Des données homogènes aux données hétérogènes en biologie des systèmes

Emmanuelle Becker

UNIVERSITÉ DE RENNES

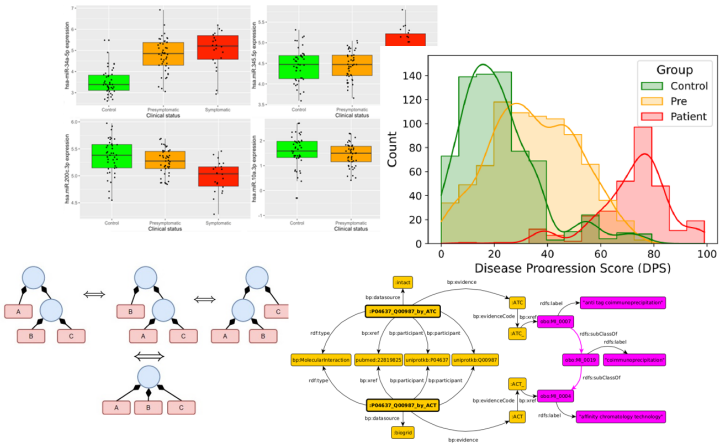
École Doctorale MathStic – Mention Informatique

Équipe de recherche DYLISS
Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA)

Des données homogènes aux données hétérogènes en biologie des systèmes

Emmanuelle BECKER

Habilitation à Diriger des Recherches
Présentée et soutenue publiquement le 14 décembre 2022



Composition du jury :

Anaïs BAUDOT	Directrice de recherche (MMG, Marseille), rapportrice
Christine BRUN	Directrice de recherche (TAGC, Marseille), examinatrice
Alessandra CARBONE	Prof. Sobonne Univ. (LCQB, Paris), examinatrice
Olivier DAMERON	Prof. Univ. Rennes (IRISA, Rennes), examinateur
Elisa FROMONT	Prof. Univ. Rennes (IRISA, Rennes), examinatrice
Alejandro MAASS	Prof. Univ. Chile (CMM, Chili), rapporteur
Anne SIEGEL	Directrice de recherche (IRISA, Rennes), examinatrice
Patricia THEBAULT	Prof. Univ. Bordeaux (LABRI, Talence), rapportrice

Abstract

From homogeneous to heterogeneous data in systems biology

by Emmanuelle BECKER

Biological systems involve a large number of different entities, each functionning in a coordinated manner with the others. Their understanding is crucial and can be approached at different scales, from the molecular to the systemic one. The observation of all these entities in different contexts and at different scales generates a “tsunami of data", posing complex and interesting computational problems. My work focuses on the development of methods for knowledge generation and knowledge extraction from these massive data.
The manuscript is organized in three axes. The first axis deals with methods to identify interpretable, robust and replicable signatures in high dimensional unimodal data. The second axis proposes the development of a new approach to integrate multimodal data (miRNA + MRI), and to identify disease progression scores. Finally, the third axis also deals with the integration of heterogeneous data, but with a systemic approach, i.e. taking into account the known relationships between entities. The work presented illustrates the complexity of extracting information from existing databases, despite the constant efforts of the bioinformatics community to structure and unify the available information.

Résumé

Des données homogènes aux données hétérogènes en biologie des systèmes

par Emmanuelle BECKER

Les systèmes biologiques font intervenir un grand nombre d'entités différentes et chacune fonctionne de façon coordonnée avec les autres. Leur compréhension est cruciale et peut être abordée à différentes échelles, de l'échelle moléculaire à l'échelle systémique. L'observation de toutes ces entités dans des contextes et à des échelles différentes génère un « tsunami de données », posant des problèmes informatiques complexes et intéressants. Mes travaux portent sur le développement de méthodes de génération de connaissances et d'extraction de connaissances à partir de cette masse de données.

Le manuscrit est organisé en trois axes. Le premier axe traite de méthodes visant à identifier des signatures interprétables, robustes et réplicables dans des données unimodales de grande dimension. Le second axe propose le développement d'une nouvelle approche pour intégrer des données multimodales (miARN + IRM), et identifier des scores de progression de maladies à partir de celles-ci. Enfin, le dernier axe traite également d'intégration de données hétérogènes, mais avec une approche systémique, c'est à dire en prenant en compte les relations connues entre entités. Les travaux présentés illustreront la complexité de l'extraction d'information de qualité des bases de données existantes, malgré les efforts constants de la communauté bioinformatique pour structurer et unifier l'information disponible.

Acknowledgement

Je remercie tout d'abord les membres de mon jury d'avoir accepté de relire et évaluer ce document. Je sais que leurs agendas sont bien chargés, et j'apprécie à sa juste valeur le temps consacré à la relecture de ce document. Merci donc à mes trois rapportrices et rapporteurs : Anaïs Baudot, Patricia Thébault et Alejandro Maass, ainsi qu'aux examinateurs et examinatrices : Alessandra Carbone, Christine Brun, Olivier Dameron, Anne Siegel et Elisa Fromont.

Cette habilitation à diriger des recherches présente des orientations et projets auxquels j'ai participé depuis ma soutenance de doctorat en 2007. Ces travaux sont le fruit de nombreuses collaborations avec des collègues, des étudiant·e·s ou des ami·e·s que je tiens tout particulièrement à remercier.

Je remercie tout d'abord Raphaël Guérois, qui a eu l'audace de recruter une informaticienne de formation dans une unité de biologie structurale. J'ai énormément appris à son contact, tant sur la biologie structurale que sur l'encadrement des doctorant·e·s. Je souhaite également remercier Christine Brun et Alain Guénoche, qui m'ont initié à la biologie des systèmes et ont su me transmettre leur enthousiasme à étudier ces questions. Ces années à Marseille ont été très formatrices et je garde de très bons souvenirs de nos discussions. Je remercie mes équipes de recherche de l'IRSET, où j'ai eu la chance de travailler pendant mes premières années comme maîtresse de conférences à l'Université de Rennes. Enfin, je remercie l'équipe Dyliss de l'IRISA qui me permet de m'épanouïr scientifiquement et de développer mes projets depuis 4 années maintenant : je souhaite à tout scientifique une équipe de recherche comme la nôtre !

J'ai eu la chance d'encadrer des étudiantes et étudiants qui m'ont aidé à progresser dans mes recherches. Je pense à Méline Wery, Virgilio Kmetzsch et Camille Juigné que j'ai eu l'opportunité de co-encadrer comme doctorante ou doctorant, et qui m'ont apporté énormément scientifiquement et humainement. Je pense aussi aux étudiants ou étudiantes de licence et master comme Sofiane, Paul, Marine, Estelle, Victoria, Thomas, Anaëlle, Hugo, Fanny, Camille, Quentin, Marc, Emmanuel et Nancy.

Certaines personnes de mon entourage scientifique ont été déterminantes dans ma capacité à m'imaginer défendre un jour cette habilitation à diriger des recherches. Je pense à Anne Siegel et Olivier Dameron : merci pour votre confiance, elle m'a beaucoup motivée et aidée ! Merci également à Claire Lemaître qui partage mon bureau et avec qui je m'étais lancée dans l'aventure : se renseigner et réfléchir aux démarches à deux, c'est beaucoup plus motivant !

Enfin, je remercie ma famille qui me soutient indéfectiblement et depuis toujours. Ce manuscrit est pour vous Benoît, Capucine, Blanche et Anatole. Et pour toi maman.

Contents

List of Abbreviations

ADN	Acide desoxyribonucléique
ARN	Acide ribonucléique
AUC	Area under the curve
BDD	Base de données
CI	Confidence interval
DFT	Démence frontotemporale
DPS	Disease progression score
HUPO-PSI	Human Proteome Organization Proteomics Standards Initiative
IDM	Interaction detection method
IPP	Interaction protéine-protéine
MI	Molecular Interaction
miARN	Micro-Acide ribonucléique
RNA-seq	RNA sequencing
ROC	Receiver operating characteristic
SLA	Sclérose laterale Amyotrophique
SPARQL	SPARQL Protocol and RDF Query Language
VAE	Autoencodeur variationnel

Chapter 1

La science du vivant : de science des données à science des connaissances

1.1 Massification des données des sciences du vivant

Récemment, Paul Nurse, biochimiste britannique et prix Nobel de médecine en 2001 avec Leland H. Hartwell et Tim Hunt pour leur découverte de la régulation du cycle cellulaire, publiait une tribune dans Nature intitulée « *Biology must generate ideas as well as data* », dans laquelle il soulignait que, souvent, il se rendait à un séminaire scientifique et se sentait noyé sous les données : « *Rather often, I go to a research talk and feel drowned in data. Some speakers seem to think they must unleash a tsunami of data if they are to be taken seriously* » (Nurse, 2021).
Le reste de ce manuscrit se concentre sur le développement de méthodes de génération de connaissances à partir de cette masse de données.

1.1.1 Le séquençage ou le début de la massification des données

Avec l'avènement de la génomique à haut débit, les spécialistes des sciences de la vie commencent dès les années 2000 à s'attaquer à des ensembles de données massives et à relever des défis en matière de manipulation, de traitement et de transfert d'informations qui étaient autrefois l'apanage des astronomes et des physiciens (Blake and Bult, 2006; Cannata et al., 2005). C'est cette réalité qui fait dire à Vivien Marx, en phrase d'accroche à sa tribune dans Nature en 2013 : « *Biologists are joining the big data club* » (Marx, 2013). De fait, d'après les estimations de Zachary Stephens dans son article « *Big Data : Astronomical or genomical ?* », en comparant la génomique à trois autres grands générateurs de données massives (l'astronomie, YouTube et Twitter), la génomique est soit égale, soit le plus exigeant des domaines analysés en termes d'acquisition, de stockage, de distribution et d'analyse des données (Stephens et al., 2015).

1.1.2 Des données interdépendantes à différentes échelles

L'accélération de l'acquisition des données en génomique par les techniques de séquençage à très haut débit est principalement responsable du tsunami de données actuel. L'accessibilité des techniques de séquençage et ses adaptations successives ont mené au développement de différentes techniques permettant de recueillir de l'information à haut débit à différentes échelles (figure 1.1) : l'échelle de la séquence, des transcrits (Wang et al., 2009a), des mécanismes de régulation de la transription (Lemmens et al., 2006; Schmidt et al., 2009), des protéines (Jiang and English, 2002;

Oda et al., 1999; Ong et al., 2002), des interactions protéique (Gavin et al., 2002; Ho et al., 2002; Ito et al., 2001; Uetz et al., 2000)... De plus, avec l'apparition des techniques de séquençage de cellules uniques et leur commercialisation, l'hétérogénéité cellulaire peut être explorée (Klein et al., 2015; Macosko et al., 2015; Picelli et al., 2013; Svensson et al., 2018). Les données ainsi disponibles à différentes échelles sont souvent analysées séparément. Pourtant, chacun des niveaux interagit avec les autres. Cette interdépendance entre données complexifie grandement leur analyse (Stephens et al., 2015).

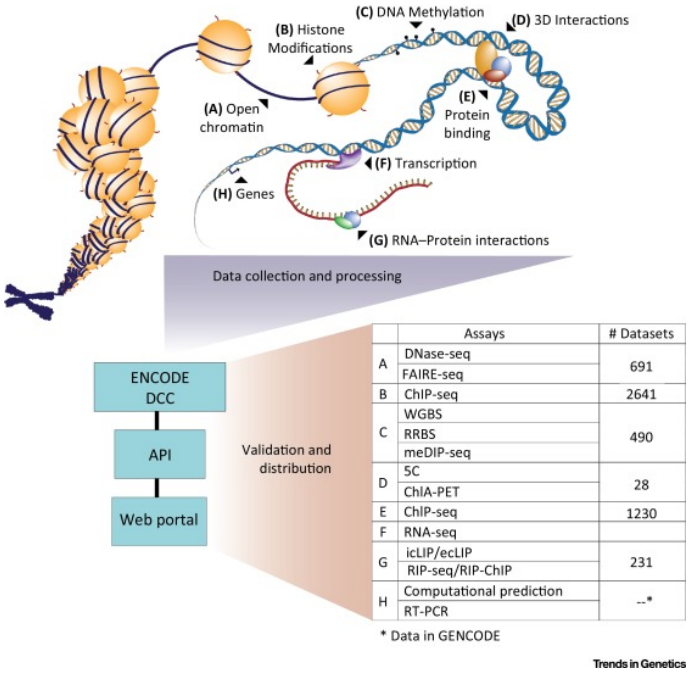


FIGURE 1.1: Données disponibles dans le projet ENCODE. Les données disponibles le sont à différents niveau : conformation de la chromatine, interactions protéine-ADN, niveau d'expression des transcrits... Figure extraite de (Diehl and Boyle, 2016)

1.1.3 Dimension des données

Les technologies de séquençage et criblage à haut débit ont été mises au point pour mesurer toutes les molécules d'intérêt dans un échantillon en une seule expérience (par exemple, le génome entier, les quantités de métabolites, la composition du microbiome), de manière non ciblée. L'exploration n'étant pas ciblée, les données ainsi collectées concernent un très grand ensemble de variables, et le nombre de réplicas collectés est généralement très inférieur au nombre de ces variables. Ce problème, connu sous les noms de « *large p small n* », ou « malédiction de la dimensionnalité », limite la pertinence des statistiques classiques, développées aux cours du 20ème siècle pour de petits ensembles de données. Ainsi, les données sont nombreuses, mais leur analyse est complexe (Stephens et al., 2015).

1.1.4 Profilages, signatures et biomarqueurs

Le lien entre information -omique disponible et phénotype observé reste complexe à mettre en évidence. Aussi, on recherche fréquemment des **profils** ou des **signatures**

d'une pathologie ou d'un autre état particulier, en ayant recours à des analyses statistiques comparant un état de référence et un autre état, objet de l'étude.

Tout d'abord, on pourra parler de *profilage* dans le cadre d'analyses prospectives visant à identifier l'ensemble des éléments -omiques dont la quantité varie entre les deux conditions étudiées ; les profilages permettent donc d'apprécier les effets globaux d'une condition, d'un traitement, etc... en privilégiant la sensibilité.

Dans un second temps, on parlera de *signature* ou de *biomarqueur* dans le cadre de l'identification d'un ou de plusieurs éléments dont les variations quantitatives ont un potentiel prédictif validé, que ce soit en terme de diagnostic, de pronostic, ou concernant la prédiction d'une réponse thérapeutique (Chibon, 2013). Ainsi, une *signature*, composée de *biomarqueurs*, impose une spécificité et un potentiel prédictif des variations quantitatives.

1.1.5 Méthodes pour l'extraction de profils

La technologie de séquençage la plus utilisée est celle étudiant quantitativement l'expression des transcrits (RNAseq). Pour traiter ces données d'expression, différents outils ont été développés pour identifier les transcrits différentiellement exprimés à partir de tables de comptage d'analyses RNAseq.

La majorité des méthodes font l'hypothèse que la distribution des comptages de chaque gène suit une loi binomiale négative, qui est une alternative robuste à la loi de Poisson lorsque la variance est plus importante que la moyenne. Les méthodes actuelles peuvent être divisées en méthodes paramétriques et non-paramétriques. Parmi les méthodes paramétriques, les outils les plus connus sont `edgeR` (Robinson et al., 2010a), `DESeq` (Anders and Huber, 2010) et `DESeq2` (Love et al., 2014a), `EBseq` (Leng et al., 2013), et `Cuffdiff2` (Trapnell et al., 2013). Les méthodes non-paramétriques sont moins utilisées, les outils les plus connus étant `SAMseq` (Li and Tibshirani, 2013) et `NOIseq` (Tarazona et al., 2011).

De nombreuses revues ont été consacrées à l'évaluation des performances des différentes méthodes, en particulier pour les méthodes les plus populaires comme `edgeR` et `DESeq2`. Du fait du faible écart de performance entre les méthodes, d'une étude comparative à la suivante, le classement des méthodes fluctue (Li et al., 2022; Liu et al., 2021; Schurch et al., 2016; Seyednasrollah et al., 2015; Soneson and Delorenzi, 2013). Cependant, toutes ces études soulignent le manque de spécificité des résultats, et des performances dégradées lorsque le nombre de réplicas est faible.

1.2 Intégration à l'échelle systémique

1.2.1 Biologie des systèmes

La **biologie des systèmes** peut être définie comme la modélisation informatique et mathématique des systèmes biologiques complexes. Elle se concentre sur les interactions au sein des systèmes biologiques, en utilisant une approche holistique plutôt que réductioniste, dans le but d'identifier des propriétés émergentes qu'on ne saurait identifier *via* une approche réductioniste. Ainsi, les **réseaux biologiques** sont par définition au fondement même de la biologie des systèmes, et leur mise en évidence est essentielle.

1.2.2 Différents réseaux biologiques

Le vivant s'étudiant et s'observant à différentes échelles, les réseaux biologiques se définissent également à plusieurs échelles. Ainsi, selon la sémantique associée aux entités et celle associée aux interactions entre entités, on peut définir différents type de réseaux, dont les plus fréquents sont :

- les réseaux de co-expression de gènes (entités : gènes, interactions : co-expression), qui peuvent être mis en évidence avec les méthodes populaires telles que `WGCNA` (Langfelder and Horvath, 2008) ou `PCIT` (Reverter and Chan, 2008) ;
- les réseaux de régulation de l'expression des gènes (entités : gènes, interactions : activation ou inhibition d'un gène par un autre), qui peuvent être mis en évidence par des approches expérimentales haut débit basées sur une immunoprécipitation de la chromatine suivie d'une hybridation sur puce (`ChIP-CHIP`, Lemmens et al., 2006) ou d'un séquençace (`ChIP-Seq`, Schmidt et al., 2009), ou de nombreuses approches bioinformatiques inférant des liens entre gènes à partir de données d'expression (pour une revue, voir Delgado and Gómez-Vela, 2019) ;
- les réseaux d'interactions protéine-protéine (entités : protéines, interactions : interaction physique entre 2 protéines), qui peuvent être mises en évidence par diverses approches expérimentales haut débit comme le double-hybride (Ito et al., 2001; Uetz et al., 2000) ou la purification par affinité en tandem couplée à la spectrométrie de masse (Gavin et al., 2002; Ho et al., 2002) ;
- les réseaux métaboliques (entités : protéines et métabolites, interactions : réactions biochimiques et contrôle de ces réactions), principalement construits par une analyse de la littérature ou des approches expérimentales bas-débit, ou reconstruits pour des espèces moins étudiées par analogie avec d'autres réseaux métaboliques (Belcour et al., 2020; Karimi et al., 2021).

Récemment, des réseaux hétérogènes ont été introduits, qui tentent de réconcilier les différentes échelles du vivant (figure 1.2) : les différents types de réseaux sont représentés chacun par un réseau monoplex, et les différentes couches sont mises en relation les unes avec les autres *via* des relations de couplage entre entités des différentes couches (Valdeolivas et al., 2019).

1.2.3 Structuration des connaissances sur les réseaux biologiques

Il existe plusieurs formalismes permettant de structurer les connaissances actuelles sur les réseaux biologiques, dont certains prennent en compte les différentes échelles du vivant. Comme souvent dans le domaine des sciences de la vie, l'information symbolique est structurée *via* l'utilisation d'**ontologies**. L'objectif premier d'une ontologie est d'expliciter et de modéliser un ensemble de connaissances dans un domaine donné en regroupant un ensemble de concepts décrivant complètement ce domaine. Ces concepts sont liés les uns aux autres par des relations taxinomiques (hiérarchisation des concepts) d'une part, et sémantiques d'autre part. Il existe des ontologies dédiées à la description des interactions entre entités biologiques.

