5).
DOI: 10.1038/nbt.4091.

Hanani, Menachem & David C. Spray (2020). "Emerging importance of satellite glia in nervous system function and dysfunction". *Nature Reviews Neuroscience*, Vol.21 (9).
DOI: 10.1038/s41583-020-0333-z.

Harris, Melissa L. *et al.* (2010). "Sox proteins in melanocyte development and melanoma". *Pigment Cell & Melanoma Research*, Vol.23 (4).
DOI: 10.1111/j.1755-148X.2010.00711.x.

Harty, Breanne L. & Kelly R. Monk (2017). "Unwrapping the unappreciated: recent progress in Remak Schwann cell biology". *Current Opinion in Neurobiology*, Vol.47.
DOI: 10.1016/j.conb.2017.10.003.

Hassan, Ayesha *et al.* (2021). "Systemic Options for Malignant Peripheral Nerve Sheath Tumors". *Current Treatment Options in Oncology*, Vol.22 (4).
DOI: 10.1007/s11864-021-00830-7.

Heroux, Michael A. *et al.* (2018). *Toward a Compatible Reproducibility Taxonomy for Computational and Computing Sciences*. Tech. rep. Albuquerque, NM, and Livermore, CA (United States): Sandia National Laboratories (SNL).
DOI: 10.2172/1481626.

Heumos, Lukas *et al.* (2023). "Best practices for single-cell analysis across modalities". *Nature Reviews Genetics*, Vol.24 (8).
DOI: 10.1038/s41576-023-00586-w.

Hie, Brian *et al.* (2019). "Efficient integration of heterogeneous single-cell transcriptomes using Scanorama". *Nature Biotechnology*, Vol.37 (6).
DOI: 10.1038/s41587-019-0113-3.

Ho, Jonathan D *et al.* (2022). "Isolated, Nonsyndromic Mucocutaneous Plexiform Neurofibromas: A Systematic Review of the Clinicopathologic Features". *The American Journal of dermatopathology*.
DOI: 10.1097/DAD.0000000000002322.

Høland, Maren *et al.* (2023). "Transcriptomic subtyping of malignant peripheral nerve sheath tumours highlights immune signatures, genomic profiles, patient survival and therapeutic targets". *eBioMedicine*, Vol.97.
DOI: 10.1016/j.ebiom.2023.104829.

Holland, Christian H. *et al.* (2020). "Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data". *Genome Biology*, Vol.21 (1).
DOI: 10.1186/s13059-020-1949-z.

Holoch, Daniel & Raphaël Margueron (2017). "Mechanisms Regulating PRC2 Recruitment and Enzymatic Activity". *Trends in Biochemical Sciences*, Vol.42 (7).
DOI: 10.1016/j.tibs.2017.04.003.

Huang, Qianhui *et al.* (2021). “Evaluation of Cell Type Annotation R Packages on Single-Cell RNA-Seq Data”. *Genomics, Proteomics & Bioinformatics*, Vol.19 (2).
DOI: 10.1016/j.gpb.2020.07.004.

Imbard, Apolline *et al.* (2015). “NF1 single and multi-exons copy number variations in neurofibromatosis type 1”. *Journal of Human Genetics*, Vol.60 (4).
DOI: 10.1038/jhg.2015.6.

Iribar, Haizea *et al.* (2016). “Does Schwann cell dedifferentiation originate dermal neurofibromas?” *Experimental Dermatology*, Vol.25 (11).
DOI: 10.1111/exd.13110.

Jacks, Tyler *et al.* (1994). “Tumour predisposition in mice heterozygous for a targeted mutation in Nf1”. *Nature Genetics*, Vol.7.
DOI: 10.1038/ng0794-353.

Jessen, K. R. & R. Mirsky (2016). “The repair Schwann cell and its function in regenerating nerves”. *Journal of Physiology*, Vol.594 (13).
DOI: 10.1113/JP270874.

Jessen, Kristjan R. & Peter Arthur-Farraj (2019). “Repair Schwann cell update: Adaptive reprogramming, EMT, and stemness in regenerating nerves”. *Glia*, Vol.67 (3).
DOI: 10.1002/glia.23532.

Jessen, Kristjan R. & Rhona Mirsky (2005). “The origin and development of glial cells in peripheral nerves”. *Nature Reviews Neuroscience*, Vol.6 (9).
DOI: 10.1038/nrn1746.

— (2008). “Negative regulation of myelination: Relevance for development, injury, and demyelinating disease”. *Glia*, Vol.56 (14).
DOI: 10.1002/glia.20761.

— (2019a). “Schwann cell precursors; multipotent glial cells in embryonic nerves”. *Frontiers in Molecular Neuroscience*, Vol.12.
DOI: 10.3389/fnmol.2019.00069.

— (2019b). “The success and failure of the Schwann cell response to nerve injury”. *Frontiers in Cellular Neuroscience*, Vol.13.
DOI: 10.3389/fncel.2019.00033.

Jessen, Kristjan R. *et al.* (2015). “Schwann cells: Development and role in nerve repair”. *Cold Spring Harbor Perspectives in Biology*, Vol.7 (7).
DOI: 10.1101/cshperspect.a020487.

Jiang, Chunhui *et al.* (2021). “Tumorigenesis in neurofibromatosis type 1: role of the microenvironment”. *Oncogene*, Vol.40 (39).
DOI: 10.1038/s41388-021-01979-z.

Jiang, Chunhui *et al.* (2023). “Basement membrane proteins in extracellular matrix characterize NF1 neurofibroma development and response to MEK inhibitor”. *Journal of Clinical Investigation*, Vol.133 (12).
DOI: 10.1172/JCI168227.

Jo, Alice *et al.* (2014). “The versatile functions of Sox9 in development, stem cells, and human diseases”. *Genes and Diseases*, Vol.1 (2).
DOI: 10.1016/j.gendis.2014.09.004.

Johnson, W. Evan *et al.* (2007). “Adjusting batch effects in microarray expression data using empirical Bayes methods”. *Biostatistics*, Vol.8 (1).
DOI: 10.1093/biostatistics/kxj037.

Joost, Simon *et al.* (2020). “The Molecular Anatomy of Mouse Skin during Hair Growth and Rest”. *Cell Stem Cell*, Vol.26 (3).
DOI: 10.1016/j.stem.2020.01.012.

Julian, Lisa M. *et al.* (2017). “Direct reprogramming with SOX factors: masters of cell fate”. *Current Opinion in Genetics & Development*, Vol.46.
DOI: 10.1016/j.gde.2017.06.005.

Kalucka, Joanna *et al.* (2020). “Single-Cell Transcriptome Atlas of Murine Endothelial Cells”. *Cell*, Vol.180 (4).
DOI: 10.1016/j.cell.2020.01.015.

