An investigation of tumor heterogeneity based on computational approaches

Magali Richard

UNIVERSITE GRENOBLE-ALPES

Mémoire

présenté en vue de l'obtention d'une

Habilitation à Diriger des Recherches

Laboratoire Recherche Translationnelle et Innovation en Médecine et Complexité (TIMC)

École doctorale EDISCE (Ecole Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement)

An investigation of tumor heterogeneity based on computational approaches

Par : Magali RICHARD

Rapporteurs et rapportrice:

Laurence CALZONE, Ingénieure, Institut Curie
Olivier FRANCOIS, PR, Grenoble-INP, laboratoire TIMC
Franck PICARD, DR CNRS, ENS Lyon, laboratoire LBMC

Examinateurs et examinatrices:

Isabelle GUYON, PR, Univ Paris-Saclay, laboratoire LRI INRIA
Philippe JUIN, DR INSERM, Univ. Nantes, laboratoire CRCI2NA
François PARCY, DR CNRS, CEA Grenoble, laboratoire LPVC
Céline VALLOT, DR CNRS, Institut Curie

Date de soutenance : 10 Octobre 2022

Abstract

The goal of this HDR thesis is to summarize my research activity over the past five years and to present my future scientific prospects. My focus will be on studying tumor heterogeneity through computational and interdisciplinary methods. Cancer is a heterogeneous disease, with each tumor evolving as an autonomous multicellular system. The tumor ecosystem consists of cells from different origins and identities that dynamically interact with each other. It is difficult to observe and quantify the heterogeneity in tumor composition, a key factor contributing to cancer progression. Our current inability to estimate heterogeneity in a given sample has hindered our understanding of its function during oncogenic processes. A review of the computational approaches used to estimate tumor heterogeneity and to study its functional implications in cancer progression will be presented in the first three chapters of the thesis. Concepts exposed will relate to my own work in the field of oncology, as a computational biologist. I will also explore the broad field of collaborative work and problem of algorithms evaluation, featuring some of my contributions to this area. I will conclude by discussing my scientific perspectives and my personal journey as a young independent researcher, including challenges associated with the position.

Résumé

L'objectif de cette thèse HDR est de résumer mon activité de recherche au cours des cinq dernières années et de présenter mes perspectives scientifiques futures. Je me concentrerai sur l'étude de l'hétérogénéité tumorale à l'aide de méthodes informatiques et interdisciplinaires. Le cancer est une maladie hétérogène, chaque tumeur évoluant comme un système multicellulaire autonome. L'écosystème tumoral est constitué de cellules d'origines et d'identités différentes qui interagissent dynamiquement les unes avec les autres. Il est difficile d'observer et de quantifier l'hétérogénéité de la composition tumorale, un facteur clé contribuant à la progression du cancer. Notre incapacité actuelle à estimer l'hétérogénéité dans un échantillon donné a entravé notre compréhension de sa fonction au cours des processus oncogènes. Une revue des approches computationnelles utilisées pour estimer l'hétérogénéité tumorale et pour étudier ses implications fonctionnelles dans la progression du cancer sera présentée dans les trois premiers chapitres de la thèse. Les concepts exposés seront liés à mes propres travaux dans le domaine de l'oncologie, en tant que biologiste computationelle. Je vais ensuite explorer le vaste domaine du travail collaboratif et de l'évaluation des algorithmes, en incluant certaines de mes contributions dans ce domaine. Je conclurai en évoquant mes perspectives scientifiques et mon parcours personnel en tant que jeune chercheuse indépendante.

Remerciements

Je souhaite remercier tout.e.s mes collègues, collaborateurs et collaboratrices qui ont contribué autant que moi aux résultats présentés dans ce manuscrit. Merci également aux équipes 'support' des différents laboratoire dans lesquelles j'ai travaillé.

Merci à Alona, Clémentine, Elise et Florence pour la relecture de dernière minute !

Merci à Laurence, Olivier et Franck d'avoir accepté de rapporter ce travail, et à Isabelle, Philippe, François et Céline d'être examinateurs.

Merci à Florent et les enfants, de me tenir occupée quand je ne suis pas sur l'ordinateur.

Et enfin, merci à Nicolas Travers pour ce template[1].

[1] https://chewbii.com/latex-template/

Forewords

A journey from experimental to computational genetics

My research work is devoted to the study of heterogeneity in living systems. What are the genetic and non-genetic determinants of the observed heterogeneity in differentiation, behavior and proliferation of living cells? During my career, I have tried to answer these fundamental questions through complementary approaches in experimental biology, biostatistics and computational biology.

Initially trained in biology, I was immersed during my master and PhD in experimental genetics and cell biology. I developed an expertise in fundamental genetics, molecular biology, and biochemistry, leading to the first description of the evolutionary conserved EMC complex, essential for the maturation of transmembrane receptors in eukaryotes. This work was published in PNAS (2013) and eLife (2018). During my postdoc, I implemented probabilistic and quantitative genetic approaches (experimental and computational) to study the genetic regulation of cellular behavior in *S. cerevisiae*. This original approach allowed me to discovered new mechanisms underlying the non-deterministic effects of natural genetic variations, published in Plos Genetics (2016) and twice in Molecular System Biology (2018).

For my postdoc, I joined a multidisciplinary team made up of biologists, computer scientists and statisticians to learn how to develop biostatistical approaches to analyse high-throughput omic data. In parallel, I obtained a university degree in applied statistics. I wanted to be able to use computational approaches to raise hypothesis from experimental data and to test it back at the bench. It was difficult for me to carry out both wet experiments and the development of dedicated statistical methods in parallel, but I persisted, and finally this period allowed me to develop a unique multidisciplinary expertise. It was a founding factor in the construction of my scientific vision and enabled me to make key discoveries by combining experimental acquisitions at high-throughput (deep DNA sequencing) with mathematical modelling, predictions and subsequent experimental validations [Salignon et al., 2018, Richard et al., 2018] . Thanks to this dual background, I can read biology papers in a creative way, I know the pitfalls of experimental approaches and I am able to design precise statistical methods to answer fundamental biology questions. My double expertise, as biologist and bio-statistician, was key to my recruitment to the CNRS in 2018 on a tenure position, and place me in a privileged position to conduct innovative research in computational biology.

Guided by a desire to study complex multicellular systems differentiation, I became interested in the evolution of tumor as a complex heterogeneous ecosystem. I joined the team 'Methods and Algorithm for Genomics (MAGe)', at the TIMC laboratory, to tackle this question through the development of novel dedicated computational approaches. TIMC is a multidisciplinary lab that gathers scientists and clinicians towards the use of quantitative science for understanding normal and pathological processes in biology and healthcare. In the past three years, I have recruited my own independent group (currently composed of one PhD student, one postdoc and internships). My group has developed, with key collaborators, methods and pipelines to study inter and intra-tumor heterogeneity, some of which were published in high-profile journals, such as Plos Computational Biology (2020) and BMC bioinformatics (2020 and 2021). I am currently the coordinator of a national consortium (ITMO cancer AVIESAN), aiming at studying spatial heterogeneity in pancreatic cancer. In parallel, I started to collaborate with Isabelle Guyon. Inspired by her work in organizing competitions in Artificial Intelligence, I independently proposed a series of Health Data Challenges, and led European funded consortiums dedicated to the organization of competitions and trainings (EIT Health HADACA and COMETH). I now serve as director at the board of ChaLearn (non-profit organization dedicated to the organization of data challenges).

Contents

Abstract 1

Remerciements 2

Forewords 3

Contents 4

1 General introduction 9

1.1 Towards a definition of cancer heterogeneity . . . 9
1.2 Motivations to study cancer heterogeneity . . . 12
1.3 Methodological and computational challenges . . . 13
1.4 Use cases of this thesis : lung and pancreatic cancers . . . 14
1.4.1 Non-small cells lung cancers . . . 14
1.4.2 Pancreatic adeno-carcinoma . . . 14
1.5 Overview of the thesis . . . 15

2 Estimation of inter-tumor heterogeneity 17

2.1 From high-throughput molecular information to personalized analysis . . . 18
2.2 The PenDA method . . . 19
2.2.1 Background . . . 19
2.2.2 Principles of the PenDA method . . . 20
2.2.3 Comparison of PenDA with other individual-based methods . . . 21
2.3 Application of PenDA to non-small cells lung cancers . . . 22
2.4 A generalization of the approach – *on going* . . . 26
2.4.1 Use case 1: a pan-cancer analysis . . . 26
2.4.2 Use case 2: individual metabolism regulation of glioblastoma . . . 27

3 Estimation of intra-tumor heterogeneity 29

3.1 Benchmarking unsupervised deconvolution approaches . . . 30
3.1.1 Cell-types heterogeneity quantification from DNA methylation . . . 31
3.1.2 Providing guidelines for cell-type heterogeneity deconvolution . . . 35
3.2 Development of a single-cell reference based PDAC deconvolution method – *on going* 35
3.2.1 Identification of 14 specific PDAC cell-types . . . 36
3.2.2 Generation of unified integrated gene markers and reference cell-types . . . 38

3.2.3 Deconvolution of PDAC samples using new robust cell-types markers and profiles . . . 40
3.3 Method development multi-omic integration and estimation of tumor functional heterogeneity– *prospects* . . . 42
3.3.1 Multi-omic based deconvolution of intra-tumor heterogeneity and patient classification . . . 42
3.3.2 Estimation of tumor functional heterogeneity at the single tumor level . . . 43

4 A functional interpretation of intra-tumor heterogeneity 49
4.1 Relationship between tumor functional heterogeneity and cancer evolution . . . 49
4.1.1 An evolutionary view of pancreatic adenocarcinoma . . . 50
4.1.2 Relationship between intra-tumor heterogeneity and somatic mutations. . . . 50
4.1.3 A systematic study of intra-tumor heterogeneity and (epi)genomic landscape – *prospects* . . . 51
4.2 A causal link between heterogeneity, environment and outcome . . . 53
4.2.1 Introduction to mediation analysis . . . 53
4.2.2 Development of a new method to perform high-dimension mediation analysis 54
4.2.3 Identification of molecular mechanisms by which tumor heterogeneity influences disease outcome – *prospects* . . . 55

5 Algorithms evaluation and collaborative science 57
5.1 Introduction to data challenges, a new avenue for collaborative science . . . 58
5.2 Feedbacks on data challenge organization . . . 60
5.2.1 Unsupervised deconvolution of methylation data. . . . 61
5.2.2 Multiomic integration. . . . 61
5.2.3 Towards a user-friendly online tool for clinicians. . . . 64
5.3 Towards a continuous benchmark . . . 65
5.3.1 Codabench, a novel benchmarking platform . . . 65
5.3.2 The next international challenge & benchmark – *prospects* . . . 67

6 General conclusion 69
6.1 Scientific perspectives . . . 69
6.1.1 Scientific contributions to cancer biology . . . 69
6.1.2 Multidisciplinary approach for a better healthcare . . . 70
6.1.3 Next challenges in computational oncology . . . 70
6.2 Personal considerations on my early career . . . 71
6.2.1 Pitfalls of computational biology and bioinformatics . . . 71
6.2.2 Navigating in a multidisciplinary environment . . . 72
6.2.3 Slow science : doing less but better? . . . 73

7 Curriculum Vitae **75**

7.1 Training and Appointments . . . 75
7.2 Fellowship and Awards . . . 75
7.3 Publications, softwares and conferences . . . 75
7.3.1 Research articles . . . 75
7.3.2 Review articles in refereed journals and books . . . 76
7.3.3 Open-access softwares . . . 77
7.3.4 Speaking engagements . . . 77
7.4 Fundings . . . 78
7.5 Expertise, editorial and scientific activities . . . 79
7.6 Scientific collaborations . . . 79
7.7 Supervision and mentoring . . . 79
7.8 Administrative responsibilities . . . 79
7.9 Teaching . . . 79

List of figures **79**

Bibliography **82**

Introduction to cancer heterogeneity

In this chapter, I will attempt to define cancer heterogeneity and introduce the motivations and the challenges associated to the study of cancer heterogeneity. Then I will present our two cancer use-cases. I will finish by an overview of the thesis content.

1.1 Towards a definition of cancer heterogeneity

Cancer is a disease caused by the accumulation of alterations (genomic, epigenomic, metabolic...). During oncogenesis, tumoral cells evolve heterogeneously due to various mechanisms including environmental cues and cell-to-cell interactions. Eventually, immune evasion and metastases are observed. Literally, the word *heterogeneous* defines something composed of elements of different kind, and thus *heterogeneity* refers to a quality of being heterogeneous. Applied to cancer, the word heterogeneity can have several meanings, which we will detail below.

If we consider only cancer cells, the basis of heterogeneity can be define as genomic and epigenomic variations that will lead to

- Intra-tumor heterogeneity: sub-clonal composition, i.e. tumor are heterogeneous as they are composed of different sub-clones of cancer cells, Figure 1.1a.
- Inter-tumor or intra-patient heterogeneity : this heterogeneity accounts for spatial and temporal heterogeneity within a patient (metastatic versus primary tumor), Figure 1.1b.
- Inter-tumor or inter-patient heterogeneity : a cancer affecting a given organ will be different between each patient, Figure 1.1c.

This view is centered on cancer cells, but intra-tumor heterogeneity is also commonly used to define the cellular composition of the tumor mass, including the tumor micro-environment (Figure 1.2). The tumor micro-environment defines all the cells present in a solid tumor that are not cancer cells (stromal and immune components). The level of heterogeneity of a tumor is sometimes defined as the level of purity of the tumor, high-purity corresponding to a tumor with a large quantity of cancerous cells, low-purity corresponding to a tumor with a large proportion of non-cancerous cells present in the micro-environment (Figure 1.3).

Inter-tumor heterogeneity is often used for patient classification and stratifications but efforts are made to also integrate the concept of intra-tumor heterogeneity within existing models. For instance, a recent review on primary liver cancer discussed the importance of accounting for both intra and inter-tumor hetereogeneity to better understand the disease behaviour (Figure 1.4) [Liu et al., 2018].

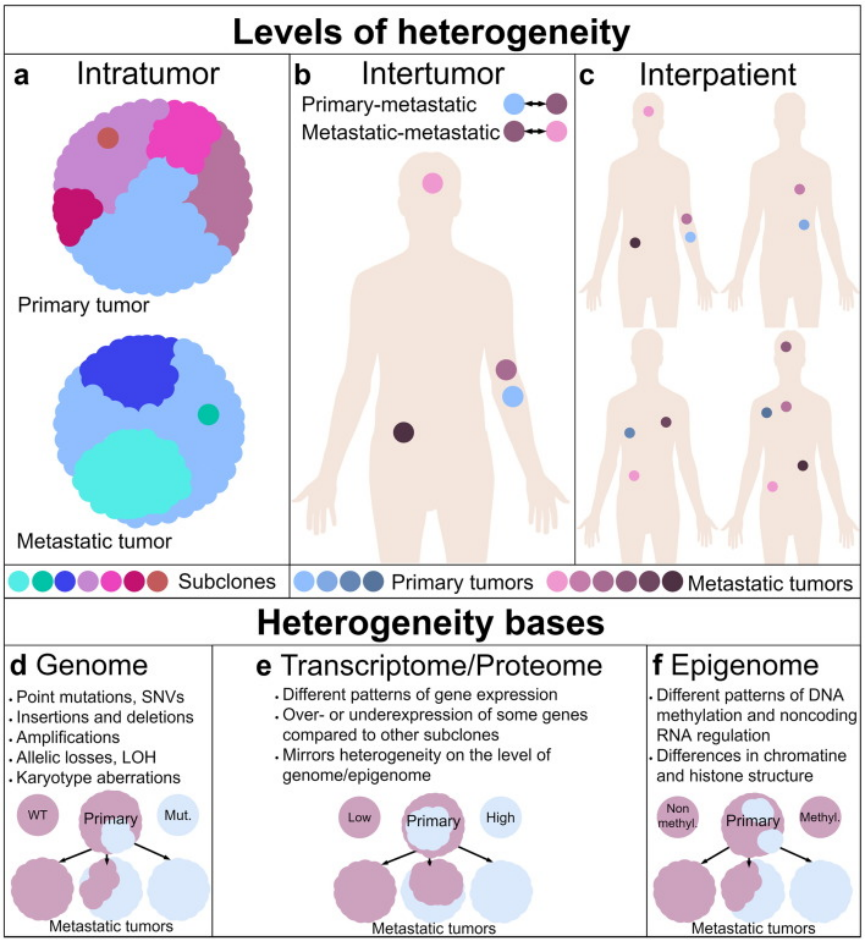

Figure 1.1: **Scales of tumor heterogeneity**
From [Grzywa et al., 2017].

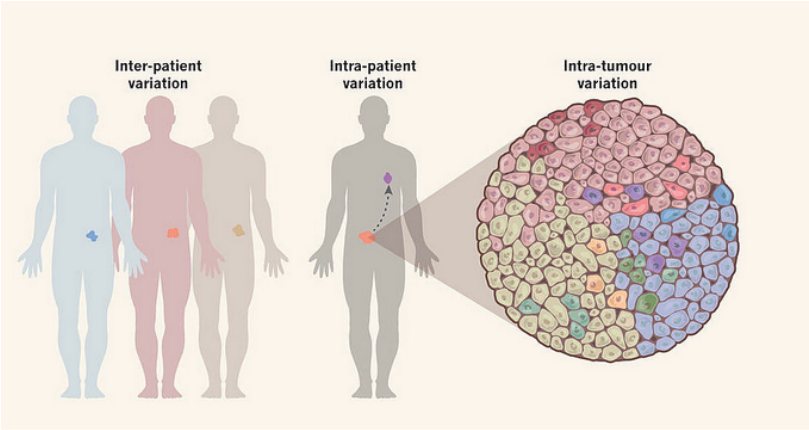

Figure 1.2: **The genetic characteristics of cancers vary between patients, between primary and metastatic tumours in a single patient, and between the individual cells of a tumour.**
Wang et al. present a single-cell, whole-genome sequencing technique that will allow a better understanding of genetic heterogeneity within individual tumours. From[Fox and Loeb, 2014].

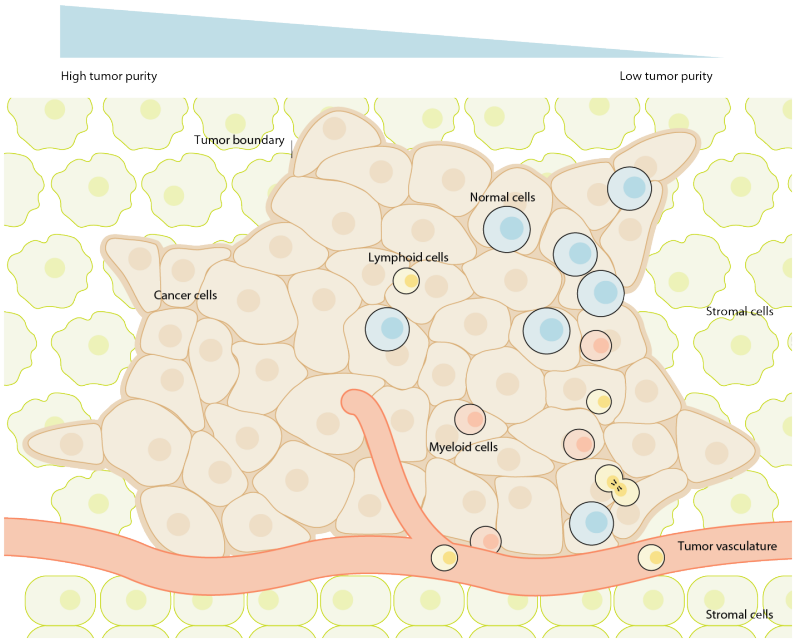

Figure 1.3: **Scales of tumor heterogeneity**
A tumor is a complex ecosystem composed of various cell types which show heterogeneous spatial distributions. The cell types within a tumor generally contain cancer cell clones, normal cells that have not been transformed, stromal cells, immune cells, and endothelial cells. From [Ren et al., 2018].

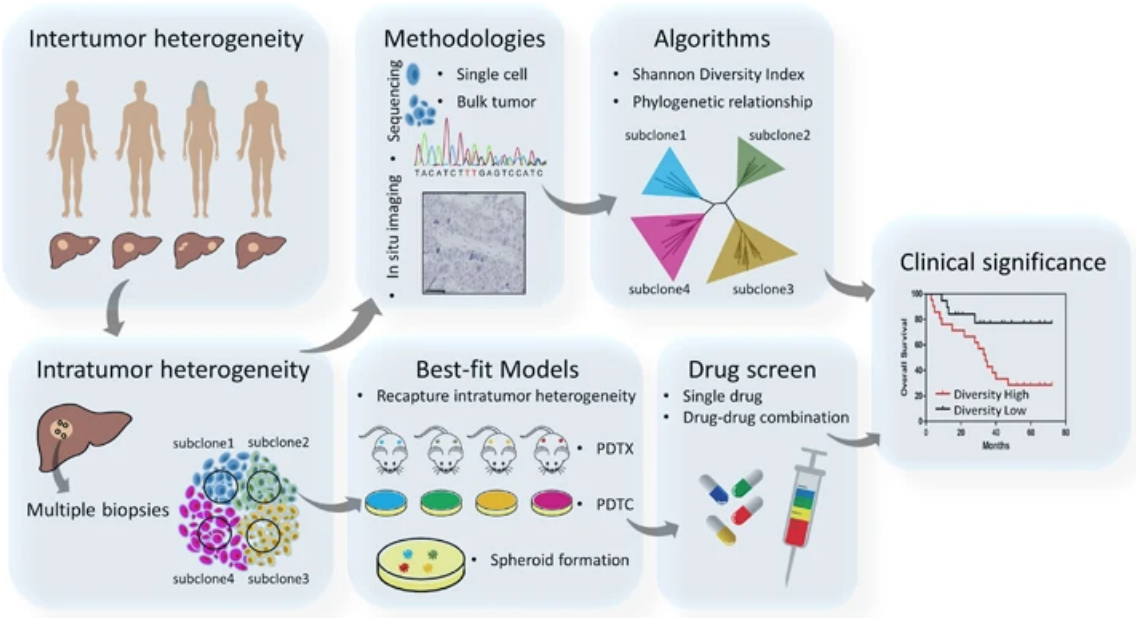

Figure 1.4: **Intra and inter-tumor heterogeneity in Primary Liver Cancer**
A schematic diagram of understanding, recapturing intra-tumor heterogeneity of PLC on best-fit models and their applications in drug screen and prognosis. PLC heterogeneity includes inter-tumor and intra-tumor heterogeneity. From [Liu et al., 2018].

In my research projects, I study both inter-tumor heterogeneity and intra-tumor heterogeneity. In this manuscript, I will rely on the definition provided by Ren *et al.* in their review "Understanding tumor ecosystems by single-cell sequencing: promises and limitations" [Ren et al., 2018].

"Cancer is known for its heterogeneity, at the inter- and intra-tumor levels. Within a tumor, different spatial sites have different composition of cancer cell clones, which results in spatial heterogeneity. As cancer cells evolve, temporal variations also arise during the course of cancer genesis and progression, causing temporal heterogeneity. In addition to cancer cells, tumors are also infiltrated with stromal, immune, and other cell types. The diversity of these cells forms the basis of the heterogeneity of the tumor microenvironments. The complex and dynamic nature of cancer heterogeneity within tumors is analogous to ecosystems. Thorough understanding of the composition, interactions, dynamics, and operating principles of tumor ecosystems is key to understanding cancer evolution and the emergence of drug resistance."

1.2 Motivations to study cancer heterogeneity

Molecular subtyping and classification of cancer patients has been mainly achieved by population stratification approaches, which provided a rough estimate of inter-patient heterogeneity. In many cases, this work has led to personalized therapies, with successful therapeutic applications, for example for acute leukemia [Druker et al., 2006] or for breast cancer [Heiser et al., 2012]. Targeted therapies used in mainstream clinical oncology include trastuzumab for HER2-expressing breast cancers, vemurafenib for BRAF-mutated melanomas, and immune checkpoint inhibitors to treat tumors with microsatellite instability. However, these population approaches are limited because they do not take into account the individual specificity of each patient. How can we understand the pathology of a patient who responds poorly to targeted therapy? How can we improve the management of a patient with a poor prognosis?

Indeed, cancer cell populations exhibit genetic, epigenetic, transcriptional and phenotypical heterogeneity, among patients, as well as within patients. Except for genetic variations, diversity in cancer cells strongly relies on macro and micro environmental cues. These cues trigger inheritable modifications that will influence tumor cells proliferation towards the development and the maintenance of an autonomous multicellular system [Lloyd et al., 2016]. Environmental cues can be separated into two main classes: extrinsic factors, such as patient history and drug treatment, and intrinsic factors, such as tumor heterogeneity, including different cancer cells and various tumor micro-environment compositions. Tumor micro-environment is a dynamic environment constituted of non-cancerous cells, such as fibroblasts, blood vessels, and immune cells [Lambrechts et al., 2018]. The tumor micro-environment pro- and anti-tumor activity during tumorigenesis is complex and remains to be explained [Connor and Gallinger, 2021]. Tumor heterogeneity triggers a large variety of changes and interactions that are not observed in normal tissue, and it has been demonstrated to have a major impact on cancer cells phenotype, such as growth and division [Marusyk et al., 2020], resistance to therapy [Hirata and Sahai, 2017] and metastasis potential [Karnoub et al., 2007]. Overall, intra-tumor heterogeneity is critical for disease outcome [Meacham and Morrison, 2013].

At the molecular level, tumor composition is difficult to assess and quantify, as it is hidden inside the bulk molecular profiles of the samples (averaged profile from millions of cells), with all cells present in the tumor (and not only cancer cells) contributing to the recorded signal. From a biomedical perspective, this heterogeneity affects patient classification (based on bulk molecular data) and is therefore a key parameter to account for in biomarkers detection and personalized treatment, such as targeted immuno therapies [Puleo et al., 2018]. Yet, computational studies aiming to study cancer biology often fail to integrate the underlying intra-tumor heterogeneity, giving only a partial picture of the real biological processes at play. Accordingly, accounting for tumor composition has not yet been included in patient care and clinical practice. Deciphering tumor heterogeneity is methodologically highly challenging (see next paragraph), but solving this issue will open new avenues towards a better understanding of the mechanisms by which tumors can evolve within an organism. In addition, it will offer leads to predict the patient response to treatment, notably in the context of personalized therapies.

1.3 Methodological and computational challenges

The hopes of precision medicine rely on our capacity to measure high throughput molecular information for each patient and to integrate this information for personalized diagnosis and treatment. Such challenging perspectives will be only possible with the concomitant development of efficient and robust methodological tools that allow the identification of molecular defects at the individual level. Currently, it exists many population-based methods, like DESeq2 [Love et al., 2014] or edgeR [Robinson et al., 2010], that robustly infer differentially expressed genes between two sets of samples, like for example between normal and tumorous tissues obtained from a cohort of cancerous patients. However, very few tools allow to take a reliable decision on gene deregulation in individual – e.g. tumorous - sample.

Historically, tumor composition has been studied using immunocytochemistry or flow cytometry approaches, which are experimental methods to estimate cell type heterogeneity. They rely on a small set of molecular markers and are therefore limited by the number of cell types that can be simultaneously quantified. Taking advantage of the large amount of bulk omic data publicly available, a wide number of supervised and unsupervised algorithms have been recently developed to estimate tumor composition [Cantini et al., 2019, Avila Cobos et al., 2020]. Overall, despite intensive recent methodological developments, estimated cell proportions from published algorithms are not always consistent with each other and do not provide sufficient robustness [Li et al., 2020, Sturm et al., 2019]. Single-cell sequencing has contributed to the development of a new type of supervised methods, taking advantage of single-cell profiles to infer cell-type specific molecular profiles [Wang et al., 2019, Steen et al., 2020]. Unfortunately, these methods only capture a subset of living cells, do not account for cell-to-cell interactions, and are still too costly for routine clinical practice. The lack of confidence in tumor composition estimation has hampered our understanding of cancer as a multicellular system, particularly with respect to tumor initiation, evolution and response to therapy. Consistently, a causal inference of the molecular mechanisms structuring the relationship between the tumor composition, the environmental cues (therapeutic treatment), and the outcome (patient survival) is still lacking.

Multi-omics approaches (multiple measurements of all molecular events of different types from the same sample) are powerful means to address heterogeneity problems. For instance, combining gene expression and DNA methylation (DNAm, a non-heritable chemical modification of the DNA sequence that regulates gene expression) captures different properties of cellular states while reducing the impact of experimental and biological noise [Cantini et al., 2021]. Taking advantage of this wealth of information provided by multi-omics data presents inherent challenges: high-dimensionality, missing data, different signal-to-noise ratios, and interpretability of the models. Therefore, multi-omics data integration aiming to solve precise biological questions requires dedicated mathematical methods, which are yet to be developed [Tarazona et al., 2020]. These methodological developments can only arise from multi-disciplinary approaches that engage a truly collaborative research community, including clinicians, biostatisticians and bioinformaticians.

1.4 Use cases of this thesis : lung and pancreatic cancers

1.4.1 Non-small cells lung cancers

Lung cancer are generally classified into three main histologies: non-small cell cancers, small cell cancers and carcinoids. Non-small cell cancers represent 85% of cases, and are themselves divided into three several classes, among which the two most common which derive from epithelial cells: lung adenocarcinoma (LUAD) and lung squamous cells carcinoma (LUSC). LUAD has a very high rate of somatic mutations and genomic rearrangements, which makes it particularly difficult to identify cancer-driving mutations [Collisson et al., 2014]. LUSC derives from squamous cells and has a high rate of genetic deregulation [Hammerman et al., 2012]. LUAD and LUSC are globally very different, for example the genetic alterations of LUSC resemble other squamous cell carcinomas more than those of LUAD, and the therapeutic targets are not the same between those two lung cancers [Campbell et al., 2016].

We developed our methods using "The Cancer Genome Atlas" (TCGA)[1]. The TCGA is a program launched in 2006 jointly between two American institutes: the NCI (National Cancer Institute) and the NHGRI (National Human Genome Research Institute). Today, the TCGA represents a huge database containing thousands of omics data of different molecular types, on 33 different cancer types, as well as matched healthy tissues. The choice of lung cancer was mainly dictated by a local collaboration between our team and that of pathologist Elisabeth Brambilla from Grenoble University Hospital, responsible for a large cohort of lung cancer data [Rousseaux et al., 2013], that we used as a validation cohort.

1.4.2 Pancreatic adeno-carcinoma

Pancreatic adenocarcinoma (PDAC) is a very aggressive and invasive tumoral lesion affecting the pancreas. This malignancy is asymptomatic until reaching an advanced stage, often liver metastasis [Paik et al., 2012], that makes it disadvantageous against treatments. Despite the significant efforts made in this field, the PDAC remains the 4th commonest lethal cancers with a median of

[1] https://portal.gdc.cancer.gov/

relative survival rate of 6 month and a 5-year survival below 8%. PDAC incidence increases regularly in Western countries and is expected to become the second leading cause of cancer-related mortality in 2025 [Ferlay et al., 2016]. Although the genetic lesions are common (such as KRAS, TP53, SMAD4, CDKN2A) [Biankin et al., 2012], epigenetic modifications are also driving a remodeling of the transcriptomic landscape, leading to heterogeneous shapes and behaviour of cancerous cells [Sausen et al., 2015, Lomberk et al., 2018].

The development of fast and cost-effective technologies for high-throughput sequencing has triggered the generation of multi-omics data repositories, such as the International Cancer Genome Consortium (ICGC) [Hudson (Chairperson) et al., 2010] and the The Cancer Genome Atlas (TCGA) [Weinstein et al., 2013]. These public datasets include bulk somatic genomic alteration, gene expression and DNA methylation data, as well as clinical data from the analyzed cohorts for PDAC cancers. However, high quality dedicated datasets with corresponding ground truth were also required to benchmark our algorithms. Our team, together with our collaborators J. Cros (AP-HP) and Y. Blum (IGDR), generated some of the datasets that were used to generate the results presented in this manuscript.

1.5 Overview of the thesis

In this thesis, I present the work I have done over the past five years and the scientific questions that I intend to pursue over the next five years. For the sake of clarity, I decided to tackle only on the scientific question of cancer heterogeneity. My previous research work can be found as published articles. Each chapter of the manuscript focuses on a specific objective, and includes a state-of-the art, current projects and prospects in this area.

Introduction – Tumor heterogeneity: definition, interest and challenges

Part 1 – Estimation of inter-tumor heterogeneity

Part 2 – Estimation of intra-tumor heterogeneity

Part 3 – A functional interpretation of intra-tumor heterogeneity

Part 4 – Algorithms evaluation and collaborative science

General conclusion – Scientific perspectives and personal considerations

In the manuscript, I will also discuss my working conditions of a young researcher, with more personal considerations on the expectations of our professional environment, the freedoms of the start of a career, and the difficulties and frustrations encountered. One of the main pitfalls when one becomes an independent researcher is the lack of time to properly carry out the many and varied missions that fall to us. Writing the HDR thesis is a perfect example of a time-consuming task, which can be interesting and useful if done well, with enough time to think and gain perspective on the work presented. Unfortunately, I did not have the time necessary to make a comprehensive dissertation on all the scientific issues related to cancer heterogeneity. I thus made the choice to focus on the contextualization of my research and the development of the links between my different projects. This is why I decided not to rewrite and paraphrase work already published in scientific articles through a peer review process. In some sections, I have copied and pasted parts

of my published articles, which will be identified like this:

> This piece is a quote from a work already published in a research article.

This work would not have been possible without a key partnership with cancer biologists and pathologists, that provide biological materials and clinical input (J. Cros, Beaujon Hospital, E. Brambilla – CHU Grenoble, S. Rousseaux and S. Khochbin – IAB). The strategic points of our projects have also been strengthened by on-going collaborations with external partners: data modelisation (D. Jost – CNRS), multi-omics data integration (Y. Blum – CNRS), multimodal network analysis (N. Thierry-Mieg – CNRS), evolution theory (A. Frenoy – U. Grenoble), mediation analysis (O. François – U. Grenoble), data challenges and machine learning (M. Blum – CNRS and I. Guyon – INRIA).

Computational estimation of inter-tumor heterogeneity

In this chapter, I will present a novel computational approach to define and quantify inter-tumor heterogeneity, at the patient level.

Recent advances in high-throughput sequencing technologies allow access to a tremendous source of precise molecular information at the single-individual, single-tissue or even at the single-cell levels: from cohorts of patients with a particular disease, to population of animals evolving in different controlled environments or single-cell cultures in various experimental conditions. However, analyses of these datasets are still mainly performed at the population-level by inferring average differential regulation between two conditions (healthy vs disease, wild-type vs mutants, etc.) and single-sample information is usually discarded. This is in part imputable to the lack of generic, robust computational tools that can give a reliable decision on differential regulation in individual sample. For example, such tools would be particularly useful in precision medicine to integrate patient-specific genomic information for personalized diagnosis and treatment.

In 2020, I published a novel method named PenDA[1], designed to perform differential analysis of gene expression at the individual level. PenDA detects gene deregulation as a perturbation of the local ordering of gene expression. Using a realistic benchmark of simulated lung tumors and a detailed parameter analysis, we showed that PenDA achieves very high specificity and sensitivity and is robust to batch effect. In particular, we demonstrated that PenDA outcompetes existing individual- or population-based approaches in terms of sensitivity at fixed low false discovery rate. This new validated method is directly relevant to translational medicine.

Based on the individual information of deregulation given by PenDA, we characterized two new molecular histologies for lung adenocarcinoma cancers, that are strongly correlated to prognosis. In particular, we identified 37 new biomarkers associated to bad prognosis and validated them on two independent cohorts. In this section, I will present the main results of the PenDA paper.

[1] Article : PenDA, a rank-based method for personalized differential analysis: Application to lung cancer
DOI: 10.1371/journal.pcbi.1007869

2.1 From high-throughput molecular information to personalized analysis

General medicine still largely relies on detecting diseases after the apparition of symptoms and on curing them with generic treatments. However, many studies have highlighted how the natural genetic or genomic diversities observed in a population, as well as patient history, or environment exposure, may strongly affect diseases risks, prognoses and responses to treatments [Lu et al., 2014, Battle et al., 2014]. This is particularly critical for cancer, where each individual tumor may be viewed as an independent disease, with specific and variable responses to generic therapeutic treatments [**?**]. Recently, thanks to the development of cheap and robust next-generation sequencing techniques, getting better insights into inter-individual heterogeneities was made possible by the analyses of large cohorts of patients. This led to the identification of individual molecular signatures or biomarkers associated with better prognosis, or better response to targeted treatment [Rousseaux et al., 2013, Lawrence et al., 2014, Hoadley et al., 2014]. This new knowledge paves the way to precision and personalized medicine where the genetic, genomic, and molecular information of each patient will be integrated to develop personalized diagnosis and treatment [Lu et al., 2014, Evans and Relling, 1999]. However, such challenging perspectives will be only possible with the concomitant development of efficient and robust methodological tools that allow the identifications of molecular defects or deregulation patterns at the individual level. Many statistical or bioinformatic methods do already exist to identify deregulated genes at the population level. For example, in the context of gene expression, fold-change methods like DESeq2 [Love et al., 2014], edgeR [Robinson et al., 2010] or limma [Ritchie et al., 2015] are designed and standardly used to identify genes that are differentially expressed in average between two groups of patients [Mutch et al., 2002].

While valuable to detect consistent typical deregulation patterns, such analyses do not provide precise information at the individual level. In addition, these global methods are usually very sensitive to batch effects that, without corrections, may lead to false discoveries or to confound important subpopulation effects [Goh et al., 2017]. Prior application of normalization routines to the investigated samples are used to mitigate such technical biases, but improper normalization may still perturb the biological signal [Evans et al., 2018, Li et al., 2015].

2.2 The PenDA method

2.2.1 Background

> Novel methods, robust to technical interference, are therefore needed to capture specific, individual data. Few promising techniques already allow to extract interpretable information from personalized omics data (see [Vitali et al., 2019] for a review). Rankcomp [Wang et al., 2015, Li et al., 2019] uses pairs of genes with a stable, relative order in a reference dataset to infer deregulated genes in individual samples [Guan et al., 2016, Qi et al., 2016]. This method, based on ranking, avoids the problem of normalization between samples, but results in very high false discovery rates (above 20%, see Materials and Methods). Alternative methods, like DEGseq [Wang et al., 2010], NOISeq [Tarazona et al., 2015] or Gfold [Feng et al., 2012], exploit paired samples from the same patient (one control versus one malignant) to perform differential analysis. However, such matched samples are usually rare (for example, in the case of cancer, a single sample from the tumorous biopsy is usually available for one patient).