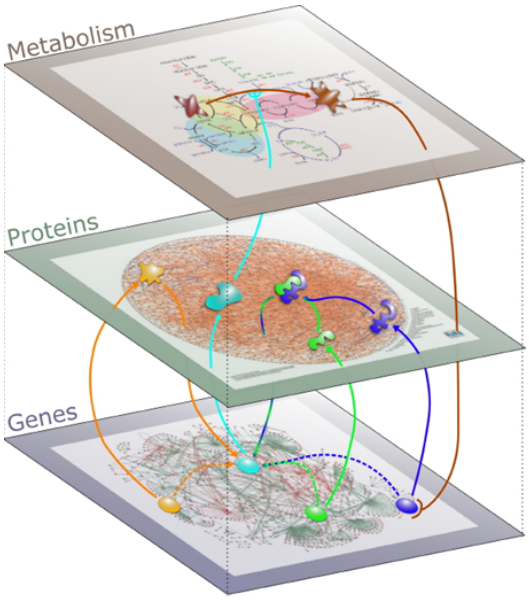


FIGURE 1.2: Représentation d'un réseau biologique hétérogène. La couche *Metabolism* représente un réseau métabolique, la couche *Proteins* représente un réseau d'interactions protéine-protéine, et la couche *Genes* représente un réseau de régulation génétique. Les arêtes reliants les différentes couches représentent les relations de couplage. Par exemple, un gène peut être transcrit en protéine, ou une protéine peut contrôler l'expression d'un gène ou catalyser une réaction chimique, ou encore la présence d'une molécule produite par une réaction chimique peut activer l'expression d'un gène, ce qui conduira à la production de la protéine correspondante.

Différents vocabulaires contrôlés et ontologies spécifiques, telles que `Gene Ontology (GO)` (Ashburner et al., 2000; The Gene Ontology Consortium, 2019) ou `Molecular Interactions (MI)` (Sivade Dumousseau et al., 2018) ont été introduites dans les années 2000-2010 et ont connu un succès considérable. Des ontologies plus transversales ont vu le jour, comme l'ontologie `BioPAX` (Demir et al., 2010b) ou l'ontologie `PSI-MI` (Sivade Dumousseau et al., 2018). D'autres formats ont été introduits pour des domaines plus spécifiques, comme `SBML` ou `CellML` pour la simulation de systèmes biologiques (Clerx et al., 2020; Keating et al., 2020), `SBGN` pour la visualisation (Le Novère et al., 2009)... La figure 1.3, extraite de Demir et al., 2010b, présente les interdépendances entre ces différents vocabulaires et formats en 2010. De nos jours, l'initiative `COMBINE` (pour *COmputational Modeling in BIology NEtwork*) poursuit ce travail de coordination, et d'autres formalismes comme `SBOL` (Baig et al., 2020) ou `NeuroML` (Gleeson et al., 2010) ont rejoint l'écosystème des formats en modélisation des systèmes biologiques. Les deux paragraphes suivants détaillent les ontologies `BioPAX` et `PSI-MI` qui jouent un rôle important dans certains travaux présentés dans ce manuscrit.

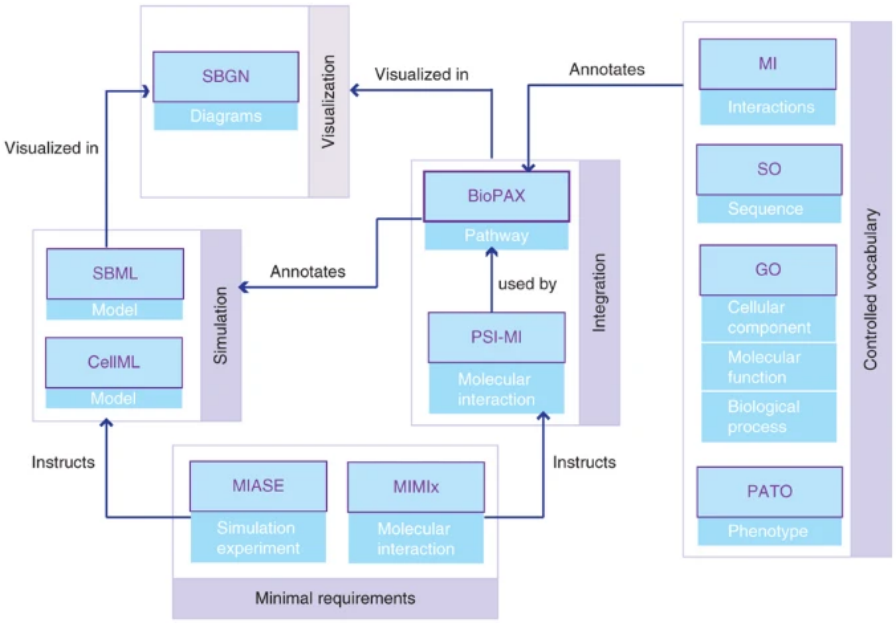


FIGURE 1.3: Vocabulaires contrôlés, ontologies et formats liés pour la visualisation et la simulation en biologie des systèmes. L'ontologie `BioPAX` est centrale à la structuration des connaissances en biologie des systèmes. Figure extraite de (Demir et al., 2010b).

L'ontologie `BioPAX`

L'ontologie d'échange de voies biologiques `BioPAX` [1] est un formalisme bien établi pour représenter les voies biologiques aux niveaux moléculaire et cellulaire, y compris les interactions (Demir et al., 2010b). Dans l'ontologie `BioPAX`, les quatre classes de premier niveau sont : `Pathway`, `Interaction`, `Physical Entity` et `Gene`. Deux classes sont particulièrement importantes.

- La classe `Interaction` représente les relations biologiques entre deux ou plusieurs entités (interactions moléculaires, contrôles et conversions).
- La classe `Physical Entity` englobe les petites molécules, les protéines, l'ADN, l'ARN et les complexes composé de ces éléments.

La figure 1.4-A illustre les différentes classes et sous-classes de l'ontologie `BioPAX`, et présente l'exemple d'un processus décrit suivant le formalisme de `BioPAX` (1.4-D).

Toutes les principales bases de données de référence sur les voies métaboliques sont disponibles au format `BioPAX` : `Reactome` (Gillespie et al., 2021), `BioCYC` (Karp et al., 2005; Romero et al., 2005), `PANTHER` (Mi et al., 2013; Thomas et al., 2022, 2003) ... et l'initiative *PathwayCommons* intègre 22 banques de données différentes et permet leur import au format `BioPAX` (Rodchenkov et al., 2020a). La librairie JAVA `Paxtools` est dédiée à l'extraction et la manipulation des fichiers au format `BioPAX` (Demir et al., 2013).

[1] http://www.biopax.org/release/biopax-level3-documentation.pdf

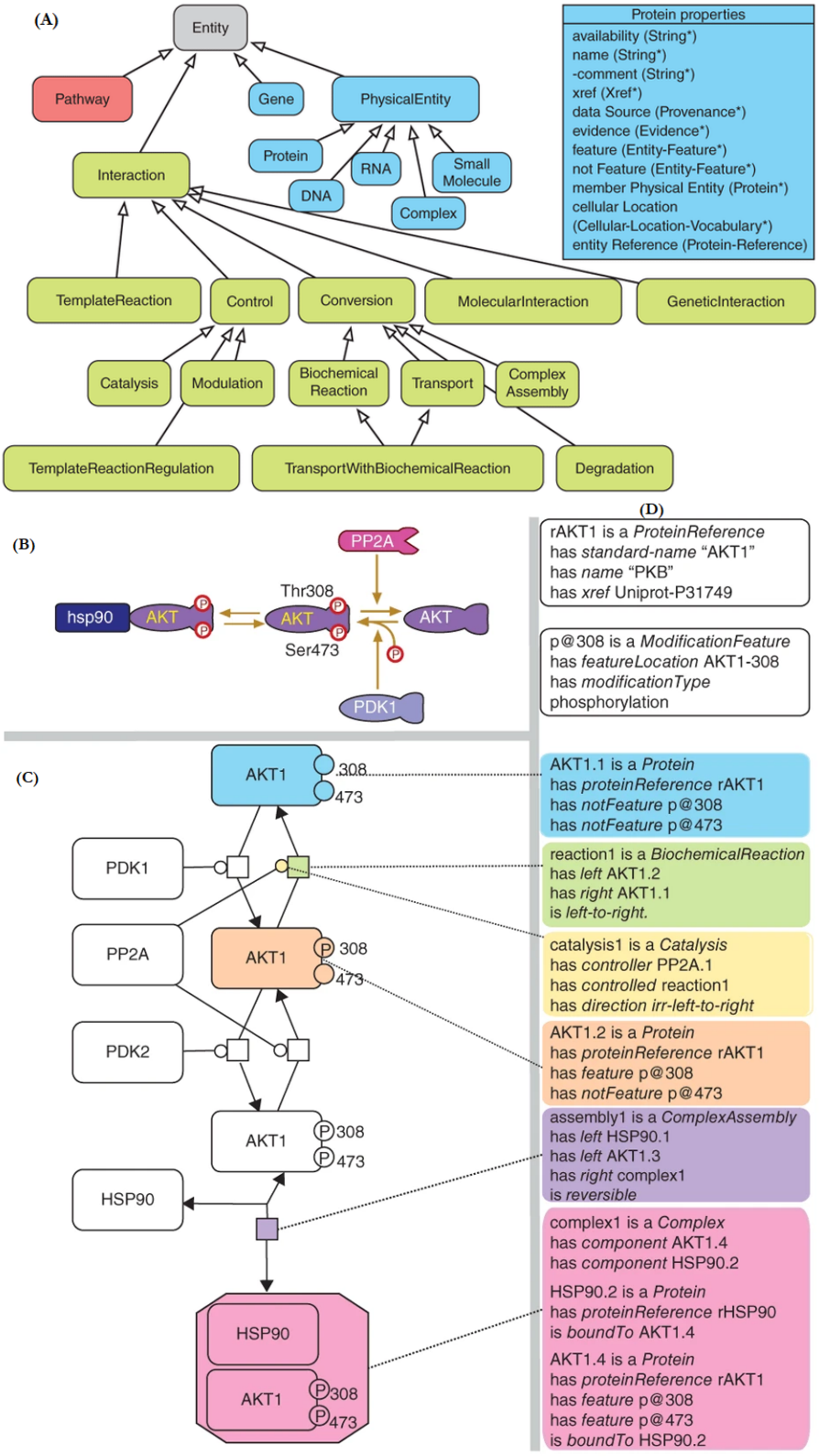


FIGURE 1.4: (A) Classes du format `BioPAX`. En vert, la classe `Interactions` et ses sous-classes. En bleu, les classes `Gene` et `PhysicalEntity` et ses sous-classes. En bleu à droite, les différentes propriétés d'une protéine. (B) La voie AKT visualisée avec Biocarta ; (C) représentation `SBGN` du même processus, et (D) formalisation correspondante avec le langage `BioPAX`. Figures extraite de (Demir et al., 2010b).

L'ontologie `PSI-MI`

Dans le cadre de l'initiative HUPO Proteomics Standards, une ontologie a été introduite pour unifier l'annotation des interactions protéine-protéine (Hermjakob et al., 2004; Kerrien et al., 2007; Sivade Dumousseau et al., 2018). La figure 1.5 présente un brève description des éléments de ce format. Parallèlement, un vocabulaire contrôlé structuré `MI`, pour *Molecular Interactions*, a été développé pour annoter de façon homogène les différentes techniques expérimentales permettant l'identification d'une interaction entre protéines.

Toutes les principales bases de données d'interactions entre protéines peuvent être téléchargées aux différents formats proposés par l'initiative HUPO Proteomics Standards (`miXML` ou `miTab`).

1.3 Le défi de la reproductibilité

Depuis les années 2000, plusieurs articles sont devenus célèbres pour avoir suscité la réflexion et le débat scientifique autour des questions de reproductibilité : « *Why Most Published Research Findings Are False* » (Ioannidis, 2005), « *1,500 scientists lift the lid on reproducibility* » (Baker, 2016), ou encore la couverture « *How science goes wrong* » de The Economist en 2013.

Dans son étude de 2016 (Baker, 2016), Monya Baker a notamment intérrogé 1 500 scientifiques de disciplines différentes. Parmi eux, 90% estiment que la science subit actuellement une crise de la reproductibilité, plus de 70% des chercheurs affirment avoir été incapables de reproduire l'expérience scientifique d'un autre chercheur et plus de la moitié affirment avoir échoué à reproduire l'une de leur propre expérience. D'après les scientifiques interrogés, cette crise de la reproductibilité est multi-factorielle et due à de mauvaises pratiques (reporting sélectif, faiblesses méthodologiques, fraude), un manque de transparence (non disponibilité des données brutes, des codes d'analyse), un manque de formation, d'encadrement ou de relecture, ou encore de facteurs extérieurs (variabilité des réactifs, manque de chance)... La pression de publication est également citée par plus de 80% des répondants comme facteur de non-reproductibilité.

De plus en plus de scientifiques cherchent ainsi à améliorer leurs pratiques pour améliorer la reproductibilité de leurs résultats. Cela passe par un meilleur accès aux données (initiative FAIR par exemple), aux codes (github), mais aussi par la collecte de nouveaux jeux de données afin de répondre aux mêmes questions scientifiques et par l'évaluation systématique et sincère des biais méthodologiques.

1.3.1 Champ lexical de la reproductibilité

Le champ lexical associé à la reproductibilité est vaste (reproductibilité, réplicabilité, répétabilité, robustesse, généralisabilité, réutilisabilité...) et la définition de certains terme a évolué au cours des dernières années.

Définitions de Claerbout et Karrenbach (1992)

La définition initiale de la **reproductibilité** a été introduite par Jon Claerbout et Martin Karrenbach en 1992, et est encore assez fréquemment utilisée dans la littérature

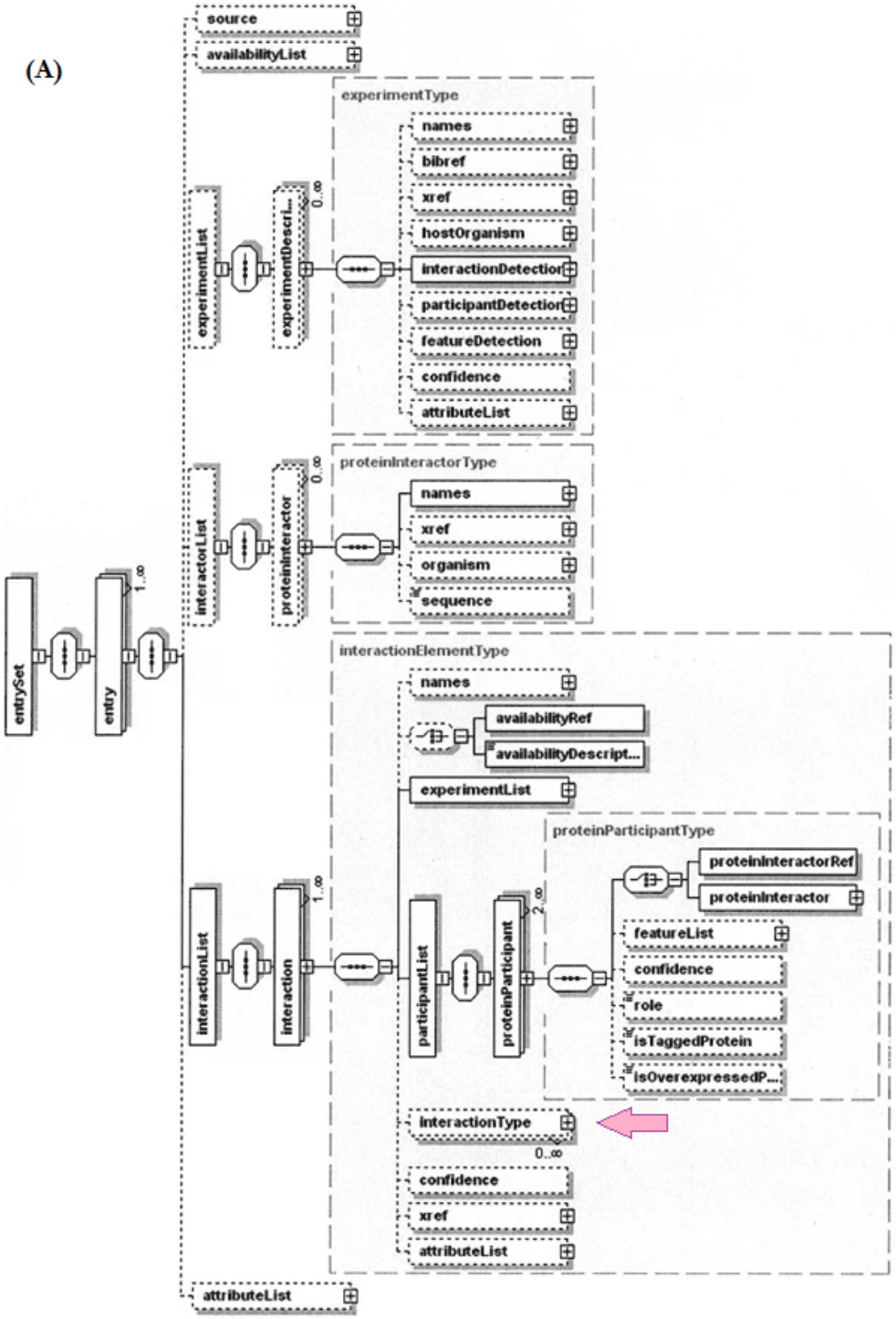


FIGURE 1.5: Format de représentation et d'échange des interactions entre protéines. Figures extraite de (Hermjakob et al., 2004). Le type d'interaction (flèche rose) est lié à un vocabulaire contrôlé hiérarchisé décrivant les méthodes expérimentales de détection des interactions protéine-protéine.

informatique (Claerbout J, 1992). Pour Jon Claerbout et Martin Karrenbach, la reproductibité d'une analyse est liée à la mise à disposition de tout le matériel (données + code) permettant de reproduire cette analyse : « *[...] an author attaches to every figure caption a pushbutton or a name tag usable to recalculate the figure from all its data, parameters and programs. This provides a concrete definition of reproducibility in computationaly oriented research.* ». L'enjeu principal est donc la transparence des analyses et des résultats.
Par la suite, ce concept a été repris et étendu par d'autres. Ainsi, en 2006, Roger Peng et ses collaborateurs introduisent la définition de la **réplicabilité** qui est liée à la collecte de nouvelles données pour traiter la même question, avec obtention de résultats cohérents : « *Scientific evidence is strengthened when important results are replicated by multiple independent investigators using independent data, analytical methods, laboratories and instruments* ». (Peng et al., 2006)

Définitions de l'Association for Computer Machinery (ACM)

En 2016, l'ACM (Association for Computer Machinery, association internationale dont la mission consiste à développer et soutenir la recherche scientifique et l'innovation informatique), a lancé une initiative autour de la reproductibilité [2], et a souhaité à cette occasion clarifier les différentes notions. Les différentes définitions retenues sont plus précises mais opposées aux définitions précédentes. Ainsi, pour l'ACM :

- **Reproductibilité** $\longrightarrow$ La mesure peut être obtenue avec la précision indiquée par une équipe différente, un système de mesure différent, dans un lieu différent et lors de plusieurs essais. Pour les expériences informatiques, cela signifie qu'un groupe indépendant peut obtenir le même résultat en utilisant des données et codes indépendants.
- **Réplicabilité**. $\longrightarrow$ La mesure peut être obtenue avec la précision indiquée par une équipe différente utilisant la même procédure de mesure, le même système de mesure, dans les mêmes conditions d'exploitation, au même endroit ou à un endroit différent, lors d'essais multiples. Pour les expériences informatiques, cela signifie qu'un groupe indépendant peut obtenir le même résultat en utilisant les données et codes de l'auteur.

Ainsi, les définitions retenues par l'ACM sont opposées aux définitions précédemment utilisées, et toutes les disciplines scientifiques ne se sont pas emparées de ces nouvelles définitions (Barba, 2018; Heroux et al., n.d.). L'évolution sémantique de ces termes au cours des dernières années reste très dépendante des disciplines, et des tensions autour de ces différents concepts existent. La table 1.1, adaptée de Barba, 2018, donne un aperçu de l'usage des différents termes selon les disciplines en 2018. On remarque que la bioinformatique suit principalement les définitions de Claerbout et Karrenbach, tandis que l'informatique suit celles de l'ACM.