Kastriti, Maria Eleni *et al.* (2022). “Schwann cell precursors represent a neural crest-like state with biased multipotency”. *The EMBO Journal*, Vol.41 (17).
DOI: 10.15252/embj.2021108780.

Katkar, Gajanan & Pradipta Ghosh (2023). “Macrophage states: there’s a method in the madness”. *Trends in Immunology*.
DOI: 10.1016/j.it.2023.10.006.

Keefe, Christopher R. *et al.* (2023). “Facilitating bioinformatics reproducibility with QIIME 2 Provenance Replay”. *PLOS Computational Biology*, Vol.19 (11).
DOI: 10.1371/journal.pcbi.1011676.

Kehrer-Sawatzki, Hildegard *et al.* (2017). “Emerging genotype–phenotype relationships in patients with large NF1 deletions”. *Human Genetics*, Vol.136 (4).
DOI: 10.1007/s00439-017-1766-y.

Kerr, Norbert L. (1998). “HARKing: Hypothesizing After the Results are Known”. *Personality and Social Psychology Review*, Vol.2 (3).
DOI: 10.1207/s15327957pspr0203_4.

Kershner, Leah J. *et al.* (2022). “Multiple Nf1 Schwann cell populations reprogram the plexiform neurofibroma tumor microenvironment”. *JCI Insight*, Vol.7 (18).
DOI: 10.1172/jci.insight.154513.

Kidd, Grahame J. *et al.* (2013). “Biology of Schwann cells”. *Handobook of clinical neurology*.
DOI: 10.1016/B978-0-444-52902-2.00005-9.

Kim, Daehwan *et al.* (2019). “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype”. *Nature Biotechnology*, Vol.37 (8).
DOI: 10.1038/s41587-019-0201-4.

Kim, Yang-Min *et al.* (2018). “Experimenting with reproducibility: a case study of robustness in bioinformatics”. *GigaScience*, Vol.7 (7).
DOI: 10.1093/gigascience/giy077.

Kochat, Veena *et al.* (2021). “Enhancer reprogramming in PRC2-deficient malignant peripheral nerve sheath tumors induces a targetable de-differentiated state”. *Acta Neuropathologica*, Vol.142 (3).
DOI: 10.1007/s00401-021-02341-z.

Koczkowska, Magdalena *et al.* (2018). “Genotype-Phenotype Correlation in NF1: Evidence for a More Severe Phenotype Associated with Missense Mutations Affecting NF1 Codons 844–848”. *The American Journal of Human Genetics*, Vol.102 (1).
DOI: 10.1016/j.ajhg.2017.12.001.

Koczkowska, Magdalena *et al.* (2020). “Clinical spectrum of individuals with pathogenic NF1 missense variants affecting p.Met1149, p.Arg1276, and p.Lys1423: genotype-phenotype study in neurofibromatosis type 1”. *Human Mutation*, Vol.41 (1).
DOI: 10.1002/humu.23929.

Kohlmeyer, Jordan L. *et al.* (2023). “CDK4/6-MEK Inhibition in MPNSTs Causes Plasma Cell Infiltration, Sensitization to PD-L1 Blockade, and Tumor Regression”. *Clinical Cancer Research*, Vol.29 (17).
DOI: 10.1158/1078-0432.CCR-23-0749.

Kohrs, Friederike E. *et al.* (2023). “Eleven Strategies for Making Reproducible Research and Open Science Training the Norm at Research Institutions”. *eLife*.
DOI: 10.7554/eLife.89736.

Korsunsky, Ilya *et al.* (2019). “Fast, sensitive and accurate integration of single-cell data with Harmony”. *Nature Methods*, Vol.16 (12).
DOI: 10.1038/s41592-019-0619-0.

Köster, Johannes & Sven Rahmann (2012). "Snakemake-a scalable bioinformatics workflow engine". *Bioinformatics*, Vol.28 (19).
DOI: 10.1093/bioinformatics/bts480.

Kreso, Antonija & John E. Dick (2014). "Evolution of the Cancer Stem Cell Model". *Cell Stem Cell*, Vol.14 (3).
DOI: 10.1016/j.stem.2014.02.006.

Kucenas, Sarah (2015). "Perineurial glia". *Cold Spring Harbor Perspectives in Biology*, Vol.7 (6).
DOI: 10.1101/cshperspect.a020511.

Kurtzer, Gregory M. *et al.* (2017). "Singularity: Scientific containers for mobility of compute". *PLOS ONE*, Vol.12 (5).
DOI: 10.1371/journal.pone.0177459.

Kyaw, Wunna *et al.* (2023). "ENTRAIN: integrating trajectory inference and gene regulatory networks with spatial data to co-localize the receptor–ligand interactions that specify cell fate". *Bioinformatics*, Vol.39 (12).
DOI: 10.1093/bioinformatics/btad765.

La Manno, Gioele *et al.* (2018). "RNA velocity of single cells". *Nature*, Vol.560 (7719).
DOI: 10.1038/s41586-018-0414-6.

Lafzi, Atefeh *et al.* (2018). "Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies". *Nature Protocols*, Vol.13 (12).
DOI: 10.1038/s41596-018-0073-y.

Lavie, Dor *et al.* (2022). "Cancer-associated fibroblasts in the single-cell era". *Nature Cancer*, Vol.3 (7).
DOI: 10.1038/s43018-022-00411-z.

Le, Lu Q. *et al.* (2009). "Cell of Origin and Microenvironment Contribution for NF1-Associated Dermal Neurofibromas". *Cell Stem Cell*, Vol.4 (5).
DOI: 10.1016/j.stem.2009.03.017.

Le, Lu Q. *et al.* (2011). "Susceptible Stages in Schwann Cells for NF1-Associated Plexiform Neurofibroma Development". *Cancer Research*, Vol.71 (13).
DOI: 10.1158/0008-5472.CAN-10-4577.

Le Pennec, Erwan & Kamil Slowikowski (2023). *ggwordcloud: A Word Cloud Geom for 'ggplot2'*.
URL: https://github.com/lepennec/ggwordcloud.

Lee, William *et al.* (2014). "PRC2 is recurrently inactivated through EED or SUZ12 loss in malignant peripheral nerve sheath tumors". *Nature Genetics*, Vol.46 (11).
DOI: 10.1038/ng.3095.

Leek, Jeffrey T. *et al.* (2012). "The sva package for removing batch effects and other unwanted variation in high-throughput experiments". *Bioinformatics*, Vol.28 (6).
DOI: 10.1093/bioinformatics/bts034.

Legius, Eric *et al.* (2021). "Revised diagnostic criteria for neurofibromatosis type 1 and Legius syndrome: an international consensus recommendation". *Genetics in Medicine*, Vol.23 (8).
DOI: 10.1038/s41436-021-01170-5.