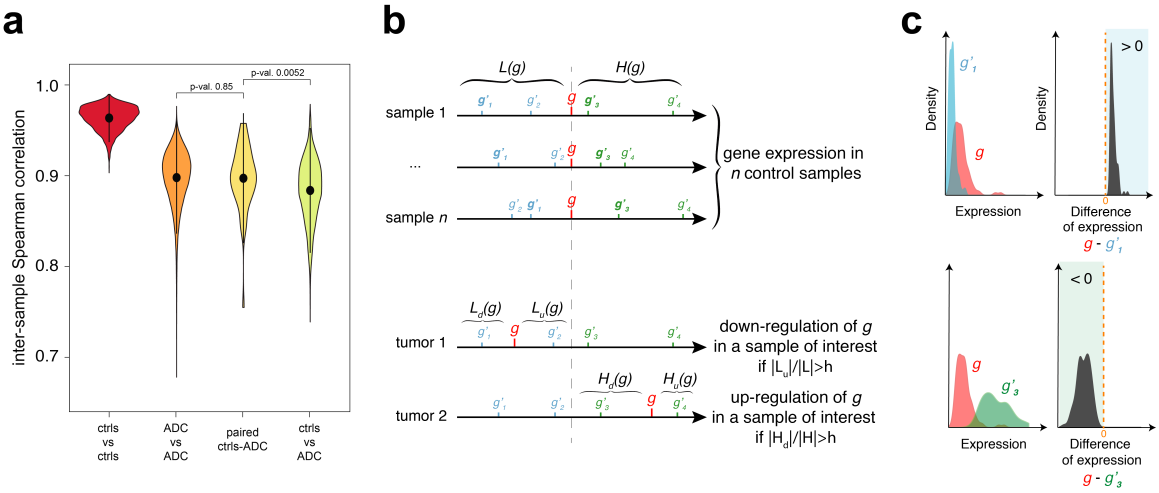

Figure 2.1: **The PenDA method**
(a) Violin-plots for the distributions of Spearman correlation between two samples taken from the TCGA database on lung adenocarcinoma: between two non-tumorous samples (ctrls vs ctrls, n=4,656 pairs), between two tumorous samples(ADC vs ADC, n=103,285), between paired normal and tumorous samples (paired ctrls-ADC, n=48), and between unpaired controls and tumors (ctrls vs ADC, n=44,135). Shown p-values correspond to Wilcoxon tests. (b) Basic scheme depicting the PenDA method. (Top) For each gene g,the algorithm infers sets of genes whose expressions are always lower (L(g)) or higher (H(g)) than that of g in a pool of control,reference samples. (Bottom) In a given individual (tumor) sample, g is viewed as deregulated i fits relative ordering with genes in the L(g) and H(g) lists is modified. (c) Examples of genes in the L (g'1,top) or H (g'3,bottom) lists of a gene g.While the individual distributions of gene expression in the control samples may overlap (left), the distribution of the difference in gene expression in controls (right) is always positive or negative for genes in L and H lists respectively. Figure from [Richard et al., 2020].

> Above all, it is not clear if the variabilities observed between paired samples are due to actual deregulation, to intrinsic inter-sample heterogeneities, or to technical biases. For example, in lung cancer, correlations between paired tumorous and normal samples are similar than between tumors of two different patients, and are only slightly higher than between a tumorous sample and an unmatched normal tissue (Figure 2.1.a).

2.2.2 Principles of the PenDA method

> PenDA is a rank-based method that allows to infer if the expression of any gene in a given sample of interest is deregulated compared to a set of reference samples (see Materials and Methods for details). The fundamental assumption behind the algorithm is that a gene is seen as deregulated in an individual sample if its local ordering compared to other genes with similar expressions is perturbed, as similarly stated by the RankComp method [Wang et al., 2015]. Briefly, PenDA starts by inferring a reference of relative ordering in control samples: for every gene g, it constructs two lists L(g) and H(g) of genes whose expression is lower and higher respectively than that of g in almost all the samples of a given reference dataset (Figure 2.1b top and c). To avoid comparison with genes having very different expression levels and to increase sensitivity of the method, lists L(g) and H(g) are then limited to the subset of l genes whose expression in control samples are closest to g. Finally, for a given sample of interest, PenDA scans every gene g to determine if it might be up- or down-regulated in that sample. This step is performed by considering the number of genes Lu(g) (respectively Hd(g)) in L(g) (resp. H(g)) in the studied case whose relative ordering to g has changed compared to controls (Figure 2.1b bottom).
>
> If the proportion of such genes with a modified order ($|Lu(g)|/|L(g)|$ or $|Hd(g)|/|H(g)|$) exceeds a given threshold h, the gene g is detected as deregulated. It has to be noted that a change of ordering between g and a gene g' of L(g) and H(g) might be caused by the deregulation of g' and not necessary by that of g. To limit the consequences of this effect on the detection of deregulation, PenDA iteratively applies the previous scheme until convergence by excluding at each iteration the current set of deregulated genes from every L and H lists. In the cases where the L(g) or H(g) lists are empty, we used the percentile method to evaluate the deregulation of g.

> The PenDA method is available as an R package at https://github.com/bcm-uga/penda. The penda vignette runs the PenDA pipeline on the samples of interest.It takes as an input two dataframes corresponding to the reference dataset of control samples and the dataset to investigate. It first filters for genes whose expressions are very low in every samples. Then, it computes the L and H lists from control samples for a given list size l. Finally, in every sample, it runs the iterative process to infer gene deregulation based on a user-defined threshold h. Optionally, the PenDA package offers the possibility to find the optimal set of parameters (in particular h) best adapted to: (i) the input data and (ii) a user-defined specific maximal false-discovery rate . Typically, on a standard personal computer (1 core of 3.6 GHz CPU), construction of L and H lists takes 10 sec CPU time for 18,000 genes and 98 controls. Downstream analysis of gene deregulation is slower and requires 2 min CPU time per analyzed sample.

2.2.3 Comparison of PenDA with other individual-based methods

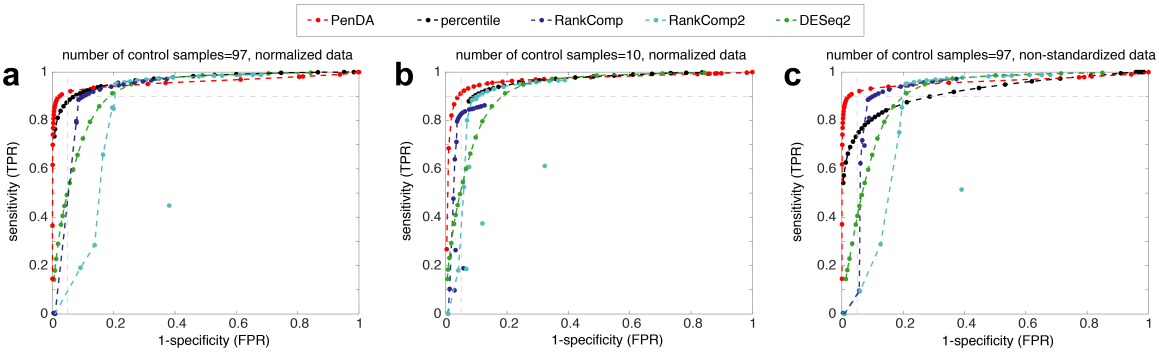

Figure 2.2: **Comparison with other methods**
(a) ROC curves on the same simulated dataset (normalized data, 97 control samples) as used in Fig 2 for PenDA, a simple percentile-based method, 2 versions of RankComp and DESeq2. (b) As in (a) but reference pool was composed by only 10 control samples. (c) As in (a) but data were not normalized. Figure from [Richard et al., 2020].

> We next sought to compare PenDA with other existing methods that also allow personalized diagnosis of gene deregulation. Using the same set of 10 simulations introduced before, we generated ROC curves for alternative methods (Figure 2.2): 2 versions of the rank-based method RankComp [Wang et al., 2015, Li et al., 2019], a simple percentile method based on outlier detection and DESeq2 [Love et al., 2014], the popular algorithm for detecting differential expression at the population level but used here on an individual basis.

> We observed that PenDA outperforms these methods, in particular in the limit of high specificity (FPR $\leq$ 5 %) where PenDA could reach very high sensitivity (TPR $\geq$ 90%) even for a limited number of control samples (Figure 2.2b). Surprisingly, outcomes of the RankComp methods were very dependent on the number of control samples and even led to better results for smaller control datasets. Note that basing our definition of deregulation on relative rankings limits the sensitivity of PenDA (and RankComp) to batch or normalization effects compared to the percentile method (Figure 2.2c), DESeq2, thanks to its internal normalization routine, being also robust.

2.3 Application of PenDA to non-small cells lung cancers

> We then applied the method to two large cohorts of patients from The Cancer Genome Atlas (TCGA) associated with lung cancer, one of the most common form of cancer in the world today. We evaluated the performances of PenDA on two large cohorts of patients from The Cancer Genome Atlas (TCGA) project representing two of the most common types of non-small-cell lung cancers: lung adenocarcinoma (ADC, 50%) and lung squamous cell carcinoma (SQCC, 40%) [Chen et al., 2014]. Personalized differential analysis was performed on the normalized gene expression data (RNA-seq) of 455 ADC cases and 473 SQCC cases. In addition to general statistics, like the number of deregulated genes per tumor, we showed that deregulated genes exhibit a cancer-type-specific commitment towards up- or down-regulation. In particular, we isolated genes with specific deregulation patterns, like genes that are up-regulated in all tumors or genes that are expressed but never deregulated in any tumors. Given their specificities, these genes are likely to be of interest in therapeutic research. We applied hierarchical clustering to classify the 455 ADC and 473 SQCC samples together, using a subset of 875 genes defined in a previous independent study (based on RNA-seq counts) as lung cancer subtypes classifiers (Classification to Nearest Centroid, [George et al., 2018]). We clustered samples with a distance based on inter-sample Pearson correlations computed from the PenDA differential expression matrix (Figure 2.3a). We observed a clear separation between ADCs and SQCCs groups, thereby validating our methodological approach.

> We could identify one main SQCC class and three ADC subclasses. The majority of ADC patients clustered into 2 subclasses (class II and III), that were not distinguishable in the clustering analysis performed by George et al on different lung cancers, using the same classifier genes [George et al., 2018]. We compared the three ADC subclasses obtained with our approach with the six ADC genomic subtypes previously identified by Chen et al, using a multiplatform-based approach on the TCGA-LUAD dataset [Chen et al., 2017]. Class II ADC patients are mainly associated with AD1, AD2 and AD3 subtypes, whereas the majority of class III ADC patients is distributed among AD4 and AD5 subtypes (Figure 2.3b). Similarly, class II and class III ADC patients did not directly relate to the integrated ADC molecular subtypes defined by the pioneer work of The Cancer Genome Atlas Research Network [Collisson et al., 2011]. Thus, clustering ADC according to their individual deregulation profiles identified new ADC subclasses. This demonstrates that personalized analysis using PenDA method brings new insights into histology classification.

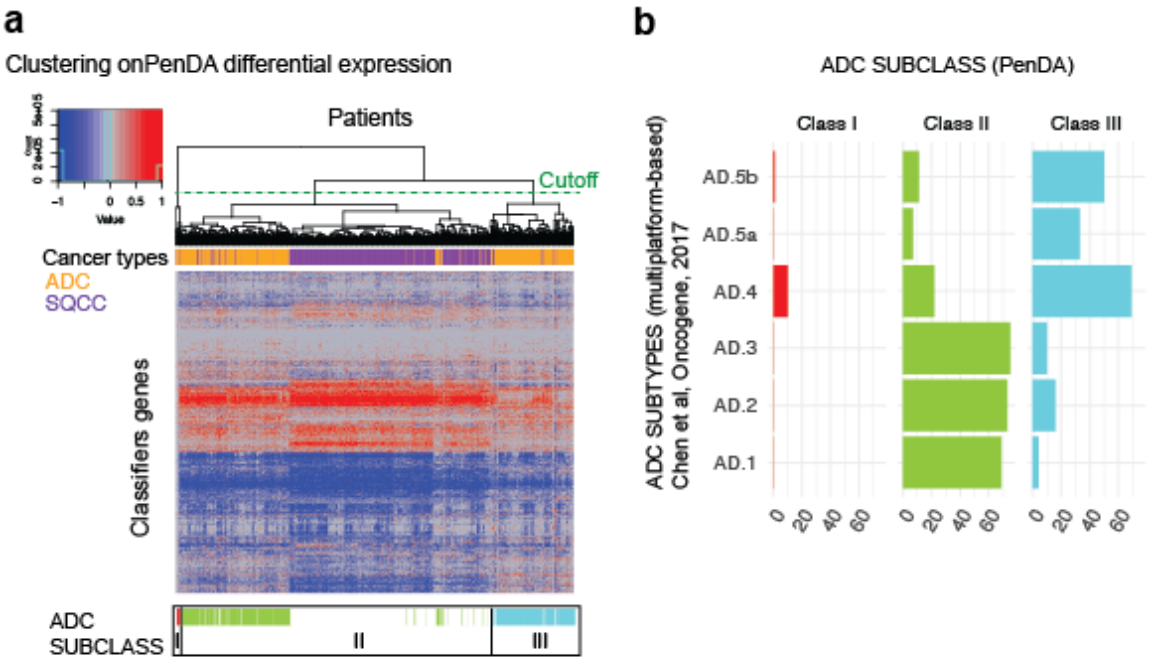

Figure 2.3: **Genetic deregulations efficiently classify cancer histologies**
(a) Heatmap of PenDA differential expression matrix applied to aspecific set of classifier genes (n =875) in TCGA non-small-cell lung cancers: ADC (orange) and SQCC (purple). Two hierarchical clustering analyses were performed: using Euclidean distance to sort genes and using Pearson correlation-based distance to classify patients, with a complete linkage function in both cases. ADC subclasses (color-coded, class Ito III) are defined according to the dendrogram cutoff n=3groups (cutting section =green dashed line). (b) Graphicalrepresentati on of the contingencytable between ADC subtypes (Chen et al,) and ADC subclasses (PenDA analysis). Each bar plot represents the total number of patients in each cell of the table. Adapted from [Richard et al., 2020].

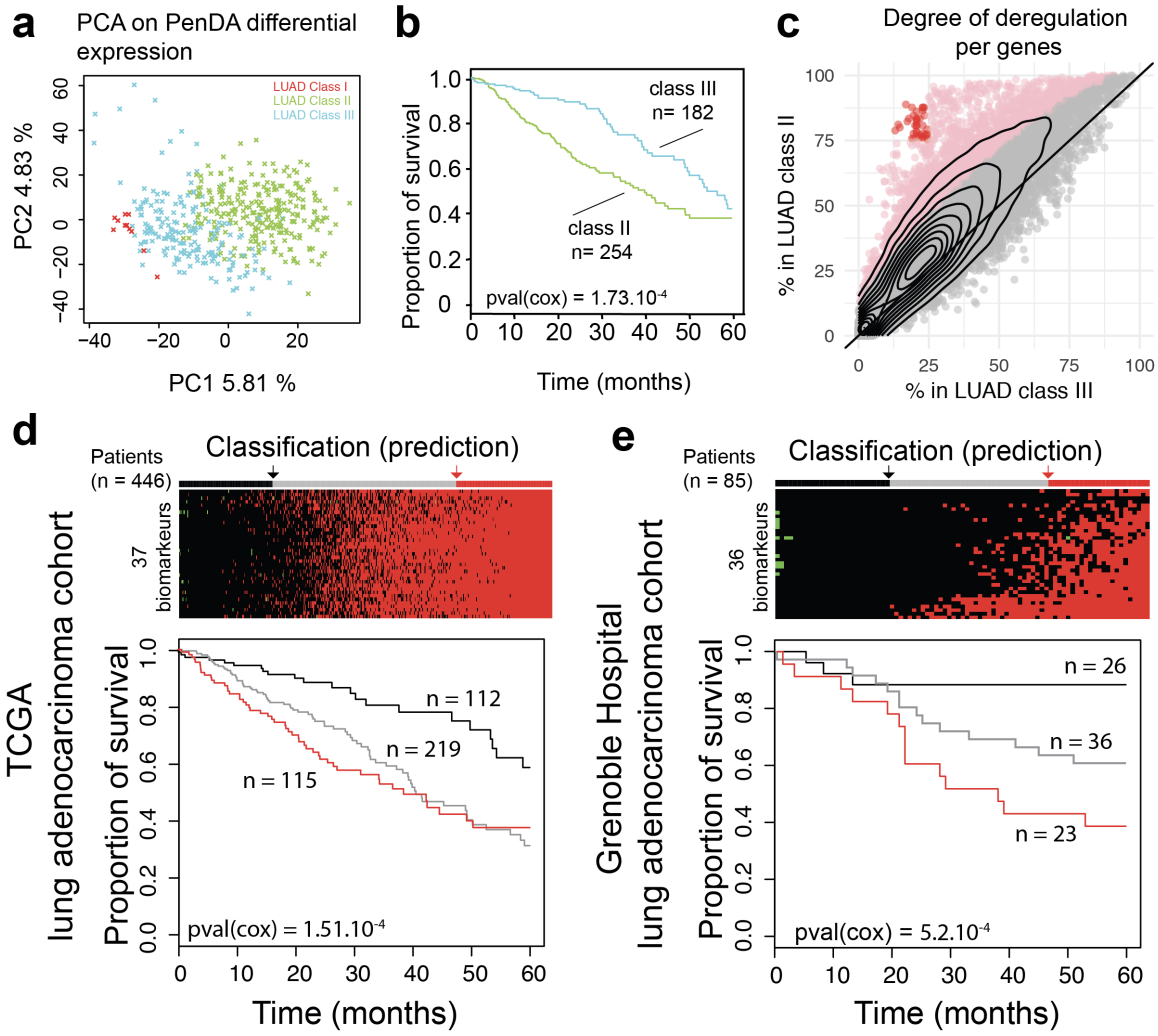

Figure 2.4: **Upregulation of 37 genes in adenocarcinoma is as trong predictor of poor prognosis** (a) Principal Component Analysis on ADC cohort. Each cross represents an individual sample. The color of the dots represents the three subclasses defined in Fig 6. (b) Survival of ADC patients classified according to the 2main subtypes (classes II and III). (c) The percentage of deregulated patients within the ADC class II(y-axis) or the ADC class III (x-axis). Each dot corresponds to one gene. The contour lines correspond to the density of genes. Pink dots indicate genes with a significant higher proportion of deregulation in the class II(proportion test, pvalue < 0.05 after Bonferroni correction for multiple testing). Red dots define 37 genes highly deregulated (> 75%) in the class II group and lowly deregulated(< 25%) in the class III group. (d) (Top) Classification of ADC TCGA-LUAD built on the total number of up-regulated genes among the subset of 37 classifiers defined in (c). Patients are separated into 3discrete groups: a group with alow upregulation (black, score < 4), a group with intermediate deregulation (gray, 4 ≤ score < 34) and a group with most genes upregulated (red, 34 ≤ score). (Bottom) Survival of patients according to these 3groups. (e) As in (d) but for ADC Grenoble Hospital patients. Patients are separated into 3discrete groups: a group with alow upregulation (black, score ≤ 0), a group with intermediate deregulation (gray, 0 ¡ score ¡ 15) and a group with most genes upregulated (red, 15 ≤ score). Figure from [Richard et al., 2020].

"We then wondered what defined these novel ADC subclasses. First, we asked whether this segmentation into three classes was specific to the classifier genes chosen to perform the hierarchical clustering. We performed a principal component analysis on ADC cohort only using the corresponding PenDA differential expression matrix for all genes (Figure 2.4a). The first two principal components of the analysis nicely discriminated classes I,II and III. We then focused on the two major groups: class II and class III. We performed a Cox survival analysis on these two groups (Figure 2.4b) and observed that the class III patients have a better 5-year survival prognosis than class II patients (cox p-value =0.00104). In order to better understand the molecular differences between class II and class III patients, we analyzed the pattern of deregulation of all genes in each class (Figure 2.4c). In class II, we observed a significant augmentation in the proportion of tumors where a given gene was detected as deregulated. In total, 13% of the genes (n =2432) were significantly more often deregulated in class II compared to class III patients (one-sided proportion test).

We verified that the cancer stages, gender, and age were evenly distributed in class II and class III patients (chi square test pvalue =0.2133, p-value =1, and p-value =0.2133, respectively) and that the shift in genetic deregulation was detectable independently of stages, gender and age. This indicated that this adenocarcinoma classification was not correlated with any of these putative confounding factors.

We decided to specifically study the 37 genes displaying the most extreme differences between the two classes, i.e. the genes deregulated in more than 75% of class II patients and in less than 25% of class III patients (red dots on Figure 2.4c). Since all these genes are committed toward up-regulation in class II patients, we tested if the up-regulation of these genes would be a good predictor of cancer survival. We added up the level of individual deregulation of the 37 genes (values equal to -1, 0 or 1, for each gene) to quantify the total deregulation score associated with those genes.

Then we defined three groups using the 1st and the 3rd quantile of the score distribution. Analysis of the 5-years survival curve in the ADC LUAD-TCGA dataset showed a significant difference between groups, with a worst prognosis for patients that display up-regulation of most of the genes (score 34, Fig. 7d). To validate our selected set of 37 genes as robust biomarkers, we applied the PenDA method on expression data (Affymetrix Human Genome U133 Plus 2.0 Array) of an independent adenocarcinoma cohort from the Grenoble Hospital (85 patients, GSE30219 [Rousseaux et al., 2013])."

> We then investigated the 5-year survival curve of the three groups predicted using 36 genes (all genes were analyzed in the Grenoble Hospital cohort, except FAM72D not measured by the array). Coherently with the results observed in TCGA-LUAD ADC cohort, patients up-regulated for many genes (score *geq* 15) have a worst prognosis (cox p-value = 5.2.10-4, Figure 2.4e). Thus, using the PenDA method, we identified 37 biomarkers predicting a bad outcome when all up-regulated. Altogether, these results suggest that PenDA method is a powerful approach to discover new biomarkers in cancer.

These applications to lung cancer demonstrate how a deep analysis of PenDA results can be complementary to standard population-based approaches by giving a precise individual knowledge on gene regulation for a specific biological or medical question. Moreover, while initially developed for studying differential gene expression, our method is general and can be applied to study other types of deregulation based on different omics data. To promote open-access, reproducible research and to ease knowledge transfer between research and clinics, we developed a user-friendly R package and corresponding tutorials for an easy utilization even by non-experts.

2.4 A generalization of the approach – *on going*

Since the publication of the PenDA article, we explored several use cases, from which I will present here two examples. In parallel, we are now working on a PenDA implementation suited to other types of omic data, such as DNA methylation data (not presented in this thesis).

2.4.1 Use case 1: a pan-cancer analysis

We systematically run the PenDA methods on all TCGA cancers displaying more than 10 control samples to perform a pan-cancer analysis. We compared PenDA individual total number of gene deregulation versus the percentage of genes deferentially expressed within the cancer population, computed by DEseq2 [Love et al., 2014] (figure 2.5).

First, we observed highly variable deregulation profiles depending on the cancer cohort studied. Some cancers showed a very low proportion of deregulated genes, such as esophageal cancer (ESCA) with a median of 2% deregulated genes or rectal cancer (READ) with a median 1% deregulation. For those two cancers, the PenDA results are in agreement with population-level number of differentially expressed genes (DESeq2 approach). Interestingly urothelial cancer (BLCA) displays a 2.5% median dysregulation with PenDA approach compared to a 14% populational differential expression computed with DESeq2. Overall, the consistency between PenDA and DESeq2 results varies. The cancer with the highest proportion of deregulation is the same for PenDA and DESeq2: colorectal carcinoma (COAD) with 37% median for Penda, 29% at the p-value threshold of 0.05 for DEseq2. These results deserve more in-depth investigations in order to determine the link between the deregulations detected by PenDA (individual approach) and by DEseq2 (population approach), as well as the influence of the selected parameters on the simulations. Indeed, the number of deregulated genes detected by PenDA is directly dependent on the threshold chosen during

the simulations, and it is therefore difficult to draw conclusions.

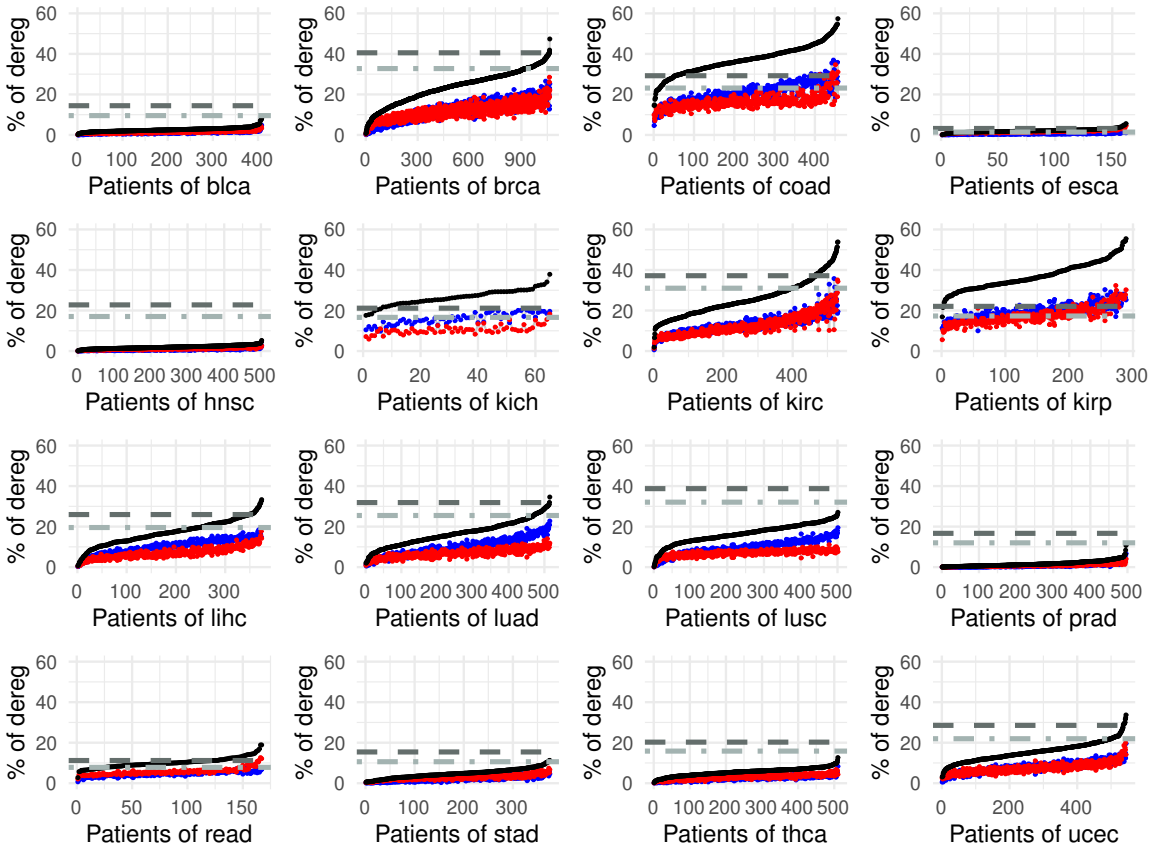

Figure 2.5: **Proportion of deregulated genes in each TCGA cancer type**
For each cancer, patients are ordered according to their total number of deregulated genes. The black dots indicate the total proportion of deregulated gene, the blue one the proportion of down-regulated genes and the red one the proportion of upregulated genes. The dark gray line corresponds to the proportion of genes detected as deregulated in a DESeq2 population approach (cases VS controls, with a p-value threshold of 0.05). The light gray line correponds to the proportion of genes detected as deregulated in a DESeq2 population approach (cases VS controls, with a p-value threshold of 0.01). Figure from C. Decamps PhD thesis.

2.4.2 Use case 2: individual metabolism regulation of glioblastoma

As part of a collaboration with Annabelle Ballesta, a researcher at Institut Curie, we look for gene pathway specifically deregulated in single patients affected by glioblastoma. The objective of Ballesta's team is to develop a mathematical model of the genetic deregulation of several gene networks, at the individual level. Their data is composed of the RNAseq sequencing of 20 samples, corresponding to cell cultures of different glioblastoma lines. Glioblastoma is an extremely heterogeneous brain cancer, with huge variability between patients [Vollmann-Zwerenz et al., 2020]. Moreover, the provenance of the original cells is still debated within the scientific community [Ceccarelli et al., 2016]. The first difficulty for the application of PenDA was therefore the absence of suitable control samples; indeed, Ballesta's study involves no control and healthy brain samples

are difficult to obtain. After consultation with expert biologists, we finally chose a cohort of cultured astrocytes as a reference [Lundin et al., 2018]. The second difficulty resides in the fact that the studied control cohort consists of only 12 samples, in reality three replicates of four different lines. In addition, astrocytes do not constitute a perfect reference because we cannot speak of healthy cells from which glioblastoma cells would come.

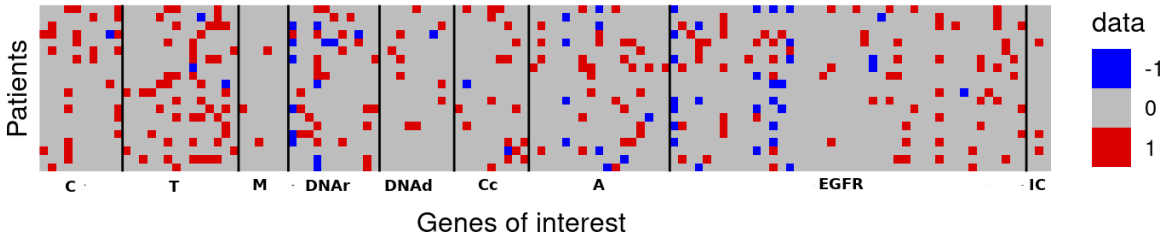

Figure 2.6: **PenDA gene deregulation of each glioblastoma derived cell line**
The 20 patient-derived cell lines are represented on lines. Each column corresponds to a gene of interest provided by our collaborators, the categories are separated by vertical lines: C, the genes associated with the CLOCK transcription factor and the circadian clock, T the genes associated with transport, M the genes associated with metabolism, DNAr the genes associated with DNA repair, DNAd the genes associated with the DNA damage response, Cc the genes associated with the cell cycle, A the genes associated with Apoptosis, EGFR for the EGFR gene network, an important growth receptor in cancer and IC for genes involved in the immune system. A blue box implies under-expression of the gene for the given patient, red over-expression, gray no change. Figure from C. Decamps PhD thesis.

Given the low number of controls, we chose to implement a cross-validation strategy to ensure the robustness of the PenDA results. We performed ten independent analyses, each time using only ten of the twelve randomly selected controls so that they were different between each analysis. At each round, the two remaining controls were mixed with the analyzed tumors in order to estimate the number of false positives. At the end of this process, we obtained for each analysis about 35% deregulated genes, this rate of deregulation corresponding to the expectations of biologists. The results were then combined two by two between each analysis to compare the proportion of commonly deregulated genes.

We therefore retained the deregulated genes common to the ten PenDA analyses, which gives us an average of 30% deregulation per sample. The aim of the analysis was to examine the genes belonging to pathways of interest for Annabelle Ballesta, the deregulation of which varies between cell lines (Figure 2.6). These results are now used in patient-specific mechanistic models.

Computational estimation of intra-tumor heterogeneity

In this chapter, I will review existing computational approaches used to quantify intra-tumor heterogeneity. I will also present novel methods that our group developed to solve these scientific challenges, in the context of pancreatic adenocarcinoma.

A tissue is a mixture of cells, which are all defined by a specific molecular signature (transcriptome, methylome, proteome, metabolome...). Bulk measurements on a biological sample are the sum of the molecular signatures of each cell present, to which is added variability due to technical and experimental noise. Since the development of high-throughput sequencing technologies, cancer research has focused on characterizing global genetic and epigenetic changes that contribute to the disease. However, these studies often neglect the fact that tumours are constituted of cells with different identities and origins (cell heterogeneity). Quantification of tumor heterogeneity is of utmost interest as the multiple components of a tumor are key factors in explaining tumor progression and response to therapy. Inference of tumor composition from bulk measurements has so far been approached using cell-type deconvolution methods (see an illustration figure 3.1). This approach relies on two critical parameters : (i) the cell proportion matrix W (i.e. the true proportion of each cell-type in the biological samples) and (ii) the molecular specific cell-type profiles matrix H (the theoretical pure molecular profile of each cell-type).

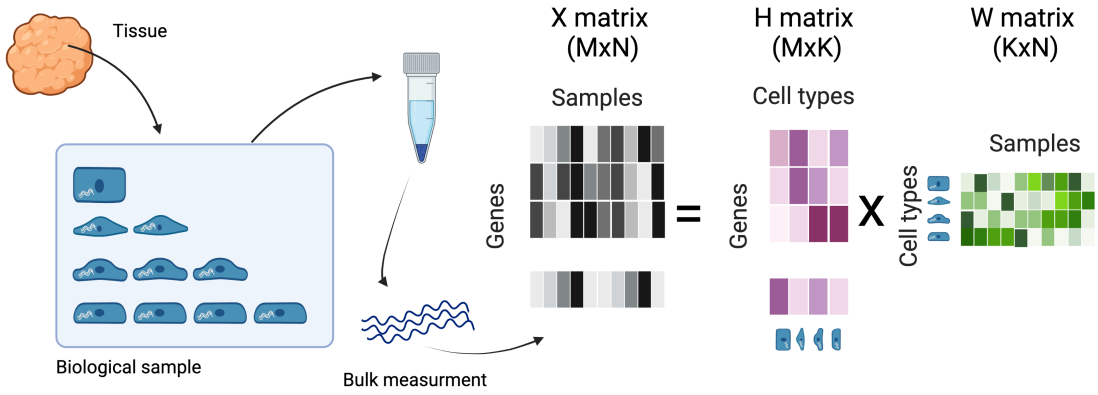

Figure 3.1: **Cell-type deconvolution problem, the bioinformatician point-of-view.** An example of deconvolution problem applied to transcriptomic bulk data. We assume that the observed matrix X is composed of a mixture of K cell-types, present in different proportion in each sample (matrix W). Therefore, X can be described as a linear combination of cell-type specific molecular profiles (matrix H).

Most of the deconvolution methods aiming to quantify the tumor cell-type composition intend to infer W from X. If H is know *a priori*, this corresponds to supervised (or reference-based) deconvolution. It can be solved using linear regression or least square approaches. If W and H are jointly inferred from X, it corresponds to unsupervised deconvolution (or reference-free), usually solved by matrix factorization approaches, such as Non-negative Matrix Factorization (NMF) or Independant Component Analysis (ICA), that enable the representation of a complex high-dimension dataset into a low-dimension subspace

The critical part of unsupervised deconvolution approaches lies in the biological interpretation of these estimated components, the detection of rare events and the reproducibility of these approaches between similar datasets [Cantini et al., 2015, Cantini et al., 2019]. On the other hand, supervised approaches strongly depend on the quality of the reference profiles used for deconvolution [Avila Cobos et al., 2020]. These references are prone to errors because they are based on the false assumption that all cell-types of interest in the sample are identified and characterized with precision, which in turn leads to failure in estimating the clonal distribution of cancer cells and the true proportion of tumor micro-environment cells. Alternative single-cell based deconvolution approaches are promising, but they are currently limited by the lack of consensus methods to infer single-cell identity and to construct biologically relevant reference profiles, in particular because of different noise structures. Despite current efforts to build a human single cell atlas encompassing tumor cellular heterogeneity (e.g. [Massalha et al., 2020]), single-cell sequencing is still in its infancy in PDAC [Han et al., 2021], with heterogeneous datasets that are difficult to integrate into a consensus study (personal observations).

In this chapter, I present :

- A benchmark of reference-free deconvolution algorithms.
- The development of a novel single-cell based deconvolution approach.
- Prospective developments of new methods to accurately quantify intra-tumor heterogeneity.

3.1 Benchmarking unsupervised deconvolution approaches

While reference-free algorithms have been developed to infer cell-type proportions, a comparative evaluation of the performance of these methods was still lacking. The use of unsupervised methods is not trivial and data scientists are left with little to no guidelines to deconvolve bulk samples. How to normalize or transform the data before analysis? Should one apply feature selection? What deconvolution algorithm should be used? How to interpret infered components?

In 2020, we published guidelines for cell-type heterogeneity quantification from DNA methylation data [HADACA consortium et al., 2020]. When geneticists historically performed association between epigenetic variation and phenotypic traits, cell-type proportions were only considered as confounding factors, their inference were not the main objective, but rather an intermediate step that can contribute to reducing false positive associations [McGregor et al., 2016]. In contrast, in our paper we compared reference-free deconvolution methods with the estimation of cell-type proportions as the main objective, as they are directly related to tumorigenesis.

Our manuscript compared three software packages that infer cell-type proportions based on

methylation data. We evaluated key factors affecting performance of deconvolution pipelines. We examined to what extent cell-type proportions can be accurately inferred when accounting for measured confounding factors. We determined how feature selection impacts algorithms' performance at inferring cell-type proportions. We also tested several methods for selecting appropriate number of constituent cell-types and ask how sensitive the results are to the variation in cell-type number. Based on these, we provided a framework to estimate intra-tumor composition, accounting for confounding factors. Our main results are presented below, in the form of quotations from our article.