Les différentes dimensions de la reproductibilité

Le projet **The Turing Way**[3] est un projet collaboratif ouvert dont l'objectif est de fournir toutes les informations dont les scientifiques ont besoin pour s'assurer que

[2] en associant des badges à certains articles publiés par la societé
[3] https://the-turing-way.netlify.app/welcome.html

C&K	ACM	Equiv.
Traitement du signal	Microbiologie, Immunologie	Sciences politiques
Calcul scientifique	Informatique	Economie
Econométrie		
Epidémiologie		
Etudes cliniques		
Médecine		
Physiologie		
Statistiques		
Bio-informatique		

TABLE 1.1: Utilisation des termes reproductibité et réplicabilité selon les disciplines, en 2018. Pour Claerbout et Karrenbach (**C&K**, colonne de gauche), pour l'ACM (**ACM**, colonne du milieu) ou pour certaines disciplines où les termes reproductibilité et réplicabilité sont utilisés de manière équivalente (colonne de droite).

les projets sur lesquels ils travaillent sont faciles à reproduire et à réutiliser (Community, 2022). Leurs définitions de la reproductibilité, sous la forme d'une table croisée, est parfaitement adaptée à la fois aux disciplines où la variabilité principale tient aux fluctuations d'échantillonnage et aux méthodes d'acquisition des données, ainsi qu'aux disciplines plus méthodologiques. Ainsi, ils différencient :

- **reproductibilité** $\longrightarrow$ Un résultat est reproductible lorsque les mêmes analyses effectuées sur les mêmes données produisent le même résultat.
- **réplicabilité** $\longrightarrow$ Un résultat est réplicable lorsque la même analyse effectuée sur différents ensembles de données produit des résultats qualitativement similaires.
- **robustesse** $\longrightarrow$ Un résultat est robuste lorsque le même ensemble de données est analysé par deux codes différents pour répondre à la même question de recherche et qu'une réponse qualitativement similaire ou identique est produite.
- **généralisabilité** $\longrightarrow$ La combinaison de résultats reproductibles et robustes nous permet de former des résultats généralisables.

Les définitions de The Turing Way sont donc plus proches de celles de Claerbout et Karrenbach que de celles de l'ACM, tout en étant plus précises et en permettant de différencier les différentes dimensions de la reproductibilité. Dans ce document, ce sont ces définitions qui seront utilisées.

1.3.2 Aspects liés au jeux de données

Dans le domaine de la bioinformatique, une première dimension des problèmes de reproductibilité est liée à la récolte de données indépendantes et à l'accès aux données elles-mêmes.

Aspects liés à la récolte de données indépendantes : réplicabilité

Historiquement, et avant l'informatisation des données et des analyses, la réplicabilité était un enjeu existant et lié à l'acquisition d'une seconde série de données,

totalement indépendante de la première série, et avec laquelle les analyses effectuées donnent des résultats cohérents. L'aquisition de cette seconde série de données, parfois appelée cohorte de réplication, est rendue nécessaire afin de minimiser les risques de premier et de second ordre liés à toute analyse statistique. Ces analyses comptent en effet des faux positifs (erreur de premier ordre), dont on cherche à contrôler le taux, mais aussi des faux négatifs (erreur de second ordre), plus ou moins nombreux selon la puissance du test statistique utilisé.

Les résultats obtenus sur une cohorte de réplication sont renforcés lorsque l'étude scientifique a fait l'objet d'un pré-enregistrement. Cela signifie qu'en amont de la collecte des données, le plan d'analyse complet (tailles de cohortes, seuils statistiques, paramètres des algorithmes...) a été établi et ne sera pas modifié lors de l'analyse. Ainsi, le pré-enregistrement des études permet de séparer clairement les études exploratoires, grâce auxquelles on génère de nouvelles hypothèses (*hypothesis-generating*), des études de testant des hypothèses précises (*hypothesis-testing*). Cette séparation claire des études selon leur objet est un élément important de réponse à la crise de la reproductibilité.

Aspects liés à l'accès aux données : reproductibilité

L'accès aux données brutes de la recherche, à des fins d'analyse de reproductibilité comme à des fins d'analyse automatique par des machines, a été considérablement amélioré par la promotion des principes FAIR. Ces principes sont issus de travaux collectifs de chercheurs, d'éditeurs, de sociétés savantes, d'universités, de bibliothécaires et d'archivistes, et sont publiés dans la revue Nature en 2016 (Wilkinson et al., 2016). Les quatre lettres du mot *fair* (juste, équitable) permettent de synthétiser ces principes :

- **F**indable : les données doivent être faciles à trouver, aussi bien par des humains que par des machines. Pour ce faire, les (méta)données doivent avoir un identificateur unique et pérenne et être enregistrées ou indexées dans un dispositif permettant de les rechercher.
- **A**ccessible : les données FAIR ne sont pas obligatoirement des données ouvertes mais doivent dans tous les cas être récupérables par leur identifiant en utilisant un protocole standard de communication, et dans tous les cas les (méta)données sont disponibles à des conditions connues, grâce à des licences claires.
- **I**nteroperable : l'interopérabilité implique l'utilisation de métadonnées contextuelles précises, et de contenu et de formats respectant les grands standards internationaux.
- **R**eusable : La réutilisation (libre, conditionnelle ou payante) doit être facilitée par l'utilisation de standards communs, grâce à des bases de données rassemblant des données vérifiées et bien décrites, directement (ré)utilisables pour la recherche ou d'autres usages.

1.3.3 Aspects liés aux analyses

Les problèmes de reproductibilité et de réplicabilité concernent également la recherche méthodologique, c'est à dire le développement et l'évaluation de nouvelles techniques d'analyse de données, même si la littérature à ce sujet est moins

abondante.
Dans le cadre de la bioinformatique, on trouve depuis 2010 quelques publications traitant de la non-réplicabilité des méthodes, que l'on définira comme la non-réplicabilité des bonnes performances de la méthode (Buchka et al., 2021; Jelizarow et al., 2010; Ullmann et al., 2022). Ce manque de réplicabilité est souvent dû à une évaluation trop optimiste (consciemment ou inconsciemment) de leur méthode par les auteurs de publication *via* l'optimisation du choix des caractéristiques des données, l'optimisation des paramètres des algorithmes, ou encore le choix des méthodes auxquelles se comparer (Jelizarow et al., 2010; Ullmann et al., 2022).

1.4 Aperçu général du manuscrit

Le manuscrit est organisé en 5 chapitres.

Dans le chapitre 2, je présenterai certains de mes travaux sur les données unimodales illustrant le continuum entre les analyses de profilage et la recherche d'une signature unimodale robuste, reproductible, et interprétable biologiquement.

Dans le chapitre 3, je présenterai une nouvelle approche pour l'identification de score de progression de maladie à partir de données multimodales. Cette nouvelle approche est particulièrement adaptée à la faible taille des cohortes, qui pose un problème aigu de dimensionnalité aux approches classiques. Pour intégrer les données multimodales, les interdépendances connues ou potentielles entre données ne sont pas modélisées.

Dans le chapitre 4, il sera également question d'intégration de données multimodales mais cette fois avec une approche systémique, c'est à dire en prenant en compte les relations connues entre entités de même nature ou de nature différentes. Les travaux présentés illustreront la complexité à extraire l'information pertinente des bases de données existantes, malgré un effort constant de la communauté bioinformatique à structurer les informations et les rendre disponibles, les ontologies et formats développés étant extrêmement complexes à manipuler.

Enfin, le chapitre 5 résumera les travaux présentés, et proposera des pistes de recherche à court terme, moyen terme et long terme en lien avec chaque thématique abordée.

Chapter 2

Données unimodales, du profilage à la recherche de signatures robustes

Les travaux présentés dans cette section concernent l'identification de signatures dans des données -omiques classiques, typiquement de type transcriptomique. Cela concerne principalement mon activité de recherche à l'IRSET, où j'étais en charge d'analyser des données -omiques. Ces travaux m'ont permi d'identifier des défis plus méthodologiques et ont donc ensuite alimenté un axe de recherche sur le volet informatique que j'ai développé à l'IRISA, visant à analyser la pertinence des signatures extraites sous plusieurs angles : pertinence biologique, reproductibilité et robustesse.

2.1 Profilage dans des données unimodales

Profilage de l'expression des transcrits (gènes, miARNs)

Parmi les technologies de séquençage développées depuis 10 ans, le séquençage de l'ARN reste le plus utilisé. On pourra prendre comme illustration de ceci les 13 466 citations[1] de l'article introduisant la technique (Mortazavi et al., 2008).
La technologie du RNA-seq permet de quantifier dans une condition donnée le niveau d'expression de chaque transcrit (Mortazavi et al., 2008; Wang et al., 2009b). Elle reflète l'état d'une population de cellules, et peut être utilisée pour comparer une population étudiée à une population contrôle. Par la suite, cette technologie a été adaptée pour étudier les miARNs (Hafner et al., 2008), ou plus récemment les transcrits à l'échelle de la cellule unique, ce qui ouvre de nouveaux champs d'application, notamment dans le domaine médical (Kolodziejczyk et al., 2015; Wagner et al., 2016).

Pour illustrer le *profilage* de l'expression des transcrits, je présente un cas d'étude typique, dans lequel j'ai investigué les effets d'une molécule utilisée en chimiothérapie, le 5-fluoruracil, dont les mécanismes moléculaires sont encore largement inconnus. Dans ce travail, dont le schéma général est présenté figure 2.1, on compare le niveau d'expression de transcrits, ici des longs ARNs codants et non-codants, chez la levure dans trois conditions : croissance de levures sur un milieu riche (YPD), croissance des levures sur un milieu riche contenant du 5-fluoruracil (YPD + 5FU), croissance de levures mutées pour un gène particulier (rrp6).

[1] chiffre au 03/10/2022 d'après Google Scholar

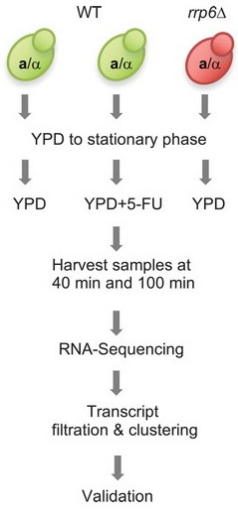


FIGURE 2.1: Schéma expérimental global. Trois conditions de levures diploïdes sont testées : milieu riche (YPD), milieu riche avec ajout de 5-fluoruracil (YPD+5FU), souches mutées pour le gène `rrp6`. Les prélèvements ont lieu à $t = 40$ min et $t = 100$ min. Un séquençage des transcrits est réalisé, suivi d'une analyse statistique permettant d'identifier les transcrits différentiellement exprimés.

Le pipeline de l'analyse de profilage est composé de trois étapes très classiques dont les concepts principaux sont présentés ci-dessous.

1. Mapping sur le génome de référence.
   Les lectures ont été mappées sur le génome de référence *S. cerevisiae* S288C en utilisant TopHat version 2.0.6 (Kim et al., 2013).

2. Analyse différentielle.
   Pour identifier les transcrits différentiellement exprimés, nous nous sommes concentrés sur les transcrits dont l'abondance n'était pas trop faible. Une telle étape de filtrage est recommandée car les analyses différentielles pour les transcrits ayant une très faible abondance ne sont pas solides (Bourgon et al., 2010; Rau et al., 2013). Les comptes ont été normalisés (Robinson and Oshlack, 2010), et l'analyse différentielle a été réalisée avec `edgeR` (Robinson et al., 2010a). L'expression différentielle a été déterminée entre les conditions et les points de temps. Les transcrits avec une valeur p corrigée inférieure à 0.05 ont été sélectionnés comme étant significativement différentiellement exprimés.

3. Clustering et enrichissement des termes de la Gene Ontology.
   Différentes stratégies ont été employées pour déterminer la méthode de clustering optimale. Ceci nous a permis d'identifier la meilleure approche et le nombre de clusters optimal (k = 5). L'enrichissement des clusters en termes Biological Process de Gene Ontology (GO) a été calculé à l'aide de GoTermFinder, et les termes GO dont la valeur p corrigée était inférieure à 0.01 ont été considérés comme significatifs.

Les résultats sont présentés classiquement sous forme de clusters de co-expression de transcrits différentiellement exprimés. Ainsi, au sein d'un cluster, on trouve des transcrits différentiellement exprimés dont les variations d'expressions ont le même

profil dans les différentes conditions. La figure 2.2 présente les cinq clusters identifiés. Dans cette analyse, on identifie 5697 transcrits différentiellement exprimés sur 9786, soit 58% de transcrits différentiellement exprimés. Ces résultats sont cohérents avec les résultats attendus, car la déplétion du gène `rrp6` ou l'ajout de 5-fluoruracil perturbent fortement le fonctionnement de la cellule.

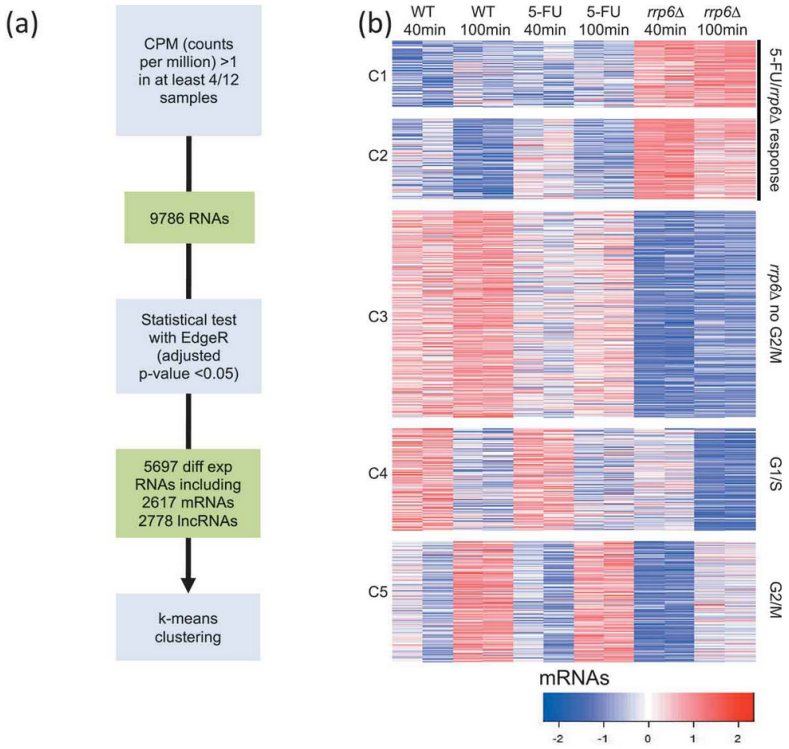


FIGURE 2.2: Schéma expérimental global d'une étude de profilage. Trois conditions de levures diploïdes sont testées : milieu riche (YPD), milieu riche avec ajout de 5-fluoruracil (YPD+5FU), souches mutées pour le gène `rrp6`. Les prélèvements ont lieu à $t = 40$ min et $t = 100$ min. Un séquençage des transcrits est réalisé, suivi d'une analyse statistique permettant d'identifier les transcrits différentiellement exprimés.

Parmi ces transcrits différentiellement exprimés, il n'est pas possible de distinguer ceux ayant trait au mécanisme de réponse au 5-FU de ceux participant à la réponse globale de la cellule en cas de stress. Le profilage de l'expression des gènes, parce qu'il privilégie la sensibilité de l'analyse, permet donc d'apprécier l'impact global à l'échelle de la cellule d'une condition donnée, mais ne permet pas d'en déduire facilement et finement les mécanismes sous-jacents. L'ensemble des transcrits différentiellement exprimés ne constitue pas ici une *signature*, qui impose une spécificité et un potentiel prédictif des variations quantitatives.

2.2 Remise en question des signatures unimodales sous différents angles

Le profilage de données -omiques, et plus particulièrement le profilage de l'expression des gènes, est un outil très largement utilisé pour caractériser un état particulier

et, *via* des clusterings et des analyses d'enrichissement, identifier de possibles mécanismes cellulaires liés. L'analyse permettant d'identifier les éléments du profil repose sur une analyse statistique d'approche fréquentiste, dans laquelle on sélectionne classiquement les éléments avec une valeur p corrigée inférieure à 0.05 (Love et al., 2014b; McCarthy et al., 2012; Robinson et al., 2010b). Cependant, le profilage de données -omiques reste largement à visée exploratoire, et les ensembles d'éléments -omiques identifiés comme différentiels doivent être consolidés pour aboutir à l'identification d'une signature.

2.2.1 Potentiel prédictif : du profilage au biomarqueur

Un premier angle de consolidation peut être l'investigation du potentiel prédictif des différents éléments, en utilisant des approches de classification à partir des éléments différentiels. L'objectif est de tester la capacité des éléments, seuls ou en combinaison, à discriminer les différents états étudiés. Cette investigation est rendue nécessaire car les analyses de profilage bénéficient de réplicas biologiques permettant d'estimer des distributions et d'identifier de légers écarts de centrage de ces distributions, alors que le potentiel prédictif repose sur une valeur unique (ou des valeurs uniques dans le cas d'ensemble d'éléments).

Dans le cadre du travail de doctorat de Virgilio Kmetzsch, co-encadré avec Olivier Colliot (équipe ARAMIS, INRIA et ICM), nous avons cherché à identifier une signature à partir de miARNs circulants qui permette de caractériser la progression de la démence front-temporale (DFT) et de la sclérose latérale amyotrophique (SLA), deux maladies neuro-dégénératives actuellement sans traitement. Il existe des formes familiales de ces maladies, liées à la transmission de mutations dans certains gènes connus. Les porteurs de ces mutations développeront tôt ou tard une DFT et/ou une SLA, sans que l'on puisse prédire l'évolution de la phase pré-symptomatique à la phase symptomatique. L'identification d'une signature serait donc importante pour mieux caractériser l'état des patients pré-symptomatiques, et évaluer l'impact de potentiels traitements préventifs. Les miARNs dans le plasma sanguin sont une cible de choix pour identifier une signature car :

1. leur prélèvement se fait par prise de sang, et est donc non-invasif ;
2. les miARNs sont exceptionnellement stables dans le plasma sanguin (Arroyo et al., 2011; Mitchell et al., 2008) ;
3. les formes sporadiques et familiales d'ALS peuvent être liées à des mutations de la protéine TDP-43, qui joue un rôle important dans la bio-génèse des miARNs (Sreedharan et al., 2008).

À l'aide de données issues d'une cohorte mise en place par Isabelle Le Ber (ICM et Hopital de Paris - Pitié Salpétrière) au sein du consortium PREV-DEMALS [2], Virgilio Kmetzsch a eu accès à des données de séquençage de miARNs prélévés dans le plasma sanguin de 22 patients, 46 individus pré-sympotomatiques, c'est-à-dire présentant la mutation caractéristique des patients mais n'ayant pas encore développé de symptomes, et 43 individus témoins issus des familles des patients.

[2] PREV-DEMALS (https:// clinicaltrials. gov/ Identifier:NCT02590276) est une étude multicentrique nationale centrée sur les porteurs de la mutation C9orf72. Entre 2015 et 2017, 111 personnes ont été suivies avec le même protocole dans quatre centres hospitaliers français (Paris, Limoges, Lille et Rouen).

Une analyse différentielle classique des miARNs a permi de mettre en évidence 4 miARNs différentiellement exprimés, comme illustré figure 2.3 [28].