Lehmann, Nathalie (2021). "Development of bioinformatics tools for single-cell transcriptomics applied to the search for signatures of symmetric versus asymmetric division mode in neural progenitors". PhD thesis. IBENS - Institut de biologie de l'ENS Paris.
URL: https://theses.hal.science/tel-04089445v1.

Leonelli, Sabina (2018). "Rethinking Reproducibility as a Criterion for Research Quality". *Emerald Publishing Limited*.
DOI: 10.1108/S0743-41542018000036B009.

Li, Stephen *et al.* (2020). "New insights into the neurofibroma tumor cells of origin". *Neuro-Oncology Advances*, Vol.2 (Supplement_1).
DOI: 10.1093/noajnl/vdz044.

Liao, Chung Ping *et al.* (2018). “Contributions of inflammation and tumor microenvironment to neurofibroma tumorigenesis”. *Journal of Clinical Investigation*, Vol.128 (7).
DOI: 10.1172/JCI99424.

Liao, Chung-Ping *et al.* (2016). “The role of nerve microenvironment for neurofibroma development”. *Oncotarget*, Vol.7 (38).
DOI: 10.18632/oncotarget.11133.

Liao, Chung-Ping *et al.* (2019). “Krox20 in epithelial and glial stem cells and their niches”. *Advances in Stem Cells and their Niches*.
DOI: 10.1016/bs.asn.2019.04.002.

Love, Michael I. *et al.* (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. *Genome Biology*, Vol.15 (12).
DOI: 10.1186/s13059-014-0550-8.

Lu, Yuanjun *et al.* (2020). “Epigenetic regulation in human cancer: the potential role of epi-drug in cancer therapy”. *Molecular Cancer*, Vol.19 (1).
DOI: 10.1186/s12943-020-01197-3.

Luecken, Malte D. & Fabian J. Theis (2019). “Current best practices in single-cell RNA-seq analysis: a tutorial”. *Molecular Systems Biology*, Vol.15 (6).
DOI: 10.15252/msb.20188746.

Luecken, Malte D. *et al.* (2022). “Benchmarking atlas-level data integration in single-cell genomics”. *Nature Methods*, Vol.19 (1).
DOI: 10.1038/s41592-021-01336-8.

Lun, Aaron T. L. *et al.* (2019). “EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data”. *Genome Biology*, Vol.20 (1).
DOI: 10.1186/s13059-019-1662-y.

Madanikia, Sara Aileen *et al.* (2012). “Increased risk of breast cancer in women with NF1”. *American Journal of Medical Genetics Part A*, Vol.158A (12).
DOI: 10.1002/ajmg.a.35550.

Magallón-Lorenz, Miriam *et al.* (2021). “Chromosomal translocations inactivating CDKN2A support a single path for malignant peripheral nerve sheath tumor initiation”. *Human Genetics*, Vol.140 (8).
DOI: 10.1007/s00439-021-02296-x.

Magallón-Lorenz, Miriam *et al.* (2023). “Deep genomic analysis of malignant peripheral nerve sheath tumor cell lines challenges current malignant peripheral nerve sheath tumor diagnosis”. *iScience*, Vol.26 (2).
DOI: 10.1016/j.isci.2023.106096.

Mandl, Maximilian M. *et al.* (2024). “Raising awareness of uncertain choices in empirical data analysis: A teaching concept toward replicable research practices”. *PLOS Computational Biology*, Vol.20 (3).
DOI: 10.1371/journal.pcbi.1011936.

Maniatis, Tom & Robin Reed (2002). “An extensive network of coupling among gene expression machines”. *Nature*, Vol.416.
DOI: 10.1038/416499a.

Margueron, Raphaël & Danny Reinberg (2011). “The Polycomb complex PRC2 and its mark in life”. *Nature*, Vol.469 (7330).
DOI: 10.1038/nature09784.

Martin, Enrico *et al.* (2019). “Non-cytotoxic systemic treatment in malignant peripheral nerve sheath tumors (MPNST): A systematic review from bench to bedside”. *Critical Reviews in Oncology/Hematology*, Vol.138.
DOI: 10.1016/j.critrevonc.2019.04.007.

Martin, Enrico *et al.* (2020). “Prognostic Significance of Immunohistochemical Markers and Genetic Alterations in Malignant Peripheral Nerve Sheath Tumors: A Systematic Review”. *Frontiers in Oncology*, Vol.10.
DOI: 10.3389/fonc.2020.594069.

Marx, Vivien (2019). “A dream of single-cell proteomics”. *Nature Methods*, Vol.16 (9).
DOI: 10.1038/s41592-019-0540-6.

Mautner, Victor-Felix. *et al.* (2008). “Assessment of benign tumor burden by whole-body MRI in patients with neurofibromatosis 1”. *Neuro-Oncology*, Vol.10 (4).
DOI: 10.1215/15228517-2008-011.

Mayes, Debra A. *et al.* (2011). “Perinatal or adult Nf1 inactivation using tamoxifen-inducible PlpCre each cause neurofibroma formation”. *Cancer Research*, Vol.71 (13).
DOI: 10.1158/0008-5472.CAN-10-4558.

Mayor, Roberto & Eric Theveneau (2013). “The neural crest”. *Development*, Vol.140 (11).
DOI: 10.1242/dev.091751.

Mazuelas, Helena *et al.* (2022). “Modeling iPSC-derived human neurofibroma-like tumors in mice uncovers the heterogeneity of Schwann cells within plexiform neurofibromas”. *Cell Reports*, Vol.38 (7).
DOI: 10.1016/j.celrep.2022.110385.

Mazuelas, Helena *et al.* (2024). “Unbalancing cAMP and Ras/MAPK pathways as a therapeutic strategy for cutaneous neurofibromas”. *JCI Insight*.
DOI: 10.1172/jci.insight.168826.

McInnes, Leland *et al.* (2018). “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. *arXiv*.
DOI: 10.48550/arXiv.1802.03426.

Mereu, Elisabetta *et al.* (2020). “Benchmarking single-cell RNA-sequencing protocols for cell atlas projects”. *Nature Biotechnology*, Vol.38 (6).
DOI: 10.1038/s41587-020-0469-4.

Merkel, Dirk (2014). “Docker: lightweight linux containers for consistent development and deployment”. *Linux journal*, Vol.2014 (239).
URL: https://dl.acm.org/doi/10.5555/2600239.2600241.

Messersmith, Lynn & Kevin Krauland (2019). *Neurofibroma*. StatPearls Publishing, Treasure Island (FL).
URL: https://www.ncbi.nlm.nih.gov/books/NBK539707/.