3.1.1 Cell-types heterogeneity quantification from DNA methylation

> We compare three software packages that infer cell-type proportions based on methylation data: RefFreeEWAS, MeDeCom and EDec [Houseman et al., 2016, Lutsik et al., 2017, Onuchic et al., 2016]. For our comparisons, we rely on simulations where real methylation profiles of different cell-types are mixed in differing proportions. While some of the methods include series of steps that may be considered a pipeline, the simulations focus on comparing the core deconvolution step shared by all the three methods (e.g., Stage 1 of EDec) that solves a convolution equation that contains two key variables: (i) the cell-type proportions within the samples, and (ii) the average methylation profiles of constituent cell-types. The main outcome of this core deconvolution step are estimates of cell-type proportions and of the methylation profiles of constituent cell-types, which are needed to characterize the constituent cell-types and quantify tumor heterogeneity. Because accurate references for cell-type specific methylation profiles are sparse, especially for solid tissues and cancer cell-types, we further assume that reference data for constituent cell-types is not available, which excludes reference-based methods from our comparative analysis [Teschendorff et al., 2017, Zheng et al., 2018].
>
> We here evaluate key factors affecting performance of deconvolution pipelines. We examine to what extent cell-type proportions can be accurately inferred when accounting for measured confounding factors. We determine how feature selection impacts algorithms' performance at inferring cell-type proportions. We study performances variability according to the randomly selected initialization of local optimization involved in solving deconvolution equation. We also test several methods for selecting appropriate number of constituent cell-types and ask how sensitive the results are to the variation in cell-type number. We apply MeDeCom, EDec (Stage 1, the core deconvolution step), and RefFreeEWAS to estimate heterogeneity within simulated tumorous tissues. Simulations are encoded in a matrix X of size MxN, where M represents the number of CpG probes and N represents the number of samples.

> All these software packages perform various types of non-negative matrix factorization to infer cell-type proportions (matrix W of size KxN, with K as the putative number of cell-types) and cell-type-specific methylation profiles (matrix H of size MxK) by solving $X = HW$, or rather by minimizing, under various constraints (that vary between the three tested algorithms), the error term: $X - HW_2$.
> We simulate D with 5 cell-types (K = 5): 2 cancer-like cells (lung epithelial and mesenchymal), healthy epithelial cells (lung epithelial), immune cells (T lymphocytes), and stromal cells (fibroblasts). These simulations mainly depend on a parameter $\alpha 0$, which controls the diversity of the generated samples: when $\alpha 0$ is small (1), the simulated proportions of the K cell-types are diverse among samples and as $\alpha 0$ increases, the variability decreases to the point at which proportions are the same for all samples. Finally, we simulate the effect of confounding factors on these mixtures by using a regression model of methylation data computed from real lung cancer clinical datasets.

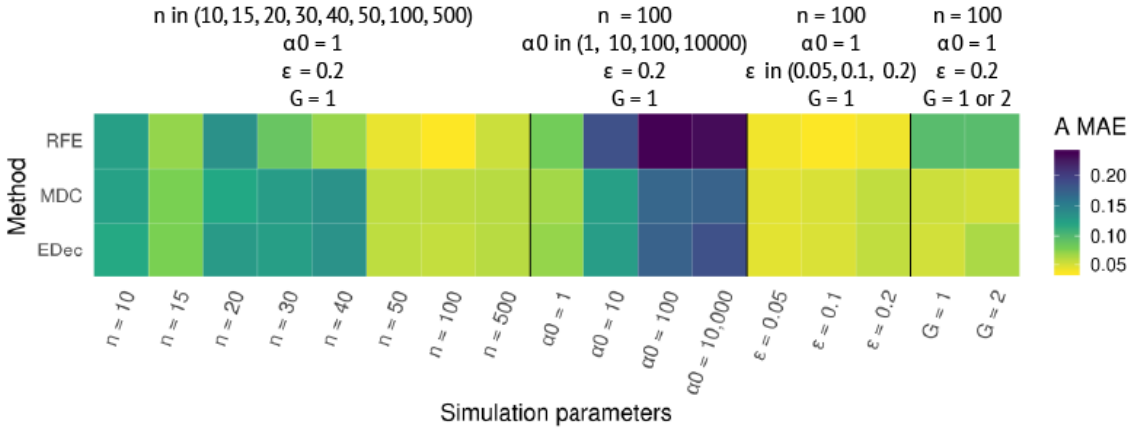

Figure 3.2: **Performance of the 3 deconvolution methods for different parameter settings.** Heatmap of method performance ('A MAE': Mean Absolute Error on estimated A, the matrix of cell proportions). RFE stands for RefFreeEWAS, MDC for MeDeCom and EDec for EDec stage 1. All algorithms were run on 10 D matrices corresponding to 10 different realizations of the random ϵ- controlled process on one D matrix computed from one simulated A matrix, each time, with the following parameters n (number of samples), $\alpha 0$ (inter-sample variation in mixture proportion), ϵ (magnitude of random noise applied on D) and G (the cell profiles used for simulations). Mean MAE corresponds to the average error of the three methods (computed for each parameter set). A random A matrix was used for testing the effect of G1 and G2, another random A matrix was used for testing the effect of magnitude. Testing the effect of n and 0 required independent simulation of A each time. As a consequence, the four simulations corresponding to the set of parameters n = 100, $\alpha 0$ =1, ϵ = 0.2, G = 1 have different results, because these simulations are based on different randomly simulated A matrices. Figure from [HADACA consortium et al., 2020].

To evaluate the methods performance, we use Mean Absolute Error (MAE) as a metric to compare inferred individual cell-type proportions to the ground truth. First, we tested the effect of altering four simulation parameters on the methods performance (Figure 3.2): (1) the number of simulated samples (N, ranging from 10 to 500), (2) the inter-sample variation in mixture proportions ($\alpha 0$, from 1 to 10,000), (3) the magnitude of random noise added to the mixture component (ϵ, from 0.05 to 0.2) and (4) the set of K cells profiles used to simulate complex tissues (termed as the cell background, G).

As expected, increasing the sample size improves the performance of all methods (Figure 3.2 columns A to H). Increasing inter-sample proportion variability also substantially improves performance of all methods (Figure 3.2 columns I to L). Average error (mean error across the three methods) is 0.074 ($\alpha 0 = 1$, column I) when inter-sample variation is large, increases to 0.147 ($\alpha 0 = 10$, column J) when variation is moderate, and reaches 0.194 ($\alpha 0 = 100$, column K) when variation is almost zero 3.2). By contrast, the performances of the three methods are neither sensitive to changes of the cell background (Figure 3.2 columns P to T) nor to variations in the magnitude of the random noise applied during simulations (Figure 3.2 columns M to O).

In this first direct comparison, the three deconvolution methods account for all 23,381 probes corresponding to a subset of the Illumina 27k and 450k DNA methylation probes, with no specific filtering. To run the algorithms, we used the following functions and parameters: RefFreeEWAS::RefFreeCellMix (5 cell-types, 9 iterations), EDec::run_edec_stage_1 (5 cell-types, all probes kept as informative loci, maximum iterations = 2000), and MeDeCom::runMeDeCom (5 cell-types, lambdas in 0, 0.00001, 0.0001, 0.001, 0.01, 0.1), maximum iterations = 300, 10 random initializations, number of cross-validation folds = 10). Under these not-optimized conditions (i.e. with no pre-processing steps), we observe that all methods provide comparable performance, each algorithm performing best under specific conditions and parameter settings. RefFreeEwas performs best for 9 out of 20 different parameter settings, MeDeCom for 8, and EDec for 3 conditions (lowest MAE on estimated A). Error obtained with EDec is on average 8% larger than the error obtained with RefFreeEwas and 2% larger than MeDeCom.

These results suggest that the differences between the tested algorithms are minor when default parameters are used and no filters are applied on the provided DNA methylation probes. The main variations in performance are related to simulation parameters, such as sample size (n) or inter-sample proportion variability ($\alpha 0$).

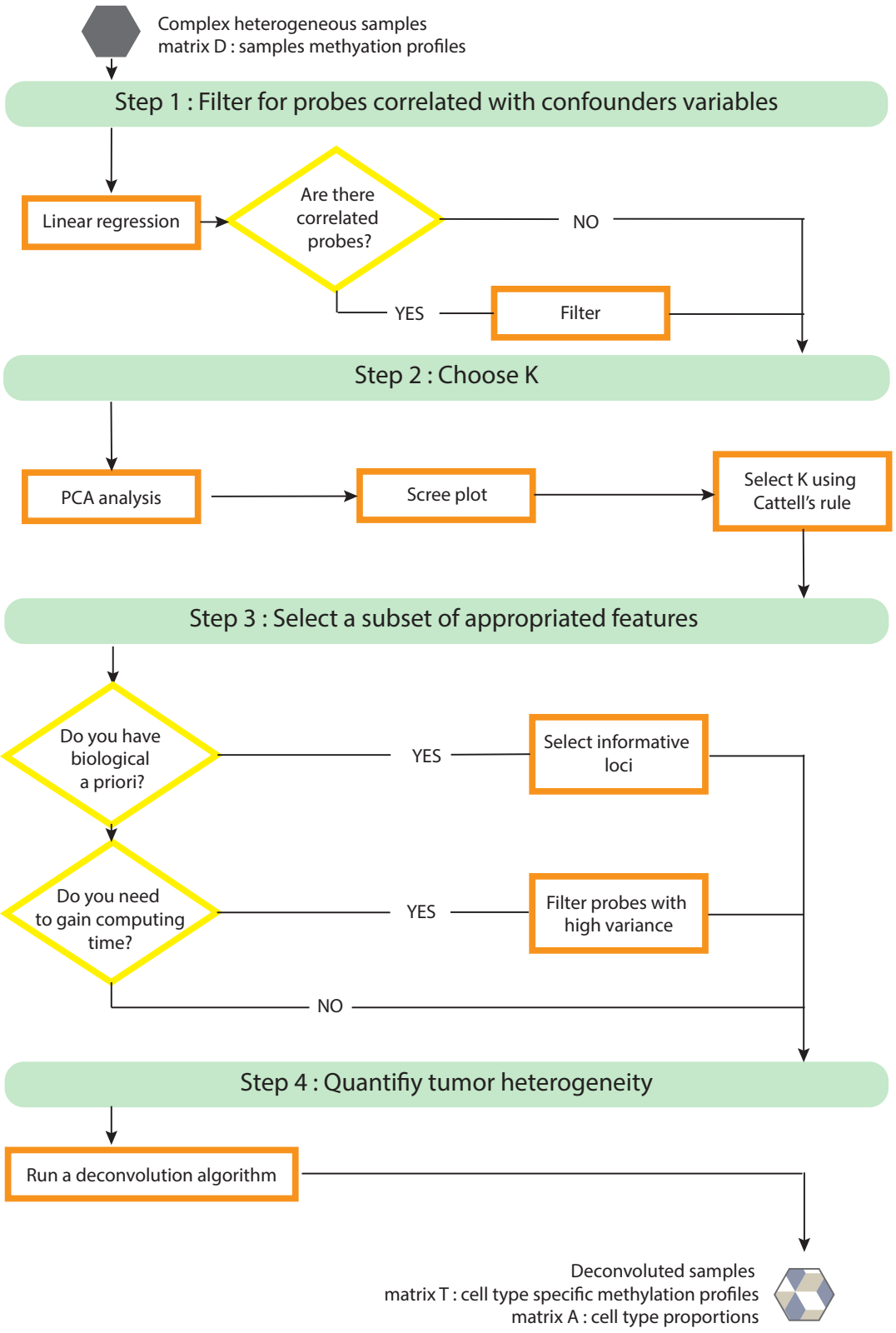

Figure 3.3: **Recommendations and benchmarking pipeline.**
Figure from [HADACA consortium et al., 2020].

3.1.2 Providing guidelines for cell-type heterogeneity deconvolution

> Based on lessons learned from the simulation experiments, we developed a benchmark pipeline to estimate cell-type proportions that addresses the presence of confounders and other key factors affecting performance of deconvolution algorithms (Figure 3.3). We anticipate that this benchmark pipeline will help catalyze wide adoption of deconvolution methods and accelerate improvement of deconvolution pipelines by (1) helping validate other deconvolution pipelines by demonstrating concordant results; (2) serving as a benchmark for demonstrating improved performance of other pipelines; (3) providing a starting point ("toolkit") for development of new pipelines.
> We note that the benchmark pipeline is not experimentally validated nor it is systematically compared as a whole against more complex pipelines that include expression data (e.g., all stages of EDec pipeline). In our experience, no deconvolution pipeline can be expected to provide accurate solutions when applied "out of the box" to a new tumor type. Tuning and validation are required in the context of each tumor type, using resources and information that may be tumor-type specific. In that sense, deconvolution may be thought of as a computational modeling approach that goes hand-in-hand with experimentation.

3.2 Development of a single-cell reference based PDAC deconvolution method – *on going*

Molecular analysis and classification based on gene expression landscapes of pancreatic adenocarcinoma (PDAC) is complexified by the intrinsic heterogeneity of this cancer. Indeed, as any solid cancer, PDAC are composed of the 'tumoral mass' (predominantly epithelial cells) which is surrounded by a microenvironment composed of 'stroma' cells (fibroblasts, pericytes, endothelial and immune cells), the stromal cells giving support, nutrients and sometimes resistance/metastatic potential to neoplastic cells. Tumor mass is also surrounded by normal epithelial cells. Two (transcriptomic) tumor cells subtypes have already been characterized by microdissection of bulk tumors: a consensus classical subtype, and a basal-like non-differentiated subtype, with worse outcome [Moffitt et al., 2015, Maurer et al., 2019, Chan-Seng-Yue et al., 2020]. In addition, the stromal component in PDAC is often very dense and fibrotic, displaying a high heterogeneity associated with prognostic relevance [Puleo et al., 2018]. PDAC intra-tumor heterogeneity is a major pathological feature that can confer aggressiveness and chemoresistance [Gutiérrez et al., 2021]. For instance, the desmoplasia reaction is orchestrated by proliferation of stromal cells and the accumulation of theirs products conferring heterogeneity and plasticity to the tumor [Whatcott et al., 2015]. On the other side, hypovascularity can block the access of drugs to the tumor center (often anti-angiogenic drugs) [Katsuta et al., 2019]. Therefore, the tumor heterogeneity is a critical issue that blocks advancement in the research area. Trying to understand and finely quantify tumor composition can pave the way to more efficient precision medicine.

Traditional experimental methods that quantify intra-tumor heterogeneity (immunocytochemistry, flow cytometry) in clinical practice rely on a small set of molecular markers and are therefore limited by the number of cell-types that can be simultaneously quantified. On the other hand, single-cell profiling is a promising technology that can detect and quantify an unlimited number of cell-types at high-resolution, provided that corresponding cell-type markers are correctly defined, or that we can identify new cell-types. However, it is costly, not applicable to Formalin-Fixed Paraffin-Embedded (FFPE) tissue, and not easily scalable in routine clinical practice, as compared to bulk transcriptomic data. Besides, cell identity assignment based on single-cell RNA-sequencing (scRNA-seq) is not straightforward, owing not only to technical and biological noise of biological data [Kim et al., 2015] but also to the weak reliability of pre-established literature markers [Zhang et al., 2019]. A promising approach to accurately quantify heterogeneity in PDAC relies on the recent emergence of bulk deconvolution algorithms based on single-cell reference profiles [Newman et al., 2019, Wang et al., 2019, Dong et al., 2020, Tsoucas et al., 2019, Du et al., 2019, Jew et al., 2020]. One of the main limitations of these approaches is the accuracy of the single-cell based profiles, which can strongly impair the quantification and the biological interpretation of the inferred tumor composition [Chen et al., 2019].

In the project I present here, we built an integrative set of PDAC cell-type specific gene markers, based on a dedicated pre-established gene markers curation and subsequent analysis of PDAC recent [Peng et al., 2019, Moncada et al., 2020, Chan-Seng-Yue et al., 2020, Lin et al., 2020] single-cell RNA-seq datasets. After several steps of quality control, filtration, annotation and data integration, we launched a systemic identification of integrative cell-types specific gene markers. We then intend to use these markers to revise our current understanding of cell-type heterogeneity in PDAC using single-cell based bulk deconvolution approaches.

3.2.1 Identification of 14 specific PDAC cell-types

We generated a curated database of gene-markers using a large number of publications describing different pancreatic cell-types in normal and/or tumor tissues. Our curated database follows a hierarchical organisation with three embedded level of granularity (nodes 1 to 3, see Figure 3.4A). For each PDAC single-cell dataset, we performed single-cell labelling using our curated database of PDAC gene-markers. At each Node, we performed a linear dimensional reduction (PCA) of the single-cell population, followed by a graph-based clustering (Seurat default clustering approach [Stuart et al., 2019]). We then applied gene set enrichment analysis (GSEA) on the average gene expression of each cluster using gene markers from our curated dataset. Cells forming clusters with a significant enrichment for one cell-type (p-value < 0.1) were assigned with the corresponding identify. Cells belonging to clusters with no significant p-values were re-clustered for a second round of GSEA analysis.

In the first Node, we decomposed the pool of cells into two main compartments (Stroma/Immune and Epithelium [Maurer et al., 2019]) using 214 Stroma/immune and 142 epithelial markers. Next, at Node 2, epithelial cells were divided into normal cells (Copy Number Variation < 0.02) and cancer cells (CNV ≥ 0.02). To label normal cells, we used gene markers of acinar, endocrine (α, β, γ and δ) and ductal cells from the study of Baron *et al* [Baron et al., 2016], merged with the ones from the Enrichr database [Chen et al., 2013], to which we also added the ductal marker *PROM1*

described in Enge et al [Enge et al., 2017]. Concerning the cancer cells classification markers, we considered classic and basal-like robust gene markers by filtering markers published in at least two different studies [Bailey et al., 2016, Nicolle et al., 2017, Collisson et al., 2011, Moffitt et al., 2015, Puleo et al., 2018]. Cancerous epithelial cells were assigned to either classic or basal subtypes, and non-assigned cells were labelled as uncharacterized (Unchar_cancer).

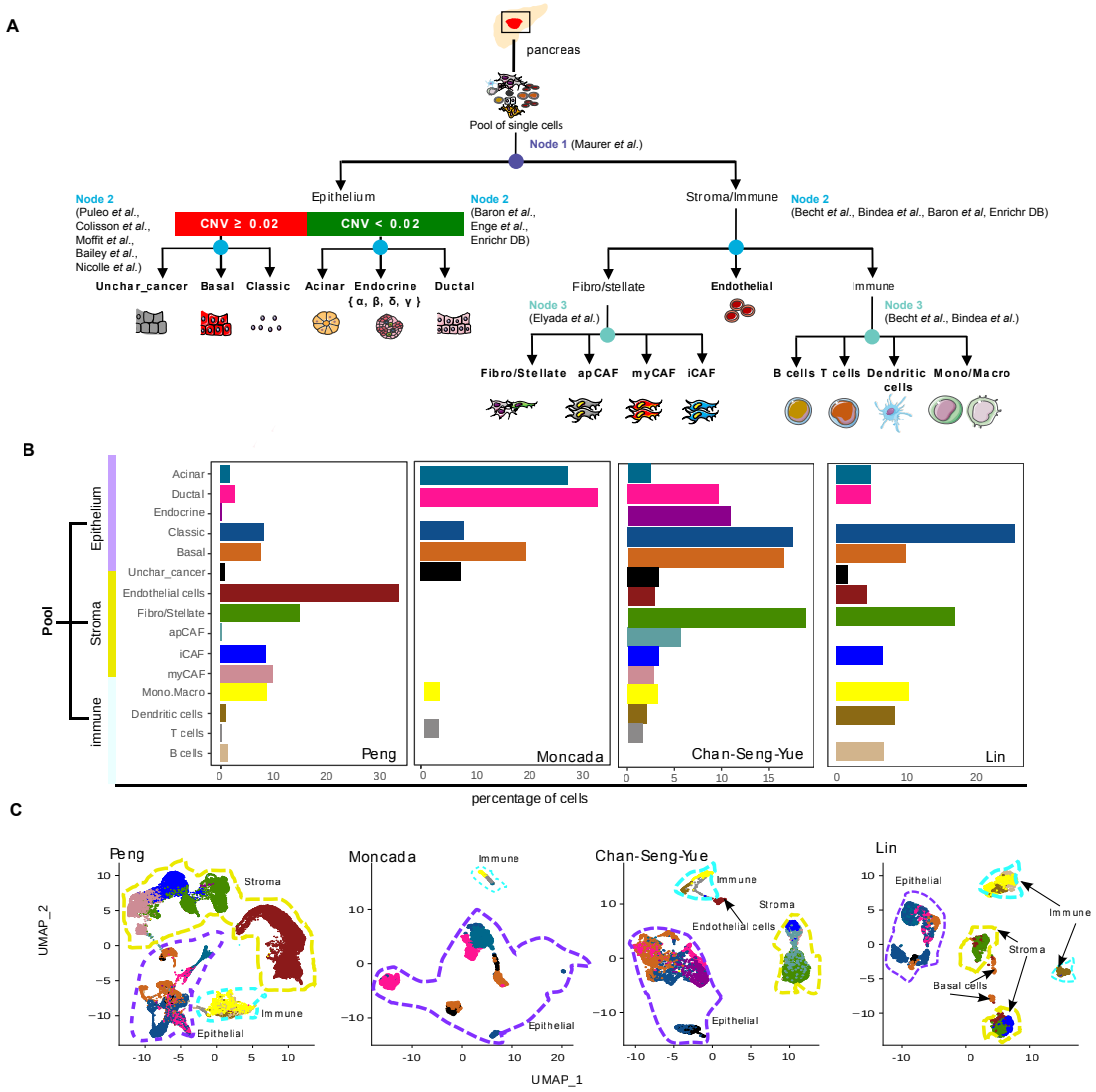

Figure 3.4: **Identification of high-resolution PDAC cell-types.**
A. Decision Tree summarizing the hierarchical nodes of cell-type identification, at different granularity levels. For each node, a cell clustering was performed followed by a GSEA test using our curated literature marker database. apCAF: antigen-presenting CAFs, iCAF: inflammatory CAF and myCAF : myofibroblastic CAFs. **B.** Barplots showing the distribution of cell-type percentage after cell-type assignation across the four different datasets. Cell-type are grouped into 3 high-level functional families of cells: epithelium (purple), stroma (yellow) and immune (light blue). **C.** UMAP projection where cell are colored according to the 14 characterized cell-types after datasets integration : Peng (26590 identified cells, i.e. 75% of the total identified cells among the 4 datasets), Moncada (1089 identified cells, i.e. 3% of the total), Chan-Seng-Yue (4842 identified cells, i.e. 14% of the total) and Lin (2978 identified cells, i.e. 8% of the total). Dashed colors lines represent high-level functional families of cells (see panel B).

Node 2 Stromal/Immune cells were divided in three subcategories, Immune [Becht et al., 2016, Bindea et al., 2013], Fibro_Stellate [Becht et al., 2016, Baron et al., 2016, Chen et al., 2013] and endothelial cells [Becht et al., 2016]. At Node 3, gene markers of apCAF, iCAF and myCAF populations were retrieved from Elyada et al gene-markers [Elyada et al., 2019] and used to give a specific identity to a subset of cancerous fibroblasts. Non-assigned cells at this step were considered as normal Fibro_Stellate cells. Finally, Node 3 Immune cells were decomposed into 4 subpopulations: lymphocyte T, lymphocyte B, Dendritic cells and Monocyte/Macrophage, using genes markers define by the union of 2 different publications [Becht et al., 2016, Bindea et al., 2013]. This hierarchical 2-steps strategy enabled the identification of 15, 14, 7 and 11 coherent cell-types in Peng, Chan-Seng-Yue, Moncada and Lin dataset, respectively (Figure 3.4B-C).

3.2.2 Generation of unified integrated gene markers and reference cell-types

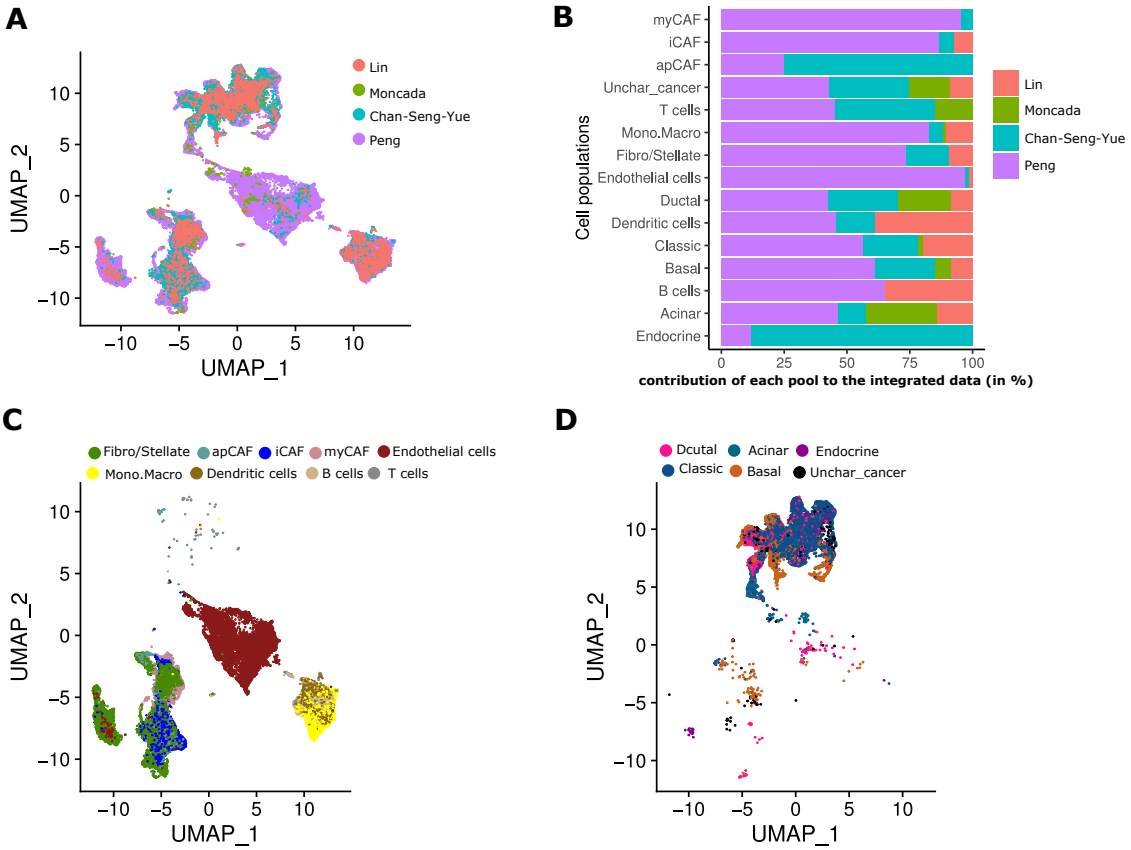

Figure 3.5: **Single cell dataset integration.**
A. UMAP representation of the pool of integrated cells (Seurat integration) according to their dataset of origin. **B.** Barplots representing the contribution of each dataset to each cell-type. **B.** UMAP projection of the pool of integrated cells according to their cell identify, split in 2 plots according Node 1 main compartments : stroma/immune (**C**) and epithelial cells **D**).

We then integrated the four dataset into a merged meta-analysis, in order to identify gene markers specific to the 14 cell-types previously identified. To correct for technical differences

and batch effect, we apply a dedicated data integration approach. We tested two different established integration methods: Harmony [Korsunsky et al., 2019] and Seurat [Hao et al., 2021], and finally selected the Seurat integration method, that displayed the highest Local Inverse Simpson's Index (LISI) score and present the advantage of providing corrected gene expression outputs [Luecken et al., 2022] (Figure 3.5A). We observed that all cell-type are not evenly distributed among the dataset of origin, with a high disequilibrium for endothelial cells and myCAF cells, which predominantly originate from Peng's dataset, whereas Endocrine cells almost all come from Chan-Seng-Yue's dataset (Figure 3.5B). Overall, we observed a efficient clustering of the three main compartments stroma, immune (Figure 3.5C) and epithelial (Figure 3.5D), when all integrated single cell data are projected into a low dimension space using UMAP.

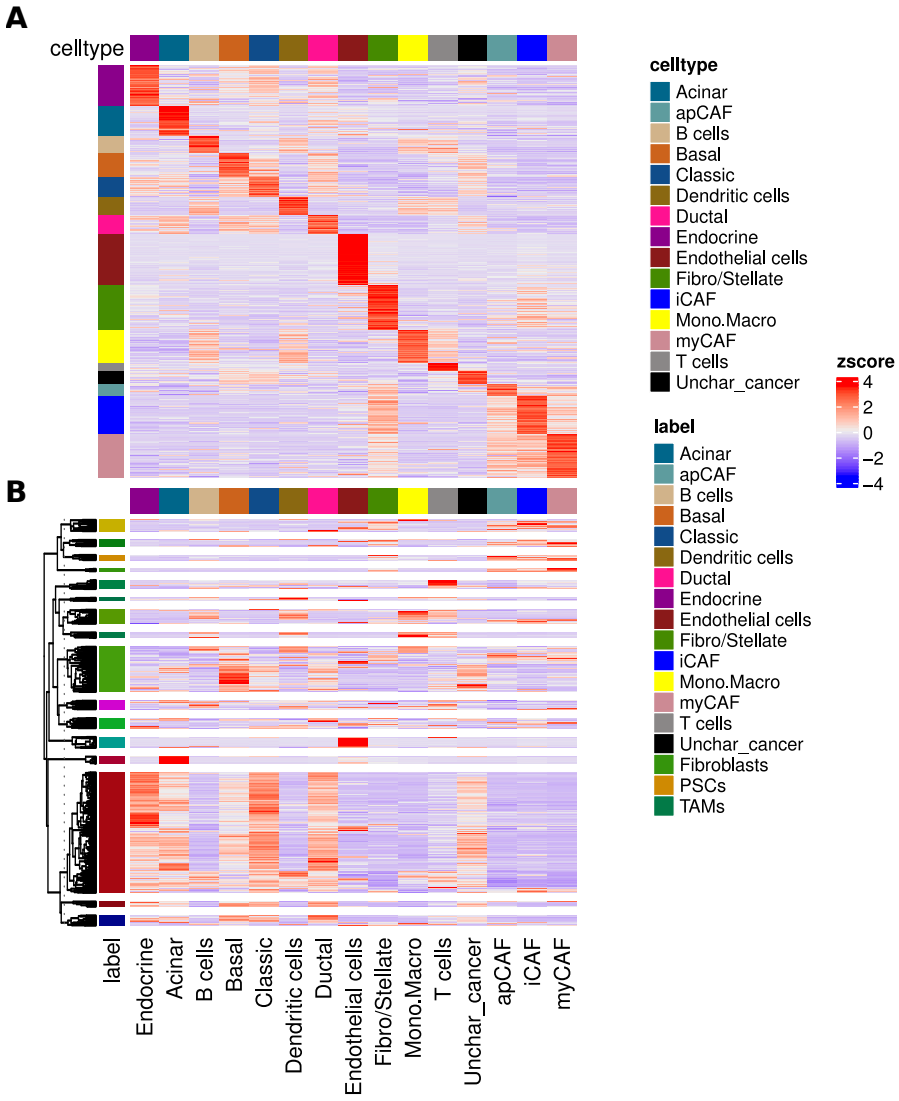

Figure 3.6: **Integrative gene markers of 14 PDAC cell-types.**
A. Heatmap of averaged expression of all cell-types, after data integration, for our robust integrative gene markers. **B.** Heatmap of averaged expression of all cell-types, after data integration, for our curated literatature markers.

Finally we extracted from this integrated dataset a robust and integrative cell-type gene signature. First we selected 70% of cells constitutive of each cell-type compartment as a training set. Then we tested for differential expression between pairs of groups to identify specific integrated markers using the scran method [Lun et al., 2016] combined with a bootstrapping approach (from which we only kept markers found in all iteration steps). We observed that robust integrative gene markers (Figure 3.6A) display a specific cell-type expression, as opposed to curated literature markers (Figure 3.6B).

3.2.3 Deconvolution of PDAC samples using new robust cell-types markers and profiles

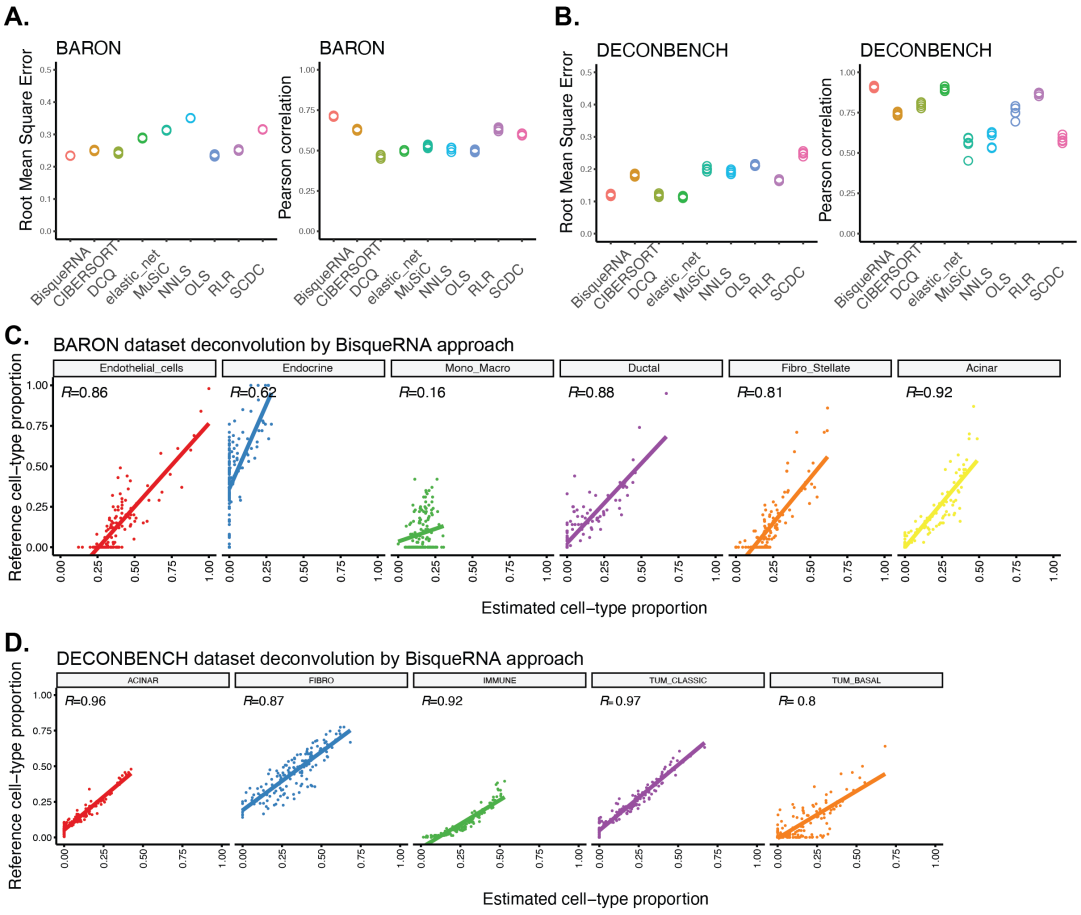

Figure 3.7: **Benchmark of deconvolution using the new single-cell references**
A. Root mean square error and pearson correlation of deconvolution methods for 5 independant simulations on pseudo-bulk healthy pancreas (from baron dataset). **B.** Root mean square error and pearson correlation of deconvolution methods for 5 independant simulations on in silico simulation of cancer pancreas (from deconbench dataset). **C.** Correlations between real cell-types proportion and BisqueRNA estimated cell-type proportion for pseudo-bulk healthy pancreas (one representative simulation). **D.** Correlations between real cell-types proportion and BisqueRNA estimated cell-type proportion for in silico simulation of cancer pancreas (one representative simulation)..

We then performed a comprehensive and quantitative evaluation of several methods using our new PDAC cell-type markers and profiles (average expression accross single-cells of a given cell type). Using root-mean-square error (RMSE) and pearson correlation, we evaluated the performances of reference-based methods (ordinary least square (OLS) [Chambers et al., 1990] , DCQ [Altboum, 2014], elastic net [Friedman et al., 2010], robust least regression [Ripley et al., 2022] (RLR) , non-negative-least-square (NNLS) [Stokkum, 2012] and CIBERSORT [Newman et al., 2015]), and single-cell reference based methods (BisqueRNA [Jew et al., 2020], MuSiC [Wang et al., 2019], and SCDC [Dong et al., 2020]). We used two different benchmark datasets, one with pseudo-bulk mixtures generated with known composition from single cells healthy pancreas (BARON dataset [Baron et al., 2016]) and *in silico* simulations from pure cell lines including pancreatic cancer cells (DECONBENCH dataset [Decamps et al., 2021]). We observed that BisqueRNA method provides the best correlation and the lowest RMSE on both benchmark datasets (Figure 3.7). We then applied this method on bulk transcriptomes of TCGA-PAAD [Weinstein et al., 2013] dataset (Figure 3.8) and obtained unprecedented high-resolution cell-type proportion matrix, correlated with previous Basal vs Classic classification ([Moffitt et al., 2015]). Our next goal is to study this high-granularity cell-type composition and to seek for association with survival data.