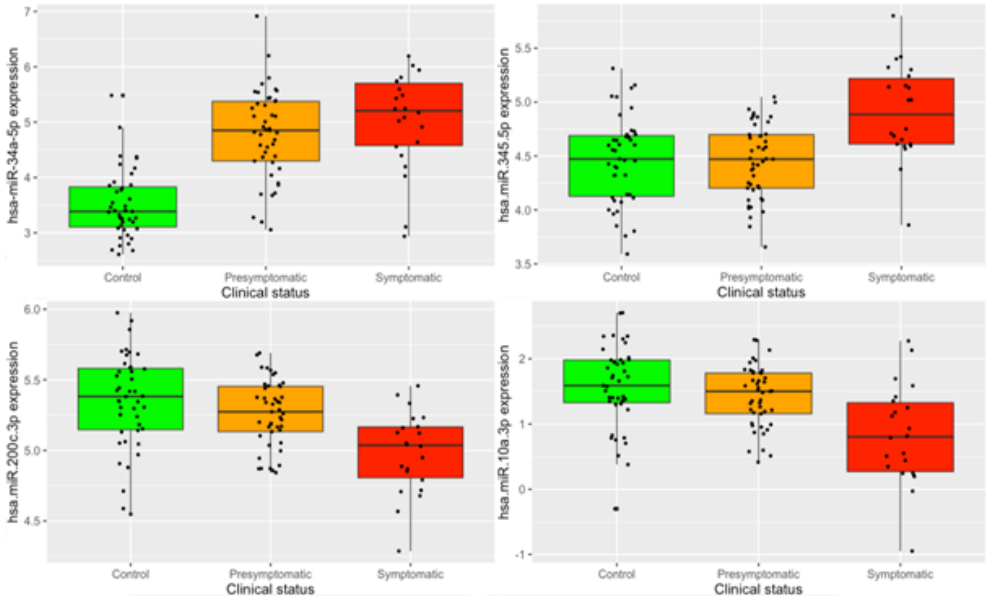


FIGURE 2.3: Boxplots avec les niveaux d'expression, pour chaque groupe clinique, des quatre miRNAs identifiés comme différentiellement exprimés. Il existe une différence claire dans l'expression du miR-34a-5p entre les contrôles et les individus mutés (présymptomatiques et symptomatiques). Figure extraite de [28].

Virgilio Kmetzsch a ensuite investigué le potentiel prédictif de ces quatre miARNs afin de déterminer s'ils pouvaient ou non être considérés comme une signature. Pour cela, il a implémenté des classifieurs basés sur une régression logistique. Les niveaux d'expression des 4 miRNAs de la signature ont été utilisés comme variables pour paramétrer trois modèles de classification binaire pour chaque comparaison : témoins *vs.* individus présymptomatiques, témoins *vs.* patients et individus présymptomatiques *vs.* patients. Dans le respect des bonnes pratiques (Poldrack et al., 2020), une stratégie de validation croisée stratifiée a été choisie (figure 2.4) : la boucle de validation croisée interne sert à déterminer l'hyperparamètre (coéfficient de régularisation L2), et la boucle de validation croisée externe permet d'évaluer la performance du modèle à l'aide de l'aire sous la courbe (ROC AUC [3]). Les résultats sont présentés sur la figure 2.5. Pour la classification des témoins *vs.* présymptomatiques, l'aire sous la courbe ROC est de 0.90 (IC 90% : $0.83 - 0.95$). Pour la classification des témoins *vs.* patients, l'aire sous la courbe ROC est également de 0.90 (IC 90% : $0.82 - 0.97$). Enin, pour la distinction entre pré-symptomatiques et patients, les résultats sont légèrements moins bons puisque l'aire sous la courbe ROC est de 0.80 (IC 90% : $0.67 - 0.90$).

Au final, l'ajout de cette étape de classification, postérieure à l'identification de gènes différentiellement exprimés, nous permet de valider le potentiel prédictif des 4 miARNs [28]. Cependant, l'analyse présente un biais potentiel, puisque l'identification des gènes différentiellement exprimés est faite sur le même ensemble

[3] area under the receiver operating characteristic curve

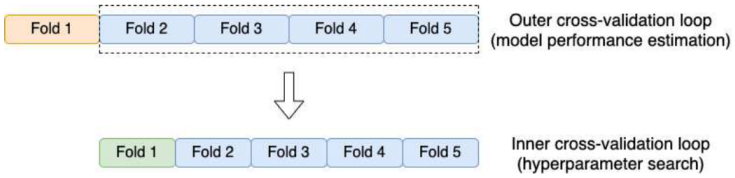


FIGURE 2.4: Schéma de validation croisée stratifiée, avec une validation croisée externe et interne à cinq folds. La boucle de validation croisée externe divise l'ensemble des données en cinq folds, utilisant quatre plis comme données d'apprentissage et un comme données de test (représenté en orange) à chacune de ses itérations. La boucle interne de validation croisée à 5 plis divise les données de formation en quatre plis pour le paramétrage du modèle et un pli de validation (représenté en vert) à chacune de ses itérations. Afin de préserver la proportion de témoins, de sujets présymptomatiques et de patients dans chaque fold, une stratification par rapport au statut clinique a été effectuée.

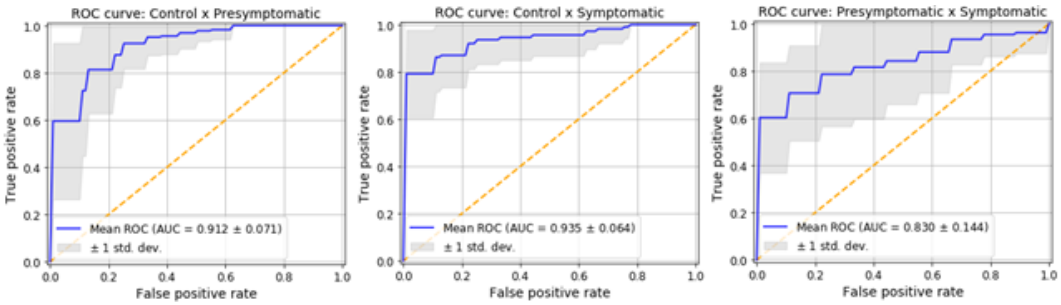


FIGURE 2.5: Courbes ROC (receiver operating characteristic) pour chaque classification deux-à-deux : témoins *vs.* pré-symptomatiques, témoins *vs.* patients et pré-symptomatiques *vs.* patients, obtenues *via* une régression logistique utilisant l'expression des 4 miARNs de la signature comme variables (miR-34a-5p, miR-345-5p, miR-200c-3p et miR-10a-3p). Les intervalles de confiance sont représntés en gris et ont été obtenus en identifiant le 5ème et le 95ème percentile à partir de 2000 échantillons bootstrap.

de données que celui utilisé pour valider le potentiel prédictif par classification. Ce point sera l'objet des sections 2.2.3 et 2.2.4.

2.2.2 Interprétabilité biologique

Le second angle de consolidation exploré concerne l'interprétabilité biologique des éléments identifiés. L'étude de cette interprétabilité est intéressante car la sensibilité des approches de profilage haut-débit fait qu'elles identifient des élements pouvant être directement liés au phénotype étudié, mais aussi plus largement des éléments très indirectement liés au phénotype d'intéret.

Dans les travaux présentés dans [28], nous avons investigué la possibilité d'utiliser les éléments de la signature pour estimer la progression d'une maladie chez des patients pré-symptomatiques, afin de tester l'hypothèse que la comparaison des valeurs d'expression d'un petit ensemble de miARNs chez des témoins,

pré-symptomatiques, et patients pouvait nous renseigner sur la progression de la maladie chez les pré-symptomatiques (Kmetzsch et al., 2021).

Ainsi, à partir de la signature de 4 miARNs identifiée précédemment (2.2.1), nous avons utilisé le classifieur témoins *vs.* patients paramétré précedemment (2.2.1), entraîné par régression logistique avec les niveaux d'expression des patients et des témoins, puis testé avec les niveaux d'expression des individus présymptomatiques. En effet, 4 individus parmi les pré-sympotmatiques avait été identifiés par les médecins en phase de transition entre le stade pré-symptomatique et le stade symptomatique (phase prodromique). L'objectif est ici d'évaluer si les scores obtenus par ces 4 individus prodromaux étaient tous supérieurs à 0.50, indiquant une plus grande similitude avec le groupe « patient ».
Les scores obtenus par les 4 individus prodromaux sont bien supérieurs à 0.50 : 0.54, 0.75, 0.80 et 0.82. Ce résultat positif doit toutefois être nuancé au vu de la distribution des scores de tous les sujets présymptomatiques ($n = 45$), puisqu'on constate qu'une large fraction de ceux-ci ont un score supérieur à 0.50. La distribution semble bi-modale (figure 2.6), avec un premier groupe plus proche des témoins ($n = 23$, score $s < 0.50$), et un groupe de taille équivalente plus proche des patients ($n = 22$, score $s > 0.60$). Ainsi, un suivi longitudinal serait nécessaire afin d'évaluer si les autres individus de score élevé développeront des symptômes dans un futur proche.

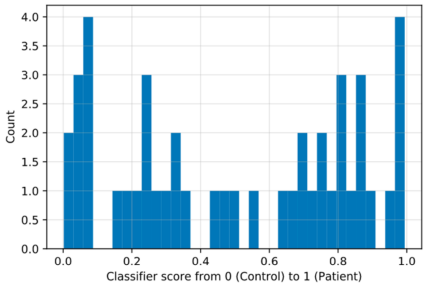


FIGURE 2.6: Scores de probabilités des pré-symptomatiques (n=45) calculés à partir de la régression logistique entre individis témoins et patients. Un score au voisinage de 0 signifie que le profil de miARNs de l'individu pré-symptomatique est proche de celui des témoins ; un score au voisinage de 1 signifie que le profil de miARNs de l'individu pré-symptomatique est proche de celui des patients.

2.2.3 Robustesse et généralisabilité des signatures identifiées

L'une des limitations identifiées dans la validation du potentiel prédictif (section 2.2.1) est que les miARNs différentiellement exprimés ont été calculés avec l'ensemble des données, de ce fait les folds de test de la validation croisée ont également été utilisés dans la sélection des variables de nos modèles de classification, ce qui peut améliorer les performances de prédiction. Pour estimer ce biais possible, nous avons incorporé la sélection de variable dans le processus de validation croisée stratifié : les miARNs différentiellement exprimés ont été calculés avec `edgeR` en utilisant uniquement les données d'entraînement de la boucle de validation croisée externe (quatre folds sur cinq) à chaque itération. La validation croisée emboîtée a été répétée 100 fois avec différentes divisions de folds. Les résultats sont présentés sur la figure 2.7.

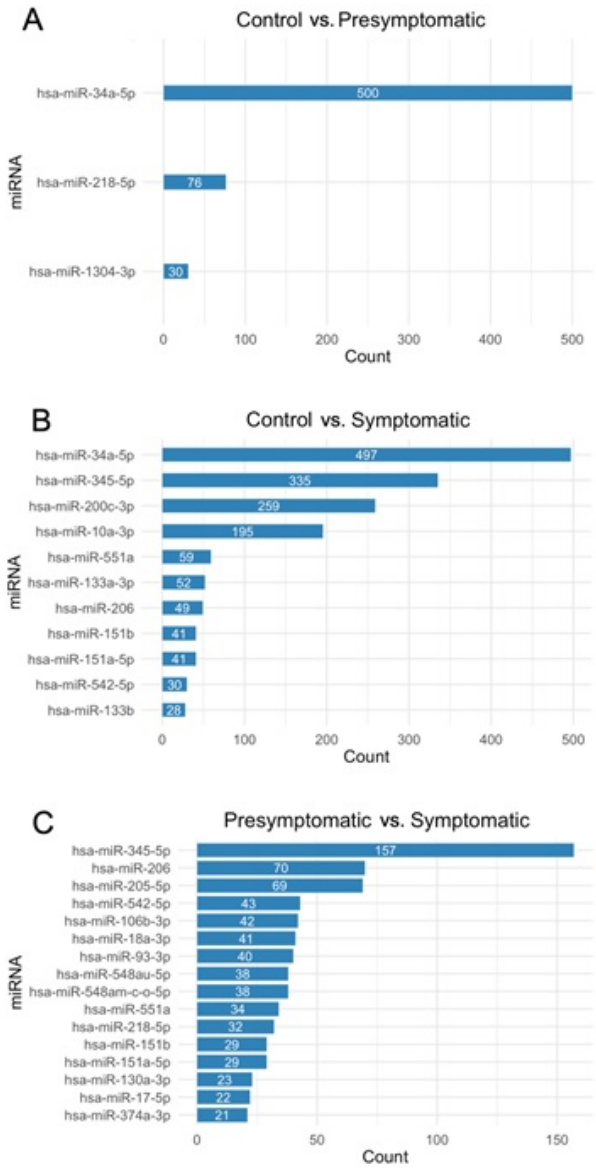


FIGURE 2.7: Nombre de fois où chaque miRNA a été trouvé différentiellement exprimé par `edgeR`, lors d'une validation croisée stratifiée répétée pour 100 divisions de folds différentes. Dans chaque étape de la boucle externe de validation croisée, quatre des cinq plis ont été utilisés pour identifier les miRNAs différentiellement exprimés. Étant donné qu'une boucle externe est constituée de 5 folds et que nous avons effectué 100 répétitions, 500 analyses différentielles ont été calculées pour chaque comparaison : contrôle vs. présymptomatique (A), contrôle vs. symptomatique (B), et présymptomatique vs. symptomatique (C).

On constate que mir-34a-5p est identifié comme différentiellement exprimé dans 100%(500/500) des comparaisons témoins *vs.* pré-symptomatiques, 99.4%(497/500) des comparaisons témoins *vs.* patients et 31.4%(157/500) des comparaison pré-symptomatique *vs.* patients. D'autres miARNs sont également fréquemment identifiés comme différentiels, certains faisant partie de notre signature identifiée sur le jeu de données complet comme mir-345-5p, mir-10a-3p, mir-200c-3p, mais d'autres n'ayant pas été identifiés comme tels sur le jeu de données complet comme mir-206

(2nd miARN trouvé le plus fréquemment différentiellement exprimé dans la comparaison pré-symptomatique *vs.* patients).

Le potentiel prédictif, mesuré par l'aire sous la courbe ROC, varie en fonction des comparaisons effectuées, comme présenté sur la figure 2.8. Comme attendu, les performances de classifications sont moins bonnes lorsque la sélection de variable est incorporée au processus de validation croisée stratifiée, ce qui confirme la présence d'un biais dans le calcul du potentiel prédictif décrit à la section 2.2.1.

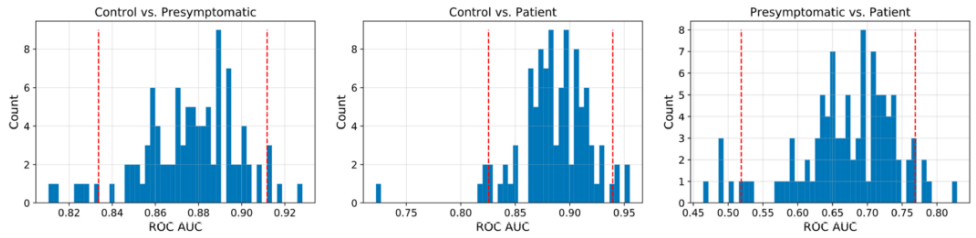


FIGURE 2.8: ROC AUC obtenue pour 100 divisions de fold différentes. Le calcul des variables prédictives (miARNs différentiellement exprimés) est fait uniquement sur 4 folds de la boucle de validation externe. Les lignes rouges indiquent le 5ème et le 95ème percentile, pour former un intervalle de confiance empirique à 90% : $[0.83, 0.91]$ pour la comparaison témoins *vs.* pré-symptomatiques, $[0.83, 0.94]$ pour la comparaison témoins *vs.* patients, et $[0.52, 0.77]$ pour la comparaison pré-symptomatiques *vs.* patients.

2.2.4 Réplicabilité des signatures identifiées

Enfin, le dernier angle de consolidation est l'angle de la réplicabilité, *ie.* la capacité, à partir d'un jeu de données indépendant, à produire des résultats cohérents avec les résultats de l'analyse exploratoire.

Dans le cadre du travail de doctorat de Virgilio Kmetzsch, après la première analyse ayant identifié une signature de 4 miARNs, nous avons fait le choix de tester dans deux cohortes indépendantes notre signature ainsi que toutes les signatures de miRNAs circulants identifiées dans la littérature pour la même pathologie. Afin de nous prémunir des biais potentiels, le protocole d'analyse complet à été préenregistré avant la collecte des données (https://osf.io/4pw8f). La publication associée à ce travail est [35].

Chacune des deux cohortes de validation est homogène et centrée sur les porteurs de mutation dans les gènes C9orf72 ou GRN. À cette fin, nous avons sélectionné toutes les études publiées qui ont identifié des miRNAs plasmatiques comme biomarqueurs potentiels de la DFT et/ou de la SLA par une requête Pubmed :

```
(microRNA[Title] OR microRNAs[Title] OR miR[Titre] OR miRNA[Title])
  AND
(serum[Title] OR circulating[Title] OR plasma[Title])
  AND
(ALS[Title] OR FTD[Title] OR amyotrophic[Title] OR frontotemporal[Title]
OR (neurodegenerative[Title] AND (frontotemporal[Title/Abstract]
OR amyotrophic[Title/Abstract])))
  NOT mice [Title/Abstract] NOT mouse [Title/Abstract]
```

```
NOT extracellular vesicles [Title]
NOT review [PT] NOT meta-analysis [PT] NOT (comment[PT])
```

Cette recherche a donné 19 résultats[4]. Deux articles ont été exclus parce qu'il s'agissait d'études de synthèse (Brennan et al., 2019; Grasso et al., 2015), un a été écarté parce qu'il était axé sur les niveaux de protéines (Freischmidt et al., 2021), et un a été exclu parce qu'il était axé sur un seul miARN provenant d'exosomes (Xu et al., 2018). Notre sélection finale comportait donc 15 articles, contenant 16 signatures suceptibles de concerner les porteurs de la mutation C9orf72, et 5 signatures susceptibles de concerner les porteurs de la mutation PGRN, certaines publications contenant plusieurs signatures, et certaines signatures pouvant s'appliquer aux deux cohortes (De Felice et al., 2014; Denk et al., 2018; Dobrowolny et al., 2021; Freischmidt et al., 2015, 2014; Grasso et al., 2019; Kmetzsch et al., 2021; Magen et al., 2021; Piscopo et al., 2018; Raheja et al., 2018; Sheinerman et al., 2017; Soliman et al., 2021; Takahashi et al., 2015; Tasca et al., 2016; Waller et al., 2017). Les résultats de ces différentes études se recoupent très peu, sans que l'on puisse déterminer si cela provient de fluctuations d'échantillonnage car les cohortes sont de faible taille, ou du fait que la majorité des études sont ciblées sur certains miARNs choisis après étude de la littérature. En effet, seules deux études ont testé tous les miARNs connus (Kmetzsch et al., 2021; Magen et al., 2021). Au total, 65 miARNs différents composent les 16 signatures testées pour la cohorte C9orf72 (DFT/ALS), et 30 miARNs composent les 5 signatures testées pour la cohorte PGRN (DFT uniquement). Les articles sélectionnés, les miARNs associés, les maladies, les types de cohortes, le nombre de patients et les méthodes d'analyse sont présentés dans le table 2.1, issue de [35].

Les résultats que nous présentons dans [35] montrent que la majorité des miARNs identifiés en lien avec l'ALS ou la DFT sont bien retrouvés différentiellement exprimés (35/65, p-valeur corrigée < 0.05), tandis que pour les miARNs spécifiquement en lien avec la DFT, les résultats sont moins bons (5/30, p-valeur corrigée < 0.05). De même, nous montrons que le potentiel discriminant des différentes signatures est plus important pour les signatures en lien avec ALS-DFT (2.9-gauche) que pour les signatures en lien avec la DFT (2.9-droite). Deux miARNs semblent particulièrement intéressants : mir-34a-5p, et mir-206. Il est intéressant de noter que ces 2 miARNs étaient parmi les mieux classés dans l'analyse de généralisation présentée dans la section 2.2.3.

Au delà de cet exemple particulier, cette étude de réplication souligne l'importance d'effectuer des études globales de réplication à partir de cohortes indépendantes. En effet, même dans le cas le plus favorable, une fraction importante des miARNs ne sont pas retrouvés différentiellement exprimés, et le potentiel prédictif des différentes signatures proposées est souvent décevant (< 0.5, c'est à dire moins bon que le pur hasard).