Meyer, Anders (2020). “Review and update in the diagnosis of peripheral nerve sheath tumors”. *Current opinion in neurology*, Vol.33 (5).
DOI: 10.1097/WCO.0000000000000857.

Miettinen, Markku M. *et al.* (2017). “Histopathologic evaluation of atypical neurofibromatous tumors and their transformation into malignant peripheral nerve sheath tumor in patients with neurofibromatosis 1—a consensus overview”. *Human Pathology*, Vol.67.
DOI: 10.1016/j.humpath.2017.05.010.

Mikolajewicz, Nicholas *et al.* (2022). “Multi-level cellular and functional annotation of single-cell transcriptomes using scPipeline”. *Communications Biology*, Vol.5 (1).
DOI: 10.1038/s42003-022-04093-2.

Miller, Shyra J. *et al.* (2009). “Integrative genomic analyses of neurofibromatosis tumours identify SOX9 as a biomarker and survival gene”. *EMBO Molecular Medicine*, Vol.1 (4).
DOI: 10.1002/emmm.200900027.

Mirsky, Rhona *et al.* (2002). “Schwann cells as regulators of nerve development”. *Journal of Physiology-Paris*, Vol.96 (1-2).
DOI: 10.1016/S0928-4257(01)00076-6.

Missarova, Alsu *et al.* (2023). “Sensitive cluster-free differential expression testing”. *BioRXiv*.
DOI: 10.1101/2023.03.08.531744.

Mo, Juan *et al.* (2020). “Humanized neurofibroma model from induced pluripotent stem cells delineates tumor pathogenesis and developmental origins”. *Journal of Clinical Investigation*, Vol.131 (1).
DOI: 10.1172/JCI139807.

Mo, Juan *et al.* (2022). “Neurofibromin and suppression of tumorigenesis: beyond the GAP”. *Oncogene*, Vol.41 (9).
DOI: 10.1038/s41388-021-02156-y.

Mohamad, Teddy *et al.* (2021). “Toward understanding the mechanisms of malignant peripheral nerve sheath tumor development”. *International Journal of Molecular Sciences*, Vol.22 (16).
DOI: 10.3390/ijms22168620.

Monk, Kelly R. *et al.* (2015). “New insights on schwann cell development”. *Glia*, Vol.63 (8).
DOI: 10.1002/glia.22852.

Moore, Amanda R. *et al.* (2020). “RAS-targeted therapies: is the undruggable drugged?” *Nature Reviews Drug Discovery*, Vol.19 (8).
DOI: 10.1038/s41573-020-0068-6.

Moses, Lambda & Lior Pachter (2022). “Museum of spatial transcriptomics”. *Nature Methods*, Vol.19 (5).
DOI: 10.1038/s41592-022-01409-2.

Müller-Dott, Sophia *et al.* (2023). “Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities”. *Nucleic Acids Research*, Vol.51 (20).
DOI: 10.1093/nar/gkad841.

Najm, Matthieu *et al.* (2024). “Representation and quantification of module activity from omics data with rROMA”. *npj Systems Biology and Applications*, Vol.10 (1).
DOI: 10.1038/s41540-024-00331-x.

Napolitano, Filomena *et al.* (2022). “Genotype-Phenotype Correlations in Neurofibromatosis Type 1: Identification of Novel and Recurrent NF1 Gene Variants and Correlations with Neurocognitive Phenotype”. *Genes*, Vol.13 (7).
DOI: 10.3390/genes13071130.

Neerven, S. G.A. van *et al.* (2013). “Schwann cell migration and neurite outgrowth are influenced by media conditioned by epineurial fibroblasts”. *Neuroscience*, Vol.252.
DOI: 10.1016/j.neuroscience.2013.08.009.

Nerve tissue, synapses, and neurotransmitters (2022). https://www.amboss.com/us/knowledge/nerve-tissue-synapses-and-neurotransmitters. Accessed: 2023-11-03.

Nguyen, Hai C. T. *et al.* (2023). “Benchmarking integration of single-cell differential expression”. *Nature Communications*, Vol.14 (1).
DOI: 10.1038/s41467-023-37126-3.

Nguyen, Rosa *et al.* (2014). “Benign whole body tumor volume is a risk factor for malignant peripheral nerve sheath tumors in neurofibromatosis type 1”. *Journal of Neuro-Oncology*, Vol.116 (2).
DOI: 10.1007/s11060-013-1293-1.

Niehues, Anna *et al.* (2024). “A multi-omics data analysis workflow packaged as a FAIR Digital Object”. *GigaScience*, Vol.13.
DOI: 10.1093/gigascience/giad115.

Nix, J. Stephen *et al.* (2018). “Malignant Peripheral Nerve Sheath Tumors Show Decreased Global DNA Methylation”. *Journal of Neuropathology & Experimental Neurology*, Vol.77 (10).
DOI: 10.1093/jnen/nly076.

Nobeyama, Yoshimasa *et al.* (2022). “Abnormal peripheral blood cell counts in neurofibromatosis type 1”. *Scientific Reports*, Vol.12 (1).
DOI: 10.1038/s41598-022-23739-z.

Oberhofer, Angela *et al.* (2022). "Tracing the Origin of Cell-Free DNA Molecules through Tissue-Specific Epigenetic Signatures". *Diagnostics*, Vol.12 (8).
DOI: 10.3390/diagnostics12081834.

Pacot, Laurence *et al.* (2024). "Identification of potential common genetic modifiers of neurofibromas: a genome-wide association study in 1333 patients with neurofibromatosis type 1". *British Journal of Dermatology*, Vol.190 (2).
DOI: 10.1093/bjd/ljad390.

Palazzo, Alexander F. & Eliza S. Lee (2015). "Non-coding RNA: what is functional and what is junk?" *Frontiers in Genetics*, Vol.6.
DOI: 10.3389/fgene.2015.00002.

Palomo-Irigoyen, Marta *et al.* (2020). "HuR/ELAVL1 drives malignant peripheral nerve sheath tumor growth and metastasis". *Journal of Clinical Investigation*, Vol.130 (7).
DOI: 10.1172/JCI130379.

Parrinello, Simona & Alison C. Lloyd (2009). "Neurofibroma development in NF1 - insights into tumour initiation". *Trends in Cell Biology*, Vol.19 (8).
DOI: 10.1016/j.tcb.2009.05.003.

Parrinello, Simona *et al.* (2010). "EphB signaling directs peripheral nerve regeneration through Sox2-dependent Schwann cell Sorting". *Cell*, Vol.143 (1).
DOI: 10.1016/j.cell.2010.08.039.