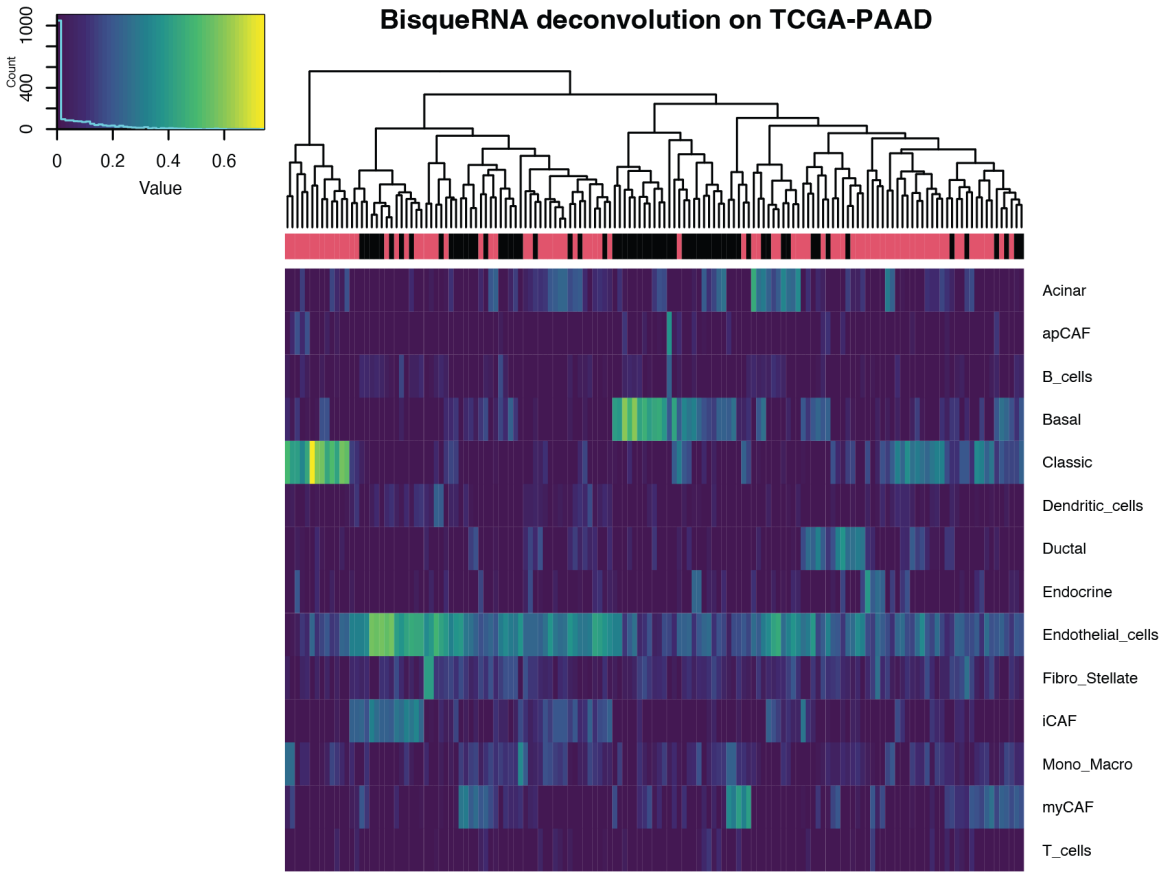

Figure 3.8: **Application to TCGA pancreatic adenocarcinoma.**
Heapmat of BisqueRNA deconvolution of TCGA-PAAD dataset. Red represent previously classified "Classic" samples, black represent previously classified "Basal" samples (Moffit classification [Moffitt et al., 2015]).

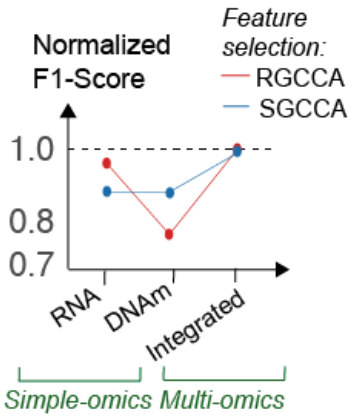

Figure 3.9: **Preliminary results on multiomic integration.**
We performed Support Vector Machine (SVM) PDAC subtypes classification (based on high Basal-like cancer cells and high immune cells content) after RGCCA and SGCCA dimensionality reduction. Using F1-score to evaluate classification, we observed that multi-omics integration outcompetes single-omic dimensionality reduction.

3.3 Method development multi-omic integration and estimation of tumor functional heterogeneity– *prospects*

Parallel to the method we are currently developing to precisely characterize cell-type composition of PDAC using single-cell data, we intend to explore several avenues of reference-free methodological approaches :

- Multi-omic based deconvolution and machine learning based classification of tumors according to intra-tumor heterogeneity (section 3.3.1).
- Inference of functional heterogeneity using a multimodal network based approach (section 3.3.2).

3.3.1 Multi-omic based deconvolution of intra-tumor heterogeneity and patient classification

Though benchmark studies [Chauvel et al., 2020, Rappoport and Shamir, 2018] indicate that integration of multiple omics shows an improvement of data clustering performance compared to single-omic approaches, a recent study found that none of the integrative clustering methods tested were able to accurately classify liver cancer subtypes [Pierre-Jean et al., 2020]. We have recently started to investigate how integrative multi-omics dimensionality reduction, embedded within classifiers, would improve PDAC classifications (Figure 3.9). We compared classifications based on single-omic, multi-omic concatenation (combination of gene and methylation features in a single matrix) and joint multi-bloc dimensionality reduction (jDR), that has been proven to perform well for multiomic integration [Cantini et al., 2021]. 4 out of 6 best-performing methods were based on a joint multi-omic approach. We expect that multi-omic integration will be instrumental to overcome current deconvolution methodological challenges link the the use of reference-free methods.

Multi-omic based tumor heterogeneity classification

Now we want to explore how multi-omic dimensionality reduction, embedded within machine learning based classifiers, will improve classification of tumor, including micro-environment subtypes. We next plan to explore the following ideas:

(i) optimize sparsity, regularization and penalization parameters of the jDR methods to improve

classification performances;

(ii) test other classifiers such as linear regression with lasso penalty and the biological relevance of associated selected features;

(iii) extend our preliminary analysis to multi-class prediction using Adjusted Rand Index (ARI) and deep learning based approach such as DeepCCA;

(iv) implement moCluster [Meng et al., 2016], a fast (consensus PCA EM-algorithm) multiblock analysis enabling variable selections, in order to identify clinical grade classifiers;

(v) explore Similarity Network Fusions [Wang et al., 2014], that gave promising results in performing classification of individuals in a recent benchmark [Pierre-Jean et al., 2020].

Multi-omic integration embedded in deconvolution algorithms

Latest advances in deconvolution algorithms quantitatively inferring tumor composition rely on single-omic approaches (mainly based on transcriptomes or methylomes). We also want to test how integration of complex multiomic datasets will improve the performances of deconvolution algorithms. Feature selection is a key and delicate step to improve performances of deconvolution methods since it may also discard relevant biological information. We expect that using different types of omic data should improve the quality of tumor heterogeneity quantification by (i) removing the bias specific to each type of data and (ii) better identifying the relevant features in each block using joint information provided by both data types. Moreover, most of the commonly used deconvolution algorithms are tested on *in silico* simulated datasets based on matrix products, which does not account for high complexity contained in real datasets (i.e. constitutive biological noise).

First, we will perform this feature integration using multi-block statistical approaches such as 2-way PCA [Pagès, 2014], sparse GCCA [Tenenhaus et al., 2014], sparse PLS [Lê Cao et al., 2008] and regularized GCCA [Tenenhaus et al., 2017], before running standard unsupervised deconvolution algorithms (NMF or ICA based) on simulation. We will use row-correlation, column-correlation and the Mean Absolute Error (MAE) metrics to evaluate accuracy of prediction between estimates and ground truth. Preliminary results indicate that jDR outperforms simple concatenation of single-omic. When we take single-omic concatenation as reference, we observe that the integrative methods tested always display a relative decrease in estimation errors (MAE) and a relative increase in row or column correlations. We next plan to use cross-validation to optimize sparsity and regularisation parameters of integrative multi-block methods, and to test different NMF algorithms (such as alternating least square approaches) to optimize the computing time. Then we will use single-cell RNA-seq specific PDAC profiles i) to perform multi-omic feature selection based on biological a priori [Singh et al., 2019] using cell-type-specific gene markers and methylation probes located in corresponding promoters, and ii) to develop a novel multi-omic supervised deconvolution method relying on least square regression models.

3.3.2 Estimation of tumor functional heterogeneity at the single tumor level

Despite great promise, conventional computational approaches to quantify cellular heterogeneity from mixtures of cells have experienced difficulties in delivering robust and biologically relevant estimations. Rather than focusing on cell-types, we will interrogate heterogeneity at the

phenotypic level, inferring and quantifying which biological function are differentially-regulated in each tumor, as compared to healthy counterparts. This relies on the hypothesis that patients share functional modules (i.e activation or inhibition of biological pathways) that can be identified and quantified as groups of deregulated genes [Hanahan and Weinberg, 2011]. These functional modules can be composed of genes expressed in different cell-types (e.g. ligand-receptor pathways) and will reflect the heterogeneity in biological functions, at the single tumor level. For instance, functional modules containing *GZMA* will be a signature of T lymphocyte function, and their weight in each sample will reflect the strength of the corresponding cytotoxic activity [Rooney et al., 2015, Steele et al., 2020]. Hereafter, *Tumor functional heterogeneity* will encompass the global intra-tumor heterogeneity in biological functions, which results from the activity of all cancer sub-clones, of the different cancer cell subtypes and of the tumor micro-environment cells.

We intend to address the following question: which biological functions are acquired *in situ* by the tumor? By quantifying the functional changes within a tumor, we will go beyond the cellular composition of a given sample, to focus on the biological properties of the malignant and non-malignant cells, shaped by dynamic cell-cell interactions and cellular plasticity .

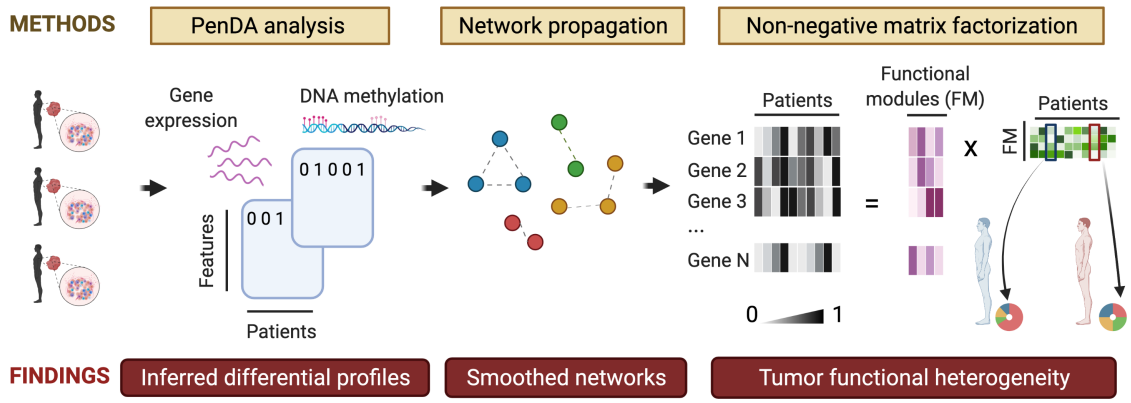

Figure 3.10: **An assessment of tumor functional heterogeneity at the single tumor level.**

Overall, a major reason for the failure of the theoretical inference of cellular heterogeneity is based on the fact that methods do not consider the true biological behavior of a tumor ecosystem: cells interact with each other and are plastic among time. Instead of trying to infer heterogeneity at the cellular level, we will focus on the functional modules structuring tumor heterogeneity (Figure 3.10). The functional modules will be defined as groups of (deregulated) genes involved in a shared biological function. As these functional modules are defined on bulk samples, they can contain deregulated genes with different cell-type of origin and identify the appearance of specific cell-cell interactions (e.g. over-expression of specific ligand-receptor pairs). Furthermore, they present the advantage of reflecting the plasticity of tumor cells, as cells of the same type can activate different modules according to their biological state (e.g. which genes were specifically deregulated at the time the sample was collected?).This work will rely on the hypothesis that patients share recurrent functional modules (i.e. cancer arises from a disruption of major biological pathways [Hanahan and Weinberg, 2011, Vanunu et al., 2010]) which relates to tumor heterogeneity. By inferring these functional modules, it is then possible to estimate the functional

heterogeneity of the tumor, at the level of a single sample. Here, we aim to combine multimodal information sources to leverage the identification of functional modules structuring tumor heterogeneity. First, we will perform a personalized differential analysis [Richard et al., 2020] to detect deregulated genes in each sample (see PenDA method, Chapter 2 of this manuscript). Second, we will enrich this differential information with the *a priori* knowledge of the topology of multilayer networks in humans [Didier et al., 2015, Hofree et al., 2013]. The use of *prior* biological knowledge on network interactions and differential analysis should solve the problem of interpretability of unsupervised algorithms without having to resort to the definition of reference profiles, which are currently limiting for the use of supervised methods.

A new method to infer tumor functional heterogeneity (preliminary results).

Several studies used gene networks as *prior* knowledge to simplify the statistical use of high dimension data while contributing to the biological robustness of the downstream analysis. The integration of the information contained in the network makes it possible to focus on the regulation of biological functions (regulatory networks, signaling pathways, protein complexes). To test the strength of our hypothesis, we performed simulations of pseudo-bulk transcriptomic samples using a public PDAC single-cell dataset [Peng et al., 2019]. Differential gene expression for each sample was generated by the PenDA method (using healthy GTEX pancreatic data as reference). Each gene was subsequently represented by a binary variable (0 = non-deregulated, 1 = upregulated). We then projected the upregulation profiles onto a human interaction network from Pathways Commons [Cerami et al., 2011]. We used a network propagation method to spread the influence of each gene's upregulations observed over their neighbors in the network [Hofree et al., 2013, Morvan et al., 2017]. The resulting profiles (Matrix X, Fig. 3.11A) reflect the strength of gene upregulation along a continuous range [0,1] that account for underlying biological functions. Heterogeneity was then estimated by unsupervised deconvolution (non-negative matrix factorization, NMF) solving the following optimisation problem: $minimise\|X - WH\|$ (Fig. 3.11A). The output W corresponds to the weights of Hidden Components in each patient, whereas the H corresponds to the gene upregulation profiles of each Hidden Component. To evaluate the performance of our method, we calculated the Mean Absolute Error (MAE) between the estimated W matrix and the ground truth (corresponding to the real weights of each cell-type in the simulations). We observed that the estimation error was lower for the *Upregulation + Network* approach, as compared to NMF applied on the *Normalized counts* of the pseudo-bulk matrix or on the *Upregulated* PenDA matrix (Fig. 3.11B). Each Hidden Components (matrix H) can be decomposed in several meaningful functional modules. As an illustration, we applied hierarchical clustering on the H matrix (selecting for monocytes and macrophages marker genes [Becht et al., 2016, Bindea et al., 2013]). We were able to identify two clusters of genes corresponding to potential functional modules specifically upregulated in two different Hidden Components Fig. 3.11C). These components HC3 and HC4 have different distribution of weights (matrix W) among patients, that can be interpreted as *TF-Het* (Fig. 3.11D). These very promising results demonstrate the validity of our hypothesis: different tumors share functional modules that can be quantified as a proxy of tumor heterogeneity. The upcoming challenge will be to determine the most relevant functional modules to accurately quantify *TF-Het* at the single sample level.

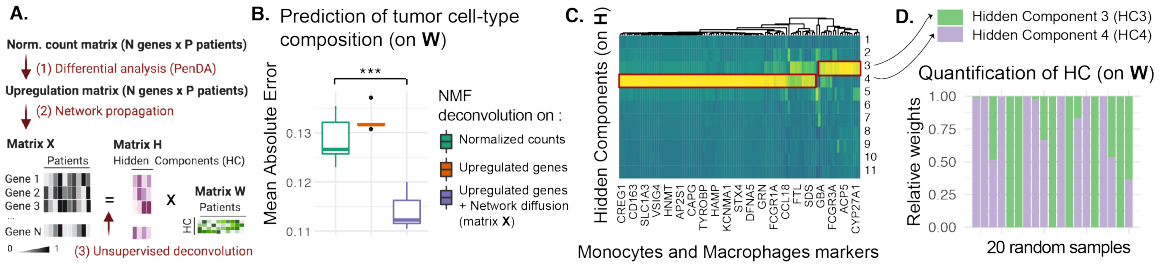

Figure 3.11: **A new method to infer tumor functional heterogeneity**
A. Scheme of the pipeline. **B**. Mean Absolute Error (MAE) between tumor heterogeneity estimation (matrix W) and real simulated cell-types proportions. 5 independent realistic simulations were performed, using each time 11 cell-types including cancer sub-clones and tumor microenvironment cells. *** stands for wilcox.test $p - val < 0.001$. **C**. Heatmap of the upregulation profiles of 11 Hidden Components (matrix H) for Monocytes/Macrophages marker genes (yellow = high network-smoothed-upregulation). Possible functional modules inferred by hierarchical clustering are highlighted in red **D**. Relative weights of Hidden Components 3 and 4 in a subset of the sample population (matrix W).

Development of a robust multimodal pipeline to estimate tumor functional heterogeneity at the single tumor level

Inference of functional modules using differential analysis and network propagation. Our first goal is to develop a robust method to quantitatively infer functional modules building on our preliminary results. Regarding the network smoothing process, we will investigate: (i) the normalization process [Morvan et al., 2017], that has been shown to impact high-dimension statistical methods; (ii) the order of the neighbours to consider (order 1 versus higher orders); (iii) the optimal number of Hidden Components to infer and the cut-off to identify the functional modules of each component (we will identify clusters of deregulated genes by applying hierarchical clustering on the features of the inferred H matrix); (iv) the different networks to consider (Common Pathways, BioGRID, HumanNet, STRING, etc.). Then we will investigate the effects of different methods for unsupervised deconvolution (ICA, NMF) and the associated regularization constraints used to infer the combination of functional modules present in the samples and their proportions in the patients. Quality of the quantified heterogeneity will be assessed using different metrics between estimation and ground truth (when available): MAE, RMSE (root-mean-square error), inter and intra-sample correlation.

Multi-omics integration of RNA and DNAm for differential analysis. Our goal is to strengthen our estimation of genetic deregulation by combining molecular information from RNA and DNA methylation (DNAm) data. As different deregulated genes or methylation probes can participate in the deregulation of the same pathways, multi-omics integration should identify biologically relevant deregulation while reducing noise of the signal. First, we will extend our PenDA framework to DNAm analysis by focusing on local ordering of differentially methylated DNAm probes. Integration will be based on biological *prior* knowledge on relationships between genes and DNAm probes (probes localized within the promoter or the coding region of a gene will be matched with this gene). We will apply feature selection using mutual information carried by omic data. We will test: (i) early integration (where omic data are concatenated); (ii) late integration (analysis of each

omic is processed individually first); and (iii) selection of the relevant features by joint dimensionality reduction [Cantini et al., 2021].

Building informative multimodal network features. Combining pathways, complexes and co-expression networks through multiplex networks (i.e., multi-layer networks sharing the same nodes but with different kinds of edges) is a powerful means to integrate different sources of biological knowledge. However, this is challenging because it requires identifying which features are informative and how they are connected. We will develop a framework to account for multiple heterogeneous layers during the network diffusion step. First, we will try to merge the different layers into a monoplex network. Second, we will test multiplex embedding (such as MultiVERSE [Pio-Lopez et al., 2021]) to integrate the possibility to jump from one layer to another during the process.

Biological relevance of inferred functional modules and associated quantification of heterogeneity. To investigate the biological relevance of functional modules (does it really reflect the cell states present in the tumor?), we will perform $\chi 2$ contingency tests on the enrichment of each functional module in each Hidden Component. The biological relevance of inferred functional modules will be studied using overrepresentation analysis (ORA), such as G-profiler of GSEA. We will investigate the co-occurence of these functional modules in our single cell data, and test if such profiles are compatible with the cell states observed, i.e., how functional modules are activated in individual cells? Finally, we will stratify the patients using the W weight matrix and we will analyse the link between functional tumor heterogeneity and survival using cox models.

A functional interpretation of tumor heterogeneity

In this chapter, I will present prospective projects dedicated to the fonctional study of intra-tumor heterogeneity, in terms of evolutionary impact and causal relationships.

Current research on personalized oncology is expanding, in particular thanks to the reduction in sequencing costs allowing the generation of multi-omics profiles at the scale of individual tumors and individual cells. However, efforts in understanding the extensive heterogeneity of tumors were so far largely limited to cancer cells because of a lack of methods to study these cells together with their environment. Thanks to the work presented in chapters 2 and 3, we can now unlock this obstacle and extend our knowledge of cancers as complex ecosystems, accounting for all gene deregulations and related biological functions present within a tumor. This allows us to investigate how intra-tumor heterogeneity functionally impacts tumorigenesis. Using the PDAC use-case, I intend to particularly study :

- The cancer evolution in the light of tumor functional heterogeneity. This should allow us to enrich recent models of PDAC development, with potential rapid applications in clinical management, especially regarding immmunotherapies (section 4.1).
- The causality link between intra-tumor heterogeneity and disease outcome. Identifying molecular mediators of tumor functional heterogeneity represents exciting opportunities to construct mechanistic models of carcinogenesis integrating this complex ecosystem (section 4.2)

4.1 Relationship between tumor functional heterogeneity and cancer evolution

The time course of cancer remains elusive because observation of cancer dynamics in its native environment is almost impossible. Mutations in oncogenes are often used to infer the evolution (lineages and clonality) of cancer cells at a given moment, with little considerations for the effect of the tumor micro-environment on the process. However, interactions of cancer cells with the micro-environment catalyze molecular changes that might confer a selective advantage to cancer cellular clones, through genetics and non-genetic mechanisms (see reviews of PDAC evolution: [Makohon-Moore and Iacobuzio-Donahue, 2016, Connor and Gallinger, 2021]), with consequences on the invasion and spread of tumor cells [Hayashi et al., 2021]. In this part, our objective will be to unravel the relationship between tumor heterogeneity and the (epi)genomic landscape observed in a sample at a given time point, to build a comprehensive view of cancer development within its environment (Figure 4.1)

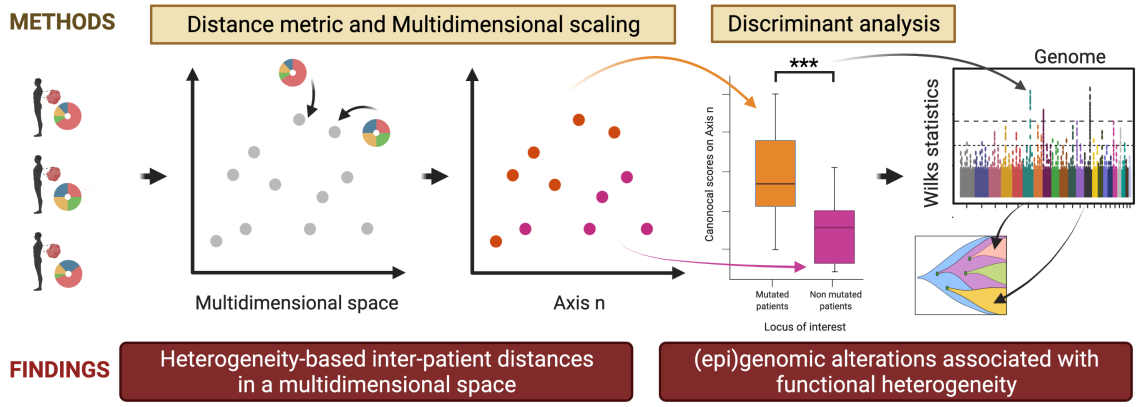

Figure 4.1: **An analysis of cancer evolution in the light of tumor heterogeneity. .**

4.1.1 An evolutionary view of pancreatic adenocarcinoma

The conceptual model of cancer as an evolutionary problem is not new and was first described in the early 1970s [Nowell, 1976]. Tumor evolution is based on changes in cell fate, resulting from somatic mutations, variations in copy number, chromosomal rearrangements and epigenetic modifications. The time course of cancer is based on clonal evolution (asexual reproduction of unicellular organisms) and competition between different cells of the same organism, in which local adaptation to the environment is crucial. The nature and sequence of genetic events defining some common cancers have been characterized in detail over the past three decades, with little consideration for the effect of tumor heterogeneity on the process. The variations in the tumor micro-environment include metabolic changes (such as oxygen and nutrients), but also changes in cell types and cell states that can influence the growth of cancer cells, which in turn shape and modulate the tumor micro-environment. Phenotypic adaptations to these pressures help cancer cells to survive during metastasis, which is a major clinical problem for patients. The PDAC genomic landscape indicates that core pathways are targeted by somatic alterations, as well as multiple degrees of structural variation [Waddell et al., 2015, Biankin et al., 2012, Jones et al., 2008]. The contribution of tumor heterogeneity to PDAC initiation and clonal expansion remains to be characterized, although recent studies indicate that tumor micro-environment has a significant impact on the phenotype of PDAC cancer cells: (i) xenotransplanation of human organoid PDACs in murine pancreatic ducts has recently shown that the microenvironment can influence the molecular subtypes of cancer cells [Miyabayashi et al., 2020]; and (ii) CAF secretion of TGFβ has also been shown to change the phenotype of cancer cells in cell cultures [Ligorio et al., 2019].

4.1.2 Relationship between intra-tumor heterogeneity and somatic mutations.

To confirm the reciprocal relationship between intra-tumor heterogeneity and cancer evolution (initiation, invasion and dissemination), we developed a pipeline enabling to associate binary variations (genetic mutations versus *wild-type*) to quantitative phenotype distribution (e.g. the weight of a functional module within the tumors cohort). We investigated if the most frequent PDAC somatic mutations were associated with tumor cell type heterogeneity in the TCGA cohort. PDAC initiation is mainly driven by alteration of four core signaling pathways (*KRAS*, *CDKN2A*, *TP53* and *SMAD4*) [Makohon-Moore and Iacobuzio-Donahue, 2016], but low frequency mutations in driver

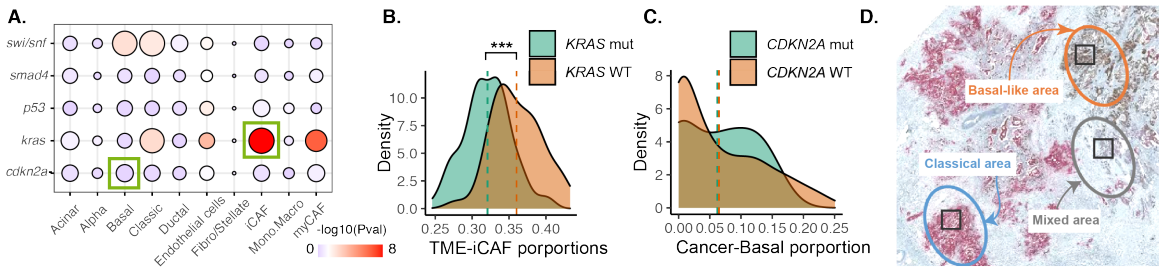

Figure 4.2: **Relationship between somatic mutations and tumor heterogeneity.** **A**. Statistical t-tests where performed to detect association between genetic mutations in the 5 key driver genes and estimated 10 cell types proportions on 149 PDAC patients (significant p-value correspond to different cell type proportions between mutated versus non-mutated patient). The size of the circles corresponds to the Kantorovitch distance between cell type distributions of mutated versus non-mutated patients (small circle = equal distributions, large circle = difference between distributions). Green squares represent two examples illustrated in panels B and C. **B**. iCAF proportion distribution between *KRAS* mutated versus non-mutated. Doted lines represent the mean of each subgroup. *** stands for student $p - val < 0.05$. **C**. Cancer-basal-like cell proportion distribution between *CDKN2A* mutated versus non-mutated. Doted lines represent the mean of each subgroup. **D**. Spatial heterogeneity in PDAC immunohistostaining. Red IHC: classical-like marker. Brown IHC: basal-like marker.

genes were also observed, as well as mutations reprogramming the epigenomic landscape (e.g the *SWI/SNF* complex). After deconvolution of 10 cell types proportions using the MuSiC single-cell based deconvolution approach [Wang et al., 2019], we observed significant differences in averaged cell type proportions between mutant and wild-type samples (considering five key driver genes), for a subset of mutations/cell types combinations (Figure 4.2A). For instance, patients carrying a *KRAS* mutation display a lower proportion of iCAFs (inflammatory Cancer Associated Fibroblasts) than the patients with no alteration in *KRAS* (Figure 4.2B). Interestingly, we observed that in some cases, somatic mutations can induce a shift in a given cell type distribution without changing the mean of this cell type proportion. This is the case for *CDKN2A* mutations, that trigger an increase in cancer-basal-like cells in a sub-population of patients (Figure 4.2C). These preliminary results confirm previous observations on the relationship between tumor heterogeneity and somatic mutations present in cancer cells. Moreover, these observations confirm the need to consider tumor heterogeneity as a multivariate quantitative phenotype, in order to be able to detect subtle or partial associations between somatic events and tumor composition. It remains to be determined how exactly the relationship between tumor heterogeneity and (epi)genomic variation is structured. Finally, using immunohistostaining we observed that some tumors display spatial heterogeneity, with some sub-compartments presenting different phenotypes of cancer cells, which could be associated with different intra-tumor heterogeneity (Figure 4.2D).Deciphering the relationships between intra-tumor heterogeneity and tumor evolution in these spatial compartments is still an ongoing challenge.

4.1.3 A systematic study of intra-tumor heterogeneity and (epi)genomic landscape – *prospects*

Correlation between genomic changes and intra-tumor heterogeneity. First, we will take advantage of the rapidly growing public repository of cancer genomic data [Raphael et al., 2017,

Australian Pancreatic Cancer Genome Initiative et al., 2016, Puleo et al., 2018] to associate genomic mutational profiles and genomic alterations with intra-tumor heterogeneity at the individual scale [Sakamoto et al., 2020]. Genomic alteration studies face important challenges: (i) not all identified mutations are biologically relevant, most mutations being passenger mutations that do not contribute to tumorigenesis and ecosystem evolution [Vogelstein and Kinzler, 2015], and (ii) mutation rates differ between patients, with the mutational burden itself having phenotypic relevance [Lawrence et al., 2014]. Here we will consider the relative abundance of each cell type as a multivariate quantitative phenotype describing the functional heterogeneity of each tumor. We will develop methods to detect significant associations between genomic mutations and intra-tumor heterogeneity. To do so, we intend to generalize a quantitative genetic approach we previously published [Chuffart et al., 2016] which enables mapping genomes for alterations modifying the statistical properties of quantitative traits. The principle is to construct a phenotypic space where the coordinate of each tumor on the k-th axis corresponds to the weight of the k-th Hidden Component in the tumor (i.e. the quantitative phenotype). We will test the association between intra-tumor heterogeneity and the mutational genomic profiles by canonical discriminant analysis on the phenotypic data, using a binary representation of each mutation (0 or 1) as a discriminating factor. Next, we will define a linkage score using Wilks' lambda statistic (suitable for multivariate analyzes) to isolate meaningful mutations. We will screen all mutations identified in each sample and study the intra-tumor heterogeneity patterns associated with common or rare mutations to identify if specific intra-tumor heterogeneity will promote the emergence of recurrent genomic landscapes and/or the acquisition of specific somatic alterations.

Correlation between epigenomic instability and intra-tumor heterogeneity. As some PDAC tumors exhibit epigenetic changes rather than genome alterations, and because epigenetic changes are more plastic than gene mutations, our next goal will be to study the association of intra-tumor heterogeneity with DNA methylation (DNAm) modifications. We will use the PenDA approach to compute differentially methylated probes (DMP) and differentially methylated regions (DMR). We will then detect significant associations between epigenomic variations and intra-tumor heterogeneity. These results will be analysed in the light of current biological knowledge to infer epigenomic landscapes associated with particular intra-tumor heterogeneity.

Relationship between tumor evolution and heterogeneity when accounting for spatial variations. PDAC staining performed by our collaborator J. Cros revealed different types of spatial intra-tumor heterogeneity (Figure 4.2D), with about 30% of tumors characterized by intermediate phenotype and prognostic (unpublished data). Current molecular patient classifications do not reflect these spatial properties, which clinical implication is currently unknown. We will perform the same association approach on the ACACIA PDAC dataset (containing multiple microdissections of the same tumor, at different spatial locations) to study if spatial heterogeneity of cancer sub-clones will affect the intra-tumor heterogeneity, and vice-versa. These results will be complementary to emerging spatial transcriptomic profiling approaches [Hwang et al., 2020, Moncada et al., 2020], that address the question of the co-occurrence of distinct transcriptional programs in complex tissues.

4.2 A causal link between heterogeneity, environment and outcome

Intra-tumor heterogeneity is central for both the trajectory of a tumor and the patient outcome [Marusyk et al., 2020], but the causal molecular changes that mediate this outcome are still unknown. Indeed, when an effect is observed by statistical association between an external exposure (**E**, i.e. tumor heterogeneity) and a patient outcome (**Y**, i.e. survival), one or more intervening variables (**M**, i.e. gene expression and/or epigenetic changes) can mediate this effect. This mediated effect is called indirect effect, as opposed to the direct effect of **E** on **Y** (unexplained by the intervening variables **M**) [Richiardi et al., 2013]. Statistical mediation analysis is a technique of choice to infer these causality relationships. However, multimodal high-dimension mediation is difficult for different reasons: correction for multiple testing, correction for confounder effects, accounting for interaction between mediators, and multimodal data integration [Blum et al., 2020]. Our goal will be to develop a new multi-omics mediation analysis framework to unravel the pathways that link tumor heterogeneity, response to environmental cues (pharmaceutical treatments) and the development of this autonomous ecosystem (disease outcomes) (Figure 4.3).

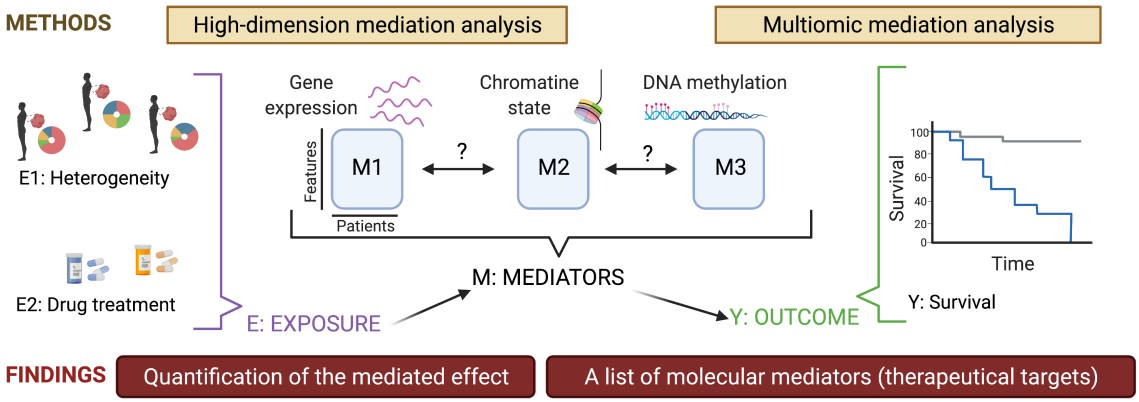

Figure 4.3: A rigorous causal analysis of the effect of tumor functional heterogeneity

4.2.1 Introduction to mediation analysis

Currently, there is no framework to infer the causal mechanisms underlying the effect of tumor heterogeneity on disease outcome. The molecular changes that mediate this outcome are still unknown, though these molecular changes are key factors to understand the foundations of disease susceptibility. Inferring the causal relationships that link **E** to **Y** can be addressed using high-dimension mediation analysis. Mediation analysis is a statistical approach for understanding the causal structure between **E** and **Y** through the inclusion of mediator variables. Mediation analysis has been primarily focused on univariate mediation where only one mediator is considered [Baron and Kenny, 1986]: (i) the effects of **E** on **M** and of **M** on **Y** are statistically tested; (ii) the significance values obtained at the first step are combined; and (iii) the indirect effect of the mediator is estimated (Sobel test or Average Causal Mediated Effects (ACME) [Sobel, 1982, Imai et al., 2010]). Generalizing mediation analysis techniques developed for one

mediator to high-dimension is not straight forward, as it involves many pitfalls: correction for multiple testing and controlling for the false discovery rate (FDR), reverse causation, correction for confounding effects (E-Y confounders, as well as E-M and M-Y confounders), accounting for interaction between mediators, and multimodal data integration when using different types of variables, such as genomic and epigenomic data [Blum et al., 2020, Zeng et al., 2021]. To date, there is no real consensus on an optimal combination of models and methods for multimodal high-dimension mediation analysis. We want to test the extent to which the effect of tumor heterogeneity on disease outcome is explained (or not) by the molecular features of the tumor (i.e. gene expression and DNAm).

4.2.2 Development of a new method to perform high-dimension mediation analysis

Our collaborator O. François (TIMC), recently developed a high-dimension mediation analysis method (HDMA) to study the indirect effect of DNAm in the pathway between exposure and outcome (considering maternal smoking as exposure and birth weight as outcome, using the mother-child cohort EDEN). First, the HDMA method uses latent factors mixed models (LFMM [Caye et al., 2019]) for estimating hidden confounders both in the association analysis of exposure and in the association analysis of outcomes.

$$M = Xa_1^T + U_1 V_1^T + E_1 \ \textbf{(1)} \text{ and } Y = Xa_2^T + Mb^T + U_2 V_2^T + E_2 \ \textbf{(2)}$$

In equation **(1)**: M is defined as DNA methylation profiles (beta-normalized values), X represents the effects of exposure, a_1 contains the effect sizes of exposure on DNAm levels, U_1 is a matrix formed of K latent factors estimated simultaneously with a_1, V_1 contains the loadings associated with the latent factors, and E_1 is a matrix of residual errors. The K latent factors represent unobserved confounders, which could be cell types of tissue samples, clinical variables (such as gender or age) or various batch effects. In equation **(2)**: Y represents the health outcome, a_2 contains the effect sizes of exposure on the outcome, b contains the effect sizes of marker levels on the outcome, U_2 are latent factors from a latent factor regression model, V_2 are the corresponding loadings, and E_2 is a matrix of errors. For each marker j, a significance value, Px (resp. Py) is computed for the test of a null effect size for exposure on DNAm (resp. for the DNAm on outcome). Then, HDMA combines the significance values Px and Py computed at each DNAm marker by using a new procedure called the max-squared test: $P = max(Px, Py)^2$. This max-squared test evaluates the null-hypothesis that either the effect of exposure on DNAm or the effect of DNAm on outcome is null.

We used the Moffitt classification of PDAC samples [Moffitt et al., 2015] as a proxy of tumor heterogeneity of 150 PDAC samples (TCGA cohort [Weinstein et al., 2013]) and applied the HDMA method to identify DNAm mediators **M** of tumor heterogeneity (**E**, exposure) on PDAC clinical grade (**Y**, outcome) (Figure 4.4). We identified 39 DNAm probes mediating the effect of **E** on **Y**. Some associated genes were already known to be correlated with cancer survival. These results demonstrate the feasibility of our approach. A systematic causal analysis of the effects of intra-tumor heterogeneity on disease outcome remains to be conducted.