[4] recherche effectuée sur le portail PubMed (https://pubmed.ncbi.nlm.nih.gov/), le 10 mars 2022

TABLE 2.1: Selected studies investigating circulating microRNA expression (from serum or plasma) of patients with FTD or ALS.

Article	Disease	Cohort	Patients, No. (discovery/replication)	Presymtomatic carriers, No.	Method of analysis	Dysregulated miRNAs
(Grasso et al., 2019)	FTD	Sporadic	10/48 split of same cohort	-	qRT-PCR of 752 miRNAs	miR-663a, miR-502-3p, miR-206
(Piscopo et al., 2018)	FTD	Sporadic	54	-	qRT-PCR of 9 miRNAs linked with apoptosis	miR-127-3p
(Denk et al., 2018)	FTD	Sporadic	48	-	qRT-PCR of 96 miRNAs identified in preliminary study	let-7b-5p, let-7g-5p, miR-106a-5p, miR-106b-5p, miR-18b-5p, miR-223-3p, miR-26a-5p, miR-26b-5p, miR-301a-3p, miR-30b-5p, miR-146a-5p, miR-15a-5p, miR-22-3p, miR-320a, miR-320b, miR-92a-3p, miR-1246
(Kmetzsch et al., 2021)	FTD, ALS	Genetic (*C9orf72*)	22	45	RNA-sequencing of 2576 miRNAs	miR-34a-5p, miR-345-5p, miR-200c-3p, miR-10a-3p
(Sheinerman et al., 2017)	FTD, ALS	Unspecified	For each disease, 25/25 split of same cohort	-	qRT-PCR of 37 brain-enriched miRNAs	miR-9/let-7e, miR-7/miR-451, miR-335-5p/let-7e (FTD) and miR-206/miR-338-3p, miR-9/miR-129-3p, miR-335-5p/miR-338-3p (ALS)
(Magen et al., 2021)	ALS	Mixed sporadic and genetic (*C9orf72*)	126/122 split of same cohort	-	RNA-sequencing of 125 miRNAs identified in longitudinal study	miR-181a-5p, miR-181b-5p
(Soliman et al., 2021)	ALS	Mixed sporadic and genetic (unspecified mutation)	30	-	qRT-PCR of 7 miRNAs involved in ALS	miR-206, miR-143-3p, miR-142-3p
(Dobrowolny et al., 2021)	ALS	Mixed sporadic and genetic (unspecified mutation)	13/23	-	RNA-sequencing followed by qRT-PCR	miR-151a-5p, miR-199a-5p, miR-423–3p
(Raheja et al., 2018)	ALS	Mixed sporadic and genetic (*C9orf72*, SOD1)	23	-	qRT-PCR of 191 miRNAs identified on prior study	miR-29b-3p, miR-320c, miR-34a-5p, miR-29c-3p, miR-320a, miR-22-3p, miR-1, miR-133a-3p, miR-191-5p, miR-144-5p, miR-320b, miR-423-3p, miR-192-5p, miR-133b, miR-194-5p, miR-7-1-3p, miR-19a-3p, miR-425-5p, miR-145-5p, miR-144-3p
(Waller et al., 2017)	ALS	Sporadic	27/23	-	qRT-PCR of 750 miRNAs	miR-206, miR-143-3p, miR-374b-5p
(Tasca et al., 2016)	ALS	Sporadic	14	-	qRT-PCR of 9 muscle-specific, inflammatory, or angiogenic miRNAs	miR-206, miR-133a, miR-133b, miR-27a
(Takahashi et al., 2015)	ALS	Sporadic	16/48 split of same cohort	-	Microarrays, followed by qRT-PCR of 9 miRNAs	miR-4649-5p, miR-4299
(Freischmidt et al., 2015)	ALS	Sporadic	18/20	-	Microarrays of 1733 miRNAs, followed by qRT-PCR of 2 miRNAs	miR-1234-3p, miR-1825
(Freischmidt et al., 2014)	ALS	Separate sporadic and genetic (SOD1, FUS, *C9orf72*)	9/13 (genetic), 14 (sporadic)	18	Microarrays of 1733 miRNAs and qRT-PCR of 4 miRNAs	miR-4745-5p, miR-3665, miR-1915-3p, miR-4530
(De Felice et al., 2014)	ALS	Sporadic	10	-	qRT-PCR of 1 miRNA	miR-338-3p

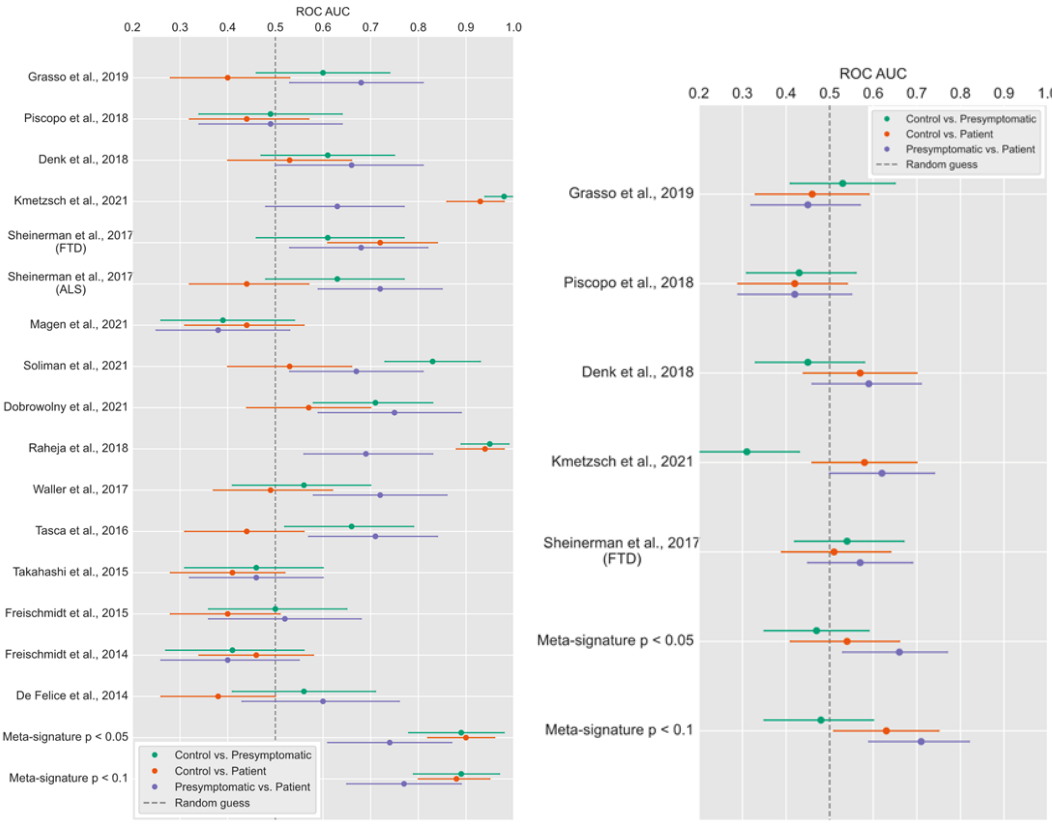


FIGURE 2.9: (gauche) Performances d'un ensemble de classifieurs (régression logistique) pour la cohorte C9orf72, en utilisant les 16 signatures de miRNA identifiées dans la littérature et deux méta-signatures obtenues à partir des analyses d'expression différentielle. (droite) Performances d'un ensemble de classifieurs (régression logistique) pour la cohorte PGRN, en utilisant les 5 signatures de miRNA identifiées dans la littérature et deux méta-signatures obtenues à partir des analyses d'expression différentielle. Les résultats de l'AUC ROC et les intervalles de confiance à 90% obtenus avec 2000 échantillons bootstrap sont calculés conformément à (Kmetzsch et al., 2021).

2.3 Ce que nous avons appris dans ce chapitre

Le *profilage* de données -omique reste largment exploratoire et les ensembles d'éléments -omiques identifiés comme différentiels doivent être consolidés en vue d'être utiles pour le diagnotic et de constituer une *signature*.

La consolidation de ces profilages passe par différents aspects :

1. l'investigation du potentiel prédictif des éléments du profil, seuls ou en combinaison, en prenant garde aux biais potentiels si l'évaluation du potentiel prédictif est faite sur le même ensemble que celui ayant identifié les éléments du profil ;

2. l'interprétatibilité biologique des élements identifiés ;

3. une consolidation statistique permettant de répondre au défi de la reproductibilité évoqué dans la section 1.3. Dans le meilleur des cas, une analyse préenregistrée et une cohorte de validation indépendante sont nécessaires pour

confirmer un résultat observé lors d'une étude exploratoire, et s'assurer ainsi de sa réplicabilité.

⟶ **Méthodes :** analyses différentielle dans des données homogènes (unimodales) ; plan d'analyses pré-enregistrés ; re-échantillonnage ; classification ; régressions.

⟶ **Domaines d'application :** biologie fondamentale (données de type transcriptomique /protéomique chez l'homme ou la levure) ; biologie marine (données homogènes : transcriptomique type miARNs) ; maladies neurodégénératives et scores de progression de maladie (données transcriptomiques type miARNs).

⟶ **Collaborations :** Mickaël Primig et Fatima Smagulova dans le cadre de mon travail à l'IRSET (IRSET, Inserm) ; Julien Bobe et Violette Thermes au Laboratoire de Physiologie et Génomique des Poissons (LPGP - INRAe) dans le cadre du projet européen PhénoMiR ; Olivier Colliot à l'Institut du Cerveau (ICM - Inserm, CNRS, AP-HP - Hôpital Pitié-Salpêtrière) et Inria (équipe-projet Aramis) dans le cadre de l'Inria Project Lab (IPL) Neuromarker.

⟶ **Travaux en lien :** [35], [28], [26], [20], [15], [14], [11].

⟶ **Publications jointes en support en Annexe C :**

[35] Kmetzsch V, Latouche M, Saracino D, Rinaldi D, Camuzat A, Gareau T, Le Ber I, Colliot O and **Becker E**.
Validation of circulating microRNA signatures as biomarkers in genetic frontotemporal dementia and amyotrophic lateral sclerosis.
accepted in Annals of Clinical and Translational Neurology

[28] Kmetzsch V, Anquetil V, Saracino D, Rinaldi D, Camuzat A, Gareau T, Jornea L, Forlani S, Couratier P, Wallon D, Pasquier F, Robil N, PREV-DEMALS study group; de la Grange P, Moszer I, Le Ber I, Colliot O, **Becker E.**
Plasma microRNA signature in presymptomatic and symptomatic subjects with C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis.
J Neurol Neurosurg Psychiatry, 2021.

Chapter 3

Intégration de données multimodales pour l'estimation de scores de progression de maladie

Les travaux présentés dans cette section concernent l'identification de signatures dans des données hétérogènes voire multimodales. L'hypothèse sous-jacente est que la combinaison de données hétérogènes/multimodales permet de prendre en compte plusieurs aspects de la réalité biologique et permet ainsi d'identifier des signatures de meilleure qualité qu'avec des données unimodales (chapitre 2). Cela concerne principalement mon activité de recherche dans le cadre d'une collaboration entre Dyliss et l'Institut du Cerveau (ICM - Paris, dans le cadre de l'IPL Neuromarker et du co-encadrement de la thèse de Virgilio Kmetzsch), visant à analyser conjointement des données de type -omique et des données d'imagerie médicale de type IRM/T1.

3.1 Intégration précoce *via* l'utilisation d'autoencodeurs variationnels

3.1.1 Scores de progression de maladie dans le cas de la DFT/SLA

Au cours des dernières années, de nombreuses approches ont été développées pour la modélisation de la progression des maladies à partir de données, comme les modèles basés sur les événements (EBM) (Fonteijn et al., 2012; Venkatraghavan et al., 2019), différents algorithmes basés sur des régressions logistiques de trajectoires des biomarqueurs (Jedynak et al., 2012; Mehdipour Ghazi et al., 2021), les modèles non linéaires à effets mixtes (Koval et al., 2021; Schiratti et al., 2017), les réseaux de neurones récurrents (Mehdipour Ghazi et al., 2019), et bien d'autres encore (Aksman et al., 2019; Garbarino et al., 2019; Lorenzi et al., 2019; Marinescu et al., 2019).
La plupart de ces approches nécessitent des données longitudinales. En effet, parmi les modèles proposés, seuls les modèles basés sur les évènements (EBM) permettent d'inférer un score de progression de maladie à partir de données transversales (Fonteijn et al., 2012; Venkatraghavan et al., 2019). Ces modèles explorent la séquence temporelle dans laquelle les biomarqueurs deviennent anormaux au cours d'une maladie. Ils ont été appliqués avec succès à diverses maladies, notamment la maladie d'Alzheimer : (Archetti et al., 2019; Firth et al., 2020; Fonteijn et al., 2012; Oxtoby et al., 2018; Venkatraghavan et al., 2019; Young et al., 2014), la sclérose en plaques : (Dekker et al., 2020; Eshaghi et al., 2018), la maladie de Parkinson :(Oxtoby et al., 2021), la maladie de Huntington :(Wijeratne et al., 2021) ainsi que la DFT : (Ende et al., 2021; Panman et al., 2021) et la SLA : (Gabel et al., 2020). Cependant,

dans ces travaux, les EBMs ont été appliqués à un nombre relativement faible de variables (généralement entre 10 et 50) et on ignore si ils seraient performants dans des dimensions plus élevées.

Malgré l'importance reconnue de l'estimation de la progression des maladies neurodégénératives, la recherche s'est concentrée principalement sur les pathologies de prévalence élevée. Les solutions existantes sont donc inadéquates pour modéliser les maladies rares avec des données transversales de grande dimension, pour trois raisons principales.

1. Des données longitudinales sont nécessaires pour la grande majorité des approches. Cependant, certaines maladies neurodégénératives comme la DFT et la SLA sont des maladies à évolution lente dans la phase présymptomatique, ce qui empêche la collecte de données longitudinales significatives.

2. La plupart des méthodes publiées bénéficient de grands échantillons, qui ne sont pas disponibles pour les pathologie de faible prévalence.

3. Enfin, il n'est pas évident que les modèles basés sur les événements, seules méthodes adaptées aux données transversales, puissent être appliqués de manière robuste aux données d'expression de miARNs de grande dimension, qui comprennent des centaines de biomarqueurs.

3.1.2 Nouvelle approche proposée à partir d'autoencodeurs variationnels

Dans le cadre du travail de doctorat de Virgilio Kmetzsch, co-encadré avec Olivier Colliot (équipe INRIA Aramis et ICM), nous avons proposé une méthode d'inférence d'un score de progression de la maladie (un trait latent) basé sur des autoencodeurs variationnel (VAE) non-supervisés [31] puis supervisés [34], entraînés avec des données de neuroimagerie et de miARNs et adaptée à une dimensionnalité défavorable des données : beaucoup de variables et peu d'individus.

Notre étude a porté sur 110 individus, répartis en trois groupes : 22 patients DFT/SLA, 45 pré-symptomatiques et 43 témoins. Tous les individus disposaient de données transcriptomiques constituées des niveaux d'expression de 589 miRNAs. Cependant 91 personnes seulement (14 patients, 40 pré-symptomatiques et 37 témoins) disposaient également de données de neuro-imagerie, consistant en des volumes de matière grise extraits d'une IRM incluant 68 régions corticales d'intérêt (ROI) et 18 ROI sous-corticales, ainsi que le volume intracrânien total estimé, soit 87 caractéristiques d'imagerie. Le nombre total de variables mesurées par patient est donc de 589 (miARNs) + 87 (IRM) pour les individus avec données d'imagerie, et 589 (miARNs) pour les individus sans données d'imagerie. La DFT et l'ALS étant des pathologies à évolution lente, nous ne disposons pas de données longitudinales.

Afin de palier à la faible taille de la cohorte, le formalisme des VAEs a été sélectionné, car les VAEs sont de puissants modèles génératifs qui projettent les données dans un espace latent régularisé de faible dimension, et il a été démontré qu'ils sont efficaces dans des contextes de haute dimension et de faible taille d'échantillon (Chadebec et al., 2021) . Ces modèles ont déjà été utilisés avec des données multimodales (Antelmi et al., 2019), mais pas dans le but de déduire un score de progression de maladie (DPS pour disease progression score). Virgilio Kmetzsch a proposé deux modèles de VAEs : un modèle non supervisé dans [31], puis un modèle supervisé dans [34]. La figure 3.1 présente l'approche retenue, en trois étapes :

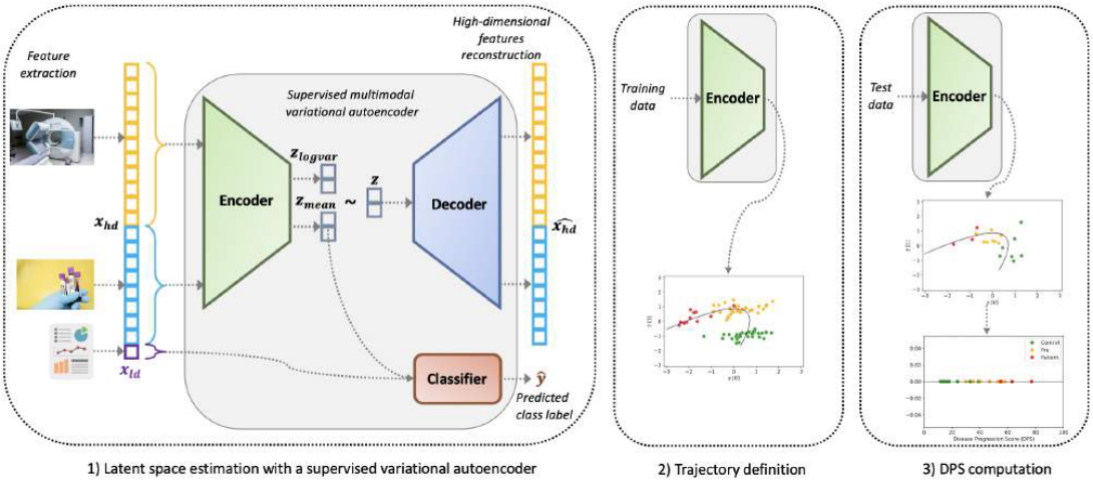


FIGURE 3.1: Nouvelle méthode proposée pour le calcul d'un score de progression de la maladie. (1) La première étape consiste à paramétrer avec un ensemble d'apprentissage un autoencodeur variationnel pour déterminer un espace latent de faible dimension. Cette étape peut inclure une branche de supervision, comme proposé dans [34], ou non, comme proposé dans [31]. (2) Définition d'une trajectoire dans l'espace latent, à l'aide des coordonnées des individus de l'ensemble d'apprentissage. Dans [31], la trajectoire est une ligne droite passant par les centroïdes des groupes témoins et patients, dans [34] c'est une courbe principale définissant une trajectoire témoins $\longrightarrow$ présymptomatiques $\longrightarrow$ patients. (3) Pour les individus de l'ensemble de test, calcul de leur coordonnées dans l'espace latent puis projection orthogonale de celles-ci sur la trajectoire précédemment déterminée afin de produire un score de progression de la maladie.

1. Etape 1, estimation d'un espace latent de faible dimension à partir d'un VAE.
   L'architecture d'un auto-encodeur variationnel est composée d'un encodeur probabiliste et d'un décodeur probabiliste. L'encodeur projette une donnée d'entrée x dans un vecteur de moyennes et un vecteur de log-variances, qui paramétrisent une distribution Gaussienne permettant d'obtenir la représentation latente z de x. A l'inverse, le décodeur renvoie la représentation latente z dans l'espace initial. Dans [31,34], nous avons utilisé un espace latent de dimension 2. La dimension de l'espace initial varie selon que l'on introduit une étape préalable de sélection de variables ou non (voir 3.2.1) (de 155 à 677). Enfin, dans [34], nous avons introduit une branche de supervision de l'encodeur à cette étape (voir 3.2.2).