Parsa, Cameron F. *et al.* (2001). "Spontaneous Regression of Optic Gliomas: Thirteen Cases Documented by Serial Neuroimaging". *Archives of Ophthalmology*, Vol.119 (4).
DOI: 10.1001/archopht.119.4.516.

Patro, Rob *et al.* (2017). "Salmon provides fast and bias-aware quantification of transcript expression". *Nature Methods*, Vol.14 (4).
DOI: 10.1038/nmeth.4197.

Pekmezci, Melike *et al.* (2017). "Significance of H3K27me3 loss in the diagnosis of malignant peripheral nerve sheath tumors". *Modern Pathology*, Vol.30 (12).
DOI: 10.1038/modpathol.2017.97.

Perry, Arie *et al.* (2001). "Short Communication NF1 Deletions in S-100 Protein-Positive and Negative Cells of Sporadic and Neurofibromatosis 1 (NF1)-Associated Plexiform Neurofibromas and Malignant Peripheral Nerve Sheath Tumors". *American Journal of Pathology*, Vol.159 (1).
DOI: 10.1016/S0002-9440(10)61673-2.

Picelli, Simone *et al.* (2013). "Smart-seq2 for sensitive full-length transcriptome profiling in single cells". *Nature Methods*, Vol.10 (11).
DOI: 10.1038/nmeth.2639.

Piña-Oviedo, Sergio & Carlos Ortiz-Hidalgo (2008). "The normal and neoplastic perineurium: A review". *Advances in Anatomic Pathology*, Vol.15 (3).
DOI: 10.1097/PAP.0b013e31816f8519.

Prada, Carlos E. *et al.* (2012a). "Neurofibroma-associated macrophages play roles in tumor growth and response to pharmacological inhibition". *Acta Neuropathologica*, Vol.125 (1).
DOI: 10.1007/s00401-012-1056-7.

Prada, Carlos E. *et al.* (2012b). "Pediatric Plexiform Neurofibromas: Impact on Morbidity and Mortality in Neurofibromatosis Type 1". *The Journal of Pediatrics*, Vol.160 (3).
DOI: 10.1016/j.jpeds.2011.08.051.

Prudner, Bethany C. *et al.* (2020). "Diagnosis and management of malignant peripheral nerve sheath tumors: Current practice and future perspectives". *Neuro-Oncology Advances*, Vol.2.
DOI: 10.1093/noajnl/vdz047.

Qi, Jingcui *et al.* (2022). "Recent advances in the regulation mechanism of SOX10". *Journal of Otology*, Vol.17 (4).
DOI: 10.1016/j.joto.2022.08.003.

Radomska, Katarzyna J. & Piotr Topilko (2017). "Boundary cap cells in development and disease". *Current Opinion in Neurobiology*, Vol.47.
DOI: 10.1016/j.conb.2017.11.003.

Radomska, Katarzyna J. *et al.* (2019). "Cellular origin, tumor progression, and pathogenic mechanisms of cutaneous neurofibromas revealed by mice with Nf1 knockout in boundary cap cells". *Cancer Discovery*, Vol.9 (1).
DOI: 10.1158/2159-8290.CD-18-0156.

Rasulic, Lukas *et al.* (2021). "Diagnostic Assessment and Treatment of Peripheral Nerve Tumors". *Neurofibromas*. Cham: Springer International Publishing.
DOI: 10.1007/978-3-030-77633-6.

Ratner, Nancy & Shyra J. Miller (2015). "A RASopathy gene commonly mutated in cancer: The neurofibromatosis type 1 tumour suppressor". *Nature Reviews Cancer*, Vol.15 (5).
DOI: 10.1038/nrc3911.

Rauen, Katherine A. (2022). "Defining RASopathy". *Disease Models & Mechanisms*, Vol.15 (2).
DOI: 10.1242/dmm.049344.

Reilly, Karlyne M. *et al.* (2017). "Neurofibromatosis Type 1–Associated MPNST State of the Science: Outlining a Research Agenda for the Future". *Journal of the National Cancer Institute*, Vol.109 (8).
DOI: 10.1093/jnci/djx124.

Reinhold, Anna K. & Heike L. Rittner (2020). "Characteristics of the nerve barrier and the blood dorsal root ganglion barrier in health and disease". *Experimental Neurology*, Vol.327.
DOI: 10.1016/j.expneurol.2020.113244.

Rhodes, Steven D. *et al.* (2019). "Cdkn2a (Arf) loss drives NF1-associated atypical neurofibroma and malignant transformation". *Human Molecular Genetics*, Vol.28 (16).
DOI: 10.1093/hmg/ddz095.

Ribeiro, Sara *et al.* (2013). "Injury Signals Cooperate with Nf1 Loss to Relieve the Tumor-Suppressive Environment of Adult Peripheral Nerve". *Cell Reports*, Vol.5 (1).
DOI: 10.1016/j.celrep.2013.08.033.

Richard, Laurence *et al.* (2012). "Endoneurial Fibroblast-Like Cells". *Journal of Neuropathology & Experimental Neurology*, Vol.71 (11).
DOI: 10.1097/NEN.0b013e318270a941.

Röhrich, Manuel *et al.* (2016). "Methylation-based classification of benign and malignant peripheral nerve sheath tumors". *Acta Neuropathologica*, Vol.131 (6).
DOI: 10.1007/s00401-016-1540-6.

Saelens, Wouter *et al.* (2019). "A comparison of single-cell trajectory inference methods". *Nature Biotechnology*, Vol.37 (5).
DOI: 10.1038/s41587-019-0071-9.

Sahai, Erik *et al.* (2020). "A framework for advancing our understanding of cancer-associated fibroblasts". *Nature Reviews Cancer*, Vol.20 (3).
DOI: 10.1038/s41568-019-0238-1.

Salzer, James L. (2015). "Schwann cell myelination". *Cold Spring Harbor Perspectives in Biology*, Vol.7 (8).
DOI: 10.1101/cshperspect.a020529.

Samuel, Sheeba & Daniel Mietchen (2024). "Computational reproducibility of Jupyter notebooks from biomedical publications". *GigaScience*, Vol.13.
DOI: 10.1093/gigascience/giad113.

Sbaraglia, Marta *et al.* (2020). "The 2020 WHO Classification of Soft Tissue Tumours: news and perspectives". *Pathologica*, Vol.113 (2).
DOI: 10.32074/1591-951X-213.

Sbidian, E. *et al.* (2016). "Neurofibromatosis type 1: neurofibromas and sex". *British Journal of Dermatology*, Vol.174 (2).
DOI: 10.1111/bjd.13966.

Schaefer, Inga-Marie & Christopher D.M. Fletcher (2015). “Malignant Peripheral Nerve Sheath Tumor (MPNST) Arising in Diffuse-type Neurofibroma”. *American Journal of Surgical Pathology*, Vol.39 (9).
DOI: 10.1097/PAS.0000000000000447.