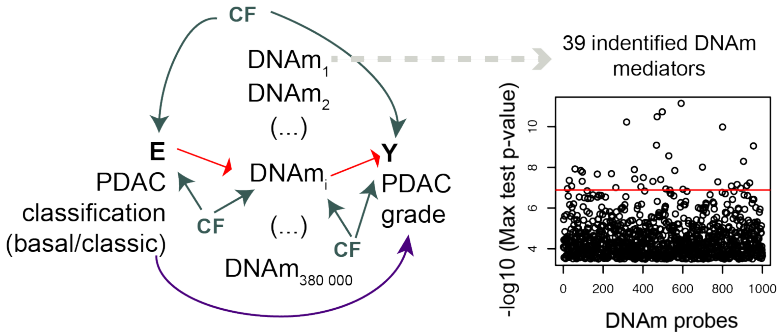

Figure 4.4: **The causal inference path.**

CF stands for confounding factors. Exposure: Basal or Classic. 380,000 potential DNAm mediators were tested.

4.2.3 Identification of molecular mechanisms by which tumor heterogeneity influences disease outcome – *prospects*

A causal analysis of DNAm in the pathway between intra-tumor heterogeneity and patient survival. We will perform a thorough analysis of the direct and indirect effects of tumor heterogeneity on PDAC outcome through DNAm modification. PDAC outcomes include survival outcomes with censoring, disease free survival and clinical grade of the tumor (qualitative variable). DNAm is a well-studied epigenetic mechanism that contributes to cell type differentiation through the control of gene expression. DNAm is often disrupted in cancer and presents a potential therapeutic target [Baylin and Jones, 2016]. This work will be conducted using the public cohorts of PDAC patients from the TCGA and the ICGC. The exposure (**E**) will correspond to intra-tumor heterogeneity. We will use the HDMA method to identify DNAm markers and DMRs, calculated using comb-p, a method that combines adjacent p-values as mediators (**M**) of indirect effect of tumor heterogeneity (**E**) on disease outcome (**Y**). Proper identification of hidden confounders is a key factor in this pipeline, as cell type composition contained in bulk samples will affect the recorded signal for each potential mediator. We will build conditional realistic simulations and use the F1-score (harmonic mean of precision and sensitivity) to evaluate the ability of our LFMM model to account for hidden confounders.We will estimate the overall indirect mediated effect using the R package mediation [Imai et al., 2010]. We will account for correlation among mediators (variation in methylation impacting other probes) by using a joint mediation model. The standard deviation of the indirect effect estimate will be computed using a bootstrap approach. We will analyse these results in the light of biological knowledge on potential causal mechanisms.

A causal analysis of gene expression in the pathway between intra-tumor heterogeneity and patient survival. We will extend our HDMA approach to study the direct and indirect effect of intra-tumor heterogeneity on PDAC outcome through gene expression. This work will also be conducted using the public cohort of PDAC patients from the TCGA and the ICGC. We will develop simulation experiments to compare the performances of HDMA with other regression models and to other mediation methods [Djordjilović et al., 2019, Sampson et al., 2018, Dai et al., 2020]. We will simulate exposure, outcome and confounding factors using multivariate models. The simulation will include the correlation between those variables. We will build realistic simulations to set the vectors of effect sizes (a for exposure and b for outcome) by deducing these parameters from real datasets, which will likely differ between gene expression and DNAm. For each simulation, we will compute a list of potential mediators and compare it with the real causal markers using the F1-score. Then, similarly to the approach developed before, we will estimate the overall indirect

effect and compute its standard deviation using a bootstrap approach. We will then biologically interpret these results and the tropism of the effects (negative or positive).

Multimodal mediation analysis. We will perform multimodal mediation analysis to study both the effects of DNAm and gene expression on PDAC outcomes. To combine this multimodal data, we will use complementary approaches: joint dimension reduction such as sparse generalized canonical correlation analysis, correlation analysis, and *a priori* biological knowledge on the causal relations between these multimodal layers [Garali et al., 2018, Liu et al., 2020]. We will move from the simple 3-variables system to integrate multilayer mediators and order them to detect functional entities (i.e., what are the relationship between gene expression and DNAm mediators?). We will evaluate the indirect effect mediated by our joint multimodal analysis and compare this with our previous results.

Contribution of intra-tumor heterogeneity in the response to pharmaceutical therapeutic treatment. To study the effect of tumor heterogeneity in the causal path between pharmaceutical therapeutic treatment and PDAC outcome, we will use two different cohorts generated with our collaborator J. Cros: (i) MOSAPAC samples (containing two chemotherapy regimens as adjuvant treatment - gemcitabine and folfirinox) and (ii) a novel PDAC cohort exposed to folfirinox neo-adjuvant treatment. These data include DNAm, gene expression and standardized clinicopathological variables (including sex, age at diagnosis, preoperative assessment of clinical disease stage, tumor stage, histologic grade, adjuvant therapy and relevant outcome parameters including overall survival and disease-free survival). We will define a mediation model accounting for multiple exposures to describe the effect of pharmaceutical therapeutic treatment on PDCA outcome through tumor heterogeneity. We will identify molecular mediators of pharmaceutical exposure and measure the indirect mediated effect caused by tumor heterogeneity. This will lead to functional comprehensive models that predict the impact of tumor heterogeneity at the patient level, in response to drug exposure.

Algorithms evaluation and collaborative science

In this chapter, I will discuss the use of competitions and data challenges as a proxy for collaboratively comparing newly developed computational algorithms. In particular, I will present different workshops that I have organized over the past few years, as well as my participation in collective efforts to develop open source and open access tools facilitating crowdsourced science.

Since the recent development of high-throughput sequencing technologies, access to massive multi-omics data has revolutionized life sciences. This transformation has been accompanied by the development of computational methods accounting for the complex nature of these data: high-dimension, multimodal data integration, confounding patient history (age, gender, etc.), missing data, intrinsic and extrinsic noise. The development of open-source computational tools ensuring the reproducibility of analyzes and knowledge transfer between research and clinical settings is a major issue in the integration of big data in healthcare. To date, data scientists seriously lack efficient tools to benchmark these computational methods in an objective way. First, the development and diffusion of outstanding data analysis tools require reproducibility, good practices and infrastructure. Second, meaningful publicly available datasets suited for benchmarking are sparse, and researchers lack good quality clinical datasets containing ground truthing to properly evaluate the methods. Third, scientific benchmarking of computational methods is often unsatisfactory because researchers that develop their own method are biased towards demonstrating its superiority.

When I started to take an interest in tumor heterogeneity and deconvolution, I quickly decided to organize data challenges in order to familiarize myself with existing methods, and to facilitate exchanges with scientists in the field. These data challenges have led to fruitful long-term collaborations with oncologists (J. Cros), bioinformaticians (Y. Blum) and computer scientists in the field of artificial intelligence (I. Guyon). I led several scientific projects in this framework (benchmarks for computational methods in life sciences), but I also contributed to the development of a new open-source data challenges & benchmarking platform, as well as to the writing of a book in preparation, entitled "Competition IA and benchmarks, The science behind the competitions". The following sections present these different results, as well as a new project to be carried out in the coming years.

5.1 Introduction to data challenges, a new avenue for collaborative science

Data challenges are a fairly innovative format of collaborative workshops, where participants are invited to perform a complex task defined by the organizers. The principle is to work as a team, to solve problems, in the form of a competition. The most successful methods are ranked using metrics calculated on "benchmark" data, for which the true solution is known by the organizers. This introductory part contains mainly extracts from the following book "AI competitions and benchmarks: The science behind the contests"[1] .

> The quintessential challenge revolves around an existing quantitative standard or benchmark, and seeks to improve upon state-of-the-art. One of the more longstanding benchmark initiatives is the Critical Assessment for Structural Proteins (CASP), which asks participants to predict protein structure (folding) from protein sequence. Groups who specialize in this domain are naturally incentivized to compare their approach in the structured and objective format of a data challenge in the hope that their method out-competes other approaches and can therefore become a new standard in the field [Bender, 2016]. CASP is now recognized within the protein structure community as the *de facto* forum for assessing algorithms, and is therefore as much an incentive as a mandate for formal recognition with the community. This incentive generalizes to all specialties, including image recognition (e.g. MNIST [Madry et al., 2019]), gene identification and function prediction (e.g. RGASP [Steijger et al., 2013]) or drug binding (e.g. on going DREAM drug binding challenge).
>
> Any published AI algorithm is expected to include a formal performance comparison against state-of-the-art methods. No good data-driven approach could emerge without good quality, well curated data. This task can be cumbersome and require a great deal of work to assemble and prepare benchmark datasets. Consequently, a natural perk of a scientific data challenge is that the work involved to generate and prepare a benchmarking dataset is managed by the challenge organizers. Therefore, AI competitions offer a playground with data that are usually costly and complicated to generate. Access to these types of datasets is a strong motivation for participants aiming to develop cutting edge methodological approaches to solve a complex scientific problem.
>
> Recurrent challenges also present the advantage or keeping people on a regular schedule, as they expect the challenge to come and reserve time for it. It provides participants the opportunity to start new collaborations with people from different disciplines gravitating around the same topic.

[1] Quotation from Chapter 12: Practical issues: Proposals, grant money, sponsors, prizes, dissemination, publicity
Leader author: **Magali Richard**, with co-authors: Gustavo Stolovitsky, Justin Guinney, Yuna Blum and Adrien Pavao.
Book: AI competitions and benchmarks: The science behind the contests, in preparation, for the Springer series on Challenges, Data, and Benchmarks [approved proposal]

Figure 5.1: **The incentives for participating in a challenge.**

> Data challenges remain the best functioning way of implementing coopetitions: people compete and get credit for winning, then they share their solution publicly and the community can move together to the next step. How to incentivize participants to work on complex problems is a key feature of challenge organization (Figure 5.1). Mechanisms for engaging and disseminating a competition towards a targeted community are complex and highly dependant on the scientific field. See Figure 5.2 for a review of community engagement strategies and examples of recent competitions. Participatory benchmarking competitions generally result in scientific publications (see examples [Creason et al., 2021, HADACA consortium et al., 2020, Marbach et al., 2012, Eicher et al., 2019, Marot et al., 2021, Le et al., 2019]) which will be of use to the community. Offering authorship to competing teams provides international visibility and recognition to participants.

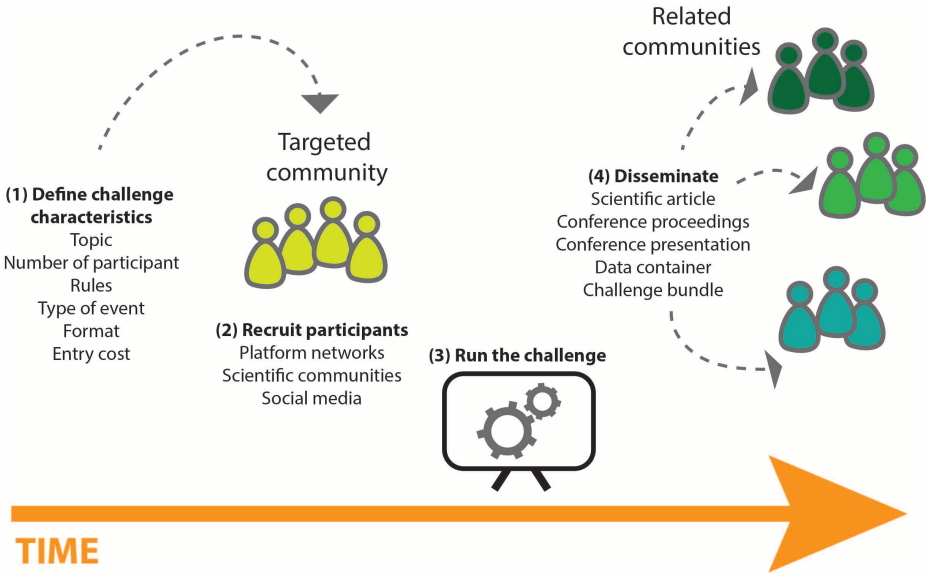

Figure 5.2: **The process of engaging a community**

5.2 Feedbacks on data challenge organization

As part of the Health data challenge program that I coordinated (mainly funded by the European EIT Health program), we organized two face-to-face data challenges and an online training. The objective of this program was to provide (i) an original framework to evaluate the new algorithms developed for the analysis of health data and (ii) innovative teaching methods to train students, scientists and health professionals to the analysis of big data in the health sciences. All these scientific events were dedicated to the quantification of intra-tumor heterogeneity using appropriate statistical methods on transcriptomic and methylation data.

- The first edition[2] was dedicated to the quantification of tumor heterogeneity using methylomic data. It led to a publication will all participants as co-authors [HADACA consortium et al., 2020].
- The second edition[3] was dedicated to he quantification of tumor heterogeneity using multiomic data (methylome and transcriptome). It led to a publication will all participants as co-authors [Decamps et al., 2021].
- The online trainig course[4] aimed to train clinicians to the use of cutting edge computational methods to quantify tumor heterogeneity through a dedicated user-friendly web interface.

For each event, around thirty participants were present, whith a high diversity of backgrounds which made it possible to mix different disciplines (medical doctors, bioinformatics, biology, mathematics, statistics, computer science, etc.), different statuses (researchers, young researchers, students, etc.) and different origins (universities and institutes from all over Europe) within the same teams. The open source platform chosen to perform the data challenges and the training course was Codalab[5] which allows to create personalized competitions and benchmarks in a fairly simple way.

[2] https://cancer-heterogeneity.github.io/data_challenge_2018.html
[3] https://cancer-heterogeneity.github.io/data_challenge_2019.html
[4] https://cancer-heterogeneity.github.io/cometh_training.html
[5] https://competitions.codalab.org/

5.2.1 Unsupervised deconvolution of methylation data.

DATA CHALLENGE 1st EDITION (December 4-10th, 2018)

- **Challenge.** This challenge focuses on estimating cell types and proportion in biological samples based on averaged DNA methylation and full patient history. The goal is to explore various statistical methods for source separation/deconvolution analysis (Non-negative Matrix Factorization, Surrogate Variable Analysis, Principal component Analysis, Latent Factor Models, ...). Participants are made aware of several pitfalls when analyzing omics data (large datasets, missing data, confounding factors...).
- **Invited speakers.** Eugene Lurie, from BCM, Houston, USA. Pavlo Lutsik, from DKFZ, Heildeberg, Germany. E. Andres Houseman, independent data scientist, USA.

The purpose of the data challenge was triple : (i) to offer participants an introduction to the unsupervised deconvolution of DNA methylation data in R, (ii) to compare the performances of the three most recent algorithms (EDec, MeDeCom and RefFreeEWAS), and (iii) to assess whether these methods could be improved by adding a step of pre-processing on the data.

During the challenge, each of the three reference-free methods was presented to the participants by its developers: E. Lurie for EDec [Onuchic et al., 2016], A. Houseman for RefFreeEwas [Houseman et al., 2016] and P. Lutsik for MeDeCom [Lutsik et al., 2017]. These three methods were based on the same deconvolution algorithms: non-negative matrix factorization (NMF), but differed on the initialization step, as well as on regularization constraints.. Given that the preprocessing of data in deconvolution issues was still largely unexplored, even though many confounding factors other than cell type have an impact on DNA methylation (genetic, biological, environmental conditions, experimental effects...), we asked participants to imagine ways to take these experimental variables into account, for example by correcting their effects or filtering out the affected probes. As a results of this work, we published guidelines [HADACA consortium et al., 2020] that are detailed Chapter 3.2 of this manuscript.

5.2.2 Multiomic integration.

DATA CHALLENGE 2nd EDITION (November 25-29th, 2019)

- **Challenge.** This challenge is dedicated to the quantification of intra-tumor heterogeneity using appropriate statistical methods on (DNA) methylome and transcriptomic data in cancer. In particular, it will focus on estimating cell types and proportion in biological samples (in vivo and in silico mixtures) for which transcriptome and/or methylome profiles have been generated. This challenge is also a unique opportunity to compare the performance of deconvolution methods between transcriptome and methylome data, which might have a great impact on clinical practice.
- **Invited speakers.** Michael Scherer from Max-Planck-Institut fur Informatik, Saarbrucken, Germany. Francisco Avila Cobos from Ghent University, Gand, Belgium. Jerome Cros from AP-HP, Paris, France.

This challenges aimed at investigation multiomic integration for deconvolution of intra-tumor heterogeneity. Following up this second challenge, we proposed DECONbench, an innovative public digital benchmarking platform, open source, and freely available for the scientific community, including both high quality benchmarking datasets and reference computational methods. The platform can be used to assess the performance of newly developed methods, which are automatically compared to the existing ones in a user-friendly fashion. The following chapter contains extracts from the published DECONbench paper [Decamps et al., 2021].

> Recent efforts have been made to objectively compare existing tools in order to guide the users. In particular, two recent benchmark studies proposed a comprehensive comparison of transcriptome-based deconvolution methods using various parameters and simulation settings [Avila Cobos et al., 2020, Jin and Liu, 2021]. In the same vein, the DREAM challenge proposed in 2019 [White et al., 2019] a data challenge dedicated to the prediction of immune cell types, showing the emerging spirit towards reproducibility and benchmarking. Although interesting, all these efforts are time-bound and cannot take into account upcoming novel methods. Moreover, the possibility to integrate different types of omic data to infer cell-type proportions is currently under-studied.
> Standardized unbiased benchmarking resources are essential to evaluate the performances of computational methods. Indeed, these resources should avoid falling into the 'self-assessment trap', in which researchers are unrealistically expected to fairly compare their own computational method with other similar algorithms [Norel et al., 2011, Buchka et al., 2021]. In addition, unbiased attempts to benchmark computational methods are often static in space and time, preventing further contributions of other scientists or the assessment of new methods developed after the publication of the benchmark [Mangul et al., 2019]. Recent collective initiatives provided formal guidelines and unified frameworks to improve unbiased performance evaluation [Marx, 2020]. For instance, the Global Alliance for Genomic and Health (GA4GH) published an open access benchmarking tool to assess germline small variant calls in human genomes [Krusche et al., 2019]. More recently, BEELINE, a uniform interface to evaluate Gene Regulatory Network inference from single-cell data, was published and made freely accessible in the form of a docker image [Pratapa et al., 2020].
> In this project, we built on a previous HADACA (Health Data Challenge consortium) benchmarking study [HADACA consortium et al., 2020] to develop a standardized benchmark framework for accurately evaluating quantification of tumor intra-heterogeneity from a multi-omic dataset.

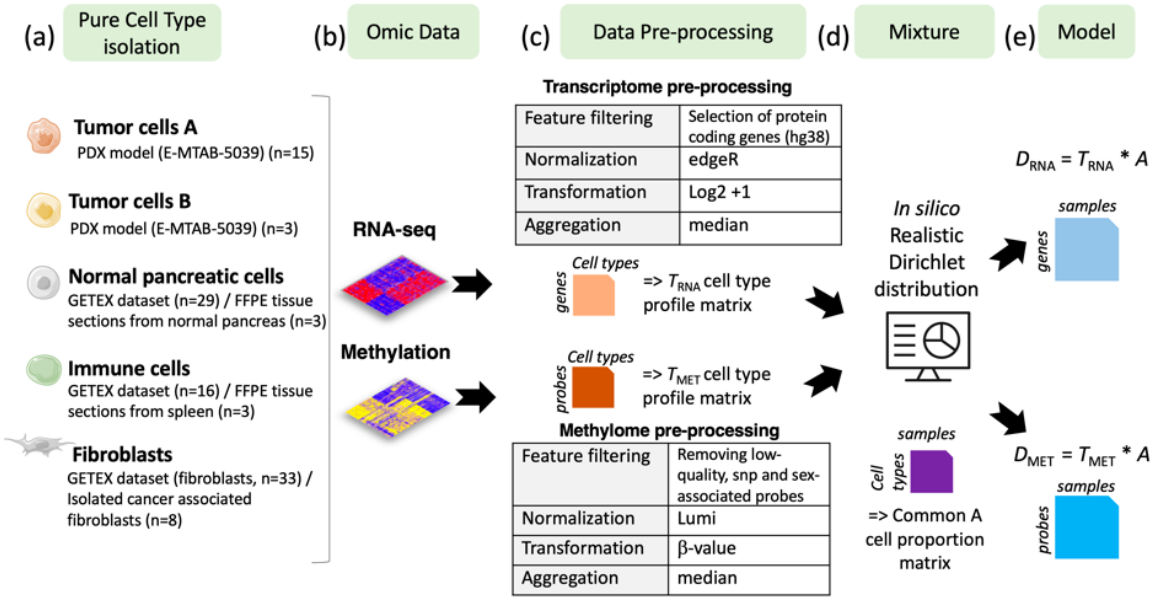

Figure 5.3: **Benchmark dataset construction**
a 5 different cell populations present in pancreatic tumors were considered. **b** Raw transcriptome and methylome profiles of these different cell populations were extracted from various sources (PDX model, tissues or isolated cells). **c** Raw cell type profile matrices were preprocessed together (Feature filtering, normalization, signal transformation, sample aggregation) to avoid any batch effect. After pre-processing, transcriptomic data are constituted of log2-transformed expression counts on 21,566 genes and methylome data of beta-values on 772,316 EPIC array CpG sites. **d** In silico Dirichlet distributions have been used based on realistic proportions defined by the anatomopathologist expertise (J. Cros). **e** Paired methylome and transcriptome of in silico mixtures from pancreatic tumors were obtained by considering D = T × A, with T the cell-type profiles (matrix of size M × K, with M the number of features and K = 5 the number of cell types) and A the cell-type proportion per patient (matrix of size K × N, with N = 30 the number of samples) common between both omics. One training set (DMET and DRNA) is accessible to the users (obtained by one realization of A). The algorithm are compared on 10 test sets (obtained from 10 other realizations of A) that are hidden on the platform. Figure from [Decamps et al., 2021].

> First, we built in silico 10 paired methylome and transcriptome benchmark datasets, using pancreatic cancer (PDAC, pancreatic adenocarcinoma) as a case study (Figure 5.3). These benchmark datasets were made realistic [...] can be used as 'truth' to evaluate computational methods quantifying tumor heterogeneity. Second, we defined Mean Absolute Error (MAE) on estimated cell-type proportions and computational time as standard performance metrics. Third, we embedded the benchmark dataset and the scoring algorithm into a web platform called DECONbench. This web platform enables continuous and crowdsourced benchmarking, by asking participants to submit source code of their algorithm [...]. Fourth, we implemented on the platform baseline methods based on some previously published deconvolution algorithms and tools.

> DECONbench is an open resource to evaluate novel computational methods in an unbiased way. It provides a private general report on the overall performances of the method submitted by any participant and offers the possibility to share all source code of the contributing methods, as well as performance evaluation on a public leaderboard [...]. This framework supports both crowdsourcing benchmarking (collaborative and competitive assessment of the methods) and continuous benchmarking (possibility to continuously integrate novel methods), two features that should contribute to the widespread community adoption of benchmarking good practices [Mangul et al., 2019, Ellrott et al., 2019]. To conclude, DECONbench is an open online benchmark framework including gold standard multi-omic benchmarking datasets, state-of-the-art baseline computational methods and it enables the submission of new methods for evaluation.

5.2.3 Towards a user-friendly online tool for clinicians.

COMETH TRAINING COURSE 1st EDITION (February 15-16th, 2021)

- **Programme** Course Format: 2-day online sessions with general lectures and practicals
  - introduction to cancer heterogeneity & to computational methods
  - interpretation and visualization of the results

Clinicians need training in statistical tools to better understand the role of big data in improving healthcare. They often lack the skills and expertise to properly choose which method to use and apply to their clinical data sets. Although it corresponds to critical unmet clinical needs, knowledge transfer between data science research and clinics has so far been ineffective, due to the lack of appropriate infrastructure and dedicated programs. We developed a training course, as well as user-friendly tools to guide clinicians in their decisions. This work in on going and should lead to a scientific publication:

- To evaluate end-to-end pipelines of deconvolution, a common benchmarking strategy uses references dataset that provide typical Gene Expression profile for some known cell types. We investigated if (and to what extent) guidelines or best practices obtained with in-silico benchmarking dataset are applicable to real-life RNA-Seq samples. All pipelines tested in the benchmark were run using gedepir, an R package we developed that simplifies the use of deconvolution tools within a complete transcriptomics analysis pipeline. It facilitates the definition of an end-to-end analysis pipeline with a set of basic functions that are connected through the pipes syntax used in magrittr, tidyr or dplyr R packages.
- We also implemented a set of useful pipelines and options of gedepir into a web-based interface named decomics. decomics is written in R-Shiny, accessible on the IFB cloud, and provides a complete access to RNAseq count analysis pipeline: preprocessing, deconvolution and results analysis such as pathway enrichment or cell type identification.

5.3 Towards a continuous benchmark

By organizing these different events, I have gained experience in data challenge and benchmark organisation, both from a scientific perspective (addressing the question of tumor heterogeneity through dedicated data challenges) and from a technical perspective (the use and the development of a digital platform). Indeed, over the past few years, in collaboration with I. Guyon (INRIA) and S. Escalera (U. Barcelona), I have contributed to developing a publicly accessible open-source benchmarking platform (Codabench) designed to support long-term benchmarks of computational methods. Codabench includes (i) a benchmarking suite containing the benchmark datasets and the methods, and (ii) a public platform allowing the creation of new competitions (data challenges) and benchmarks. The platform can be regularly updated by the scientific community, as participants can submit new computational methods and new benchmark datasets directly on the platform. The Codabench platform provides a unique single digital environment to combine available datasets and computational methods. The Codabench platform, described in our paper [Xu et al., 2022], is presented section 5.3.1.

5.3.1 Codabench, a novel benchmarking platform

> The methodology of unbiased algorithm evaluation is crucial for machine learning, and has recently received renewed attention in all data science scientific communities. Often, researchers have difficulties understanding which dataset to choose for a fair evaluation, with which metrics, under which software/hardware configurations, and on which platforms. The concept of benchmark itself is not well standardized and includes many settings. For instance, the following may be referred to as a benchmark: a set of datasets; a set of artificial tasks; a set of algorithms; one or several dataset(s) coupled with reference baseline algorithms; a package for fast prototyping algorithms for a specific task; a hub for compilation of related algorithm implementations. In addition, many benchmarks often integrate new progresses by manual verification instead of automatic submission and execution, which delays the benchmark update and requires extra human efforts. Typical examples of existing frameworks addressing such needs are inventoried in Figure 5.4, including competition platforms, repository hubs and domain specific benchmarks. Firstly, competition platforms focus on the participants and provide limited support for organizing general tasks. Famous platforms like Kaggle[a], Tianchi[b], CodaLab[c] organize many data science challenges attracting a large number of participants. However, the platform providers retain some control: the organizers do not have full flexibility and control over their competitions.
>
> [a] https://www.kaggle.com/
> [b] https://tianchi.aliyun.com/
> [c] https://codalab.lisn.upsaclay.fr/

Platform	Flexibility			Easy to use				Reproducibility
	Bundle	Result/code submit	Dataset submit	Easy creation	Open source/free	API access	Compute queue	
Kaggle	✗	✓	✗	✓	✗	✓	✓	✓
Tianchi	✗	✓	✗	✓	✗	✗	✓	✓
CodaLab	✓	✓	✗	✓	✓	✗	✓	✓
UCI	✗	✗	✓	✗	✓	✗	✗	✓
OpenML	✗	✓	✓	✓	✓	✓	✗	✓
PapersWithCode	✗	✓	✗	✓	✓	✗	✗	✓
DAWNBench	✗	✓	✗	✗	✓	✗	✗	✓
Codabench	✓	✓	✓	✓	✓	✓	✓	✓

Figure 5.4: **Comparison of various reproducible science platforms.**
'Bundle' means whether a wrap up is provided for a benchmark such that we could reuse or share. 'Result/Code/Dataset' submit means whether different submissions are supported to enable flexible tasks. 'Compute queue' means, where public or private computation resources could be provided or linked for convenient deployment. Figure from [Xu et al., 2022].

“Repository hubs such as UCI repository[a], OpenML [Vanschoren et al., 2014] and PapersWithCode[b], also play an important role for benchmarks and research. They collect large amount of datasets, methods, and results from academic papers, but reproducibility by running code in given containers (or similar ways) is not guaranteed. Besides the above-mentioned platforms, many domain specific benchmarks exist, e.g. DAWNBench [Coleman et al., 2019], KITTI Benchmark Suite [Geiger et al., 2012]. These benchmarks usually focus on a couple of closely related tasks, but are not designed to host general benchmarks. In addition, they require repetitive efforts to develop and maintain, which is not always affordable by data science teams. Thus, to facilitate benchmarking, we need a platform to allow users to flexibly and easily create benchmarks with custom evaluation protocols and custom data formats, and execution in a controlled reproducible environment, which is totally free and open-sourced.

To answer these unmet needs, we developed codabench, a meta-benchmark platform (Figure 5.5). It is designed to support general purpose benchmarks and to facilitate the organization and usage of benchmarks. codabench takes into account three types of contributors: benchmark participants, benchmark organizers and platform developers. Benchmark participants submit to different benchmarks, which are prepared and owned by different benchmark organizers. Reproducibility is required at this stage for fair benchmarking. Platform developers contribute different features to codabench to support diverse benchmarks instead of one specific benchmark, i.e. codabench is at the meta level of benchmarks. Flexibility and easiness to organize and use benchmarks are thus required at this stage.

[a] https://archive.ics.uci.edu/ml
[b] https://paperswithcode.com/

”

> Codabench realizes these features by implementing an ingestion/scoring programming paradigm, supporting multiple benchmark creation methods and API access, and using Docker to guarantee reproducibility. In , benchmarks are implemented by benchmark `bundles` which contain one or several tasks. The concept of a `task` is newly introduced, which is the minimal unit for composing a benchmark (`bundle`). A `task` consists of an "ingestion module" (including an ingestion program and input data), a "scoring module" (including a scoring program and reference data, invisible to the participant's submission), a baseline solution with sample data, and meta-data information if needed. Tasks in may be programmed in any programming language in any custom way, which are run in a docker specified by organizers.

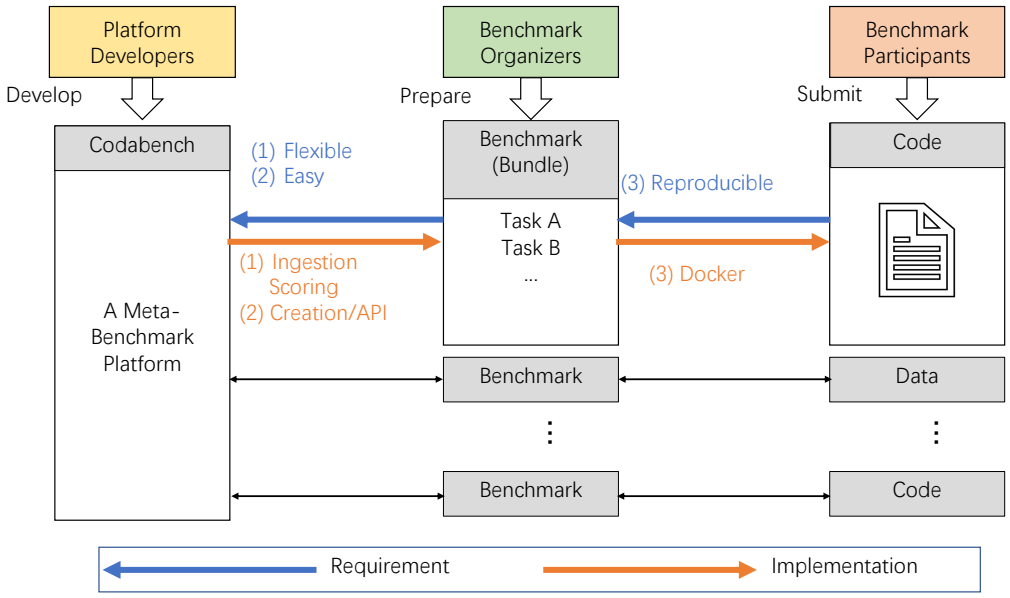

Figure 5.5: **Overview of Codabench**
A meta-benchmark platform has three types of contributors: platform developers (in yellow), benchmark organizers (in green), and benchmark participants (in red). Codabench is at the meta level to support diverse benchmarks. Each benchmark is implemented by a benchmark bundle that contains one or more tasks. Figure from [Xu et al., 2022].

5.3.2 The next international challenge & benchmark – *prospects*

Now that this digital support is running, an ambitious benchmark addressing the question of PDAC tumor heterogeneity remains to be organized.

We will first organize an international online competition (similar to a DREAM challenge). This data challenge will be built on the data-to-modeler model [Guinney and Saez-Rodriguez, 2018]. The aim of the competition will be to accurately quantify tumor heterogeneity and to stratify patients (classification task). First, we will take advantage of our unique expertise to create these

meaningful benchmark datasets: (i) ground truth will include *in silico* simulations, *in vivo* measurements of the heterogeneity by immunohistostaining and clinical observations on patients, (ii) for each sample, RNA and DNAm data will be provided. We will gather an expert steering committee to ensure that the issue raised by the competition and the design of benchmarking datasets correspond to the needs of the community. If required, we will generate complementary data to optimize the relevance of the benchmark datasets. We already have the committment of I. Guyon, S. Escalera, A. Baudot, Y. Blum and J. Cros to participate in such a committee. The performance of participants' algorithms will be evaluated using (i) Mean Absolute Error on estimated heterogeneity, and (ii) F1-score on patients' classification. As usage of identifiable patient data is legally protected, we will adhere to European and national regulations at the same time as adhering to the FAIR principles for data management and stewardship [Wilkinson et al., 2016]. We will work in sync with the global efforts for technical standardization and responsible data sharing, e.g. Global Alliance for Genomics and Health [6]. The data storage and sharing concept will comply with national and international legal requirements. We will connect with high-profile journal editors ahead of the challenge organization to discuss the possibility of publishing the competition outcome, as participatory competitions generally result in scientific publications (see [Creason et al., 2021, Marbach et al., 2012, Eicher et al., 2019, Marot et al., 2021, Le et al., 2019]), which benefit the entire community. The competition will be open to anyone, though targeted to people with a certain degree of expertise in the field. We will offer authorship to competing teams, along with participation in manuscript writing. This is a strong incentive that will provide international visibility and recognition to participants. We will organize an international conference gathering all competition participants, organizers and the related scientific community. Best performing teams will be offered the ability to present their solution during this international scientific conference. The conference will act as a facilitator of collaborations between health data scientists and health professionals, while contributing to knowledge transfer between researchers and clinicians, thus fostering innovation.

Then, to ensure that competition data can then be re-used by research scientists as gold standard for computational methods that will be developed in the future, we will turn this competition into a continuous benchmark (thanks to the specific features of the Codabench platform [Xu et al., 2022]). The creation of a continuous benchmark will help both the researchers developing methods and clinicians interested in using these methods, as it is designed to become the reference to reproducibly evaluate dedicated algorithms. All members of the engaged community will get access, with minimum effort, to the catalogue of datasets and computational methods to evaluate their work. As we believe that academic research should massively rely on the open science framework, we will encourage participants to submit their code under an open source license.

[6] https://www.ga4gh.org

General conclusion

In this concluding chapter, I will present the scientific perspectives of my research work and I will address some personal considerations about the research environment and my beginnings as a young researcher.

6.1 Scientific perspectives

6.1.1 Scientific contributions to cancer biology

My group expertise lies in the use of advanced statistical methods and bioinformatic processing of multimodal high-dimension data to investigate precise fundamental oncology questions, with potential applications in translational research. The main objective of our research projects is to provide a better understanding of tumor heterogeneity, allowing us to address new questions related to oncogenic processes.

Current research on personalized oncology is expanding, in particular thanks to the reduction of sequencing costs allowing the generation of multi-omics profiles at the scale of individual tumors and individual cells. However, efforts to understand the extensive heterogeneity of tumors have heretofore been largely limited to cancer cells due to the lack of methods to study these cells together with their environment. Through a multidisciplinary approach, we try to overcome this obstacle and expand our understanding of cancers as complex ecosystems, taking into account all cell states and functions present within a tumor.

Through a multidisciplinary approach, we develop computational methods to quantify inter- and intra-tumor heterogeneity, thus contributing to our knowledge of cancers as complex ecosystems, taking into account all genetic and non-genetic deregulations, as well as environmental effects. The analysis of the evolution of cancer in the light of tumor heterogeneity should pave the way for a better understanding of tumor initiation and propagation, and will enrich recent models of cancer development, with potential rapid applications in clinical management, in particular with regard to immunotherapies. In addition, the research project I will initiate on causality could identify for the first time molecular mediators of tumor heterogeneity, which represent exciting prospects for building mechanistic models of carcinogenesis integrating the complex tumor ecosystem. In this regard, I have an ongoing collaboration with A. Bellasta (Institut Curie), who designs mathematical methods for personalized combinations of anticancer drugs and timing (systems pharmacology). Finally, we wish to provide a unique objective resource for scientists and healthcare professionals wishing to quantify tumor heterogeneity using a multi-omics dataset generated from surgical patient samples. The benchmark platform we want to build is designed to

become indispensable for those who wish to develop and/or publish new methods to quantify tumor heterogeneity. We would contribute to define a new model of knowledge transfer between research and clinic and meet a growing demand for data training from healthcare professionals.

Our main research topic is cancer heterogeneity. However, the methods we have developed may have several applications in other fields (eg, multimodal analysis in computational biology, quantitative trait evolution, large-scale mediation analysis in epidemiology). Moreover, if the question of temporal heterogeneity is not directly covered by this thesis, recent findings indicate that it would be a very interesting phenomenon to investigate [Quek et al., 2018]. This will likely be the subject of a future grant proposal.

6.1.2 Multidisciplinary approach for a better healthcare

In order to significantly improve early diagnosis and patient care, new multidisciplinary approaches have emerged. Indeed, the field of oncology research faces significant computational challenges that can only be addressed through a multidisciplinary approach, with expertise in mathematical data modeling closely linked to crucial oncological questions. I believe that our concept-driven (as opposed to data-driven) approach meets these requirements. It benefits from a solid expertise in different fields, as it is based on a strong connection between clinical settings and health data sciences, and it is designed to meet clinical needs, especially for personalized medicine. However, besides the biological relevance of our results, their usability in clinical routine is far from guaranteed. That is why our close collaboration with J. Cros acts a safeguard to ensure that the methods developed are consistent with clinical needs. His group has the resources to further investigate the diagnostic and therapeutic potential of our findings and to develop corresponding tools applicable in clinical routine.