2. Etape 2, définition de la trajectoire.
   Une fois le modèle entraîné, tous les indidividus de l'ensemble d'apprentissage sont encodés dans l'espace latent. Une première approximation naïve de trajectoire de la maladie peut être obtenue par une ligne droite passant par le centroïde des groupes témoins et patients, comme nous l'avons fait dans [31]. Par la suite, ce modèle a été raffiné et cette ligne droite a été utilisée en initialisation d'une recherche de courbe principale définissant une trajectoire témoins $\longrightarrow$ présymptomatiques $\longrightarrow$ patients [34]. Une courbe principale est

une courbe unidimensionnelle lisse passant par le milieu de groupes de points donnés, et peut être calculée par l'algorithme décrit dans Hastie and Stuetzle, 1989.

3. Etape 3, calcul du score de progression de la maladie (DPS).
   Les individus de l'ensemble de test sont encodés dans l'espace latent, et la projection orthogonale de leur représentation latente sur la trajectoire calculée précédemment permet de déduire le score associé à l'individu.

3.2 Améliorations du modèle

3.2.1 Sélection *a priori* de variables

Certains sujets ne disposent que de données transcriptomiques (n=19), sans données d'imagerie. Dans `[31]`, nous avons montré que ces sujets pouvaient être utilisés afin de sélectionner les variables de transcriptomique les plus intéressantes (feature selection), et que cette sélection améliore les performances du VAEs. Dans cette étude, les sujets ont été divisés en deux ensembles de données : 19 sujets avec uniquement des données de miARNs, utilisés comme ensemble de découverte pour la sélection de variables (analyse différentielle, 68 miARNs avec les p-valeurs les plus faibles sélectionnés), et 91 sujets avec des données multimodales de neuro-imagerie et de miARNs, utilisés comme entrées pour le VAE.

Cette sélection *a priori* de variables améliore la performance du VAE , ce qui est illustré sur la figure 3.2, où on constate une meilleure séparation des distributions des scores inférés pour les trois groupes composant notre cohorte (patients, pré-symptomatiques et témoins), en comparant les performances du modèle dans le cas où les données de miARNs sont de dimension $n = 589$ (gauche) où $n = 68$ après sélection de certaines variables sur un ensemble indépendant de 19 individus (droite).

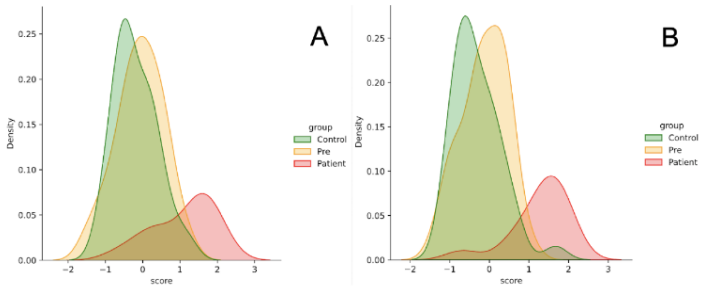


FIGURE 3.2: Distribution des scores de progression de la maladie en fonction du statut clinique, obtenues par validation croisée 5-fold avec 91 individus, en utilisant (A) les 589 miRNAs et les 87 ROIs de l'imagerie T1 et (B) 68 miRNA sélectionnés et les 87 ROIs de l'imagerie T1.

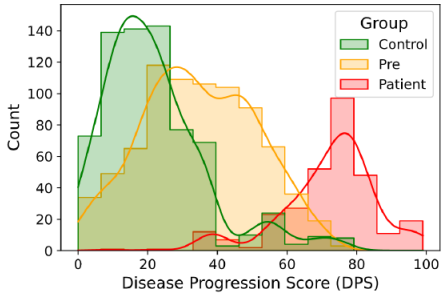


FIGURE 3.3: Distribution des scores de progression de la maladie en fonction du statut clinique, obtenues par validation croisée 5-fold avec 91 individus après introduction d'une branche de supervision, en utilisant 68 miRNA sélectionnés, 87 ROIs de l'imagerie T1, ainsi que l'âge (pour l'encodeur) et le statut clinique (pour le classifieur).

3.2.2 Introduction d'une branche de classification pour superviser le VAE

Dans [34], nous introduisons une branche de classification à l'encodeur, de façon à exploiter notre connaissance qualitative de l'état clinique des patients (témoin, pré-symptomatique ou patient). Brièvement, lors du paramétrage du VAE, nous cherchons à minimiser l'erreur définie par :

$$\mathcal{L} = \alpha_1 \cdot \mathcal{L}_r(x, \hat{x}) + \alpha_2 \cdot \mathcal{L}_{KL}(q_\phi(z|x), p(z)) + \alpha_3 \cdot \mathcal{L}_c(y, \hat{y}),$$

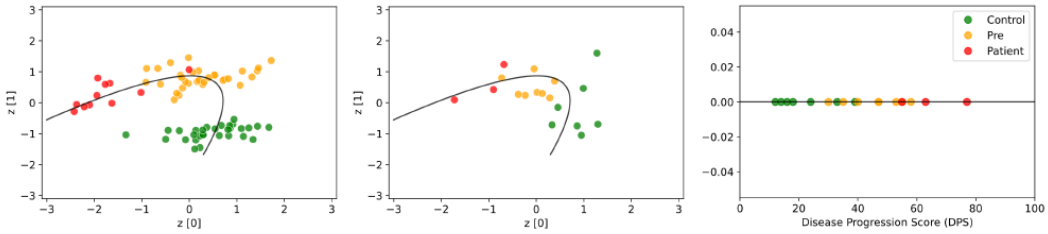
dont le premier terme $\mathcal{L}_r$ représente l'erreur de reconstruction, le second terme $\mathcal{L}_{KL}$ $\mathcal{L}_{KL}$ est la divergence de Kullback-Leibler, communément utilisée pour mesurer la différence entre deux distributions de probabilités, et $\mathcal{L}_c$ est un terme de cross-entropie qui pénalise les erreurs de classification. Les deux premiers termes sont classiques des autoencodeurs variationnels, tandis que le dernier terme est ajouté spécifiquement afin de pouvoir superviser notre VAE [34]. Chaque terme est contrôlé par un hyperparamètre α_k du modèle, de façon à ce que $\sum_{k=1}^{3} \alpha_k = 1$. Les hyperparamètres ont été fixés à $\alpha_1 = 0.2$, $\alpha_2 = 0.2$ et $\alpha_3 = 0.6$. Toutefois, nous montrons dans [34] que les résultats sont relativement stables lorsqu'on modifie la valeur des hyperparamètres (différentes combinaisons testées avec $0.1 \leq \alpha_1 \leq 0.3$, $0.1 \leq \alpha_2 \leq 0.3$ et $0.5 \leq \alpha_3 \leq 0.8$).

L'introduction de cette branche de supervision améliore la performance du VAE, ce qui est illustré sur la figure 3.3, où on constate une meilleure séparation des distributions des scores inférés pour les trois groupes composant notre cohorte (patients, prés-symptomatiques et témoins).

3.2.3 Performances du modèle proposé

Dans [34], les performances du nouveau modèle proposé ont été calculées à partir des données de la cohorte DFT/SLA sur une centaine de folds randomisés ($n = 91$ individus avec des données de miARNs et imagerie). Pour chaque fold, les données de 73 individus sont utilisées pour paramétrer le modèle, et les données de 18 individus sont utilisées pour prédire un score de progression de maladie (figure 3.4).

Afin d'évaluer la pertinence des score prédits, et en l'absence de vérité terrain, nous avons proposé d'évaluer les performances du DPS prédit en l'utilisant pour



(a) Training data projected in the latent space (b) Test data projected in the latent space (c) Computed scores for the test data

FIGURE 3.4: Les données de 73 individus sont utilisées pour paramétrer le modèle, et les données de 18 individus sont utilisés pour prédire un score de progression de maladie. (a) Coordonnées des 73 points de l'ensemble d'apprentissage dans l'espace latent et la courbe principale suivant la trajectoire témoins, prés-symptomatiques, patients. (b) Coordonnées des 18 points de l'ensemble de test, et la courbe principale précédemment déterminée. (c) Projection orthogonale des coordonnées des 18 points de l'ensemble de test sur la courbe principale pour le calcul du DPS.

faire des comparaisons 2 à 2 entre les différents groupes et en calculant l'aire sous la courbe ROC pour ces différentes comparaisons. Les courbes ROC sont présentées figure 3.5.

Les résultats obtenus sont très bons pour séparer les patients des deux autres groupes : $AUC = 0.98 \pm 0.05$ pour la comparaison témoins *vs.* patients, et $AUC = 0.96 \pm 0.07$ pour la comparaison prés-symptomatiques *vs.* patients. La comparaison entre témoins et pré-symptomatiques donne des résultats acceptables : $AUC = 0.73 \pm 0.13$. Ces résultats sont meilleurs que ceux obtenus en appliquant une stratégie similaire à partir d'un score de progression de maladie construit par un modèle basé sur des évènements (EBMs), seule autre stratégie possible en l'absence de données longitudinales (figure 3.5). Les différences de performances sont importantes, mesurée en variation de l'aire sous la courbe ROC : -0.07 pour la comparaison témoins *pré-symptomatiques*, -0.09 pour la comparaison témoins *vs.* patients, et -0.11 pour la comparaison pré-symptomatiques *vs.* patients.

3.2.4 Intérêts des différentes modalités

Afin d'évaluer l'intérêt de chacune des modalités : miARNs ($n = 589$ ou $n = 68$, selon que l'on applique une sélection de variables ou non), imagerie médicale ($n = 87$) et donnée démographique ($n = 0$ dans `[31]` et $n = 1$ lorsqu'on utilise l'âge dans `[34]`), nous avons évalué séparément chacune des modalités dans le modèle `[34]`. Les résultats sont présentés dans la table 3.1.

Brièvement, on note que les miARNs sont la modalité la plus informative pour séparer les témoins des pré-symptomatiques, tandis que les données de neuro-imagerie sont les plus informatives pour séparer les pré-symptomatiques des patients. Les données démographiques sont ici réduites à une seule variable, l'âge. On notera que le groupe des patients dans la cohorte étudiée est sensiblement plus âgé que les deux autres groupes (voir les paramètres démographiques dans `[31]`), d'où les résultats corrects de l'âge pour séparer les patients des deux autres groupes. Dans

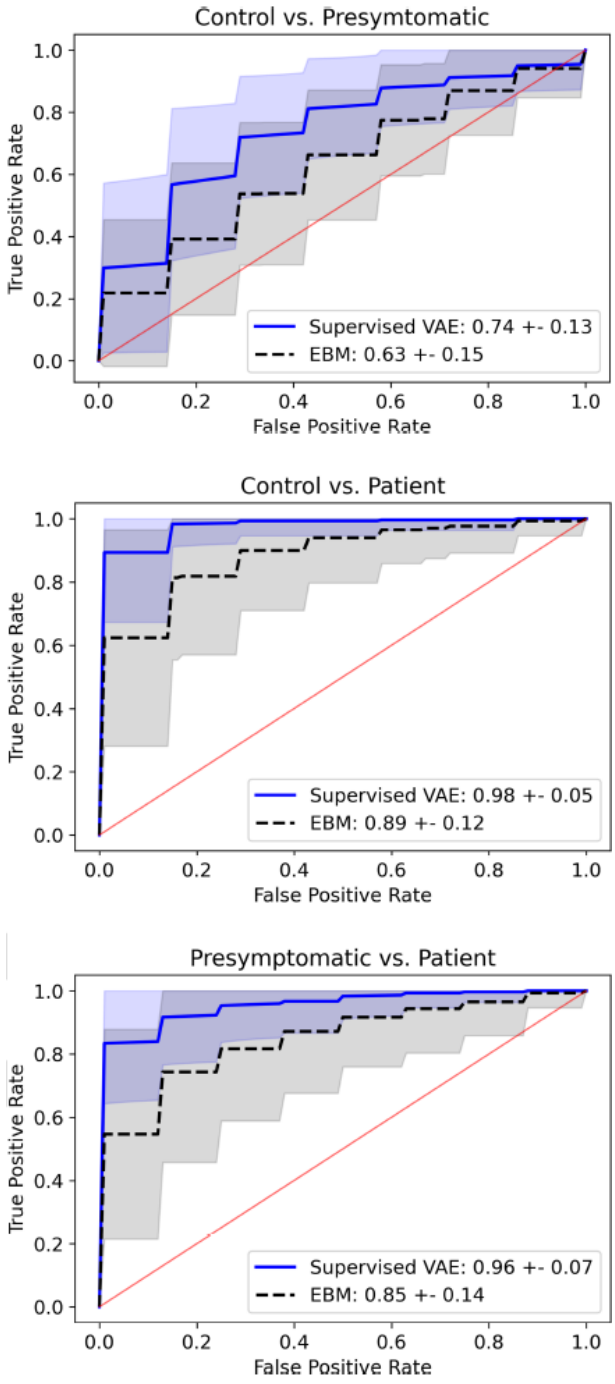


FIGURE 3.5: Courbes ROC (receiver operating characteristic) moyennes pour les comparaisons 2 à 2 entre groupes (100 splits aléatoires), pour notre VAE supervisé (courbe bleue) et pour un modèle de type EBM (pointillés noirs). Les zones colorées correspondent à une déviation standard. La ligne rouge indique le hasard.

l'ensemble, les différentes modalités semblent donc capturer des aspects complémentaires du score de progression global.

Comparaison	miRNAs	neuroimagerie	démographie (âge)
Témoin *vs.* Pré-sympto	0.73 ± 0.12	0.60 ± 0.15	0.40 ± 0.13
Témoin *vs.* Patient	0.81 ± 0.18	0.97 ± 0.04	0.81 ± 0.14
Pré-sympto *vs.* Patient	0.67 ± 0.19	0.96 ± 0.05	0.91 ± 0.09

TABLE 3.1: Moyenne et écart-type de l'aire sous la courbe ROC (cohorte DFT/ALS), calculés à partir de 100 folds différents, pour évaluer l'impact de chaque modalité unique.

3.3 Ce que nous avons appris dans ce chapitre

Il existe un besoin de définir des scores de progression de maladie qui soient adaptés aux maladies de faible prévalence et d'évolution lente. La faible prévalence limite la taille des cohortes étudiées, tandis que l'évolution lente rend le recueil de données longitudinales pour alimenter les cohortes irréaliste.

Nous avons proposé un nouveau modèle permettant d'utiliser des données multimodales transversales basé sur des auto-encodeurs variationels en 3 étapes : (1) estimation de l'espace latent, (2) définition dans l'espace latent d'une courbe définissant la progression de la maladie ; et enfin (3) estimation du DPS d'un individu par projection orthogonale de ses coordonnées dans l'espace latent sur la courbe principale de la trajectoire. Pour ce modèle, nous avons montré :

1. que dans le cas de la démence fronto-temporale (DFT) et de la sclérose latérale amyotrophique (ALS), le modèle proposé est plus performant que le modèle à base d'évènements (EBM) ;
2. que l'introduction d'une étape préliminaire de sélection de variables permet d'améliorer les performances du modèles ;
3. que l'introduction d'une branche de classification dans le VAE (VAE supervisé) permet d'améliorer les performances du modèle ;
4. que les différents types de données d'entrée (miARNs et imagerie) sont complémentaires.

$\longrightarrow$ **Méthodes :** analyse conjointe de données multimodales ; apprentissage ; classification ; autoencodeurs variationnels ; scores de progression de maladie.

$\longrightarrow$ **Domaines d'application :** maladies neurodégénératives et scores de progression de maladie (données hétérogènes : imagerie médicale + transcriptomique type miARNs).

$\longrightarrow$ **Collaborations :** Olivier Colliot à l'Institut du Cerveau (ICM - Inserm, CNRS, AP-HP - Hôpital Pitié-Salpêtrière) et Inria (équipe-projet Aramis) dans le cadre de l'Inria Project Lab (IPL) Neuromarker.

$\longrightarrow$ **Travaux en lien :** [31] [34]

⟶ **Publications jointes en support en Annexe C :**

[31] Kmetzsch V, **Becker E**, Saracino D, Anquetil V, Rinaldi D, Camuzat A, Gareau T, Le Ber I, Colliot O and the PREV-DEMALS study group.
A multimodal variational autoencoder for estimating progression scores from imaging and microRNA data in neurodegenerative diseases.
SPIE Medical Imaging 2022, 2022

[34] Kmetzsch V, **Becker E**, Saracino D, Rinaldi D, Camuzat A, Le Ber I, and Colliot O.
Disease progression score estimation from multimodal imaging and microRNA data using supervised variational autoencoders.
accepted in IEEE J Biomed Health Inform

Chapter 4

Approche systémique pour l'intégration de données unimodales et multimodales

Les travaux présentés dans cette section concernent l'intégration de données homogènes ou hétérogènes avec une approche systémique, c'est à dire en modélisant les relations connues entre les différentes variables (contrairement aux chapitres 2 et 3). Ces relations connues sont issues de bases de données de connaissances.
Lors de mon séjour post-doctoral à Marseille et avec l'aide de Christine Brun (IBDM puis TAGC) et Alain Guénoche (IML), j'ai travaillé sur l'analyse de l'interactome de différentes espèces dans le but d'analyser la topologie de cet interactome et d'en déduire des informations fonctionnelles sur les protéines qui le composent.
Plus récemment, au cours des 3 dernières années au sein de l'équipe Dyliss, j'ai été amenée à travailler sur l'extraction et l'analyse des réseaux biologiques adaptés à différentes problématiques, notamment dans le cadre des stages de master de Camille Juigné et Marc Melkonian, co-encadrés avec Gwenaël Rabut (INSERM, IGDR) et Olivier Dameron (IRISA, Dyliss), ainsi que dans le cadre de la thèse de Camille Juigné, que je co-encadre avec Florence Gondret (INRAe, UMR Pegase).

4.1 Systèmes biologiques et extraction de réseaux de haute qualité

L'intégration de données biologique homogènes ou hétérogènes peut être envisagée sous un aspect systémique, en intégrant la connaissance des relations entre entités. Cependant, extraire la connaissance adéquate des différentes banques de données et trouver un formalisme de représentation adéquat reste un sujet de recherche.

Le premier défi pour l'extraction de réseaux de qualité est celui de l'*exhaustivité* et de l'agrégation des connaissances. En effet, les données biologiques étant fragmentées dans plusieurs bases, recouper et regrouper les informations contenues dans ces bases est un premier challenge (Aranda et al., 2011; Porras et al., 2020; Schreiber et al., 2019; Villaveces et al., 2015). Le second défi est celui de l'*abstraction*. En effet, il n'existe actuellement pas de formalisme unifié pour traiter l'information biologique de nature différente ou à différentes échelles (Schreiber et al., 2019). L'abstraction choisie doit de plus être en adéquation avec la problématique biologique associée : une extraction ciblée de certaines connaissances, représentées par un formalisme *via* lequel une analyse est possible. Enfin, le dernier défi est celui de la *reproductibilité*,

pour déterminer quelles connaissances sont fiables parmi les nombreuses connaissances qui peuplent les bases de données (Alonso-López et al., 2016; Alonso-López et al., 2019; Szklarczyk et al., 2021).

4.2 Le défi de l'exhaustivité et de l'agrégation des connaissances

Les informations sur les interactions entre entités biologiques sont collectées dans de nombreuses bases de données primaires avec leur propre thématique et leur propre processus de curation. Afin de limiter l'impact de cet éparpillement de l'information, la communauté s'est organisée afin de standardiser des formats communs, tels que PSI-MI (Hermjakob et al., 2004), BioPAX (Demir et al., 2010a) etc... Cependant, l'adoption de ces formats par la communauté ne solutionne pas tous les problèmes. L'adoption de formats communs par la communauté facilite l'agrégation des connaissances en résolvant le *problème d'interopérabilité*, afin de fournir des ensembles de données plus exhaustifs. Cependant, le fait que ces interactions soient stockées dans plusieurs bases complémentaires soulève un *problème d'intégration*. Le fait que ces bases se recouvrent partiellement soulève un *problème de redondance*, qu'il faut détecter et supprimer correctement.