— (2018). “Recent advances in the diagnosis of soft tissue tumours”. *Pathology*, Vol.50 (1).
DOI: 10.1016/j.pathol.2017.07.007.

Schubert, Michael *et al.* (2018). “Perturbation-response genes reveal signaling footprints in cancer gene expression”. *Nature Communications*, Vol.9 (1).
DOI: 10.1038/s41467-017-02391-6.

Schulz, Jörg B. *et al.* (2016). “The impact of fraudulent and irreproducible data to the translational research crisis – solutions and implementation”. *Journal of Neurochemistry*, Vol.139 (S2).
DOI: 10.1111/jnc.13844.

Schwaederlé, Maria *et al.* (2014). “Cyclin alterations in diverse cancers: outcome and co-amplification network”. *Oncotarget*, Vol.6 (5).
DOI: 10.18632/oncotarget.2848.

Sharif, Saba *et al.* (2006). “Second Primary Tumors in Neurofibromatosis 1 Patients Treated for Optic Glioma: Substantial Risks After Radiotherapy”. *Journal of Clinical Oncology*, Vol.24 (16).
DOI: 10.1200/JCO.2005.03.8349.

Shurell, Elizabeth *et al.* (2016). “Characterizing the immune microenvironment of malignant peripheral nerve sheath tumor by PD-L1 expression and presence of CD8+ tumor infiltrating lymphocytes”. *Oncotarget*, Vol.7 (39).
DOI: 10.18632/oncotarget.11734.

Sikkema, Arend H. *et al.* (2009). “Kinome Profiling in Pediatric Brain Tumors as a New Approach for Target Discovery”. *Cancer Research*, Vol.69 (14).
DOI: 10.1158/0008-5472.CAN-08-3660.

Simões-Costa, Marcos & Marianne E. Bronner (2015). “Establishing neural crest identity: a gene regulatory recipe”. *Development*, Vol.142 (2).
DOI: 10.1242/dev.105445.

Simoneau, Joël *et al.* (2021). “Current RNA-seq methodology reporting limits reproducibility”. *Briefings in Bioinformatics*, Vol.22 (1).
DOI: 10.1093/bib/bbz124.

Single Cell 3’ v3.1 Dual Index Gene Expression (2023). https://www.10xgenomics.com/support/single-cell-gene-expression/documentation/steps/sequencing/sequencing-requirements-for-single-cell-3. Accessed: 2024-02-16.

Soldatov, Ruslan *et al.* (2019). “Spatiotemporal structure of cell fate decisions in murine neural crest”. *Science*, Vol.364 (6444).
DOI: 10.1126/science.aas9536.

Solovieva, Tatiana & Marianne E. Bronner (2021). “Schwann cell precursors: Where they come from and where they go”. *Cells and Development*, Vol.166.
DOI: 10.1016/j.cdev.2021.203686.

Somatilaka, Bandarigoda N. *et al.* (2022). “Malignant peripheral nerve sheath tumor: models, biology, and translation”. *Oncogene*, Vol.41 (17).
DOI: 10.1038/s41388-022-02290-1.

Somatilaka, Bandarigoda N. *et al.* (2024). “STING activation reprograms the microenvironment to sensitize NF1-related malignant peripheral nerve sheath tumors for immunotherapy”. *Journal of Clinical Investigation*, Vol.134 (10).
DOI: 10.1172/JCI176748.

Soneson, Charlotte & Mark D Robinson (2018). “Bias, robustness and scalability in single-cell differential expression analysis”. *Nature Methods*, Vol.15 (4).
DOI: 10.1038/nmeth.4612.

Spyk, Sebastian *et al.* (2011). “Neurofibromatosis type 1-associated tumours: Their somatic mutational spectrum and pathogenesis”. *Human Genomics*, Vol.5 (6).
DOI: 10.1186/1479-7364-5-6-623.

Spyra, Melanie *et al.* (2011). “Cancer Stem Cell-Like Cells Derived from Malignant Peripheral Nerve Sheath Tumors”. *PLoS ONE*, Vol.6 (6).
DOI: 10.1371/journal.pone.0021099.

Stewart, John D (2003). “Peripheral nerve fascicles: anatomy and clinical relevance”. *Muscle Nerve*, Vol.28.
DOI: 10.1002/mus.10454.

Stierli, Salome *et al.* (2018). “The regulation of the homeostasis and regeneration of peripheral nerve is distinct from the CNS and independent of a stem cell population”. *Development*, Vol.145 (24).
DOI: 10.1242/dev.170316.

Street, Kelly *et al.* (2018). “Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics”. *BMC Genomics*, Vol.19 (1).
DOI: 10.1186/s12864-018-4772-0.

Stuart, Tim *et al.* (2019). “Comprehensive Integration of Single-Cell Data”. *Cell*, Vol.177 (7).
DOI: 10.1016/j.cell.2019.05.031.

Sun, Daochun *et al.* (2021). “Stem-like cells drive NF1-associated MPNST functional heterogeneity and tumor progression”. *Cell Stem Cell*, Vol.28 (8).
DOI: 10.1016/j.stem.2021.04.029.

Sun, Shiquan *et al.* (2019). “Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis”. *Genome Biology*, Vol.20 (1).
DOI: 10.1186/s13059-019-1898-6.

Sundaram, Venkat Krishnan *et al.* (2019). “Liver X Receptors and Their Implications in the Physiology and Pathology of the Peripheral Nervous System”. *International Journal of Molecular Sciences*, Vol.20 (17).
DOI: 10.3390/ijms20174192.

Sundby, R. Taylor *et al.* (2024). “Early detection of malignant and pre-malignant peripheral nerve tumors using cell-free DNA fragmentomics Introduction”. *medRxiv*.
DOI: 10.1101/2024.01.18.24301053.

Suppiah, Suganth *et al.* (2023). “Multiplatform molecular profiling uncovers two subgroups of malignant peripheral nerve sheath tumors with distinct therapeutic vulnerabilities”. *Nature Communications*, Vol.14 (1).
DOI: 10.1038/s41467-023-38432-6.

Svensson, Valentine *et al.* (2018). “Exponential scaling of single-cell RNA-seq in the past decade”. *Nature Protocols*, Vol.13 (4).
DOI: 10.1038/nprot.2017.149.

Szukiewicz, Dariusz (2024). “CX3CL1 (Fractalkine)-CX3CR1 Axis in Inflammation-Induced Angiogenesis and Tumorigenesis”. *International Journal of Molecular Sciences*, Vol.25 (9).
DOI: 10.3390/ijms25094679.