The close interconnection between scientists who develop innovative computational methods and clinicians who need to use these methods to improve healthcare is essential for effecting a change in the clinical routine. For instance, transitioning from the use of single-parameter biomarkers (such as immunohistochemistry) to original multi-omics signatures (derived from the quantification of tumor functional heterogeneity) could completely change paradigms in PDAC care and prognosis. This requires efficient computational tools capable of fully characterizing tumors considering their heterogeneity and which will work on small biopsy samples with possible heavy contamination. Thus, our work should enhance the interpretability of each patient's molecular data, which will provide significant opportunities to improve disease diagnosis and tailor treatment accordingly.

6.1.3 Next challenges in computational oncology

A major problem facing oncology research is the lack of reproducibility of published results. Indeed, many results found in a given study are not reproducible in another independent cohort. For example, a recent pan-cancer analysis showed that genetic studies are often not reproducible, with only 40 out of 440 genes associated with cancer survival being confirmed in more than one study [Kaubryte and Lai, 2022]. The recent "Reproducibility project: Cancer Biology"

investigated replicability in preclinical research in cancer biology. This project encountered recurring obstacles preventing contributing researchers from replicating the majority of the ~200 experiments they had selected from high-impact articles [Errington et al., 2021]. Although this review focuses primarily on wet-lab experiments, the author's observation that "none of the experiments were described in sufficient detail in the original paper to allow them to be repeated [...] experiments" can easily be generalized to dry-lab experiments. The lack of reproducibility and the absence of good coding practices is a major concern when it come to interpreting and generalizing the results of a study, especially regarding the transfer from computationally-based observations to daily clinical practice. A new model for the development and integration of digital tools remains to be defined [Wiens et al., 2019], the guidelines of which are currently being discussed [Ballester and Carmona, 2021] (for example: code reproducibility, sharing and transparency [Haibe-Kains et al., 2020, Kakarmath et al., 2020]).

The extensive use of high-throughput sequencing has opened up new avenues in diagnostics and precision medicine. Patient stratifications based on omics analyzes have thus led to the development of cancer therapies that are more specific and more adapted to the patient [Mardis, 2021]. New methodological approaches are then necessary to analyze these new datasets, characterized by their large size and complexity. If we talk a lot about the 'Artificial Intelligence in Health' revolution, recent studies remain cautious about the use of machine learning in the discovery of biomarkers and diagnostic factors. Thus, a recent review analyzed 247 articles aimed at identifying prognostic factors related to cancer survival. Of these, only 6 approaches used machine learning approaches while the vast majority were based on traditional [Kaubryte and Lai, 2022] regression. In parallel, the contribution of decision support systems aimed at optimizing the choice of treatment (which therapy will be the most effective for a given patient?) are also discussed, because in the majority of cases there is no sufficiently well-annotated and good-quality multi-institution data (molecular and clinical). High-quality datasets are thus critical for the development of high-performance models [Zhang et al., 2022]. If machine learning approaches have been relatively successful in transitioning from computer vision to image-based clinical diagnosis, computational based approaches are slowly deployed in healthcare. In the years to come, one of the main challenges will be to increase the quality of available public data (sufficiently annotated, large-scale, and correctly processed molecular and clinical data) in order to be able to build robust biostatistical models, that are stable against data shifts [Zhang et al., 2022], to contribute to a better understanding of the biological processes involved.

6.2 Personal considerations on my early career

6.2.1 Pitfalls of computational biology and bioinformatics

I often have trouble defining myself as a scientist. Depending on whether you consider my initial training (wet-lab), the journals in which I mainly publish (fundamental biology), the CNRS section that recruited me (mathematics, information and physical models for the life sciences) or the CNRS institute to which I belong (computer science), I could be a geneticist, a computational geneticist, a computational biologist, a bioinformatician... Finally, I'm not sure that finding the exact definition of my research field is so important. But what I know for sure is that I work

in an interdisciplinary field, and that I develop and apply computational methods to analyze big biological data, trying to solve fundamental questions in biology.

The development of computational methods has some pitfalls, especially when it comes to reproducible science. As a young PI, with high turnover in my group, I found it particularly difficult to achieve high standards in terms of code development, data storage, repeatable workflows, and sustainability. PhD and postdoc are the main developers of the methods we use in the group, and I still haven't found an effective way to maintain their work once they leave the team, which is very frustrating. Moreover, as the software maintenance is not really recognized in the community, it is complicated to devote time to this underestimated, but fundamental task. Ultimately, this raises the question of the usefulness of our work. Is it worth developing an umpteenth method, which will not be maintained, and eventually used by no one?

I also found it particularly difficult to follow all the computational methods that are constantly being developed and published. As with wet-lab experimentation, hands-on experience in a dry lab is essential to gain good technical skills. However, unlike wet lab experiments, these skills are quickly devalued because standard methods change very fast. This is due: (i) in part to the fact that new technologies evolve very dynamically, constantly generating new types of data, with associated computational problems, and (ii) to the fact that researchers in the field generally try to develop a new method, instead of trying to use what has already been done. As a result, an accumulation of new methods flourishes, that are published in numerous journals and articles. Thus, I find it extremely challenging to keep up to date with the literature, especially if we take into account that evaluations of methods are generally biased, and that certain methods are hegemonic only because everyone uses them, regardless of their intrinsic qualities. In this complex landscape, my wish is to promote contributions to existing code, allowing us to adapt existing solutions to specific questions related to our research theme.

Finally, I think data-driven science is only relevant if it brings biologically meaning-full results. For this, it is necessary to have good collaborations for experimental validation, with partners who are experts in the biological issues at stake.

6.2.2 Navigating in a multidisciplinary environment

By definition, computational biology and bioinformatics are rooted in a multidisciplinary field, because they call on skills in biology, mathematics and computer science. Naturally, the following questions then arise: how to exchange, communicate, and build with people of different origins and expertise? It can be difficult to share knowledge with people from other scientific fields due to the language barrier (avoid specific jargon) and the culture specific to the discipline. For example, the culture in biology is different from that in bioinformatics or biostatistics, whether at the level of the hierarchical organization within the teams, the way of asking scientific questions or the strategies for promotion and publications.

Being at the interface of several disciplines teaches humility, because it is impossible to be at the forefront of all areas simultaneously. You have to accept to sometimes feel inferior or incompetent, to ask 'dumb questions' and if necessary to overcome a feeling of imposture. It is a position that is not always easy to defend at the institutional level. Indeed, with regard to funding requests,

many institutions still do not have interdisciplinary panels, often leaving project leaders at the interface of several disciplinary fields helpless. In addition, the evaluation of the research work is still based on the number of articles published, the impact factors and the position of the authors. During multidisciplinary collaborations, it is not uncommon to find 'computational' authors in intermediate positions, which can be detrimental to them later on. If it is easier to overcome these considerations when obtaining a permanent position, these problems remain pregnant for the postdocs and doctoral students that we supervise, and for whom we are responsible.

Remarkably, more and more community approaches are emerging to foster multidisciplinary collaborations and remove scientific obstacles (see for review [Lee et al., 2017] and the example of the ClViC: clinical interpretation of cancer variants [Krysiak et al., 2022]). This seems to me to be a very powerful way to advance science, particularly in cancer biology, given the fact that no single team has all the tools necessary to solve current problems in oncology.

6.2.3 Slow science : doing less but better?

In 2010 the SLOW SCIENCE MANIFESTO[1] was published, advocating more time to think, read, misunderstand, learn and fail. Currently, we are strongly encouraged by our institutions to produce science, that is to say, to constantly respond to calls for grants and to publish as much as possible. In parallel, these scientific productions are perpetually evaluated by peers, expert committees, scientific councils, and other bodies. Unfortunately, that leaves us little time to do good science. This is particularly true for young researchers, who are very inexperienced in this administrative and bureaucratic environment, environment quite different from the scientific tasks they used to carry out as postdocs. I myself felt this headlong rush, always looking for new funding opportunities. I noted that, paradoxically, what is recognized and valued by our institutions does not guarantee the quality of research work. In my case, I found the writing of reports, the management of credits and human resources, and the recruitment of non-permanent staff, extremely time-consuming. These multiple tasks took me away from my research questions and likely harmed the quality of the student supervision I carried out.

Particularly in multidisciplinary fields, we need time to discuss, exchange and understand concepts that are initially foreign to us. It is often necessary to step aside to have a new look at our work, to leave room for the imagination, and sometimes to be unproductive. Unfortunately, the fact that multidisciplinary science is not well suited to standard evaluation (i.e. publication rate and impact factor) tends to push young, unestablished scientists to do more and faster to gain recognition and legitimacy.

I often felt isolated on my research area. I started working on cancer heterogeneity only a few years ago, and so I can no longer rely on the networks and scientific communities that I had previously built during my thesis and postdoc. Unfortunately, I recently found it difficult to go to conferences in my field (lack of time to devote to it, and problems amplified during the COVID-19 pandemic period). Moreover, the scientific theme covered locally by my host laboratory are quite far from my research subject. As a result, I feel like I lack scientific interaction, stimulation and criticism. In this context, I ask myself the following questions: how to stay up to date, follow the

[1] http://slow-science.org/

news in the field and build a professional network? What to do when you are not an established researcher, in a renowned research center, with a good international reputation?

Over the past five years, I also had to adjust my work pace and style to my changing personal situation. As my family grew (kids arrived in 2017 and 2020), my sleep deprivation increased and my scientific productivity decreased. I am always faced with organizational challenges: how do I adapt to the vagaries of scientific life (deadlines, rush, intensity) with a very constrained domestic life in terms of rhythm and schedule? How can I remain sufficiently available and focused for my research team, whose management requires significant time?

Recent initiatives attempt to address the problem of research evaluation and the quest for productivity, for example the San Francisco Declaration on Research Evaluation[2] (reinventing academic career assessment), or the peer-review community initiative[3]. Despite everything, I still have the feeling that the university environment favors quantity over quality, and that my professional career has followed this injunction. I did not succeed in solving the following dilemma: how to produce quality science while meeting the expectations of my institutions (laboratory, university, CNRS). Over the next five years, I would like to engage in slower, more diverse, and more collaborative science. Promote a healthier scientific environment. Keep a small research group, being present for a real supervision of my trainees, in order to offer them enriching and pleasant moments. Carefully accept administrative responsibilities to avoid overwhelming myself.

Finally, the COVID-19 epidemic has clearly shown how disconnected science is today from society. We have seen a huge acceleration in scientific production, with the publication of more than 100,000 manuscripts on COVID-19 in 2020 [Leite and Diele-Viegas, 2021], and at the same time, an inability to communicate and share the scientific debate with the civil society. It reminded me that as scientists we have an important responsibility to society. I would like to take more time to think on how, given my position as a scientist, I can act as a citizen to improve our society and contribute to a better knowledge sharing.

[2] DORA: https://sfdora. org/read/
[3] https://peercommunityin.org/

Curriculum Vitae

In this chapter, I will define everything

7.1 Training and Appointments

2018-present **CNRS researcher** ; Permanent position at the Univ. Grenoble-Apes (UGA). France.
TIMC laboratory : Translational research, Innovation, Medicine & Complexity
Maternity leave ; from Jan. 2020 to Sept. 2020
2018 **Independent Young Researcher Chair** ; TIMC lab. Grenoble. France
2017 **Postdoctoral scholar**; in collab. w/ Daniel Jost, TIMC lab. Grenoble. France
Maternity leave ; from Aug. 2017 to Jan. 2018
2016 **University Diploma in Statistics**; Univ. Strasbourg. France
2013-2016 **Postdoctoral scholar**; in collab. w/ Gael Yvert, LBMC lab, ENS Lyon. France
2008-2012 **Ph.D in Genetics**; advisor: Jean-Louis Bessereau, IBENS lab, ENS Paris. France
2008 **M.S. in Genetics**; (rank: 1st). Univ. Denis Diderot (Paris 7). France
2007 **Research internship**; advisor: Jessica Treisman, Skirball Insitute, NYU. USA

7.2 Fellowship and Awards

2017 3-year independent research chair in data science; Univ. Grenoble Alpes
2012 1-year Ph.D fellowship; ARC (French Cancer Research Agency)
2008 3-year full Ph.D. fellowship; French ministry of research

7.3 Publications, softwares and conferences

7.3.1 Research articles

In my field, the main junior author is listed first, the main senior author last. : denotes equal contribution (co-first or co-last authorships),

✉ : indicates when I am corresponding author, students/postdocs that I supervised are underlined, Red items emphasizes for work as senior author.

⚙ : computational work, ⚗ : experimental work

Tumor heterogeneity ⚙

I develop supervised and unsupervised approaches to study intra (10, 11) and inter-tumor (9) heterogeneity.

11 Saillard, C., Delecourt, F., ... , **Richard M.**, Kermezli Y., ... , Nicolle, R., Cros, J (2021) *PACpAInt: a deep learning approach to identify molecular subtypes of pancreatic adenocarcinoma on histology slides*; submitted.

10 Decamps, C., Privé, F., Bacher, R., Jost, D., Waguet, A., HADACA consortium, Houseman EA., Lurie E., Lutsik P., Milosavljevic A., Scherer M. Blum M., & **Richard, M.,** ✉ (2021) *Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation*

deconvolution software.; BMC Bioinformatics 21, 1–15.

9 **Richard, M.,**, Decamps, C., Chuffart, F., Brambilla, E., Rousseaux, S., Khochbin, S., & Jost, D (2020) *PenDA, a rank-based method for Personalized Differential Analysis: application to lung cancer*; PLoS Comput Biol 16, e1007869.

Reproducible science,

8 Xu, Z., Escalera, S., Pavao A., **Richard, M.**, Tu, W., Yao Q., Zhao, H., Guyon, I (2022) *Codabench: Flexible, Easy-to-Use and reproducible meta-benchmark platform*; Patterns 3, 100543

7 Decamps, C., Arnaud, A., Petitprez, F., Ayadi, M., Baures, A., Armenoult, L., HADACA consortium, Nicolle, R., Tomasini, R., de Rynies, A., Cros, J., Blum, Y., & **Richard, M.** (2021) *DECONbench: a benchmarking platform dedicated to deconvolution methods for tumor heterogeneity quantification*; BMC Bioinformatics 21, 1–15.

Quantitative genetics

I made a major contribution to the first genomic characterization of the fitness (ability to reproduce) of a micro-organism in a dynamic environment (6). I conducted an in-depth study of a regulatory network in response to an external stimulus (functional analysis of genetic variants and quantitative network modeling) (5). I participated in the development of an innovative statistical method to analyze the probabilistic behavior of a cell (4).

6. Salignon, J. , **Richard, M,**, Fulcrand, E., Duplus-Bottin, H., & Yvert, G. (2018) *Genomics of cellular proliferation in periodic environmental fluctuations*; Mol Syst Biol. 2018 Mar 5;14(3):e7823
5. **Richard, M.**, Chuffart, F, Duplus-Bottin, H., Pouyet, F., Spichty, M., Fulcrand, E., Entrevan, M., Barthelaix, A., Springer, M., Jost, D., & Yvert, G. (2018) *Assigning function to natural allelic variation via dynamic modeling of gene network induction*; Molecular Systems Biology, 14, e7803.
4. Chuffart, F., **Richard, M.**, Jost, D., Burny, C., Duplus-Bottin, H., Ohya, Y., & Yvert, G. (2016) *Exploiting Single-Cell Quantitative Data to Map Genetic Variants Having Probabilistic Effects*; PLoS Genetics, 12(8), e1006213.

Genetics and cell biology

I used model organisms to discover and characterize for the first time new genes involved in cell differentiation and developmental processes. I have been interested in the intracellular mechanisms that regulate the maturation, trafficking and function of nicotinic acetylcholine receptors in the Nematode (2, 3) and photoreceptors in Drosophila (1).

3. D'Alessandro, M., **Richard, M.,**, Stigloher, C., Gache, V., Boulin, T., Richmond, JE., & Bessereau, J.-L. (2018) *CRELD1 is an evolutionarily-conserved maturational enhancer of ionotropic acetylcholine receptors*; Elife. 2018 Nov 7;7. pii: e39649.
2. **Richard, M.**, Boulin, T., Robert, V. J. P., Richmond, J. E., & Bessereau, J.-L. (2013) *Biosynthesis of ionotropic acetylcholine receptors requires the evolutionarily conserved ER membrane complex*; Proc Natl Acad Sci U S A. 2013 Mar 12;110(11):E1055-63.
1. Legent, K., Steinhauer, J., **Richard, M.**, & Treisman, J. E. (2012) *A screen for X linked mutations affecting Drosophila photoreceptor differentiation identifies Casein kinase 1α as an essential negative regulator of wingless signaling*; Genetics.

7.3.2 Review articles in refereed journals and books

2 *Practical issues: Incentives, community engagement and costs* (Competition and Benchmark Springer Book, Editors: A. Pavao, E. Viegas & I. Guyon), in submission. (invited contribution) (2021) Blum Y., Guinney, J., Pavao A., Stolovitsky, G., **Richard, M.**

1. *How does evolution tune biological noise?*; Front. Genet. 5, 1011 (2014). **Richard, M.,** & Yvert, G.

7.3.3 Open-access softwares

4 RITMIC, an R package to identify genetic deregulation correlated with tumor heterogeneity.
3 GEDIPIR (R package) and DECOMICS (ShinyApp) dedicated to unsupervised deconvolution of bulk gene expression.
2 MEDEPIR, a R package to perform methylation deconvolution of bulk samples.
1 PENDA an R Open-access package to perform personalized differential expression.

7.3.4 Speaking engagements

Invited conference and seminar presentations

5 Journée thématique RIS, virtual conference, France (2021)
4 Workshop Modeling ctDNA dynamics for detecting targeted therapy resistance, Nancy, France (2021)
3 Epigenetic and Mediation data challenge, Aussois, France (2017)
2. Computational Biology and Mathematics seminar, Grenoble, France (2015)
1. LBMC seminar, Lyon, France (2012)

Contributed conference presentations

17 Cancer Genomics, Virtual, Germany (2021)
16 CSMB, Bordeaux, France (2021)
15 EMBL symposium: Multiomics to Mechanisms: Challenges in Data Integration, Virtual, Germany (2021)
14 RECOMB, Virtual conference, Italy (2021)
13 R meetings, Grenoble, France (2019)
12 Epigenetics and Cancer, Heidelberg, Germany (2017)
11. COMPSYSBIO, Aussois , France (2017)
10. JOBIM, Lyon, France (2016)
9. Design optimization and control in system and synthetic biology, Paris, France (2015)
8. JOBIM, Lyon, France (2015)
7. R meeting, Lyon, France (2015)
6. Experimental Approaches to Evolution and Ecology using Yeast and other Model Systems, Heidelberg, Germany (2014)
5. Single cells genomics, Stockholm, Sweden (2014)
4. Young Researcher in Life Sciences, Paris, France (2012)
3. 18th International Worm Meeting, LA, USA (2011)
2. Neuronal development, synaptic function and behaviour in C. elegans, Madison, USA (2010)
1. 17th International Worm Meeting, LA, USA (2019)

7.4 Fundings

Grants just acquired		
Project title and role of the PI	Funding source, Amount, Period	Topic
CauseHet Causes and consequences of tumor heterogeneity **PI**	Agence Nationale de la Recherche (Young researcher call) **300kE** admin. amount **300kE** 2022-2026	Prospects : **Chapter 3.3.2** Estimation of tumor functional heterogeneity at the single tumor level and **chapter 4.1** Relationship between tumor functional heterogeneity and cancer evolution
THEMA Mediation analysis of tumor heterogeneity **PI**	IRGA IDEX UGA **160kE** admin. amount **160kE** 2022-2025	Prospects : **chapter 4.2** A causal link between heterogeneity, environment and outcome

On-going grants		
Project title and role of the PI	Funding source, Amount, Period	Topic
ACACIA AI on multi-omics data to study tumor heterogeneity **PI**, consortium coordinator	ITMO Cancer of Aviesan **315kE** admin. amount **218kE** 2021-2024	Prospects : **Chapter 3.3.1** Multiomic based deconvolution and classification of tumor heterogeneity

Past grants		
Project title and role of the PI	Funding source, Amount, Period	Topic
ARTICAH ARTificial Intelligence for CAncere Heterogeneity **PI**	UGA-IRS **50kE** 2021	We benchmark unsupervised method to estimate tumor heterogeneity from transcriptomic data.
COMETH Benchmarking of COputational METHods **PI**, consortium coordinator	Campus project - EIT Health **630kE** admin. amount **340kE** 2020	We generated a series of *in silico*, *in vitro* and *in vivo* benchmark dataset (mRNA and DNAm) to study tumor heterogeneity of PDAC.
LuCaH LUng Cancer Heterogeneity **PI**	IRS-IDEX U. Grenoble Alpes **11kE** 2019	We studied unsupervised deconvolution of methylomic data.
HADACA Health Data Challenges **PI**, consortium coordinator	Campus project - EIT Health **250kE** admin. amount **127kE** 2019	We organized a first winter school based on a data challenge and started to contribute the Codalab challenge platform.
Epigenetic deregulation in cancer **PI**	PEPS – CNRS INS2I **10kE** 2019	We developed the PenDA method to infer personalized differential analysis.

7.5 Expertise, editorial and scientific activities

Referee for *Nucleic Acid Research*, *BMC bioinformatics*, *Epigenetics*
2020-present **Member of the executive board of ChaLearn**; *link*, international non-profit organization dedicated to data challenges organization, USA

Organization of international scientific conference and data challenges

2021 **COMPSYSBIO winter school**, CNRS, Aussois, France
Health Data challenge at AI4Health winter school, Paris, France
2020 **COmputational METHods in Health winter school**, *founder*, online event
2019 **Health Data challenge at MEDINFO conference**, Lyon, France
2nd edition of Heath Data Challenges, *founder*, Aussois, France
2018 **1st edition of Heath Data Challenges**, *founder*, Aussois, France

7.6 Scientific collaborations

Yuna Blum & Jerome Cros (Inst. Genetics and Development of Rennes & Beaujon Hospital, Paris); tumor heterogeneity in pancreatic cancers (2 papers, 2 in prep, 2 joint grant applications)
Isabelle Guyon & Sergio Escalera (INRIA Paris-Saclay & U. Barcelona); data challenges and CodaLab digital platform (1 paper, paper in rev., 2 joint grant applications)
Carl Herrmann (U. Heidelberg, Health Data Science Unit); unsupervised deconvolution of gene expression (1 paper in prep., 1 joint grant application)
Daniel Jost (Lab. for Biology and Cell Models of Lyon); differential expression (3 papers, 1 in prep.)
Saadi Khochbin & Sophie Rousseaux (Institute for Advanced Biosciences of Grenoble); (epi)genomic deregulations in lung cancers (1 paper, 1 in prep.)

7.7 Supervision and mentoring

3 postdocs (bioinformatics). Y. Kermezli (2020-21), S. Karkar (2021), E. Amblard (2021-22)
1 Ph.D student (bioinformatics) : C. Decamps (2018-21); *co-supervised with D. Jost*
4 Engineers (computer sciences and statistics): C. Burny (2015), R. Bacher (2018), B. Afshinpour (2019), A. Arnaud (2019)
8 Master students (bioinformatics and statistics) : P. Terzian (2017), C. Decamps (2018), A. Waguet (2018), M. Jacobi (2019), F. Quinquis (2021), F. Kon-Sun-Tack (2021), F. Pittion (2022), A. Petrova (2022)

7.8 Administrative responsibilities

2020-present **Elected member of the MSTIC faculty board**; Council of the Mathematic, information and technologies department of the U. Grenoble-Alpes, France
2020-present **Quality of life at work delegate**; Laboratory TIMC-IMAG, Grenoble
2018-2019 **Organizer of the Computational Biology seminar**, *link*, Grenoble, France
2017-2018 **Founder and organizer of the R seminar series**, *link*, Grenoble, France

7.9 Teaching

2017-present **Adjunct lecturer**; Grad and undergrad, Statistics and Genetics, UGA, Grenoble
2013-2014 **Adjunct lecturer**; Undergrad, Molecular genetics, ENS, Lyon
2008-2012 **Teaching assistant**; Undergrad, Genetics and Cell Biology, Univ. Paris 6, Paris

List of Figures

1.1 **Scales of tumor heterogeneity** . . . 10
1.2 **The genetic characteristics of cancers vary between patients, between primary and metastatic tumours in a single patient, and between the individual cells of a tumour.** . . . 10
1.3 **Scales of tumor heterogeneity** . . . 11
1.4 **Intra and inter-tumor heterogeneity in Primary Liver Cancer** . . . 11

2.1 **The PenDA method** . . . 19
2.2 **Comparison with other methods** . . . 21
2.3 **Genetic deregulations efficiently classify cancer histologies** . . . 23
2.4 **Upregulation of 37 genes in adenocarcinoma is as trong predictor of poor prognosis** 24
2.5 **Proportion of deregulated genes in each TCGA cancer type** . . . 27
2.6 **PenDA gene deregulation of each glioblastoma derived cell line** . . . 28

3.1 **Cell-type deconvolution problem, the bioinformatician point-of-view.** . . . 29
3.2 **Performance of the 3 deconvolution methods for different parameter settings.** . . . 32
3.3 **Recommendations and benchmarking pipeline.** . . . 34
3.4 **Identification of high-resolution PDAC cell-types.** . . . 37
3.5 **Single cell dataset integration.** . . . 38
3.6 **Integrative gene markers of 14 PDAC cell-types.** . . . 39
3.7 **Benchmark of deconvolution using the new single-cell references** . . . 40
3.8 **Application to TCGA pancreatic adenocarcinoma.** . . . 41
3.9 **Preliminary results on multiomic integration.** . . . 42
3.10 **An assessment of tumor functional heterogeneity at the single tumor level.** . . . 44
3.11 **A new method to infer tumor functional heterogeneity** . . . 46

4.1 **An analysis of cancer evolution in the light of tumor heterogeneity.** . . . 50
4.2 **Relationship between somatic mutations and tumor heterogeneity.** . . . 51
4.3 A rigorous causal analysis of the effect of tumor functional heterogeneity . . . 53
4.4 **The causal inference path.** . . . 55

5.1 **The incentives for participating in a challenge.** . . . 59
5.2 **The process of engaging a community** . . . 60
5.3 **Benchmark dataset construction** . . . 63
5.4 **Comparison of various reproducible science platforms.** . . . 66

5.5 **Overview of Codabench** 67

Bibliography

[Altboum, 2014] Altboum, Z. (2014). Digital cell quantification identifies global immune cell dynamics during influenza infection. *Molecular Systems Biology*, 10(2):720. Publisher: John Wiley & Sons, Ltd.

[Australian Pancreatic Cancer Genome Initiative et al., 2016] Australian Pancreatic Cancer Genome Initiative, Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A.-M., Gingras, M.-C., Miller, D. K., Christ, A. N., Bruxner, T. J. C., Quinn, M. C., Nourse, C., Murtaugh, L. C., Harliwong, I., Idrisoglu, S., Manning, S., Nourbakhsh, E., Wani, S., Fink, L., Holmes, O., Chin, V., Anderson, M. J., Kazakoff, S., Leonard, C., Newell, F., Waddell, N., Wood, S., Xu, Q., Wilson, P. J., Cloonan, N., Kassahn, K. S., Taylor, D., Quek, K., Robertson, A., Pantano, L., Mincarelli, L., Sanchez, L. N., Evers, L., Wu, J., Pinese, M., Cowley, M. J., Jones, M. D., Colvin, E. K., Nagrial, A. M., Humphrey, E. S., Chantrill, L. A., Mawson, A., Humphris, J., Chou, A., Pajic, M., Scarlett, C. J., Pinho, A. V., Giry-Laterriere, M., Rooman, I., Samra, J. S., Kench, J. G., Lovell, J. A., Merrett, N. D., Toon, C. W., Epari, K., Nguyen, N. Q., Barbour, A., Zeps, N., Moran-Jones, K., Jamieson, N. B., Graham, J. S., Duthie, F., Oien, K., Hair, J., Grützmann, R., Maitra, A., Iacobuzio-Donahue, C. A., Wolfgang, C. L., Morgan, R. A., Lawlor, R. T., Corbo, V., Bassi, C., Rusev, B., Capelli, P., Salvia, R., Tortora, G., Mukhopadhyay, D., Petersen, G. M., Munzy, D. M., Fisher, W. E., Karim, S. A., Eshleman, J. R., Hruban, R. H., Pilarsky, C., Morton, J. P., Sansom, O. J., Scarpa, A., Musgrove, E. A., Bailey, U.-M. H., Hofmann, O., Sutherland, R. L., Wheeler, D. A., Gill, A. J., Gibbs, R. A., Pearson, J. V., Waddell, N., Biankin, A. V., and Grimmond, S. M. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, 531(7592):47–52.

[Avila Cobos et al., 2020] Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P., and De Preter, K. (2020). Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature Communications*, 11(1):5650. Number: 1 Publisher: Nature Publishing Group.

[Bailey et al., 2016] Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A.-M., Gingras, M.-C., Miller, D. K., Christ, A. N., Bruxner, T. J. C., Quinn, M. C., Nourse, C., Murtaugh, L. C., Harliwong, I., Idrisoglu, S., Manning, S., Nourbakhsh, E., Wani, S., Fink, L., Holmes, O., Chin, V., Anderson, M. J., Kazakoff, S., Leonard, C., Newell, F., Waddell, N., Wood, S., Xu, Q., Wilson, P. J., Cloonan, N., Kassahn, K. S., Taylor, D., Quek, K., Robertson, A., Pantano, L., Mincarelli, L., Sanchez, L. N., Evers, L., Wu, J., Pinese, M., Cowley, M. J., Jones, M. D., Colvin, E. K., Nagrial, A. M., Humphrey, E. S., Chantrill, L. A., Mawson, A., Humphris, J., Chou, A., Pajic, M., Scarlett, C. J., Pinho, A. V., Giry-Laterriere, M., Rooman, I., Samra, J. S., Kench, J. G., Lovell, J. A., Merrett, N. D., Toon, C. W., Epari, K., Nguyen, N. Q., Barbour, A., Zeps, N., Moran-Jones, K., Jamieson, N. B., Graham, J. S., Duthie, F., Oien, K., Hair, J., Grützmann, R., Maitra, A., Iacobuzio-Donahue, C. A., Wolfgang, C. L., Morgan, R. A., Lawlor, R. T., Corbo, V., Bassi, C., Rusev, B., Capelli, P., Salvia, R., Tortora, G., Mukhopadhyay, D., Petersen, G. M., Australian Pancreatic Cancer Genome Initiative, Munzy, D. M., Fisher, W. E., Karim, S. A., Eshleman, J. R., Hruban, R. H., Pilarsky, C., Morton, J. P., Sansom, O. J., Scarpa, A., Musgrove, E. A., Bailey, U.-M. H., Hofmann, O., Sutherland, R. L., Wheeler, D. A., Gill, A. J., Gibbs, R. A., Pearson, J. V., Waddell, N., Biankin, A. V., and Grimmond, S. M. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, 531(7592):47–52.

[Ballester and Carmona, 2021] Ballester, P. J. and Carmona, J. (2021). Artificial intelligence for the next generation of precision oncology. *npj Precision Oncology*, 5(1):1–3. Number: 1 Publisher: Nature Publishing Group.

[Baron et al., 2016] Baron, M., Veres, A., Wolock, S., Faust, A., Gaujoux, R., Vetere, A., Ryu, J., Wagner, B., Shen-Orr, S., Klein, A., Melton, D., and Yanai, I. (2016). A Single-Cell Transcriptomic

Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, 3(4):346–360.e4.

[Baron and Kenny, 1986] Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182.

[Battle et al., 2014] Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C. D., Beckman, K. B., Shi, J., Mei, R., Urban, A. E., Montgomery, S. B., Levinson, D. F., and Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, 24(1):14–24. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[Baylin and Jones, 2016] Baylin, S. B. and Jones, P. A. (2016). Epigenetic Determinants of Cancer. *Cold Spring Harbor Perspectives in Biology*, 8(9):a019505.

[Becht et al., 2016] Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautès-Fridman, C., Fridman, W. H., and Reyniès, A. d. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology*, 17(1):1–20. Number: 1 Publisher: BioMed Central.

[Bender, 2016] Bender, E. (2016). Challenges: Crowdsourced solutions. *Nature*, 533(7602):S62–S64. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7602 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Drug discovery and development Subject_term_id: drug-discovery-and-development.

[Biankin et al., 2012] Biankin, A. V., Waddell, N., Kassahn, K. S., Gingras, M.-C., Muthuswamy, L. B., Johns, A. L., Miller, D. K., Wilson, P. J., Patch, A.-M., Wu, J., Chang, D. K., Cowley, M. J., Gardiner, B. B., Song, S., Harliwong, I., Idrisoglu, S., Nourse, C., Nourbakhsh, E., Manning, S., Wani, S., Gongora, M., Pajic, M., Scarlett, C. J., Gill, A. J., Pinho, A. V., Rooman, I., Anderson, M., Holmes, O., Leonard, C., Taylor, D., Wood, S., Xu, Q., Nones, K., Fink, J. L., Christ, A., Bruxner, T., Cloonan, N., Kolle, G., Newell, F., Pinese, M., Mead, R. S., Humphris, J. L., Kaplan, W., Jones, M. D., Colvin, E. K., Nagrial, A. M., Humphrey, E. S., Chou, A., Chin, V. T., Chantrill, L. A., Mawson, A., Samra, J. S., Kench, J. G., Lovell, J. A., Daly, R. J., Merrett, N. D., Toon, C., Epari, K., Nguyen, N. Q., Barbour, A., Zeps, N., Australian Pancreatic Cancer Genome Initiative, Kakkar, N., Zhao, F., Wu, Y. Q., Wang, M., Muzny, D. M., Fisher, W. E., Brunicardi, F. C., Hodges, S. E., Reid, J. G., Drummond, J., Chang, K., Han, Y., Lewis, L. R., Dinh, H., Buhay, C. J., Beck, T., Timms, L., Sam, M., Begley, K., Brown, A., Pai, D., Panchal, A., Buchner, N., De Borja, R., Denroche, R. E., Yung, C. K., Serra, S., Onetto, N., Mukhopadhyay, D., Tsao, M.-S., Shaw, P. A., Petersen, G. M., Gallinger, S., Hruban, R. H., Maitra, A., Iacobuzio-Donahue, C. A., Schulick, R. D., Wolfgang, C. L., Morgan, R. A., Lawlor, R. T., Capelli, P., Corbo, V., Scardoni, M., Tortora, G., Tempero, M. A., Mann, K. M., Jenkins, N. A., Perez-Mancera, P. A., Adams, D. J., Largaespada, D. A., Wessels, L. F. A., Rust, A. G., Stein, L. D., Tuveson, D. A., Copeland, N. G., Musgrove, E. A., Scarpa, A., Eshleman, J. R., Hudson, T. J., Sutherland, R. L., Wheeler, D. A., Pearson, J. V., McPherson, J. D., Gibbs, R. A., and Grimmond, S. M. (2012). Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*, 491(7424):399–405.

[Bindea et al., 2013] Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., Angell, H., Fredriksen, T., Lafontaine, L., Berger, A., Bruneval, P., Fridman, W. H., Becker, C., Pagès, F., Speicher, M. R., Trajanoski, Z., and Galon, J. (2013). Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity*, 39(4):782–795.

[Blum et al., 2020] Blum, M. G. B., Valeri, L., François, O., Cadiou, S., Siroux, V., Lepeule, J., and Slama, R. (2020). Challenges Raised by Mediation Analysis in a High-Dimension Setting. *Environmental Health Perspectives*, 128(5):55001.

[Buchka et al., 2021] Buchka, S., Hapfelmeier, A., Gardner, P. P., Wilson, R., and Boulesteix, A.-L. (2021). On the optimistic performance evaluation of newly introduced bioinformatic methods. *Genome Biology*, 22(1):1–8. Number: 1 Publisher: BioMed Central.

[Campbell et al., 2016] Campbell, J. D., Cancer Genome Atlas Research Network, Alexandrov, A., Kim, J., Wala, J., Berger, A. H., Pedamallu, C. S., Shukla, S. A., Guo, G., Brooks, A. N., Murray, B. A., Imielinski, M., Hu, X., Ling, S., Akbani, R., Rosenberg, M., Cibulskis, C., Ramachandran, A., Collisson, E. A., Kwiatkowski, D. J., Lawrence, M. S., Weinstein, J. N., Verhaak, R. G. W., Wu, C. J., Hammerman, P. S., Cherniack, A. D., Getz, G., Artyomov, M. N., Schreiber, R., Govindan, R., and Meyerson, M. (2016). Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics*, 48(6):607–616.

[Cantini et al., 2019] Cantini, L., Kairov, U., de Reyniès, A., Barillot, E., Radvanyi, F., and Zinovyev, A. (2019). Assessing reproducibility of matrix factorization methods in independent transcriptomes. *Bioinformatics*, 35(21):4307–4313.

[Cantini et al., 2015] Cantini, L., Medico, E., Fortunato, S., and Caselle, M. (2015). Detection of gene communities in multi-networks reveals cancer drivers. *Scientific Reports*, 5(1):17386. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Biological physics;Regulatory networks Subject_term_id: biological-physics;regulatory-networks.

[Cantini et al., 2021] Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., and Baudot, A. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, 12:124.

[Caye et al., 2019] Caye, K., Jumentier, B., Lepeule, J., and François, O. (2019). LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. *Molecular Biology and Evolution*, 36(4):852–860.

[Ceccarelli et al., 2016] Ceccarelli, M., Barthel, F. P., Malta, T. M., Sabedot, T. S., Salama, S. R., Murray, B. A., Morozova, O., Newton, Y., Radenbaugh, A., Pagnotta, S. M., Anjum, S., Wang, J., Manyam, G., Zoppoli, P., Ling, S., Rao, A. A., Grifford, M., Cherniack, A. D., Zhang, H., Poisson, L., Carlotti, C. G., Tirapelli, D. P. d. C., Rao, A., Mikkelsen, T., Lau, C. C., Yung, W. K. A., Rabadan, R., Huse, J., Brat, D. J., Lehman, N. L., Barnholtz-Sloan, J. S., Zheng, S., Hess, K., Rao, G., Meyerson, M., Beroukhim, R., Cooper, L., Akbani, R., Wrensch, M., Haussler, D., Aldape, K. D., Laird, P. W., Gutmann, D. H., TCGA Research Network, Noushmehr, H., Iavarone, A., and Verhaak, R. G. W. (2016). Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*, 164(3):550–563.