4.2.1 Intégration de bases de données au format BioPAX

Les formats de données tels que BioPAX (Demir et al., 2010a) ont été introduits afin d'exprimer avec une grande finesse la réalité des connaissances sur les processus biologiques à différentes échelles, de l'échelle moléculaire à l'échelle cellulaire. Son principal avantage est d'éviter l'éparpillement d'informations dans différentes bases de données qui ne seraient pas interopérables. L'interopérabilité facilite (1) l'*échange* d'informations et donc leur agrégation, et (2) leur *utilisation* en permettant d'avoir des procédures de traitement et librairies de fonctions qui fonctionnent sur différentes bases de données. Ainsi, grâce à l'interopérabilité du format BioPAX, des millions d'interactions organisées au sein de centaines de voies sont disponibles, pour de nombreux organismes (Demir et al., 2010a). Nous assistons à une croissance rapide des sources de données au format BioPAX librement disponibles, parmi lesquelles Reactome (Fabregat et al., 2016), Panther (Mi et al., 2013; Thomas et al., 2022, 2003), BioCyc (Karp et al., 2005; Romero et al., 2005), HPRD (Peri et al., 2003)...ainsi que des librairies pour leur utilisation (Bauer-Mehren et al., 2009; Demir et al., 2013).

Dans le cadre d'un projet de recherche dédié à l'étude des voies métabolique et de leur régulation, nous avons d'abord considéré individuellement 21 bases de données sources au format BioPAX collectées sur Pathway Commons (Rodchenkov et al., 2020b). Nous avons développé une bibliothèque python open source, PAX2GRAPHML [29], qui fournit un moyen générique de filtrer les interactions et de manipuler les réseaux sous forme de graphes à partir de n'importe quelle source de données BioPAX. Elle permet un import et une exploitation simplifiés et efficaces des sources de données BioPAX. Elle fournit deux types d'abstractions discutées au chapitre 4.3 pour représenter l'information contenue dans les fichiers BioPAX au moyen de graphes. Le format choisi pour représenter ces graphes est le format GraphML, basé sur XML, qui se compose d'un langage pour décrire les propriétés structurelles d'un graphe,

et d'extensions flexibles pour ajouter des données spécifiques à une application[1]. Ce format est compatible avec de nombreuses libraires de graphe Python et R, ce qui permettra ensuite de profiter de la puissance de calcul de ces librairies pour analyser les graphes extraits (Csardi and Nepusz, 2006; Hagberg et al., 2008).

Puisqu'il existe de nombreuses sources au format BioPAX, notamment sur le site de PathwayCommons, deux schémas d'intégration sont disponibles *via* PAX2GRAPHML :

- Import de chaque source de données au format BioPAX vers un graphe en GraphML, puis fusion des graphes sélectionnés : **procédure import multiple** $\longrightarrow$ **fusion**.
  Cette procédure chronophage a été exécutée sur une machine virtuelle avec 48G de RAM et a duré plus d'une semaine. Aussi, pour faciliter la visualisation des résultats de cette procédure, les fichiers générés sont également téléchargeables sur le site PAX2GRAPHML en tant que ressources de données prêtes à l'emploi et mise à jour automatiquement à l'aide de BIOMAJ (Filangi et al., 2008), un logiciel de synchronisation et de traitement automatisé des banques de données.

- Import du fichier global de PathwayCommons pour générer un unique graphe, puis filtrage de celui-ci pour ne conserver que les sources qui nous intéressent : **procédure import unique** $\longrightarrow$ **filtrage**.
  Après import du graphe généré par PAX2GRPAHML d'après le fichier global de PathwayCommons, un filtrage peut être appliqué pour générer des fichiers GraphML ne contenant que les sources sélectionnées. Grâce à la disponibilité au téléchargement sur le site de PAX2GRPAHML du fichier GraphML de PathwayCommons global, les utilisateurs peuvent ainsi sélectionner n'importe quel sous-ensemble de bases de données, ou filtrer en fonction d'autres propriétés des sommets du graphe.

La table 4.1, issue de [29], présente les différents schémas d'agrégation et leur résultats. On distingue dans la première section (1) le fichier global de PathwayCommons, et (2 à 11) les fichiers individuels correspondant à chaque source de données au format BioPAX disponible sur PathwayCommons. La section suivante, (1) - (2) - (7), illustre la procédure import unique $\longrightarrow$ filtrage, puisqu'on part du fichier global de PathwayCommons (1), duquel on retire les sources CTD (2) et Mirtarbase (7). Enfin, la dernière section présente la procédure import multiple $\longrightarrow$ fusion.

4.2.2 Redondances dans les bases de données d'interactions protéine - protéine

Dans le cas des bases d'interactions protéine-protéine, plusieurs méta-bases de données agrègent les bases de données primaires pour fournir des ensembles de données plus exhaustifs. Ceci a été rendu possible par l'adoption du format miTab, le format d'export des interactions défini par PSI-MI (1.2.3), et l'utilisation commune de l'ontologie MI qui décrit les méthodes de détection des interactions par toutes les bases de données primaires. Cependant, les bases de données primaires sont partiellement redondantes car certaines publications rapportant des interactions sont intégrées dans plusieurs bases de données primaires. La simple agrégation peut donc introduire un biais si ces redondances ne sont pas identifiées et éliminées. Pour surmonter ces biais, les méta-bases de données s'appuient sur le

[1] http://graphml.graphdrawing.org/

sources de données	**sommets**	dont réactions	dont entités	**arêtes**
(1) PC* all sources	636,038	543,880	92,158	1,320,090
(2) CTD	84,267	58,498	25,769	174,110
(3) HumanCyc	8,165	4,160	4,005	24,633
(4) INOH	6,792	4,652	2,140	16,222
(5) Intact**	2,187	563	1,624	2,869
(6) KEGG***	4,696	3,123	1,573	14,093
(7) Mirtarbase	413,366	395,703	17,663	776,342
(8) Panther	4,198	1,862	2,336	6,870
(9) PID	14,043	8,850	5,193	22,203
(10) Reactome	37,976	16,435	21,541	67,118
(11) Reconx	10,321	6,061	4,260	39,888
(1) - (2) - (7)	159,895	93,983	65,912	391,893
(9) + (3)	19,632	11,255	8,377	30,291
(9) + (3) + (6)	24,197	14,379	9,818	36,748
(9) + (3) + (6) + (10)	53,425	23,729	29,696	89,034

TABLE 4.1: Données disponibles dans Pathway Commons (PC) et leur transformation en graphes de réactions régulées avec `PAX2GRAPHML`. Les sommets sont soit des réactions soir des entités (protéines, petites molécules...). Ces graphes peuvent être téléchargés sur le site de `PAX2GRAPHML`.
* 'PC' version 12, 09/2019. ** Intact restricted to complexes *** KEGG 07/2011 (only human, hsa* files)

format unifié `PSI-MI` et l'ontologie `MI`, mais cette stratégie ne tire pas pleinement parti de la richesse sémantique de l'ontologie `MI` pour détecter les redondances, ce qui conduit à ne pas détecter un certain nombre de redondances.

Dans le cadre du stage de Master 2 de Camille Juigné puis de Marc Melkonian, co-encadré avec Gwenaël Rabut (Institut de Génétique et Développement de Rennes), nous avons proposé une définition claire de la redondance entre entrées des différentes bases de données (*curation events*), en exploitant la richesse sémantique de l'ontologie `MI` [30].

Formellement, considérons une paire de protéines (A, B). Les bases de données primaires telles que `BioGRID` ou `IntAct` peuvent fournir plusieurs entrées (ie. *curation events*) correspondant à cette paire de protéines, qui diffèrent par la méthode de détection, l'identifiant PubMed, ou les deux. Une entrée dans ces bases de données peut donc être définie par un quadruplet

$$(A, B, M_i, P_x)$$

où A et B sont les protéines, M_i est la méthode de détection (IDM pour Interaction Detection Method), et P_x est l'identifiant Pubmed de l'article original décrivant leur interaction. Lorsque deux entrées ne diffèrent que par la méthode de détection, cela devrait signifier que l'article original a observé l'interaction plusieurs fois avec des techniques expérimentales différentes. Lorsque deux entrées ne diffèrent que par l'identifiant PubMed, cela signifie que l'interaction a été reproduite dans deux études distinctes utilisant la même méthode de détection.

Pour les méta-bases de données telles que APID, alimentées par l'agrégation d'autres bases de données, une entrée peut être définie par un quintuplet (A, B, M_i, P_x, D_a) où D_a indique la base de données primaire indexant l'interaction. Nous considérons que deux entrées E_i et E_j présentent une **redondance explicite** si et seulement si :

$$\begin{cases} E_i = & (A, B, M_i, P_x, D_a) \\ E_j = & (A, B, M_i, P_x, D_b) \end{cases}$$

tandis que E_i et E_j présentent une **redondance implicite** si et seulement si :

$$\begin{cases} E_i = & (A, B, M_i, P_x, D_a) \\ E_j = & (A, B, M_j, P_x, D_b) \\ M_i & \textit{est un ancêtre (parent direct ou indirect) de } M_j \end{cases}$$

Il faut remarquer que la redondance explicite est symétrique, tandis que la redondance implicite est antisymétrique et dans la définition précédente, E_i n'apporte aucune information supplémentaire par rapport à E_j même si on a l'impression (trompeuse) que E_i et E_j font appel à des méthodes de détection différentes. La redondance explicite correspond à ce qui est actuellement détecté par les méta-bases, et la redondance implicite est en général ignorée.

La figure 4.1 illustre les différents cas de figures. Pour un triplet (A, B, P_x) commun à deux entrées, on distinguera le cas où (violet, panel B) chaque entrée correspond à une méthode de détection distincte, et il n'y a pas de redondance mais bien 2 détections rapportées dans une même publication; (jaune, panel C) les deux entrées ont la même méthode de détection, l'information est donc explicitement redondante; (bleu, panel D) les deux entrées ont deux méthodes de détection différentes, mais l'une d'elle est un ancêtre de l'autre : ce cas de figure signifie que les bases de données primaires ont annoté une même détection avec deux termes de précision différente dans l'ontologie MI, et l'information est implicitement redondante.

Après avoir proposé une définition précise de la redondance explicite et implicite, nous avons montré que les deux peuvent être facilement détectées en utilisant les technologies du Web Sémantique [30]. Nous avons poursuivi ce travail afin d'identifier la proportion de redondances explicites et implicites présentes dans les méta-bases de données d'interactions, et de quantifier l'impact de l'élimination de ces redondances. Au sein de la méta-base de données APID, nos résultats montrent que si les redondances explicites ont été détectées par le processus d'agrégation d'APID, environ 15% des entrées APID sont implicitement redondantes et devraient être éliminées lors du processus d'agrégation.

Une analyse fine de ces redondances implicites montrent qu'elles résultent principalement de l'agrégation de bases de données primaires distinctes (redondance inter-bases, 91% et 95% des redondances implicites respectivement chez l'homme et la levure), alors qu'une faible proportion est détectée entre entrées d'une même bases de données (redondances intra-base). Les tables 4.2 et 4.3 présentent une analyse de l'origine de ces redondances implicites inter-bases, qui souligne que des redondances implicites sont observées entre quasiment tous les couples de bases de données primaires. Le couple de bases de données qui génère la plus grande partie des redondances implicites est BioGRID et IntAct, à la fois chez l'homme et chez la levure, ce qui est cohérent avec le fait que ce sont les deux bases de données sources qui contribuent le plus à APID.

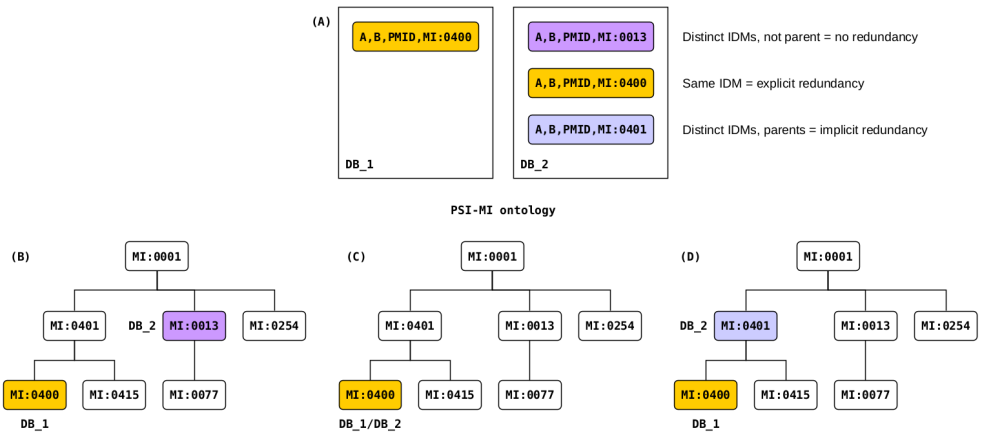


FIGURE 4.1: Illustration des différents types de redondance entre bases de données primaires. (A) Curation events provenant de deux bases primaires (DB_1 et DB_2). En fonction de la méthode de détection de l'interaction du curation event dans DB_1 et DB_2, on peut identifier qu'il n'y a pas de redondance (violet), une redondance explicite (jaune), ou une redondance implicite (bleu). Les représentations de l'ontologie Molecular Interactions associées aux différents cas de figures sont présentées dans les panels (B), (C) et (D). Figure extraite de [30].

Le fait que des redondances implicites soient observées entre des termes très différents de l'ontologie PSI-MI suggère que les différentes bases de données primaires ont des politiques différentes en matière d'annotation des méthodes de détection d'interactions (IDM), comme cela a déjà été noté pour IntAct et BioGRID (Alonso-López et al., 2016). En effet, nous avons observé qu'IntAct et DIP utilisent un large éventail d'IDMs pour les PPIs (165 pour IntAct et 89 pour DIP), tandis que BioGRID, HPRD et BioPlex en utilisent beaucoup moins (12, 3 et 1 IDM, respectivement) et des IDM plus généraux. Ainsi, les fortes divergences dans les politiques d'annotation des bases de données sont la principale source des redondances implicites entre les bases de données.

Dans l'ensemble, nous avons observé dans [30] que la redondance implicite (i) se produit entre un large éventail de termes de l'ontologie PSI-MI, quelle que soit l'espèce, (ii) résulte principalement de l'intégration de différentes bases de données primaires primaires différentes avec des politiques d'annotation différentes, et (iii) se produit pour toutes les bases de données. Par conséquent, si l'intégration de différentes bases de données PPI est nécessaire pour mieux couvrir l'interactome, une attention particulière doit être portée à la détection des les redondances implicites généralisées entre les bases de données.

Pour cela, la Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI) a mis au point la spécification PSICQUIC et des services web qui facilitent l'extraction de données de plusieurs bases de données, aident à la gestion des données et qui facilitent leur intégration, mais ne traitent pas de la détection de la redondance (Aranda et al., 2011; Toro et al., 2013). De plus, dans plusieurs (méta-)bases de données, les IPPs sont annotées par un score de confiance, qui est calculé en utilisant le nombre de preuves expérimentales indépendantes et la nature des IDMs (Villaveces et al., 2015). Pour être pertinents, ces algorithmes nécessitent des jeux de données fiables et non redondants. Par conséquent, plusieurs bases de données

BDD utilisant le PSI-MI ancetre	BDD utilisant le PSI-MI enfant	Occurrences des red. implicites	Pourcentage des red. implicites
BioGRID	IntAct	90368	84.95%
BioGRID	BioGRID	9030	8.49%
HPRD	IntAct	5309	4.99%
BioGRID	DIP	1049	0.99%
IntAct	IntAct	440	0.41%
IntAct	BioGRID	73	0.07%
DIP	BioGRID	43	0.04%
DIP	IntAct	31	0.03%
DIP	DIP	18	0.02%
IntAct	DIP	12	0.01%
Intra-database red.		9488	8.92%
Inter-database red.		96885	91.08%

TABLE 4.2: Origine des redondances implicites chez l'homme. La base de donnée (BDD) utilisant le terme le moins précis est indiquée dans la colonne de gauche.

BDD utilisant le PSI-MI ancetre	BDD utilisant le PSI-MI enfant	Occurrences des red. implicites	Pourcentage des red. implicites
BioGRID	IntAct	33426	71.52%
BioGRID	DIP	10091	21.59%
BioGRID	BioGRID	2075	4.44%
DIP	IntAct	426	0.91%
DIP	BioGRID	383	0.82%
IntAct	IntAct	244	0.52%
IntAct	BioGRID	81	0.17%
IntAct	DIP	9	0.02%
DIP	DIP	4	0.01%
Intra-database red.		2323	4.97%
Inter-database red.		44416	95.03%

TABLE 4.3: Origine des redondances implicites chez la levure. La base de donnée (BDD) utilisant le terme le moins précis est indiquée dans la colonne de gauche.

primaires ont décidé de coordonner leurs efforts de curation dans le cadre du consortium consortium IMEx afin de fournir un ensemble unique et non redondant de données d'interactions protéiques annotées de manière homogène (Orchard et al., 2012; Porras et al., 2020). Les principales bases de données du consortium IMEx coordonnent et partagent leurs efforts de curation afin de produire un ensemble de données non redondantes de preuves expérimentales d'IPP (Orchard et al., 2012). Les membres d'IMEx utilisent des règles de curation pour harmoniser leur processus d'annotation. L'unicité des entrées est assurée en imposant qu'une publication ne soit intégrée que dans une seule base de données, et toutes les données sont centralisées dans IntAct. Ce travail du consortium IMEx et le nôtre, basés sur les technologies du Web Sémantique, soulignent la nécessité d'une approche générale pour assembler des ensembles de données d'interactions non redondantes.

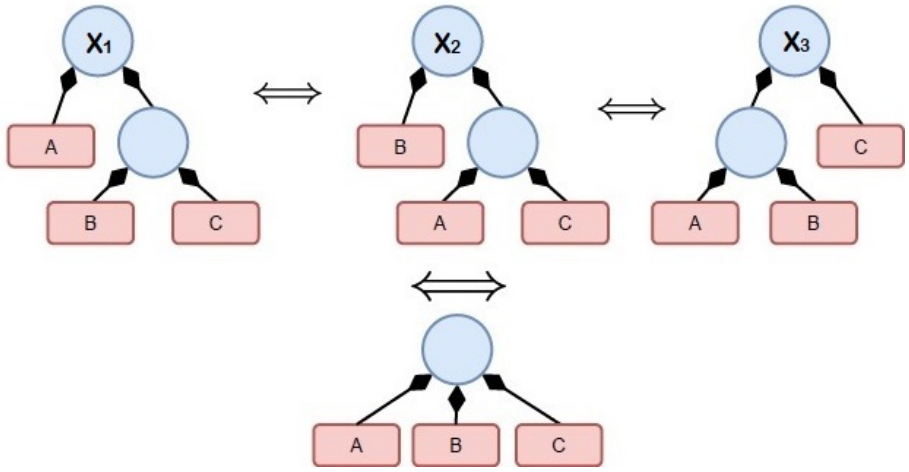


FIGURE 4.2: Différentes représentation du complexe composé de 3 partenaires A, B et C. Sur la première ligne: (gauche) un complexe dont les composants sont A et le complexe B-C ; (milieu) un complexe dont les composants sont B et le complexe A-C ; (droite) un complexe dont les composants sont le complexe A-B et C. Sur la seconde ligne, un complexe aplani dont les composants sont A, B et C.

4.2.3 Redondances dans les bases de données `BioPAX`

Nos travaux sur les bases de données au format `BioPAX` (Demir et al., 2010a), et notamment sur `Reactome` (Fabregat et al., 2016), dans la cadre de la thèse de Camille Juigné, nous ont permis d'identifier de la redondance implicite dans les données de `Reactome`. Ici, c'est une utilisation très répandue mais non-conforme aux spécifications du format `BioPAX` qui induit une redondance entre entre certains complexes, et ce point particulier sera discuté dans la partie 4.3.1.