Szymanski, Jeffrey J. *et al.* (2021). “Cell-free DNA ultra-low-pass whole genome sequencing to distinguish malignant peripheral nerve sheath tumor (MPNST) from its benign precursor lesion: A cross-sectional study”. *PLOS Medicine*, Vol.18 (8).
DOI: 10.1371/journal.pmed.1003734.

Tang, Fuchou *et al.* (2009). “mRNA-Seq whole-transcriptome analysis of a single cell”. *Nature Methods*, Vol.6 (5).
DOI: 10.1038/nmeth.1315.

Taveggia, Carla *et al.* (2010). “Signals to promote myelin formation and repair”. *Nature Reviews Neurology*, Vol.6 (5).
DOI: 10.1038/nrneurol.2010.37.

Tay, Andy (2024). “The open-science movement for sharing laboratory materials gains momentum”. *Nature*, Vol.625 (7996).
DOI: 10.1038/d41586-024-00172-y.

Theveneau, Eric & Roberto Mayor (2012). “Neural crest delamination and migration: From epithelium-to-mesenchyme transition to collective cell migration”. *Developmental Biology*, Vol.366 (1).
DOI: 10.1016/j.ydbio.2011.12.041.

Thomson, Craig S. *et al.* (2021). “WNT5A inhibition alters the malignant peripheral nerve sheath tumor microenvironment and enhances tumor growth”. *Oncogene*, Vol.40 (24).
DOI: 10.1038/s41388-021-01773-x.

Timp, Winston & Gregory Timp (2020). “Beyond mass spectrometry, the next step in proteomics”. *Science Advances*, Vol.6 (2).
DOI: 10.1126/sciadv.aax8978.

Todorov, Helena *et al.* (2020). “TinGa: fast and flexible trajectory inference with Growing Neural Gas”. *Bioinformatics*, Vol.36 (Supplement_1).
DOI: 10.1093/bioinformatics/btaa463.

Toma, Jeremy S. *et al.* (2020). “Peripheral nerve single-cell analysis identifies mesenchymal ligands that promote axonal growth”. *eNeuro*, Vol.7 (3).
DOI: 10.1523/ENEURO.0066-20.2020.

Tommaso, Paolo Di *et al.* (2017). “Nextflow enables reproducible computational workflows”. *Nature Biotechnology*, Vol.35 (4).
DOI: 10.1038/nbt.3820.

Tran, Hoa Thi Nhu *et al.* (2020). “A benchmark of batch-effect correction methods for single-cell RNA sequencing data”. *Genome Biology*, Vol.21 (1).
DOI: 10.1186/s13059-019-1850-9.

Tripathi, Surya Kant *et al.* (2022). “SOX9 as an emerging target for anticancer drugs and a prognostic biomarker for cancer drug resistance”. *Drug Discovery Today*, Vol.27 (9).
DOI: 10.1016/j.drudis.2022.05.022.

“Trust but verify” (2024). *Nature Materials*, Vol.23 (1).
DOI: 10.1038/s41563-023-01790-z.

Tsuchie, Hiroyuki *et al.* (2023). “Soft Tissue Tumor Without Neurofibromatosis Type 1 With Histopathological Suspicion of Atypical Neurofibromatous Neoplasm of Uncertain Biological Potential (ANNUBP)”. *Cureus*.
DOI: 10.7759/cureus.38187.

Tucker, T. *et al.* (2005). “Association between benign and malignant peripheral nerve sheath tumors in NF1”. *Neurology*, Vol.65 (2).
DOI: 10.1212/01.wnl.0000168830.79997.13.

Valentin, T. *et al.* (2016). “Management and prognosis of malignant peripheral nerve sheath tumors: The experience of the French Sarcoma Group (GSF-GETO)”. *European Journal of Cancer*, Vol.56.
DOI: 10.1016/j.ejca.2015.12.015.

Valeyrie-Allanore, Laurence *et al.* (2008). “Histopathologically dysplastic neurofibromas in neurofibromatosis 1: diagnostic criteria, prevalence and clinical significance”. *British Journal of Dermatology*, Vol.158 (5).
DOI: 10.1111/j.1365-2133.2008.08494.x.

Vallet, Nicolas *et al.* (2022). “Toward practical transparent verifiable and long-term reproducible research using Guix”. *Scientific Data*, Vol.9 (1).
DOI: 10.1038/s41597-022-01720-9.

Van de Sande, Bram *et al.* (2020). “A scalable SCENIC workflow for single-cell gene regulatory network analysis”. *Nature Protocols*, Vol.15 (7).
DOI: 10.1038/s41596-020-0336-2.

Vandenberk, Bert *et al.* (2011). “p16INK4a: A central player in cellular senescence and a promising aging biomarker in elderly cancer patients”. *Journal of Geriatric Oncology*, Vol.2 (4).
DOI: 10.1016/j.jgo.2011.08.004.

Vieth, Beate *et al.* (2019). “A systematic evaluation of single cell RNA-seq analysis pipelines”. *Nature Communications*, Vol.10 (1).
DOI: 10.1038/s41467-019-12266-7.

Walker, James A. & Meena Upadhyaya (2018). “Emerging therapeutic targets for neurofibromatosis type 1”. *Expert Opinion on Therapeutic Targets*, Vol.22 (5).
DOI: 10.1080/14728222.2018.1465931.

Warbey, Victoria S. *et al.* (2009). “[18F]FDG PET/CT in the diagnosis of malignant peripheral nerve sheath tumours in neurofibromatosis type-1”. *European Journal of Nuclear Medicine and Molecular Imaging*, Vol.36 (5).
DOI: 10.1007/s00259-008-1038-0.

Wassef, Michel & Raphaël Margueron (2017). “The Multiple Facets of PRC2 Alterations in Cancers”. *Journal of Molecular Biology*, Vol.429 (13).
DOI: 10.1016/j.jmb.2016.10.012.

White, Emily E. & Steven D. Rhodes (2024). “The NF1+/- Immune Microenvironment: Dueling Roles in Neurofibroma Development and Malignant Transformation”. *Cancers*, Vol.16 (5).
DOI: 10.3390/cancers16050994.

Wilkinson, Mark D. *et al.* (2016). “The FAIR Guiding Principles for scientific data management and stewardship”. *Scientific Data*, Vol.3 (1).
DOI: 10.1038/sdata.2016.18.

Willems, Stefan M. *et al.* (2010). “Running GAGs: myxoid matrix in tumor pathology revisited”. *Virchows Archiv*, Vol.456 (2).
DOI: 10.1007/s00428-009-0822-y.

Wojcik, John B. *et al.* (2019). “Epigenomic Reordering Induced by Polycomb Loss Drives Oncogenesis but Leads to Therapeutic Vulnerabilities in Malignant Peripheral Nerve Sheath Tumors”. *Cancer Research*, Vol.79 (13).
DOI: 10.1158/0008-5472.CAN-18-3704.