[Cerami et al., 2011] Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, , Anwar, N., Schultz, N., Bader, G. D., and Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(suppl_1):D685–D690.

[Chambers et al., 1990] Chambers, J., Hastie, T., and Pregibon, D. (1990). Statistical Models in S. In Momirović, K. and Mildner, V., editors, *Compstat*, pages 317–321, Heidelberg. Physica-Verlag HD.

[Chan-Seng-Yue et al., 2020] Chan-Seng-Yue, M., Kim, J. C., Wilson, G. W., Ng, K., Figueroa, E. F., O'Kane, G. M., Connor, A. A., Denroche, R. E., Grant, R. C., McLeod, J., Wilson, J. M., Jang, G. H., Zhang, A., Liang, S.-B., Borgida, A., Chadwick, D., Kalimuthu, S., Lungu, I., Bartlett, J. M. S., Krzyzanowski, P. M., Sandhu, V., Tiriac, H., Froeling, F. E. M., Karasinska, J. M., Topham, J. T., Renouf, D. J., Schaeffer, D. F., Jones, S. J. M., Marra, M. A., Laskin, J., Chetty, R., Stein, L. D., Zogopoulos, G., Haibe-Kains, B., Campbell, P. J., Tuveson, D. A., Knox, J. J., Fischer, S. E., Gallinger, S., and Notta, F. (2020). Author Correction: Transcription phenotypes of pancreatic cancer are driven by genomic events during tumor evolution. *Nature Genetics*, 52(4):463.

[Chauvel et al., 2020] Chauvel, C., Novoloaca, A., Veyre, P., Reynier, F., and Becker, J. (2020). Evaluation of integrative clustering methods for the analysis of multi-omics data. *Briefings in Bioinformatics*, 21(2):541–552.

[Chen et al., 2013] Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, 14:128.

[Chen et al., 2017] Chen, F., Zhang, Y., Parra, E., Rodriguez, J., Behrens, C., Akbani, R., Lu, Y., Kurie, J. M., Gibbons, D. L., Mills, G. B., Wistuba, I. I., and Creighton, C. J. (2017). Multiplatform-based molecular subtypes of non-small-cell lung cancer. *Oncogene*, 36(10):1384–1393. Number: 10 Publisher: Nature Publishing Group.

[Chen et al., 2019] Chen, G., Ning, B., and Shi, T. (2019). Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Frontiers in Genetics*, 10. Publisher: Frontiers.

[Chen et al., 2014] Chen, Z., Fillmore, C. M., Hammerman, P. S., Kim, C. F., and Wong, K.-K. (2014). Non-small-cell lung cancers: a heterogeneous set of diseases. *Nature Reviews Cancer*, 14(8):535–546. Number: 8 Publisher: Nature Publishing Group.

[Chuffart et al., 2016] Chuffart, F., Richard, M., Jost, D., Burny, C., Duplus-Bottin, H., Ohya, Y., and Yvert, G. (2016). Exploiting Single-Cell Quantitative Data to Map Genetic Variants Having Probabilistic Effects. *PLOS Genetics*, 12(8):e1006213. Publisher: Public Library of Science.

[Coleman et al., 2019] Coleman, C., Kang, D., Narayanan, D., Nardi, L., Zhao, T., Zhang, J., Bailis, P., Olukotun, K., Ré, C., and Zaharia, M. (2019). Analysis of DAWNBench, a Time-to-Accuracy Machine Learning Performance Benchmark. *ACM SIGOPS Operating Systems Review*, 53(1):14–25.

[Collisson et al., 2014] Collisson, E. A., Campbell, J. D., Brooks, A. N., Berger, A. H., Lee, W., Chmielecki, J., Beer, D. G., Cope, L., Creighton, C. J., Danilova, L., Ding, L., Getz, G., Hammerman, P. S., Neil Hayes, D., Hernandez, B., Herman, J. G., Heymach, J. V., Jurisica, I., Kucherlapati, R., Kwiatkowski, D., Ladanyi, M., Robertson, G., Schultz, N., Shen, R., Sinha, R., Sougnez, C., Tsao, M.-S., Travis, W. D., Weinstein, J. N., Wigle, D. A., Wilkerson, M. D., Chu, A., Cherniack, A. D., Hadjipanayis, A., Rosenberg, M., Weisenberger, D. J., Laird, P. W., Radenbaugh, A., Ma, S., Stuart, J. M., Averett Byers, L., Baylin, S. B., Govindan, R., Meyerson, M., Rosenberg, M., Gabriel, S. B., Cibulskis, K., Sougnez, C., Kim, J., Stewart, C., Lichtenstein, L., Lander, E. S., Lawrence, M. S., Getz, G., Kandoth, C., Fulton, R., Fulton, L. L., McLellan, M. D., Wilson, R. K., Ye, K., Fronick, C. C., Maher, C. A., Miller, C. A., Wendl, M. C., Cabanski, C., Ding, L., Mardis, E., Govindan, R., Creighton, C. J., Wheeler, D., Balasundaram, M., Butterfield, Y. S. N., Carlsen, R., Chu, A., Chuah, E., Dhalla, N., Guin, R., Hirst, C., Lee, D., Li, H. I., Mayo, M., Moore, R. A., Mungall, A. J., Schein, J. E., Sipahimalani, P., Tam, A., Varhol, R., Gordon Robertson, A., Wye, N., Thiessen, N., Holt, R. A., Jones, S. J. M., Marra, M. A., Campbell, J. D., Brooks, A. N., Chmielecki, J., Imielinski, M., Onofrio, R. C., Hodis, E., Zack, T., Sougnez, C., Helman, E., Sekhar Pedamallu, C., Mesirov, J., Cherniack, A. D., Saksena, G., Schumacher, S. E., Carter, S. L., Hernandez, B., Garraway, L., Beroukhim, R., Gabriel, S. B., Getz, G., Meyerson, M., Hadjipanayis, A., Lee, S., Mahadeshwar, H. S., Pantazi, A., Protopopov, A., Ren, X., Seth, S., Song, X., Tang, J., Yang, L., Zhang, J., Chen, P.-C., Parfenov, M., Wei Xu, A., Santoso, N., Chin, L., Park, P. J., Kucherlapati, R., Hoadley, K. A., Todd Auman, J., Meng, S., Shi, Y., Buda, E., Waring, S., Veluvolu, U., Tan, D., Mieczkowski, P. A., Jones, C. D., Simons, J. V., Soloway, M. G., Bodenheimer, T., Jefferys, S. R., Roach, J., Hoyle, A. P., Wu, J., Balu, S., Singh, D., Prins, J. F., Marron, J., Parker, J. S., Neil Hayes, D., Perou, C. M., Liu, J., Cope, L., Danilova, L., Weisenberger, D. J., Maglinte, D. T., Lai, P. H., Bootwalla, M. S., Van Den Berg, D. J., Triche Jr, T., Baylin, S. B., Laird, P. W., Rosenberg, M., Chin, L., Zhang, J., Cho, J., DiCara, D., Heiman, D., Lin, P., Mallard, W., Voet, D., Zhang, H., Zou, L., Noble, M. S., Lawrence, M. S., Saksena, G., Gehlenborg, N., Thorvaldsdottir, H., Mesirov, J., Nazaire, M.-D., Robinson, J., Getz, G., Lee, W., Arman Aksoy, B., Ciriello, G., Taylor, B. S., Dresdner, G., Gao, J., Gross, B., Seshan, V. E., Ladanyi, M., Reva, B., Sinha, R., Onur Sumer, S., Weinhold,

N., Schultz, N., Shen, R., Sander, C., Ng, S., Ma, S., Zhu, J., Radenbaugh, A., Stuart, J. M., Benz, C. C., Yau, C., Haussler, D., Spellman, P. T., Wilkerson, M. D., Parker, J. S., Hoadley, K. A., Kimes, P. K., Neil Hayes, D., Perou, C. M., Broom, B. M., Wang, J., Lu, Y., Kwok Shing Ng, P., Diao, L., Averett Byers, L., Liu, W., Heymach, J. V., Amos, C. I., Weinstein, J. N., Akbani, R., Mills, G. B., Curley, E., Paulauskis, J., Lau, K., Morris, S., Shelton, T., Mallery, D., Gardner, J., Penny, R., Saller, C., Tarvin, K., Richards, W. G., Cerfolio, R., Bryant, A., Raymond, D. P., Pennell, N. A., Farver, C., Czerwinski, C., Huelsenbeck-Dill, L., Iacocca, M., Petrelli, N., Rabeno, B., Brown, J., Bauer, T., Dolzhanskiy, O., Potapova, O., Rotin, D., Voronina, O., Nemirovich-Danchenko, E., Fedosenko, K. V., Gal, A., Behera, M., Ramalingam, S. S., Sica, G., Flieder, D., Boyd, J., Weaver, J., Kohl, B., Huy Quoc Thinh, D., Sandusky, G., Juhl, H., The Cancer Genome Atlas Research Network, Disease analysis working group, Genome sequencing centres: The Eli & Edythe L. Broad Institute, Washington University in St. Louis, Baylor College of Medicine, Genome characterization centres: Canada's Michael Smith Genome Sciences Centre, B. C. C. A., The Eli & Edythe L. Broad Institute, Harvard Medical School/Brigham & Women's Hospital/MD Anderson Cancer Center, University of North Carolina, C. H., University of Kentucky, The USC/JHU Epigenome Characterization Center, Genome data analysis centres: The Eli & Edythe L. Broad Institute, Memorial Sloan-Kettering Cancer Center, University of California, S. C. I., Oregon Health & Sciences University, The University of Texas MD Anderson Cancer Center, Biospecimen core resource: International Genomics Consortium, Tissue source sites: Analytical Biological Service, I., Brigham & Women's Hospital, University of Alabama at Birmingham, Cleveland Clinic, Christiana Care, Cureline, Emory University, Fox Chase Cancer Center, ILSbio, Indiana University, Indivumed, and John Flynn Hospital (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550. Number: 7511 Publisher: Nature Publishing Group.

[Collisson et al., 2011] Collisson, E. A., Sadanandam, A., Olson, P., Gibb, W. J., Truitt, M., Gu, S., Cooc, J., Weinkle, J., Kim, G. E., Jakkula, L., Feiler, H. S., Ko, A. H., Olshen, A. B., Danenberg, K. L., Tempero, M. A., Spellman, P. T., Hanahan, D., and Gray, J. W. (2011). Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature Medicine*, 17(4):500–503.

[Connor and Gallinger, 2021] Connor, A. A. and Gallinger, S. (2021). Pancreatic cancer evolution and heterogeneity: integrating omics and clinical data. *Nature Reviews Cancer*, pages 1–12. Bandiera_abtest: a Cg_type: Nature Research Journals Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Cancer genomics;Oncogenes;Pancreatic cancer;Tumour-suppressor proteins Subject_term_id: cancer-genomics;oncogenes;pancreatic-cancer;tumour-suppressor-proteins.

[Creason et al., 2021] Creason, A., Haan, D., Dang, K., Chiotti, K. E., Inkman, M., Lamb, A., Yu, T., Hu, Y., Norman, T. C., Buchanan, A., van Baren, M. J., Spangler, R., Rollins, M. R., Spellman, P. T., Rozanov, D., Zhang, J., Maher, C. A., Caloian, C., Watson, J. D., Uhrig, S., Haas, B. J., Jain, M., Akeson, M., Ahsen, M. E., Stolovitzky, G., Guinney, J., Boutros, P. C., Stuart, J. M., Ellrott, K., Zhang, H., Wang, Y., Guan, Y., Nguyen, C., Sugai, C., Jha, A., Li, J. W., and Dobin, A. (2021). A community challenge to evaluate RNA-seq, fusion detection, and isoform quantification methods for cancer discovery. *Cell Systems*, page S2405471221002076.

[Dai et al., 2020] Dai, J. Y., Stanford, J. L., and LeBlanc, M. (2020). A Multiple-Testing Procedure for High-Dimensional Mediation Hypotheses. *Journal of the American Statistical Association*, 0(0):1–16. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2020.1765785.

[Decamps et al., 2021] Decamps, C., Arnaud, A., Petitprez, F., Ayadi, M., Baurès, A., Armenoult, L., HADACA consortium, Escalera, S., Guyon, I., Nicolle, R., Tomasini, R., de Reyniès, A., Cros, J., Blum, Y., and Richard, M. (2021). DECONbench: a benchmarking platform dedicated to deconvolution methods for tumor heterogeneity quantification. *BMC bioinformatics*, 22(1):473.

[Didier et al., 2015] Didier, G., Brun, C., and Baudot, A. (2015). Identifying communities from multiplex biological networks. *PeerJ*, 3:e1525.

[Djordjilović et al., 2019] Djordjilović, V., Page, C. M., Gran, J. M., Nøst, T. H., Sandanger, T. M., Veierød, M. B., and Thoresen, M. (2019). Global test for high-dimensional mediation: Testing groups of potential mediators. *Statistics in Medicine*, page sim.8199.

[Dong et al., 2020] Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C. M., Zou, F., and Jiang, Y. (2020). SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Briefings in Bioinformatics*.

[Druker et al., 2006] Druker, B. J., Gathmann, I., Kantarjian, H., Deininger, M. W. N., Goldman, J. M., Hochhaus, A., Rousselot, P., Hughes, T., Verhoef, G., Gratwohl, A., Simonsson, B., So, C., and Larson, R. A. (2006). Five-Year Follow-up of Patients Receiving Imatinib for Chronic Myeloid Leukemia. *n engl j med*, page 10.

[Du et al., 2019] Du, R., Carey, V., and Weiss, S. T. (2019). deconvSeq: deconvolution of cell mixture distribution in sequencing data. *Bioinformatics*, 35(24):5095–5102.

[Eicher et al., 2019] Eicher, T., Patt, A., Kautto, E., Machiraju, R., Mathé, E., and Zhang, Y. (2019). Challenges in proteogenomics: a comparison of analysis methods with the case study of the DREAM proteogenomics sub-challenge. *BMC bioinformatics*, 20(Suppl 24):669.

[Ellrott et al., 2019] Ellrott, K., Buchanan, A., Creason, A., Mason, M., Schaffter, T., Hoff, B., Eddy, J., Chilton, J. M., Yu, T., Stuart, J. M., Saez-Rodriguez, J., Stolovitzky, G., Boutros, P. C., and Guinney, J. (2019). Reproducible biomedical benchmarking in the cloud: lessons from crowdsourced data challenges. *Genome Biology*, 20(1):195.

[Elyada et al., 2019] Elyada, E., Bolisetty, M., Laise, P., Flynn, W. F., Courtois, E. T., Burkhart, R. A., Teinor, J. A., Belleau, P., Biffi, G., Lucito, M. S., Sivajothi, S., Armstrong, T. D., Engle, D. D., Yu, K. H., Hao, Y., Wolfgang, C. L., Park, Y., Preall, J., Jaffee, E. M., Califano, A., Robson, P., and Tuveson, D. A. (2019). Cross-Species Single-Cell Analysis of Pancreatic Ductal Adenocarcinoma Reveals Antigen-Presenting Cancer-Associated Fibroblasts. *Cancer Discovery*, 9(8):1102–1123.

[Enge et al., 2017] Enge, M., Arda, H. E., Mignardi, M., Beausang, J., Bottino, R., Kim, S. K., and Quake, S. R. (2017). Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell*, 171(2):321–330.e14. Publisher: Elsevier.

[Errington et al., 2021] Errington, T. M., Denis, A., Perfito, N., Iorns, E., and Nosek, B. A. (2021). Challenges for assessing replicability in preclinical cancer biology. *eLife*, 10:e67995. Publisher: eLife Sciences Publications, Ltd.

[Evans et al., 2018] Evans, C., Hardin, J., and Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, 19(5):776–792.

[Evans and Relling, 1999] Evans, W. and Relling, M. (1999). Pharmacogenomics: Translating Functional Genomics into Rational Therapeutics.

[Feng et al., 2012] Feng, J., Meyer, C. A., Wang, Q., Liu, J. S., Shirley Liu, X., and Zhang, Y. (2012). GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics*, 28(21):2782–2788.

[Ferlay et al., 2016] Ferlay, J., Partensky, C., and Bray, F. (2016). More deaths from pancreatic cancer than breast cancer in the EU by 2017. *Acta Oncologica (Stockholm, Sweden)*, 55(9-10):1158–1160.

[Fox and Loeb, 2014] Fox, E. J. and Loeb, L. A. (2014). One cell at a time. *Nature*, 512(7513):143–144. Number: 7513 Publisher: Nature Publishing Group.

[Friedman et al., 2010] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22.

[Garali et al., 2018] Garali, I., Adanyeguh, I. M., Ichou, F., Perlbarg, V., Seyer, A., Colsch, B., Moszer, I., Guillemot, V., Durr, A., Mochel, F., and Tenenhaus, A. (2018). A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia. *Briefings in Bioinformatics*, 19(6):1356–1369.

[Geiger et al., 2012] Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, Providence, RI. IEEE.

[George et al., 2018] George, J., Walter, V., Peifer, M., Alexandrov, L. B., Seidel, D., Leenders, F., Maas, L., Müller, C., Dahmen, I., Delhomme, T. M., Ardin, M., Leblay, N., Byrnes, G., Sun, R., De Reynies, A., McLeer-Florin, A., Bosco, G., Malchers, F., Menon, R., Altmüller, J., Becker, C., Nürnberg, P., Achter, V., Lang, U., Schneider, P. M., Bogus, M., Soloway, M. G., Wilkerson, M. D., Cun, Y., McKay, J. D., Moro-Sibilot, D., Brambilla, C. G., Lantuejoul, S., Lemaitre, N., Soltermann, A., Weder, W., Tischler, V., Brustugun, O. T., Lund-Iversen, M., Helland, , Solberg, S., Ansén, S., Wright, G., Solomon, B., Roz, L., Pastorino, U., Petersen, I., Clement, J. H., Sänger, J., Wolf, J., Vingron, M., Zander, T., Perner, S., Travis, W. D., Haas, S. A., Olivier, M., Foll, M., Büttner, R., Hayes, D. N., Brambilla, E., Fernandez-Cuesta, L., and Thomas, R. K. (2018). Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors. *Nature Communications*, 9(1):1048. Number: 1 Publisher: Nature Publishing Group.

[Goh et al., 2017] Goh, W. W. B., Wang, W., and Wong, L. (2017). Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends in Biotechnology*, 35(6):498–507.

[Grzywa et al., 2017] Grzywa, T. M., Paskal, W., and Włodarski, P. K. (2017). Intratumor and Intertumor Heterogeneity in Melanoma. *Translational Oncology*, 10(6):956–975.

[Guan et al., 2016] Guan, Q., Chen, R., Yan, H., Cai, H., Guo, Y., Li, M., Li, X., Tong, M., Ao, L., Li, H., Hong, G., and Guo, Z. (2016). Differential expression analysis for individual cancer samples based on robust within-sample relative gene expression orderings across multiple profiling platforms. *Oncotarget*, 7(42):68909–68920. Publisher: Impact Journals.

[Guinney and Saez-Rodriguez, 2018] Guinney, J. and Saez-Rodriguez, J. (2018). Alternative models for sharing confidential biomedical data. *Nature Biotechnology*, 36(5):391–392. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5 Primary_atype: Correspondence Publisher: Nature Publishing Group Subject_term: Policy;Research data Subject_term_id: policy;research-data.

[Gutiérrez et al., 2021] Gutiérrez, M. L., Muñoz-Bellvís, L., and Orfao, A. (2021). Genomic Heterogeneity of Pancreatic Ductal Adenocarcinoma and Its Clinical Impact. *Cancers*, 13(17):4451.

[HADACA consortium et al., 2020] HADACA consortium, Decamps, C., Privé, F., Bacher, R., Jost, D., Waguet, A., Houseman, E. A., Lurie, E., Lutsik, P., Milosavljevic, A., Scherer, M., Blum, M. G. B., and Richard, M. (2020). Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation deconvolution software. *BMC Bioinformatics*, 21(1):16.

[Haibe-Kains et al., 2020] Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C. S., Broderick, T., Hoffman, M. M., Leek, J. T., Korthauer, K., Huber, W., Brazma, A., Pineau, J., Tibshirani, R., Hastie, T., Ioannidis, J. P. A., Quackenbush, J., and Aerts, H. J. W. L. (2020). Transparency and reproducibility in artificial intelligence. *Nature*, 586(7829):E14–E16. Number: 7829 Publisher: Nature Publishing Group.

[Hammerman et al., 2012] Hammerman, P. S., Lawrence, M. S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E. S., Gabriel, S., Getz, G., Sougnez, C., Imielinski, M., Helman, E., Hernandez, B., Pho, N. H., Meyerson, M., Chu, A., Chun, H.-J. E.,

Mungall, A. J., Pleasance, E., Gordon Robertson, A., Sipahimalani, P., Stoll, D., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Chuah, E., Coope, R. J. N., Corbett, R., Dhalla, N., Guin, R., He, A., Hirst, C., Hirst, M., Holt, R. A., Lee, D., Li, H. I., Mayo, M., Moore, R. A., Mungall, K., Ming Nip, K., Olshen, A., Schein, J. E., Slobodan, J. R., Tam, A., Thiessen, N., Varhol, R., Zeng, T., Zhao, Y., Jones, S. J. M., Marra, M. A., Saksena, G., Cherniack, A. D., Schumacher, S. E., Tabak, B., Carter, S. L., Pho, N. H., Nguyen, H., Onofrio, R. C., Crenshaw, A., Ardlie, K., Beroukhim, R., Winckler, W., Hammerman, P. S., Getz, G., Meyerson, M., Protopopov, A., Zhang, J., Hadjipanayis, A., Lee, S., Xi, R., Yang, L., Ren, X., Zhang, H., Shukla, S., Chen, P.-C., Haseley, P., Lee, E., Chin, L., Park, P. J., Kucherlapati, R., Socci, N. D., Liang, Y., Schultz, N., Borsu, L., Lash, A. E., Viale, A., Sander, C., Ladanyi, M., Todd Auman, J., Hoadley, K. A., Wilkerson, M. D., Shi, Y., Liquori, C., Meng, S., Li, L., Turman, Y. J., Topal, M. D., Tan, D., Waring, S., Buda, E., Walsh, J., Jones, C. D., Mieczkowski, P. A., Singh, D., Wu, J., Gulabani, A., Dolina, P., Bodenheimer, T., Hoyle, A. P., Simons, J. V., Soloway, M. G., Mose, L. E., Jefferys, S. R., Balu, S., O'Connor, B. D., Prins, J. F., Liu, J., Chiang, D. Y., Neil Hayes, D., Perou, C. M., Cope, L., Danilova, L., Weisenberger, D. J., Maglinte, D. T., Pan, F., Van Den Berg, D. J., Triche Jr, T., Herman, J. G., Baylin, S. B., Laird, P. W., Getz, G., Noble, M., Voet, D., Saksena, G., Gehlenborg, N., DiCara, D., Zhang, J., Zhang, H., Wu, C.-J., Yingchun Liu, S., Lawrence, M. S., Zou, L., Sivachenko, A., Lin, P., Stojanov, P., Jing, R., Cho, J., Nazaire, M.-D., Robinson, J., Thorvaldsdottir, H., Mesirov, J., Park, P. J., Chin, L., Schultz, N., Sinha, R., Ciriello, G., Cerami, E., Gross, B., Jacobsen, A., Gao, J., Arman Aksoy, B., Weinhold, N., Ramirez, R., Taylor, B. S., Antipin, Y., Reva, B., Shen, R., Mo, Q., Seshan, V., Paik, P. K., Ladanyi, M., Sander, C., Akbani, R., Zhang, N., Broom, B. M., Casasent, T., Unruh, A., Wakefield, C., Craig Cason, R., Baggerly, K. A., Weinstein, J. N., Haussler, D., Benz, C. C., Stuart, J. M., Zhu, J., Szeto, C., Scott, G. K., Yau, C., Ng, S., Goldstein, T., Waltman, P., Sokolov, A., Ellrott, K., Collisson, E. A., Zerbino, D., Wilks, C., Ma, S., Craft, B., Wilkerson, M. D., Todd Auman, J., Hoadley, K. A., Du, Y., Cabanski, C., Walter, V., Singh, D., Wu, J., Gulabani, A., Bodenheimer, T., Hoyle, A. P., Simons, J. V., Soloway, M. G., Mose, L. E., Jefferys, S. R., Balu, S., Marron, J. S., Liu, Y., Wang, K., Liu, J., Prins, J. F., Neil Hayes, D., Perou, C. M., Creighton, C. J., Zhang, Y., Travis, W. D., Rekhtman, N., Yi, J., Aubry, M. C., Cheney, R., Dacic, S., Flieder, D., Funkhouser, W., Illei, P., Myers, J., Tsao, M.-S., Penny, R., Mallery, D., Shelton, T., Hatfield, M., Morris, S., Yena, P., Shelton, C., Sherman, M., Paulauskis, J., Meyerson, M., Baylin, S. B., Govindan, R., Akbani, R., Azodo, I., Beer, D., Bose, R., Byers, L. A., Carbone, D., Chang, L.-W., Chiang, D., Chu, A., Chun, E., Collisson, E., Cope, L., Creighton, C. J., Danilova, L., Ding, L., Getz, G., Hammerman, P. S., Neil Hayes, D., Hernandez, B., Herman, J. G., Heymach, J., Ida, C., Imielinski, M., Johnson, B., Jurisica, I., Kaufman, J., Kosari, F., Kucherlapati, R., Kwiatkowski, D., Ladanyi, M., Lawrence, M. S., Maher, C. A., Mungall, A., Ng, S., Pao, W., The Cancer Genome Atlas Research Network, Genome sequencing centres: Broad Institute, Genome characterization centres: BC Cancer Agency, Broad Institute, Brigham & Women's Hospital/Harvard Medical School, Memorial Sloan-Kettering Cancer Center (TCGA pilot phase only), University of North Carolina at Chapel Hill, University of Southern California/Johns Hopkins, Genome data analysis centres: Broad Institute, Memorial Sloan-Kettering Cancer Center, The University of Texas MD Anderson Cancer Center, University of California Santa Cruz/Buck Institute, Baylor College of Medicine, Pathology committee, Biospecimen core resources: International Genomics Consortium, and Disease working group (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525. Number: 7417 Publisher: Nature Publishing Group.

[Han et al., 2021] Han, J., DePinho, R. A., and Maitra, A. (2021). Single-cell RNA sequencing in pancreatic cancer. *Nature Reviews Gastroenterology & Hepatology*, 18(7):451–452. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Cancer genetics;Pancreatic cancer Subject_term_id: cancer-genetics;pancreatic-cancer.

[Hanahan and Weinberg, 2011] Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674.

[Hao et al., 2021] Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P., and Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29.

[Hayashi et al., 2021] Hayashi, A., Hong, J., and Iacobuzio-Donahue, C. A. (2021). The pancreatic cancer genome revisited. *Nature Reviews Gastroenterology & Hepatology*, 18(7):469–481. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Cancer;Cancer genetics;Pancreatic cancer Subject_term_id: cancer;cancer-genetics;pancreatic-cancer.

[Heiser et al., 2012] Heiser, L. M., Sadanandam, A., Kuo, W.-L., Benz, S. C., Goldstein, T. C., Ng, S., Gibb, W. J., Wang, N. J., Ziyad, S., Tong, F., Bayani, N., Hu, Z., Billig, J. I., Dueregger, A., Lewis, S., Jakkula, L., Korkola, J. E., Durinck, S., Pepin, F., Guan, Y., Purdom, E., Neuvial, P., Bengtsson, H., Wood, K. W., Smith, P. G., Vassilev, L. T., Hennessy, B. T., Greshock, J., Bachman, K. E., Hardwicke, M. A., Park, J. W., Marton, L. J., Wolf, D. M., Collisson, E. A., Neve, R. M., Mills, G. B., Speed, T. P., Feiler, H. S., Wooster, R. F., Haussler, D., Stuart, J. M., Gray, J. W., and Spellman, P. T. (2012). Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proceedings of the National Academy of Sciences*, 109(8):2724–2729. Publisher: Proceedings of the National Academy of Sciences.

[Hirata and Sahai, 2017] Hirata, E. and Sahai, E. (2017). Tumor Microenvironment and Differential Responses to Therapy. *Cold Spring Harbor Perspectives in Medicine*, 7(7):a026781. Publisher: Cold Spring Harbor Laboratory Press.

[Hoadley et al., 2014] Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D., Niu, B., McLellan, M. D., Uzunangelov, V., Zhang, J., Kandoth, C., Akbani, R., Shen, H., Omberg, L., Chu, A., Margolin, A. A., van't Veer, L. J., Lopez-Bigas, N., Laird, P. W., Raphael, B. J., Ding, L., Robertson, A. G., Byers, L. A., Mills, G. B., Weinstein, J. N., Van Waes, C., Chen, Z., Collisson, E. A., Benz, C. C., Perou, C. M., and Stuart, J. M. (2014). Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*, 158(4):929–944.

[Hofree et al., 2013] Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nature Methods*, 10(11):1108–1115.

[Houseman et al., 2016] Houseman, E. A., Kile, M. L., Christiani, D. C., Ince, T. A., Kelsey, K. T., and Marsit, C. J. (2016). Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC bioinformatics*, 17:259.

[Hudson (Chairperson) et al., 2010] Hudson (Chairperson), T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusuda, J., Lane, D. P., Laplace, F., Lu, Y., Nettekoven, G., Ozenberger, B., Peterson, J., Rao, T., Remacle, J., Schafer, A. J., Shibata, T., Stratton, M. R., Vockley, J. G., Watanabe, K., Yang, H., Yuen, M. M. F., Knoppers (Leader), B. M., Bobrow, M., Cambon-Thomsen, A., Dressler, L. G., Dyke, S. O. M., Joly, Y., Kato, K., Kennedy, K. L., Nicolás, P., Parker, M. J., Rial-Sebbag, E., Romeo-Casabona, C. M., Shaw, K. M., Wallace, S., Wiesner, G. L., Zeps, N., Lichter (Leader), P., Biankin, A. V., Chabannon, C., Chin, L., Clément, B., de Alava, E., Degos, F., Ferguson, M. L., Geary, P., Hayes, D. N., Hudson, T. J., Johns, A. L., Kasprzyk, A., Nakagawa, H., Penny, R., Piris, M. A., Sarin, R., Scarpa, A., Shibata, T., van de Vijver, M., Futreal (Leader), P. A., Aburatani, H., Bayés, M., Bowtell, D. D., Campbell, P. J., Estivill, X., Gerhard, D. S., Grimmond, S. M., Gut, I., Hirst, M., López-Otín, C., Majumder, P., Marra, M., McPherson, J. D., Nakagawa, H., Ning, Z., Puente, X. S., Ruan, Y., Shibata, T., Stratton, M. R., Stunnenberg, H. G., Swerdlow, H., Velculescu, V. E., Wilson, R. K., Xue, H. H., Yang, L., Spellman (Leader), P. T., Bader, G. D., Boutros, P. C., Campbell, P. J., Flicek, P., Getz, G., Guigó, R., Guo, G., Haussler, D., Heath, S., Hubbard, T. J., Jiang, T., Jones, S. M.,

Li, Q., López-Bigas, N., Luo, R., Muthuswamy, L., Francis Ouellette, B. F., Pearson, J. V., Puente, X. S., Quesada, V., Raphael, B. J., Sander, C., Shibata, T., Speed, T. P., Stein, L. D., Stuart, J. M., Teague, J. W., Totoki, Y., Tsunoda, T., Valencia, A., Wheeler, D. A., Wu, H., Zhao, S., Zhou, G., Stein (Leader), L. D., Guigó, R., Hubbard, T. J., Joly, Y., Jones, S. M., Kasprzyk, A., Lathrop, M., López-Bigas, N., Francis Ouellette, B. F., Spellman, P. T., Teague, J. W., Thomas, G., Valencia, A., Yoshida, T., Kennedy (Leader), K. L., Axton, M., Dyke, S. O. M., Futreal, P. A., Gerhard, D. S., Gunter, C., Guyer, M., Hudson, T. J., McPherson, J. D., Miller, L. J., Ozenberger, B., Shaw, K. M., Kasprzyk (Leader), A., Stein (Leader), L. D., Zhang, J., Haider, S. A., Wang, J., Yung, C. K., Cross, A., Liang, Y., Gnaneshan, S., Guberman, J., Hsu, J., Bobrow (Leader), M., Chalmers, D. R. C., Hasel, K. W., Joly, Y., Kaan, T. S. H., Kennedy, K. L., Knoppers, B. M., Lowrance, W. W., Masui, T., Nicolás, P., Rial-Sebbag, E., Lyman Rodriguez, L., Vergely, C., Yoshida, T., Grimmond (Leader), S. M., Biankin, A. V., Bowtell, D. D. L., Cloonan, N., deFazio, A., Eshleman, J. R., Etemadmoghadam, D., Gardiner, B. A., Kench, J. G., Scarpa, A., Sutherland, R. L., Tempero, M. A., Waddell, N. J., Wilson, P. J., McPherson (Leader), J. D., Gallinger, S., Tsao, M.-S., Shaw, P. A., Petersen, G. M., Mukhopadhyay, D., Chin, L., DePinho, R. A., Thayer, S., Muthuswamy, L., Shazand, K., Beck, T., Sam, M., Timms, L., Ballin, V., Lu (Leader), Y., Ji, J., Zhang, X., Chen, F., Hu, X., Zhou, G., Yang, Q., Tian, G., Zhang, L., Xing, X., Li, X., Zhu, Z., Yu, Y., Yu, J., Yang, H., Lathrop (Leader), M., Tost, J., Brennan, P., Holcatova, I., Zaridze, D., Brazma, A., Egevad, L., Prokhortchouk, E., Elizabeth Banks, R., Uhlén, M., Cambon-Thomsen, A., Viksna, J., Ponten, F., Skryabin, K., Stratton (Leader), M. R., Futreal, P. A., Birney, E., Borg, A., Børresen-Dale, A.-L., Caldas, C., Foekens, J. A., Martin, S., Reis-Filho, J. S., Richardson, A. L., Sotiriou, C., Stunnenberg, H. G., Thomas, G., van de Vijver, M., van't Veer, L., Calvo (Leader), F., Birnbaum, D., Blanche, H., Boucher, P., Boyault, S., Chabannon, C., Gut, I., Masson-Jacquemier, J. D., Lathrop, M., Pauporté, I., Pivot, X., Vincent-Salomon, A., Tabone, E., Theillet, C., Thomas, G., Tost, J., Treilleux, I., Calvo (Leader), F., Bioulac-Sage, P., Clément, B., Decaens, T., Degos, F., Franco, D., Gut, I., Gut, M., Heath, S., Lathrop, M., Samuel, D., Thomas, G., The International Cancer Genome Consortium, Executive committee, Ethics and policy committee, Tissue and clinical annotation working group, Technologies working group, Bioinformatics analyses working group, Data coordination and management working group, Data release, d. t. a. p. w. g., Data coordination centre, International data access committee, Cancer genome projects: Pancreatic cancer (ductal adenocarcinoma) and ovarian cancer (serous adenocarcinoma) (Australia), Pancreatic cancer (ductal adenocarcinomas) (Canada), Gastric cancer (intestinal- and diffuse-type) (China), Renal cancer (renal cell carcinoma; focus on but not limited to clear cell subtype) (European Union/France), Breast cancer (subtypes defined by an amplification of ER+ HER ductal-type) (European Union/United Kingdom), Breast cancer (subtype defined by an amplification of the HER2 gene) (France), and Liver cancer (hepatocellular carcinoma; secondary to alcohol and adiposity) (France) (2010). International network of cancer genome projects. *Nature*, 464(7291):993–998. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7291 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Cancer genomics;Cancer therapy Subject_term_id: cancer-genomics;cancer-therapy.

[Hwang et al., 2020] Hwang, W. L., Jagadeesh, K. A., Guo, J. A., Hoffman, H. I., Yadollahpour, P., Mohan, R., Drokhlyansky, E., Van Wittenberghe, N., Ashenberg, O., Farhi, S., Schapiro, D., Reeves, J., Zollinger, D. R., Eng, G., Schenkel, J. M., Freed-Pastor, W. A., Rodrigues, C., Gould, J., Lambden, C., Porter, C., Tsankov, A., Dionne, D., Abbondanza, D., Waldman, J., Cuoco, M., Nguyen, L., Delorey, T., Phillips, D., Ciprani, D., Kern, M., Mehta, A., Fuhrman, K., Fropf, R., Beechem, J., Loeffler, J. S., Ryan, D. P., Weekes, C. D., Ting, D. T., Ferrone, C. R., Wo, J. Y., Hong, T. S., Aguirre, A. J., Rozenblatt-Rosen, O., Mino-Kenudson, M., Castillo, C. F.-d., Liss, A. S., Jacks, T., and Regev, A. (2020). Single-nucleus and spatial transcriptomics of archival pancreatic cancer reveals multi-compartment reprogramming after neoadjuvant treatment. preprint, Genomics.

[Imai et al., 2010] Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

[Jew et al., 2020] Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K. M., Sul, J. H., Pietiläinen, K. H., Pajukanta, P., and Halperin, E. (2020). Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications*, 11(1):1971.

[Jin and Liu, 2021] Jin, H. and Liu, Z. (2021). A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biology*, 22(1):102.

[Jones et al., 2008] Jones, S., Zhang, X., Parsons, D. W., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., Hong, S.-M., Fu, B., Lin, M.-T., Calhoun, E. S., Kamiyama, M., Walter, K., Nikolskaya, T., Nikolsky, Y., Hartigan, J., Smith, D. R., Hidalgo, M., Leach, S. D., Klein, A. P., Jaffee, E. M., Goggins, M., Maitra, A., Iacobuzio-Donahue, C., Eshleman, J. R., Kern, S. E., Hruban, R. H., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E., and Kinzler, K. W. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science (New York, N.Y.)*, 321(5897):1801–1806.

[Kakarmath et al., 2020] Kakarmath, S., Esteva, A., Arnaout, R., Harvey, H., Kumar, S., Muse, E., Dong, F., Wedlund, L., and Kvedar, J. (2020). Best practices for authors of healthcare-related artificial intelligence manuscripts. *npj Digital Medicine*, 3(1):134, s41746–020–00336–w.

[Karnoub et al., 2007] Karnoub, A. E., Dash, A. B., Vo, A. P., Sullivan, A., Brooks, M. W., Bell, G. W., Richardson, A. L., Polyak, K., Tubo, R., and Weinberg, R. A. (2007). Mesenchymal stem cells within tumour stroma promote breast cancer metastasis. *Nature*, 449(7162):557–563.

[Katsuta et al., 2019] Katsuta, E., Qi, Q., Peng, X., Hochwald, S. N., Yan, L., and Takabe, K. (2019). Pancreatic adenocarcinomas with mature blood vessels have better overall survival. *Scientific Reports*, 9(1):1310.

[Kaubryte and Lai, 2022] Kaubryte, J. and Lai, A. G. (2022). Pan-cancer prognostic genetic mutations and clinicopathological factors associated with survival outcomes: a systematic review. *npj Precision Oncology*, 6(1):1–10. Number: 1 Publisher: Nature Publishing Group.

[Kim et al., 2015] Kim, J. K., Kolodziejczyk, A. A., Ilicic, T., Illicic, T., Teichmann, S. A., and Marioni, J. C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nature Communications*, 6:8687.

[Korsunsky et al., 2019] Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 12 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational models;Data integration;Statistical methods Subject_term_id: computational-models;data-integration;statistical-methods.

[Krusche et al., 2019] Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., Gonzalez-Porta, M., Eberle, M. A., Tezak, Z., Lababidi, S., Truty, R., Asimenos, G., Funke, B., Fleharty, M., Chapman, B. A., Salit, M., and Zook, J. M. (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology*, 37(5):555–560. Number: 5 Publisher: Nature Publishing Group.

[Krysiak et al., 2022] Krysiak, K., Danos, A. M., Kiwala, S., McMichael, J. F., Coffman, A. C., Barnell, E. K., Sheta, L., Saliba, J., Grisdale, C. J., Kujan, L., Pema, S., Lever, J., Spies, N. C., Chiorean, A., Rieke, D. T., Clark, K. A., Jani, P., Takahashi, H., Horak, P., Ritter, D. I., Zhou, X., Ainscough, B. J., Delong, S., Lamping, M., Marr, A. R., Li, B. V., Lin, W.-H., Terraf, P., Salama, Y., Campbell, K. M., Farncombe, K. M., Ji, J., Zhao, X., Xu, X., Kanagal-Shamanna, R., Cotto, K. C., Skidmore, Z. L., Walker, J. R., Zhang, J., Milosavljevic, A., Patel, R. Y., Giles, R. H., Kim, R. H., Schriml, L. M., Mardis, E. R., Jones, S. J. M., Raca, G., Rao, S., Madhavan, S., Wagner, A. H., Griffith, O. L., and Griffith, M. (2022). A community approach to the cancer-variant-interpretation bottleneck. *Nature Cancer*, 3(5):522–525. Number: 5 Publisher: Nature Publishing Group.

[Lambrechts et al., 2018] Lambrechts, D., Wauters, E., Boeckx, B., Aibar, S., Nittner, D., Burton, O., Bassez, A., Decaluwé, H., Pircher, A., Eynde, K. V. d., Weynand, B., Verbeken, E., Leyn, P. D., Liston, A., Vansteenkiste, J., Carmeliet, P., Aerts, S., and Thienpont, B. (2018). Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature Medicine*, 24(8):1277–1289. Number: 8 Publisher: Nature Publishing Group.

[Lawrence et al., 2014] Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501. Number: 7484 Publisher: Nature Publishing Group.

[Le et al., 2019] Le, E. P. V., Wang, Y., Huang, Y., Hickman, S., and Gilbert, F. J. (2019). Artificial intelligence in breast imaging. *Clinical Radiology*, 74(5):357–366. Publisher: Elsevier.

[Lee et al., 2017] Lee, Y. J., Arida, J. A., and Donovan, H. S. (2017). The application of crowdsourcing approaches to cancer research: a systematic review. *Cancer Medicine*, 6(11):2595–2605.

[Leite and Diele-Viegas, 2021] Leite, L. and Diele-Viegas, L. M. (2021). Juggling slow and fast science. *Nature Human Behaviour*, 5(4):409–409. Number: 4 Publisher: Nature Publishing Group.

[Li et al., 2015] Li, P., Piao, Y., Shon, H. S., and Ryu, K. H. (2015). Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics*, 16(1):1–9. Number: 1 Publisher: BioMed Central.

[Li et al., 2020] Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., Li, B., and Liu, X. S. (2020). TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Research*, 48(W1):W509–W514.

[Li et al., 2019] Li, X., Cai, H., Wang, X., Ao, L., Guo, Y., He, J., Gu, Y., Qi, L., Guan, Q., Lin, X., and Guo, Z. (2019). A rank-based algorithm of differential expression analysis for small cell line data with statistical control. *Briefings in Bioinformatics*, 20(2):482–491.

[Ligorio et al., 2019] Ligorio, M., Sil, S., Malagon-Lopez, J., Nieman, L. T., Misale, S., Di Pilato, M., Ebright, R. Y., Karabacak, M. N., Kulkarni, A. S., Liu, A., Vincent Jordan, N., Franses, J. W., Philipp, J., Kreuzer, J., Desai, N., Arora, K. S., Rajurkar, M., Horwitz, E., Neyaz, A., Tai, E., Magnus, N. K., Vo, K. D., Yashaswini, C. N., Marangoni, F., Boukhali, M., Fatherree, J. P., Damon, L. J., Xega, K., Desai, R., Choz, M., Bersani, F., Langenbucher, A., Thapar, V., Morris, R., Wellner, U. F., Schilling, O., Lawrence, M. S., Liss, A. S., Rivera, M. N., Deshpande, V., Benes, C. H., Maheswaran, S., Haber, D. A., Fernandez-Del-Castillo, C., Ferrone, C. R., Haas, W., Aryee, M. J., and Ting, D. T. (2019). Stromal Microenvironment Shapes the Intratumoral Architecture of Pancreatic Cancer. *Cell*, 178(1):160–175.e27.

[Lin et al., 2020] Lin, W., Noel, P., Borazanci, E. H., Lee, J., Amini, A., Han, I. W., Heo, J. S., Jameson, G. S., Fraser, C., Steinbach, M., Woo, Y., Fong, Y., Cridebring, D., Von Hoff, D. D., Park, J. O., and Han, H. (2020). Single-cell transcriptome analysis of tumor and stromal compartments of pancreatic ductal adenocarcinoma primary tumors and metastatic lesions. *Genome Medicine*, 12(1):80.

[Liu et al., 2018] Liu, J., Dang, H., and Wang, X. W. (2018). The significance of intertumor and intratumor heterogeneity in liver cancer. *Experimental & Molecular Medicine*, 50(1):e416–e416.

[Liu et al., 2020] Liu, X., Maiorino, E., Halu, A., Glass, K., Prasad, R. B., Loscalzo, J., Gao, J., and Sharma, A. (2020). Robustness and lethality in multilayer biological molecular networks. *Nature Communications*, 11(1):6043.

[Lloyd et al., 2016] Lloyd, M. C., Cunningham, J. J., Bui, M. M., Gillies, R. J., Brown, J. S., and Gatenby, R. A. (2016). Darwinian Dynamics of Intratumoral Heterogeneity: Not Solely Random Mutations but Also Variable Environmental Selection Forces. *Cancer Research*, 76(11):3136–3144.

[Lomberk et al., 2018] Lomberk, G., Blum, Y., Nicolle, R., Nair, A., Gaonkar, K. S., Marisa, L., Mathison, A., Sun, Z., Yan, H., Elarouci, N., Armenoult, L., Ayadi, M., Ordog, T., Lee, J.-H., Oliver, G., Klee, E., Moutardier, V., Gayet, O., Bian, B., Duconseil, P., Gilabert, M., Bigonnet, M., Garcia, S., Turrini, O., Delpero, J.-R., Giovannini, M., Grandval, P., Gasmi, M., Secq, V., De Reyniès, A., Dusetti, N., Iovanna, J., and Urrutia, R. (2018). Distinct epigenetic landscapes underlie the pathobiology of pancreatic cancer subtypes. *Nature Communications*, 9(1):1978.

[Love et al., 2014] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.

[Lu et al., 2014] Lu, Y.-F., Goldstein, D. B., Angrist, M., and Cavalleri, G. (2014). Personalized Medicine and Human Genetic Diversity. *Cold Spring Harbor Perspectives in Medicine*, 4(9):a008581–a008581.

[Luecken et al., 2022] Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., and Theis, F. J. (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50. Number: 1 Publisher: Nature Publishing Group.

[Lun et al., 2016] Lun, A. T. L., McCarthy, D. J., and Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, 5:2122.

[Lundin et al., 2018] Lundin, A., Delsing, L., Clausen, M., Ricchiuto, P., Sanchez, J., Sabirsh, A., Ding, M., Synnergren, J., Zetterberg, H., Brolén, G., Hicks, R., Herland, A., and Falk, A. (2018). Human iPS-Derived Astroglia from a Stable Neural Precursor State Show Improved Functionality Compared with Conventional Astrocytic Models. *Stem Cell Reports*, 10(3):1030–1045. Publisher: Elsevier.

[Lutsik et al., 2017] Lutsik, P., Slawski, M., Gasparoni, G., Vedeneev, N., Hein, M., and Walter, J. (2017). MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. *Genome Biology*, 18(1):1–20. Number: 1 Publisher: BioMed Central.

[Lê Cao et al., 2008] Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1):Article 35.

[Madry et al., 2019] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2019). Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv:1706.06083 [cs, stat]*. arXiv: 1706.06083.

[Makohon-Moore and Iacobuzio-Donahue, 2016] Makohon-Moore, A. and Iacobuzio-Donahue, C. A. (2016). Pancreatic cancer biology and genetics from an evolutionary perspective. *Nature Reviews Cancer*, 16(9):553–565.

[Mangul et al., 2019] Mangul, S., Martin, L. S., Hill, B. L., Lam, A. K.-M., Distler, M. G., Zelikovsky, A., Eskin, E., and Flint, J. (2019). Systematic benchmarking of omics computational tools. *Nature Communications*, 10(1):1393. Number: 1 Publisher: Nature Publishing Group.

[Marbach et al., 2012] Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., DREAM5 Consortium, Kellis, M., Collins, J. J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804.

[Mardis, 2021] Mardis, E. R. (2021). The emergence of cancer genomics in diagnosis and precision medicine. *Nature Cancer*, 2(12):1263–1264. Number: 12 Publisher: Nature Publishing Group.

[Marot et al., 2021] Marot, A., Donnot, B., Dulac-Arnold, G., Kelly, A., O'Sullivan, A., Viebahn, J., Awad, M., Guyon, I., Panciatici, P., and Romero, C. (2021). Learning to run a Power Network Challenge: a Retrospective Analysis. *arXiv:2103.03104 [cs, eess]*. arXiv: 2103.03104.

[Marusyk et al., 2020] Marusyk, A., Janiszewska, M., and Polyak, K. (2020). Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance. *Cancer Cell*, 37(4):471–484. Publisher: Elsevier.

[Marx, 2020] Marx, V. (2020). Bench pressing with genomics benchmarkers. *Nature Methods*, 17(3):255–258. Number: 3 Publisher: Nature Publishing Group.

[Massalha et al., 2020] Massalha, H., Baha Halpern, K., Abu-Gazala, S., Jana, T., Massasa, E. E., Moor, A. E., Buchauer, L., Rozenberg, M., Pikarsky, E., Amit, I., Zamir, G., and Itzkovitz, S. (2020). A single cell atlas of the human liver tumor microenvironment. *Molecular Systems Biology*, 16(12):e9682. Publisher: John Wiley & Sons, Ltd.

[Maurer et al., 2019] Maurer, C., Holmstrom, S., He, J., Laise, P., Su, T., Ahmed, A., Hibshoosh, H., Chabot, J., Oberstein, P., Sepulveda, A., Genkinger, J., Zhang, J., Iuga, A., Bansal, M., Califano, A., and Olive, K. (2019). Experimental microdissection enables functional harmonization of pancreatic cancer subtypes. *Gut*, 68(6):1034–1043.

[McGregor et al., 2016] McGregor, K., Bernatsky, S., Colmegna, I., Hudson, M., Pastinen, T., Labbe, A., and Greenwood, C. M. (2016). An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome Biology*, 17(1):84.

[Meacham and Morrison, 2013] Meacham, C. E. and Morrison, S. J. (2013). Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328–337. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7467 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Cancer;Cell biology Subject_term_id: cancer;cell-biology.

[Meng et al., 2016] Meng, C., Helm, D., Frejno, M., and Kuster, B. (2016). moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. *Journal of Proteome Research*, 15(3):755–765.

[Miyabayashi et al., 2020] Miyabayashi, K., Baker, L. A., Deschênes, A., Traub, B., Caligiuri, G., Plenker, D., Alagesan, B., Belleau, P., Li, S., Kendall, J., Jang, G. H., Kawaguchi, R. K., Somerville, T. D. D., Tiriac, H., Hwang, C.-I., Burkhart, R. A., Roberts, N. J., Wood, L. D., Hruban, R. H., Gillis, J., Krasnitz, A., Vakoc, C. R., Wigler, M., Notta, F., Gallinger, S., Park, Y., and Tuveson, D. A. (2020). Intraductal Transplantation Models of Human Pancreatic Ductal Adenocarcinoma Reveal Progressive Transition of Molecular Subtypes. *Cancer Discovery*, 10(10):1566–1589. Publisher: American Association for Cancer Research Section: Research Articles.

[Moffitt et al., 2015] Moffitt, R. A., Marayati, R., Flate, E. L., Volmar, K. E., Loeza, S. G. H., Hoadley, K. A., Rashid, N. U., Williams, L. A., Eaton, S. C., Chung, A. H., Smyla, J. K., Anderson, J. M., Kim, H. J., Bentrem, D. J., Talamonti, M. S., Iacobuzio-Donahue, C. A., Hollingsworth, M. A., and Yeh, J. J. (2015). Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature Genetics*, 47(10):1168–1178.

[Moncada et al., 2020] Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J. C., Baron, M., Hajdu, C. H., Simeone, D. M., and Yanai, I. (2020). Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology*, 38(3):333–342.

[Morvan et al., 2017] Morvan, M. L., Zinovyev, A., and Vert, J.-P. (2017). NetNorM: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. *PLOS Computational Biology*, 13(6):e1005573. Publisher: Public Library of Science.

[Mutch et al., 2002] Mutch, D. M., Berger, A., Mansourian, R., Rytz, A., and Roberts, M.-A. (2002). The limit fold change model: A practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics*, page 11.

[Newman et al., 2015] Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., and Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453–457.

[Newman et al., 2019] Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D., Diehn, M., and Alizadeh, A. A. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, 37(7):773–782. Number: 7 Publisher: Nature Publishing Group.

[Nicolle et al., 2017] Nicolle, R., Blum, Y., Marisa, L., Loncle, C., Gayet, O., Moutardier, V., Turrini, O., Giovannini, M., Bian, B., Bigonnet, M., Rubis, M., Elarouci, N., Armenoult, L., Ayadi, M., Duconseil, P., Gasmi, M., Ouaissi, M., Maignan, A., Lomberk, G., Boher, J.-M., Ewald, J., Bories, E., Garnier, J., Goncalves, A., Poizat, F., Raoul, J.-L., Secq, V., Garcia, S., Grandval, P., Barraud-Blanc, M., Norguet, E., Gilabert, M., Delpero, J.-R., Roques, J., Calvo, E., Guillaumond, F., Vasseur, S., Urrutia, R., de Reyniès, A., Dusetti, N., and Iovanna, J. (2017). Pancreatic Adenocarcinoma Therapeutic Targets Revealed by Tumor-Stroma Cross-Talk Analyses in Patient-Derived Xenografts. *Cell Reports*, 21(9):2458–2470.

[Norel et al., 2011] Norel, R., Rice, J. J., and Stolovitzky, G. (2011). The self-assessment trap: can we all be better than average? *Molecular Systems Biology*, 7(1):537. Publisher: John Wiley & Sons, Ltd.

[Nowell, 1976] Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science (New York, N.Y.)*, 194(4260):23–28.

[Onuchic et al., 2016] Onuchic, V., Hartmaier, R. J., Boone, D. N., Samuels, M. L., Patel, R. Y., White, W. M., Garovic, V. D., Oesterreich, S., Roth, M. E., Lee, A. V., and Milosavljevic, A. (2016). Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types. *Cell Reports*, 17(8):2075–2086. Publisher: Elsevier.

[Pagès, 2014] Pagès, J. (2014). *Multiple Factor Analysis by Example Using R*. Journal Abbreviation: Multiple Factor Analysis by Example Using R Pages: 253 Publication Title: Multiple Factor Analysis by Example Using R.

[Paik et al., 2012] Paik, K. Y., Choi, S. H., Heo, J. S., and Choi, D. W. (2012). Analysis of liver metastasis after resection for pancreatic ductal adenocarcinoma. *World Journal of Gastrointestinal Oncology*, 4(5):109–114.

[Peng et al., 2019] Peng, J., Sun, B.-F., Chen, C.-Y., Zhou, J.-Y., Chen, Y.-S., Chen, H., Liu, L., Huang, D., Jiang, J., Cui, G.-S., Yang, Y., Wang, W., Guo, D., Dai, M., Guo, J., Zhang, T., Liao, Q., Liu, Y., Zhao, Y.-L., Han, D.-L., Zhao, Y., Yang, Y.-G., and Wu, W. (2019). Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Research*, 29(9):725–738.

[Pierre-Jean et al., 2020] Pierre-Jean, M., Deleuze, J.-F., Le Floch, E., and Mauger, F. (2020). Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. *Briefings in Bioinformatics*, 21(6):2011–2030.

[Pio-Lopez et al., 2021] Pio-Lopez, L., Valdeolivas, A., Tichit, L., Remy, , and Baudot, A. (2021). MultiVERSE: a multiplex and multiplex-heterogeneous network embedding approach. *Scientific Reports*, 11(1):8794. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational biology and bioinformatics;Data mining;Machine learning Subject_term_id: computational-biology-and-bioinformatics;data-mining;machine-learning.

[Pratapa et al., 2020] Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154. Number: 2 Publisher: Nature Publishing Group.

[Puleo et al., 2018] Puleo, F., Nicolle, R., Blum, Y., Cros, J., Marisa, L., Demetter, P., Quertinmont, E., Svrcek, M., Elarouci, N., Iovanna, J., Franchimont, D., Verset, L., Galdon, M. G., Devière, J., de Reyniès, A., Laurent-Puig, P., Van Laethem, J.-L., Bachet, J.-B., and Maréchal, R. (2018). Stratification of Pancreatic Ductal Adenocarcinomas Based on Tumor and Microenvironment Features. *Gastroenterology*, 155(6):1999–2013.e3.

[Qi et al., 2016] Qi, L., Chen, L., Li, Y., Qin, Y., Pan, R., Zhao, W., Gu, Y., Wang, H., Wang, R., Chen, X., and Guo, Z. (2016). Critical limitations of prognostic signatures based on risk scores summarized from gene expression levels: a case study for resected stage I non-small-cell lung cancer. *Briefings in Bioinformatics*, 17(2):233–242.

[Quek et al., 2018] Quek, L., David, M. D., Kennedy, A., Metzner, M., Amatangelo, M., Shih, A., Stoilova, B., Quivoron, C., Heiblig, M., Willekens, C., Saada, V., Alsafadi, S., Vijayabaskar, M. S., Peniket, A., Bernard, O. A., Agresta, S., Yen, K., MacBeth, K., Stein, E., Vassiliou, G. S., Levine, R., De Botton, S., Thakurta, A., Penard-Lacronique, V., and Vyas, P. (2018). Clonal heterogeneity of acute myeloid leukemia treated with the IDH2 inhibitor enasidenib. *Nature Medicine*, 24(8):1167–1177. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 8 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Acute myeloid leukaemia;Cancer genomics;Cancer metabolism;Translational research Subject_term_id: acute-myeloid-leukaemia;cancer-genomics;cancer-metabolism;translational-research.

[Raphael et al., 2017] Raphael, B. J., Hruban, R. H., Aguirre, A. J., Moffitt, R. A., Yeh, J. J., Stewart, C., Robertson, A. G., Cherniack, A. D., Gupta, M., Getz, G., Gabriel, S. B., Meyerson, M., Cibulskis, C., Fei, S. S., Hinoue, T., Shen, H., Laird, P. W., Ling, S., Lu, Y., Mills, G. B., Akbani, R., Loher, P., Londin, E. R., Rigoutsos, I., Telonis, A. G., Gibb, E. A., Goldenberg, A., Mezlini, A. M., Hoadley, K. A., Collisson, E., Lander, E., Murray, B. A., Hess, J., Rosenberg, M., Bergelson, L., Zhang, H., Cho, J., Tiao, G., Kim, J., Livitz, D., Leshchiner, I., Reardon, B., Van Allen, E., Kamburov, A., Beroukhim, R., Saksena, G., Schumacher, S. E., Noble, M. S., Heiman, D. I., Gehlenborg, N., Kim, J., Lawrence, M. S., Adsay, V., Petersen, G., Klimstra, D., Bardeesy, N., Leiserson, M. D., Bowlby, R., Kasaian, K., Birol, I., Mungall, K. L., Sadeghi, S., Weinstein, J. N., Spellman, P. T., Liu, Y., Amundadottir, L. T., Tepper, J., Singhi, A. D., Dhir, R., Paul, D., Smyrk, T., Zhang, L., Kim, P., Bowen, J., Frick, J., Gastier-Foster, J. M., Gerken, M., Lau, K., Leraas, K. M., Lichtenberg, T. M., Ramirez, N. C., Renkel, J., Sherman, M., Wise, L., Yena, P., Zmuda, E., Shih, J., Ally, A., Balasundaram, M., Carlsen, R., Chu, A., Chuah, E., Clarke, A., Dhalla, N., Holt, R. A., Jones, S. J., Lee, D., Ma, Y., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Tse, K., Wong, T., Brooks, D., Auman, J. T., Balu, S., Bodenheimer, T., Hayes, D. N., Hoyle, A. P., Jefferys, S. R., Jones, C. D., Meng, S., Mieczkowski, P. A., Mose, L. E., Perou, C. M., Perou, A. H., Roach, J., Shi, Y., Simons, J. V., Skelly, T., Soloway, M. G., Tan, D., Veluvolu, U., Parker, J. S., Wilkerson, M. D., Korkut, A., Senbabaoglu, Y., Burch, P., McWilliams, R., Chaffee, K., Oberg, A., Zhang, W., Gingras, M.-C., Wheeler, D. A., Xi, L., Albert, M., Bartlett, J., Sekhon, H., Stephen, Y., Howard, Z., Judy, M., Breggia, A., Shroff, R. T., Chudamani, S., Liu, J., Lolla, L., Naresh, R., Pihl, T., Sun, Q., Wan, Y., Wu, Y., Jennifer, S., Roggin, K., Becker, K.-F., Behera, M., Bennett, J., Boice, L., Burks, E., Carlotti Junior, C. G., Chabot, J., Pretti da Cunha Tirapelli, D., Sebastião dos Santos, J., Dubina, M., Eschbacher, J., Huang, M., Huelsenbeck-Dill, L., Jenkins, R., Karpov, A., Kemp, R., Lyadov, V., Maithel, S., Manikhas, G., Montgomery, E., Noushmehr, H., Osunkoya, A., Owonikoko, T., Paklina, O., Potapova, O., Ramalingam, S., Rathmell, W. K., Rieger-Christ, K., Saller, C., Setdikova, G., Shabunin, A., Sica, G., Su, T., Sullivan, T., Swanson, P., Tarvin, K., Tavobilov, M., Thorne, L. B., Urbanski, S., Voronina, O., Wang, T., Crain, D., Curley, E., Gardner, J., Mallery, D., Morris, S., Paulauskis, J., Penny, R., Shelton, C., Shelton, T., Janssen, K.-P., Bathe, O., Bahary, N., Slotta-Huspenina, J., Johns, A., Hibshoosh, H., Hwang, R. F., Sepulveda, A., Radenbaugh, A., Baylin, S. B., Berrios, M., Bootwalla, M. S., Holbrook, A., Lai, P. H., Maglinte, D. T., Mahurkar, S., Triche, T. J., Van Den Berg, D. J., Weisenberger, D. J., Chin, L., Kucherlapati, R., Kucherlapati, M., Pantazi, A., Park, P., Saksena, G., Voet, D., Lin, P., Frazer, S., Defreitas, T., Meier, S., Chin, L., Kwon, S. Y., Kim, Y. H., Park, S.-J., Han, S.-S., Kim, S. H., Kim, H., Furth, E., Tempero, M., Sander, C., Biankin, A., Chang, D.,

Bailey, P., Gill, A., Kench, J., Grimmond, S., Johns, A., Cancer Genome Initiative (APGI, A. P., Postier, R., Zuna, R., Sicotte, H., Demchok, J. A., Ferguson, M. L., Hutter, C. M., Mills Shaw, K. R., Sheth, M., Sofia, H. J., Tarnuzzer, R., Wang, Z., Yang, L., Zhang, J. J., Felau, I., and Zenklusen, J. C. (2017). Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell*, 32(2):185–203.e13.

[Rappoport and Shamir, 2018] Rappoport, N. and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, 46(20):10546–10562.

[Ren et al., 2018] Ren, X., Kang, B., and Zhang, Z. (2018). Understanding tumor ecosystems by single-cell sequencing: promises and limitations. *Genome Biology*, 19(1):1–14. Number: 1 Publisher: BioMed Central.

[Richard et al., 2018] Richard, M., Chuffart, F., Duplus-Bottin, H., Pouyet, F., Spichty, M., Fulcrand, E., Entrevan, M., Barthelaix, A., Springer, M., Jost, D., and Yvert, G. (2018). Assigning function to natural allelic variation via dynamic modeling of gene network induction. *Molecular Systems Biology*, 14(1):e7803.

[Richard et al., 2020] Richard, M., Decamps, C., Chuffart, F., Brambilla, E., Rousseaux, S., Khochbin, S., and Jost, D. (2020). PenDA, a rank-based method for personalized differential analysis: Application to lung cancer. *PLOS Computational Biology*, 16(5):e1007869.

[Richiardi et al., 2013] Richiardi, L., Bellocco, R., and Zugna, D. (2013). Mediation analysis in epidemiology: methods, interpretation and bias. *International Journal of Epidemiology*, 42(5):1511–1519.

[Ripley et al., 2022] Ripley, B., Venables, B., Bates, D. M., ca 1998), K. H. p. p., ca 1998), A. G. p. p., and Firth, D. (2022). MASS: Support Functions and Datasets for Venables and Ripley's MASS.

[Ritchie et al., 2015] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47.

[Robinson et al., 2010] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

[Rooney et al., 2015] Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., and Hacohen, N. (2015). Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell*, 160(1):48–61. Publisher: Elsevier.

[Rousseaux et al., 2013] Rousseaux, S., Debernardi, A., Jacquiau, B., Vitte, A.-L., Vesin, A., Nagy-Mignotte, H., Moro-Sibilot, D., Brichon, P.-Y., Lantuejoul, S., Hainaut, P., Laffaire, J., de Reyniès, A., Beer, D. G., Timsit, J.-F., Brambilla, C., Brambilla, E., and Khochbin, S. (2013). Ectopic Activation of Germline and Placental Genes Identifies Aggressive Metastasis-Prone Lung Cancers. *Science Translational Medicine*, 5(186).

[Sakamoto et al., 2020] Sakamoto, H., Attiyeh, M. A., Gerold, J. M., Makohon-Moore, A. P., Hayashi, A., Hong, J., Kappagantula, R., Zhang, L., Melchor, J. P., Reiter, J. G., Heyde, A., Bielski, C. M., Penson, A. V., Gönen, M., Chakravarty, D., O'Reilly, E. M., Wood, L. D., Hruban, R. H., Nowak, M. A., Socci, N. D., Taylor, B. S., and Iacobuzio-Donahue, C. A. (2020). The Evolutionary Origins of Recurrent Pancreatic Cancer. *Cancer Discovery*, 10(6):792–805.

[Salignon et al., 2018] Salignon, J., Richard, M., Fulcrand, E., Duplus-Bottin, H., and Yvert, G. (2018). Genomics of cellular proliferation in periodic environmental fluctuations. *Molecular Systems Biology*, 14(3):e7823.

[Sampson et al., 2018] Sampson, J. N., Boca, S. M., Moore, S. C., and Heller, R. (2018). FWER and FDR control when testing multiple mediators. *Bioinformatics*, 34(14):2418–2424.

[Sausen et al., 2015] Sausen, M., Phallen, J., Adleff, V., Jones, S., Leary, R. J., Barrett, M. T., Anagnostou, V., Parpart-Li, S., Murphy, D., Kay Li, Q., Hruban, C. A., Scharpf, R., White, J. R., O'Dwyer, P. J., Allen, P. J., Eshleman, J. R., Thompson, C. B., Klimstra, D. S., Linehan, D. C., Maitra, A., Hruban, R. H., Diaz, L. A., Von Hoff, D. D., Johansen, J. S., Drebin, J. A., and Velculescu, V. E. (2015). Clinical implications of genomic alterations in the tumour and circulation of pancreatic cancer patients. *Nature Communications*, 6:7686.

[Singh et al., 2019] Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., and Lê Cao, K.-A. (2019). DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055–3062.

[Sobel, 1982] Sobel, M. E. (1982). Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology*, 13:290–312. Publisher: [American Sociological Association, Wiley, Sage Publications, Inc.].

[Steele et al., 2020] Steele, N. G., Carpenter, E. S., Kemp, S. B., Sirihorachai, V., The, S., Delrosario, L., Lazarus, J., Amir, E.-a. D., Gunchick, V., Espinoza, C., Bell, S., Harris, L., Lima, F., Irizarry-Negron, V., Paglia, D., Macchia, J., Chu, A. K. Y., Schofield, H., Wamsteker, E.-J., Kwon, R., Schulman, A., Prabhu, A., Law, R., Sondhi, A., Yu, J., Patel, A., Donahue, K., Nathan, H., Cho, C., Anderson, M. A., Sahai, V., Lyssiotis, C. A., Zou, W., Allen, B. L., Rao, A., Crawford, H. C., Bednar, F., Frankel, T. L., and Pasca di Magliano, M. (2020). Multimodal Mapping of the Tumor and Peripheral Blood Immune Landscape in Human Pancreatic Cancer. *Nature cancer*, 1(11):1097–1112.

[Steen et al., 2020] Steen, C. B., Liu, C. L., Alizadeh, A. A., and Newman, A. M. (2020). Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. *Methods in Molecular Biology (Clifton, N.J.)*, 2117:135–157.

[Steijger et al., 2013] Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Hubbard, T. J., Guigó, R., Harrow, J., and Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, 10(12):1177–1184. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 12 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genome informatics Subject_term_id: genome-informatics.

[Stokkum, 2012] Stokkum, K. M. M. a. I. H. M. v. (2012). nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS).

[Stuart et al., 2019] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21. Publisher: Elsevier.

[Sturm et al., 2019] Sturm, G., Finotello, F., Petitprez, F., Zhang, J. D., Baumbach, J., Fridman, W. H., List, M., and Aneichyk, T. (2019). Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics (Oxford, England)*, 35(14):i436–i445.

[Tarazona et al., 2020] Tarazona, S., Balzano-Nogueira, L., Gómez-Cabrero, D., Schmidt, A., Imhof, A., Hankemeier, T., Tegnér, J., Westerhuis, J. A., and Conesa, A. (2020). Harmonization of quality metrics and power calculation in multi-omic studies. *Nature Communications*, 11(1):3092.

[Tarazona et al., 2015] Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., and Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, 43(21):e140.

[Tenenhaus et al., 2014] Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J., and Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics (Oxford, England)*, 15(3):569–583.

[Tenenhaus et al., 2017] Tenenhaus, M., Tenenhaus, A., and Groenen, P. J. F. (2017). Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods. *Psychometrika*.

[Teschendorff et al., 2017] Teschendorff, A. E., Breeze, C. E., Zheng, S. C., and Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics*, 18(1):105.

[Tsoucas et al., 2019] Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., and Yuan, G.-C. (2019). Accurate estimation of cell-type composition from gene expression data. *Nature Communications*, 10(1):2975.

[Vanschoren et al., 2014] Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. (2014). OpenML: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60. arXiv:1407.7722 [cs].

[Vanunu et al., 2010] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating Genes and Protein Complexes with Disease via Network Propagation. *PLOS Computational Biology*, 6(1):e1000641. Publisher: Public Library of Science.

[Vitali et al., 2019] Vitali, F., Li, Q., Schissler, A. G., Berghout, J., Kenost, C., and Lussier, Y. A. (2019). Developing a 'personalome' for precision medicine: emerging methods that compute interpretable effect sizes from single-subject transcriptomes. *Briefings in Bioinformatics*, 20(3):789–805.

[Vogelstein and Kinzler, 2015] Vogelstein, B. and Kinzler, K. W. (2015). The Path to Cancer — Three Strikes and You're Out. Archive Location: world Publisher: Massachusetts Medical Society.

[Vollmann-Zwerenz et al., 2020] Vollmann-Zwerenz, A., Leidgens, V., Feliciello, G., Klein, C. A., and Hau, P. (2020). Tumor Cell Invasion in Glioblastoma. *International Journal of Molecular Sciences*, 21(6):1932. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[Waddell et al., 2015] Waddell, N., Pajic, M., Patch, A.-M., Chang, D. K., Kassahn, K. S., Bailey, P., Johns, A. L., Miller, D., Nones, K., Quek, K., Quinn, M. C. J., Robertson, A. J., Fadlullah, M. Z. H., Bruxner, T. J. C., Christ, A. N., Harliwong, I., Idrisoglu, S., Manning, S., Nourse, C., Nourbakhsh, E., Wani, S., Wilson, P. J., Markham, E., Cloonan, N., Anderson, M. J., Fink, J. L., Holmes, O., Kazakoff, S. H., Leonard, C., Newell, F., Poudel, B., Song, S., Taylor, D., Waddell, N., Wood, S., Xu, Q., Wu, J., Pinese, M., Cowley, M. J., Lee, H. C., Jones, M. D., Nagrial, A. M., Humphris, J., Chantrill, L. A., Chin, V., Steinmann, A. M., Mawson, A., Humphrey, E. S., Colvin, E. K., Chou, A., Scarlett, C. J., Pinho, A. V., Giry-Laterriere, M., Rooman, I., Samra, J. S., Kench, J. G., Pettitt, J. A., Merrett, N. D., Toon, C., Epari, K., Nguyen, N. Q., Barbour, A., Zeps, N., Jamieson, N. B., Graham, J. S., Niclou, S. P., Bjerkvig, R., Grützmann, R., Aust, D., Hruban, R. H., Maitra, A., Iacobuzio-Donahue, C. A., Wolfgang, C. L., Morgan, R. A., Lawlor, R. T., Corbo, V., Bassi, C., Falconi, M., Zamboni, G., Tortora, G., Tempero, M. A., Gill, A. J., Eshleman, J. R., Pilarsky, C., Scarpa, A., Musgrove, E. A., Pearson, J. V., Biankin, A. V., and Grimmond, S. M. (2015). Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*, 518(7540):495–501. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7540 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Medical research Subject_term_id: medical-research.

[Wang et al., 2014] Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337.

[Wang et al., 2015] Wang, H., Sun, Q., Zhao, W., Qi, L., Gu, Y., Li, P., Zhang, M., Li, Y., Liu, S.-L., and Guo, Z. (2015). Individual-level analysis of differential expression of genes and pathways for personalized medicine. *Bioinformatics*, 31(1):62–68.

[Wang et al., 2010] Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1):136–138.

[Wang et al., 2019] Wang, X., Park, J., Susztak, K., Zhang, N. R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications*, 10(1):380.

[Weinstein et al., 2013] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120. Bandiera_abtest: a Cc_license_type: cc_y Cg_type: Nature Research Journals Number: 10 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Cancer;Genomics Subject_term_id: cancer;genomics.

[Whatcott et al., 2015] Whatcott, C. J., Diep, C. H., Jiang, P., Watanabe, A., LoBello, J., Sima, C., Hostetter, G., Shepard, H. M., Hoff, D. D. V., and Han, H. (2015). Desmoplasia in Primary Tumors and Metastatic Lesions of Pancreatic Cancer. *Clinical Cancer Research*, 21(15):3561–3568. Publisher: American Association for Cancer Research Section: Biology of Human Tumors.

[White et al., 2019] White, B. S., Gentles, A. J., Reyniès, A. d., Newman, A. M., Lamb, A., Heiser, L., Waterfall, J. J., Yu, T., and Guinney, J. (2019). Abstract 1690: A tumor deconvolution DREAM Challenge: Inferring immune infiltration from bulk gene expression data. *Cancer Research*, 79(13 Supplement):1690–1690. Publisher: American Association for Cancer Research Section: Molecular and Cellular Biology / Genetics.

[Wiens et al., 2019] Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., and Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9):1337–1340.

[Wilkinson et al., 2016] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Publication characteristics;Research data Subject_term_id: publication-characteristics;research-data.

[Xu et al., 2022] Xu, Z., Escalera, S., Pavão, A., Richard, M., Tu, W.-W., Yao, Q., Zhao, H., and Guyon, I. (2022). Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 0(0). Publisher: Elsevier.

[Zeng et al., 2021] Zeng, P., Shao, Z., and Zhou, X. (2021). Statistical methods for mediation analysis in the era of high-throughput genomics: Current successes and future challenges. *Computational and Structural Biotechnology Journal*, 19:3209–3224.

[Zhang et al., 2022] Zhang, A., Xing, L., Zou, J., and Wu, J. C. (2022). Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, pages 1–16. Publisher: Nature Publishing Group.

[Zhang et al., 2019] Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., Ping, Y., Li, F., Shi, A., Bai, J., Zhao, T., Li, X., and Xiao, Y. (2019). CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Research*, 47(D1):D721–D728.

[Zheng et al., 2018] Zheng, S. C., Breeze, C. E., Beck, S., and Teschendorff, A. E. (2018). Identification of differentially methylated cell types in epigenome-wide association studies. *Nature Methods*, 15(12):1059–1066.

List of Figures

List of Figures