Wolf, F. Alexander *et al.* (2018). “SCANPY: large-scale single-cell gene expression data analysis”. *Genome Biology*, Vol.19 (1).
DOI: 10.1186/s13059-017-1382-0.

Wratten, Laura *et al.* (2021). “Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers”. *Nature Methods*, Vol.18 (10).
DOI: 10.1038/s41592-021-01254-9.

Wu, Guanggeng *et al.* (2024). “Roles of Macrophages and Their Interactions with Schwann Cells After Peripheral Nerve Injury”. *Cellular and Molecular Neurobiology*, Vol.44 (1).
DOI: 10.1007/s10571-023-01442-5.

Wu, Jianqiang *et al.* (2008). “Plexiform and Dermal Neurofibromas and Pigmentation Are Caused by Nf1 Loss in Desert Hedgehog-Expressing Cells”. *Cancer Cell*, Vol.13 (2).
DOI: 10.1016/j.ccr.2007.12.027.

Wu, Lai Man Natalie *et al.* (2018). “Programming of Schwann Cells by Lats1/2-TAZ/YAP Signaling Drives Malignant Peripheral Nerve Sheath Tumorigenesis”. *Cancer Cell*, Vol.33 (2).
DOI: 10.1016/j.ccell.2018.01.005.

Wu, Lai Man Natalie *et al.* (2022a). “Single-cell multiomics identifies clinically relevant mesenchymal stem-like cells and key regulators for MPNST malignancy”. *Sci. Adv*, Vol.8 (44).
DOI: 10.1126/sciadv.abo5442.

Wu, Sijie *et al.* (2022b). “Single-cell transcriptomics reveals lineage trajectory of human scalp hair follicle and informs mechanisms of hair graying”. *Cell Discovery*, Vol.8 (1).
DOI: 10.1038/s41421-022-00394-2.

Xi, Nan Miles & Jingyi Jessica Li (2021). “Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data”. *Cell Systems*, Vol.12 (2).
DOI: 10.1016/j.cels.2020.11.008.

Xu, Si-Mei *et al.* (2022). “RNA polyadenylation patterns in the human transcriptome”. *Gene*, Vol.816.
DOI: 10.1016/j.gene.2021.146133.

Yamanaka, Ryuya & Azusa Hayano (2017). “Radiation-Induced Malignant Peripheral Nerve Sheath Tumors: A Systematic Review”. *World Neurosurgery*, Vol.105.
DOI: 10.1016/j.wneu.2017.06.010.

Yao, Chengjun *et al.* (2023a). “Malignant Peripheral Nerve Sheath Tumors: Latest Concepts in Disease Pathogenesis and Clinical Management”. *Cancers*, Vol.15 (4).
DOI: 10.3390/cancers15041077.

Yao, Xuan *et al.* (2023b). “Mechanism of SOX9 nuclear translocation induced by NF1 mutation promoting neurobroma proliferation and collagen secretion”. *ResearchSquare*.
DOI: 10.21203/rs.3.rs-2974925/v1.

Ydens, Elke *et al.* (2020). “Profiling peripheral nerve macrophages reveals two macrophage subsets with distinct localization, transcriptome and response to injury”. *Nature Neuroscience*, Vol.23 (5).
DOI: 10.1038/s41593-020-0618-6.

Yeoh, Stewart *et al.* (2022). “Incorporating Blood Flow in Nerve Injury and Regeneration Assessment”. *Frontiers in Surgery*, Vol.9.
DOI: 10.3389/fsurg.2022.862478.

Yim, Aldrin K.Y. *et al.* (2022). “Disentangling glial diversity in peripheral nerves at single-nuclei resolution”. *Nature Neuroscience*, Vol.25 (2).
DOI: 10.1038/s41593-021-01005-1.

Yu, Guangchuang *et al.* (2012). “clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters”. *OMICS: A Journal of Integrative Biology*, Vol.16 (5).
DOI: 10.1089/omi.2011.0118.

Yu, Lijia *et al.* (2022). “Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data”. *Genome Biology*, Vol.23 (1).
DOI: 10.1186/s13059-022-02622-0.

Zehou, Ouidad *et al.* (2013). “Chemotherapy for the treatment of malignant peripheral nerve sheath tumors in neurofibromatosis 1: a 10-year institutional review”. *Orphanet Journal of Rare Diseases*, Vol.8 (1).
DOI: 10.1186/1750-1172-8-127.

Zhang, Xiannian *et al.* (2019). “Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems”. *Molecular Cell*, Vol.73 (1).
DOI: 10.1016/j.molcel.2018.10.020.

Zhang, Xiyuan *et al.* (2020a). “The role of polycomb repressive complex in malignant peripheral nerve sheath tumor”. *Genes*, Vol.11 (3).
DOI: 10.3390/genes11030287.

Zhang, Xiyuan *et al.* (2022). “Single-cell sequencing reveals activation of core transcription factors in PRC2-deficient malignant peripheral nerve sheath tumor”. *Cell Reports*, Vol.40 (12).
DOI: 10.1016/j.celrep.2022.111363.

Zhang, Yaru *et al.* (2020b). “Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data”. *Computational and Structural Biotechnology Journal*, Vol.18.
DOI: 10.1016/j.csbj.2020.10.007.

Zhao, Xiao-Feng *et al.* (2022). “The injured sciatic nerve atlas (iSNAT), insights into the cellular and molecular basis of neural tissue degeneration and regeneration”. *eLife*, Vol.11.
DOI: 10.7554/eLife.80881.

Zheng, Grace X. Y. *et al.* (2017). “Massively parallel digital transcriptional profiling of single cells”. *Nature Communications*, Vol.8 (1).
DOI: 10.1038/ncomms14049.

Zheng, Huarui *et al.* (2008). “Induction of Abnormal Proliferation by Nonmyelinating Schwann Cells Triggers Neurofibroma Formation”. *Cancer Cell*, Vol.13 (2).
DOI: 10.1016/j.ccr.2008.01.002.

Zhu, Yuan *et al.* (2002). “Neurofibromas in NF1: Schwann cell origin and role of tumor environment”. *Science*, Vol.296 (5569).
DOI: 10.1126/science.1068452.

Ziemann, Mark *et al.* (2023). “The five pillars of computational reproducibility: bioinformatics and beyond”. *Briefings in Bioinformatics*, Vol.24 (6).
DOI: 10.1093/bib/bbad375.

Zitvogel, Laurence *et al.* (2017). “Nutrition, inflammation and cancer”. *Nature Immunology*, Vol.18 (8).
DOI: 10.1038/ni.3754.