Camille Juigné a pu identifier des complexes redondants au sein de `Reactome`. Ces redondances sont implicites car elles sont masquées par une description arborescente différente des composants des complexes, comme le montre l'illustration 4.2. Pour un complexe composé de 3 partenaires (A, B, C), nous avons trouvé au sein de `Reactome` humain des redondances sous la forme $(A, (B, C))$, ou $(B, (C, A))$, ou $((A, B), C)$, les quantités stoechiométriques des différentes partenaires étant équivalentes dans les différentes descriptions. Il s'agit donc de redondances implicites non détectées.

On notera que, comme nous le discuterons dans la section 4.3, selon les spécifications du format `BioPAX` depuis `BioPAX2.0`, la seule forme correcte pour ce complexe est sa version aplanie (A, B, C), dont l'utilisation évite ces redondances implicites. C'est donc le non-respect de ces spécifications qui entraîne cette redondance implicite dans les fichiers `BioPAX` de la banque `Reactome`.

Les travaux de Camille Juigné ont abouti à la détection de nombreuses redondances dans la base de données `Reactome` [33]. Les redondances peuvent être détectées entre des paires d'entités mais peuvent se produire entre plus de deux entités, comme l'illustre la figure 4.2 avec 3 complexes équivalents. Dans cet exemple, 3 redondances (X_1, X_2), (X_2, X_3), (X_1, X_3) sont ainsi détectées, correspondant à la taille d'une clique maximale à 3 sommets.

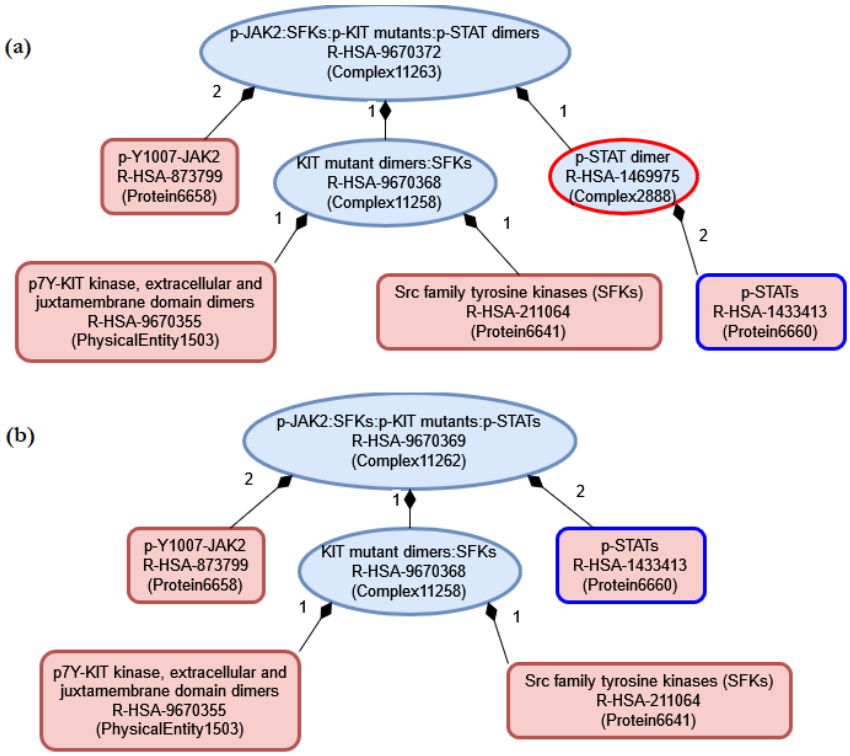


FIGURE 4.3: Différentes représentation d'un même complexe au sein de `Reactome`. (a) le participant `p-STATs R-HSA-1433413` participe au complexe principal avec une stoechiométrie de 2. (b) le participant `p-STATs R-HSA-1433413` participe au complexe principal via un sous-complexe composé de `p-STATs R-HSA-1433413` en 2 exemplaires.

Parmi les 14 987 complexes de la base `Reactome` originale, Camille Juigné a écrit une requête `SPARQL` qui a identifié 137 paires de complexes redondants, impliquant 249 complexes distincts. Ces redondances constituaient 121 cliques maximales de complexes équivalents (taille de la clique allant de 2 à 6 complexes), ce qui correspond à 128 complexes en excès. En d'autres termes, nous avons mis en évidence 121 groupes de 2 à 6 complexes redondants que nous avons identifiés en recherchant les paires de complexes redondants. Ces redondances sont explicites, car elles ne sont pas masquées par une définition arborescente des composants des complexes.

Après avoir aplani les complexes représentés récursivement, la requête a permis d'identifier 217 paires de complexes redondants impliquant 347 complexes distincts. Ils constituaient 164 cliques maximales de complexes équivalents (taille de la clique comprise entre 2 et 6), correspondant à 183 complexes en excès. La procédure d'aplanissement, en remplaçant la représentation arborescente des complexes par une représentation plane des complexes, a permis de détecter davantage de redondances. Ces redondances étaient implicites, puisqu'elles étaient masquées par une définition arborescente des composants des complexes, et sont devenues explicites avec la description plane des composants des complexes.

La figure 4.3 illustre un exemple de redondance identifiée : dans cet exemple, l'un

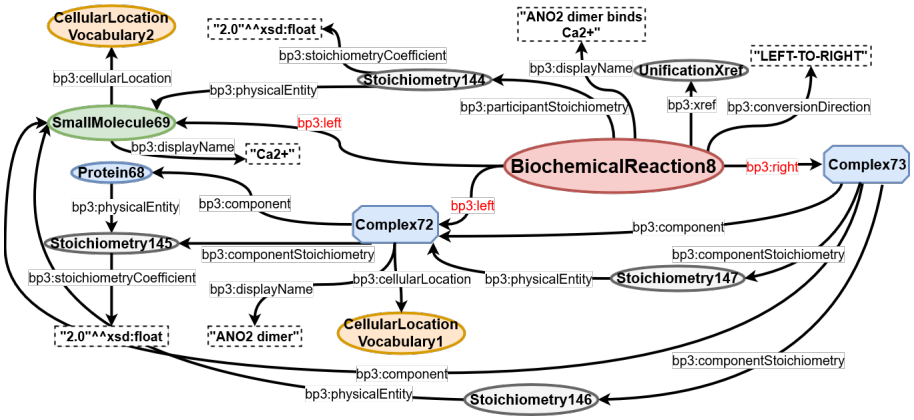


FIGURE 4.4: Sous-ensemble de relations permettant de décrire un complexe conformément à l'ontologie `BioPAX`

des participant participe au complexe principal soit en tant que participant direct en 2 exemplaires (coef. stoechiométrique = 2), soit sous forme d'un sous complexe (homo-dimère).

4.3 Le défi de l'abstraction

Des ontologies telles que `BioPAX` ont été introduites afin d'exprimer avec une grande finesse la réalité des connaissances sur les processus biologiques à différentes échelles (de l'échelle moléculaire et à l'échelle cellulaire). L'ontologie `BioPAX` est qualitative, très riche et expressive, et parvient à concilier les différentes échelles de description de l'information biologique. Cette expressivité raffinée complexifie énormément l'extraction d'informations pertinentes de données stockées au format `BioPAX`. À titre d'exemple, la figure 4.4 illustre simplement une création de complexe entre `Ca2+` et un dimère d'`ANO2`. On peut constater que l'information utile est structurée *via* de nombreuses relations qui stockent chacune une partie précise de l'information. Ainsi la manipulation de ce format de données riche et orienté vers la connaissance, qui capture finement la complexité des réseaux biologiques, nécessite des outils appropriés.

Identifier un formalisme expressif et informatiquement exploitable est un défi lorsqu'on s'intéresse à différentes échelles biologiques simultanément. Deux problématiques s'entremêlent ici :

1. l'extraction ciblée du sous-ensemble d'informations pertinentes pour un problème donné ;

2. le choix du formalisme pour représenter l'information disponible, de façon à ce que des analyses informatiques puissent analyser les propriétés du systèmes ou inférer des résultats.

La section 4.3.1 présente un outil permettant d'aborder la complexité de `BioPAX`. La section 4.3.2 présente un outil permettant d'aborder la complexité de la réalité biologique.

4.3.1 Topologie des complexes dans les données au format BioPAX

Les complexes sont définis dans BioPAX comme des entités physiques composées d'autres entités physiques : « *physical entities whose structure is comprised of other physical entities bound to each other non-covalently, at least one of which is a macromolecule (e.g. protein, DNA, or RNA)* ». La spécification BioPAX mentionne cependant explicitement que les complexes ne doivent pas être définis de manière récursive ou arborescente : « *complexes should not be defined recursively [...] i.e. a complex should not be a component of another complex. [...] Exceptions are black-box complexes (i.e. complexes in which the component property is empty), which may be used as components of other complexes because their constituent parts are unknown* ».

Cette contrainte a été introduite lors du passage de la norme BioPAX1.0 à la norme BioPAX2.0, et la justification donnée pour l'introduction de cette contrainte était que l'utilisation d'une arborescence pourrait être interprétée par certains utilisateurs comme un ordre dans l'assemblage macromoléculaire : « *The reason for keeping complexes flat is to signify that there is no information stored in the way complexes are nested, such as assembly order. Otherwise, the complex assembly order may be implicitly encoded and interpreted by some users, while others created hierarchical complexes randomly, which could lead to data loss.* ». L'introduction de cette contrainte a également un effet direct sur la topologie du graphe des relations BioPAX puisqu'elle permet de rapprocher les différentes entités participant à un complexe.

Toutes les principales bases de données de référence sur les voies métaboliques sont disponibles au format BioPAX. Parmi elles, Reactome est une base de données décrivant les voies biologiques, disponible en ligne, gratuite, et d'excellente qualité du fait d'un processus de curation efficace (Gillespie et al., 2021). Dans le cadre des travaux de thèse de Camille Juigné, nous avons noté la présence de nombreux complexes décrits de façon récursive dans l'export BioPAX de Reactome. Leur présence dans Reactome pose problème, non seulement car ils ne sont pas conformes aux spécifications BioPAX, mais aussi car ils ont un effet important sur la topologie du graphe des relations BioPAX en allongeant artificiellement les chemins entre des entités participant à un même complexe. Ainsi, dans la figure 4.5, le plus court chemin entre PCNA et DNA2 est de longueur 8, alors que ces 2 protéines participent au même complexe R-HSA-68466 (Okazaki fragment-Flap-RPA heterotrimer-dna2). Il est intéressant de noter que ces complexes invalides ne sont pas détectés par le validateur BioPAX [2] (Rodchenkov et al., 2013).

En utilisant des requêtes SPARQL, les travaux de Camille Juigné montrent que la base de données Reactome contiennent une grande fraction de complexes définis récursivement, c'est-à-dire dont les participants contiennent au moins un complexe [33]. Nous avons quantifié la fraction de complexes invalides avant de corriger ceux-ci. Les résultats sont présentés dans la table 4.4 : environ 39% des complexes Reactome humains sont invalides car définis récursivement. L'importante fraction de complexes invalides n'est pas spécifique à l'homme et se retrouve avec toutes les espèces testées présentes dans Reactome : les 13 espèces testées présentent toutes entre 30% (*Plasmodium falciparum*) et 40% (*Sus scrofa*, *Bos taurus*, *Canis familiaris*, *Gallus gallus*) de complexes invalides.
On note que cette définition arborescente des complexes peut être profonde. La figure 4.6 présente la distribution de la profondeur de définition des complexes chez

[2] https://biopax.baderlab.org/check

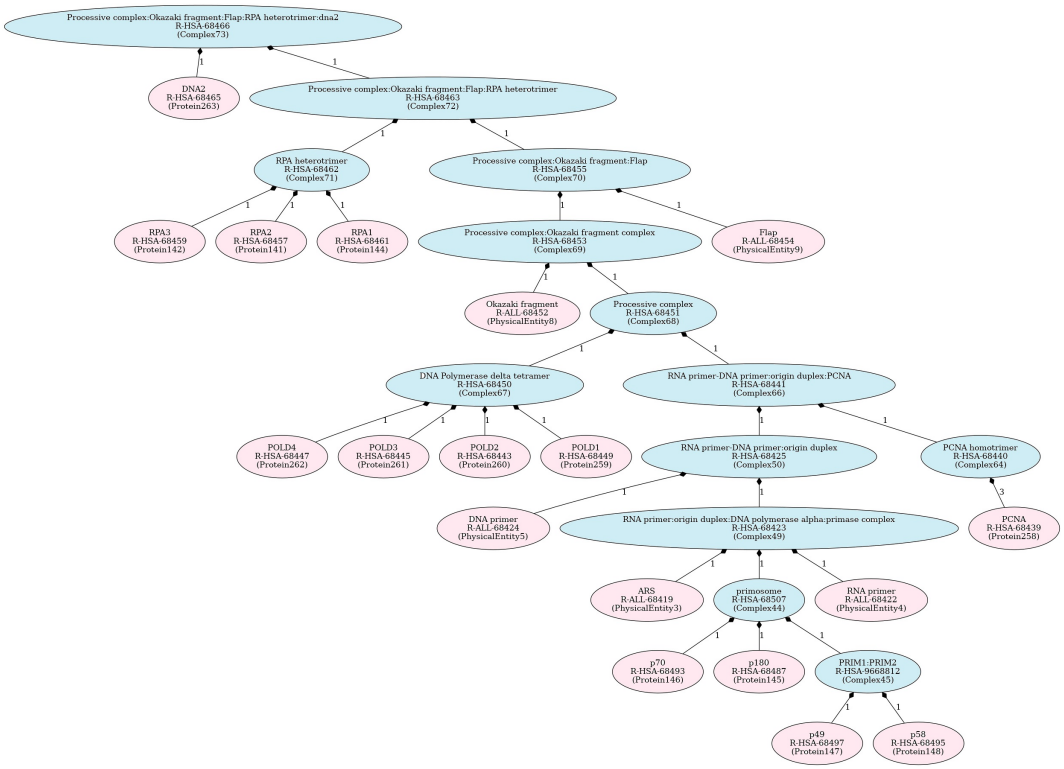


FIGURE 4.5: Complexe R-HSA-68466 défini récursivement sur 10 niveaux dans la base de données Reactomev81 chez *H. Sapiens*. Les complexes sont en bleu, les protéines et petites modélcules en rose.

l'homme. On constate la présence de complexes de profondeur 0, qui correspondent aux black-box complexes, des complexes de profondeur 1, qui sont conformes aux spécifications `BioPAX`, mais aussi de nombreux complexes non valides (tous ceux dont la profondeur est supérieure ou égale à 2). Dans la bases de données `Reactome` humaine, la profondeur maximale atteint 10 imbrications, pour un exemple, voir figure 4.5.

La correction de ces complexes invalides mène à une représentation aplanie de ces complexes : toutes les entités physiques participant au complexe sont alors explicitement listées *via* la propriété `component`. Elle a pour effet direct et immédiat d'augmenter significativement le nombre d'entités listées *via* la propriété `component` [33] . À titre d'exemple, chez l'Homme, le nombre moyen d'entités participant directement à un complexe est de 2.2 avant la correction et de 4.3 après la correction. Chez l'Homme, le plus gros complexe est constitué de 65 entités avant correction, contre 151 après. Dans [33] , nous montrons que les résultats chez les autres espèces sont similaires. (table 4.4).

4.3.2 Extraction « à façon » dans des données au format `BioPAX`

En préparation au travail de thèse de Camille Juigné, nous avons conçu l'outil `PAX2GRAPHML` afin d'extraire des sous-ensembles de données à partir de fichiers au format `BioPAX` [29] . Dans l'outil `PAX2GRAPHML`, il est possible d'extraire des données selon (i) les bases de données sources, que l'on peut choisir et combiner à façon lors de l'import, mais aussi (ii) en filtrant les graphes obtenus, en se basant sur certaines propriétés.

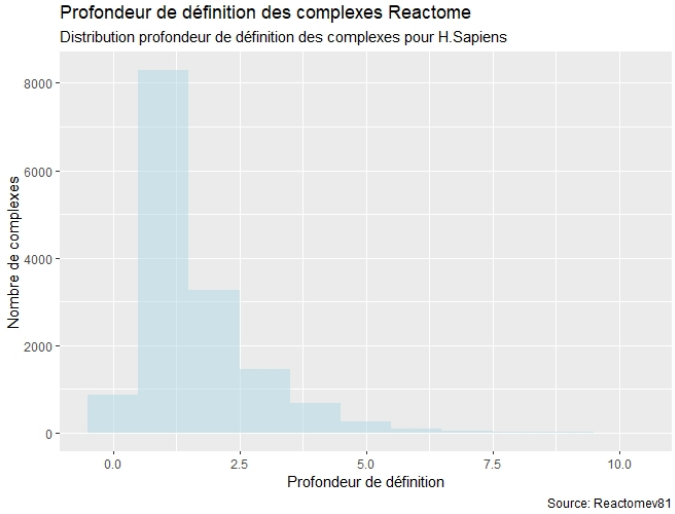


FIGURE 4.6: Profondeur de définition récursive des complexes dans la base de données Reactomev81 chez *H. Sapiens*.

Organisme	**Complexes**	**dont invalides**	**Ave. direct components**	
			before fixing	**after fixing**
homo sapiens	14987	5833 (39%)	2.2 (std 2.6) max 65	4.3 (std 8.7) max 151
mus musculus	10707	4235 (39%)	2.3 (std 2.9) max 65	4.5 (std 9.0) max 151
sus scrofa	9022	3638 (40%)	2.3 (std 2.9) max 65	4,7 (std 9,4) max 151
bos taurus	9412	3773 (40%)	2.3 (std 2.9) max 65	4.6 (std 9.2) max 147
s. cerevisiae	1662	517 (31%)	2.5 (std 3.3) max 50	6.0 (std 12.4) max 106
c. elegans	4350	1560 (36%)	2.4 (std 3.3) max 64	5.0 (std 10.4) max 149
canis familiaris	8945	3601 (40%)	2.3 (std 3.0) max 65	4.7 (std 9.4) max 151
danio rerio	8618	3391 (39%)	2.3 (std 3.0) max 65	4.7 (std 9.5) max 150
d. discoideum	2366	792 (33%)	2.4 (std 2.8) max 50	5.3 (std 9.8) max 103
d. melanogaster	5361	1955 (36%)	2.4 (std 3.0) max 64	4.8 (std 9.7) max 149
gallus gallus	8046	3244 (40%)	2.3 (std 2.8) max 65	4.7 (std 9.4) max 149
p. falciparum	875	264 (30%)	2.4 (std 3.6) max 50	5.4 (std 11.9) max 103
rattus norvegicus	9645	3780 (39%)	2.3 (std 2.9) max 65	4.5 (std 9.2) max 151

TABLE 4.4: Pour chaque organisme dans la base de données Reactome : le nombre d'entités physique de type `Complexes` présentes, le nombre et la fraction de celles-ci étant définies de façon invalide, puis le nombre moyen de participants à un complexe avant et après la procédure de correction.

L'outil `PAX2GRAPHML` abstrait l'information sous forme d'un graphe de « réactions régulées », tel que décrit par Blavy et al., 2014. Une réaction régulée est un motif centré sur un sommet représentant une réaction et associé à ses métadonnées de réaction.

Ce sommet de type `reaction` est lié à un ou plusieurs sommets de type `substrat` et `produits`. Si nécessaire, le sommet de type `reaction` peut également être lié à des sommets de type `régulateurs`, qu'il s'agisse d'`activateurs`, d'`inhibiteurs` ou de `modulateurs`. Le graphe de réactions régulées est un graphe dirigé. Les `substrats` et les `régulateurs` sont des entrées des `réactions`. Les `produits` sont les sorties des `réactions`. Tous les nœuds sont également associés à leurs propres métadonnées. La figure 4.7 présente un exemple type d'un motif de réaction régulée. Les réactions biochimiques et les régulations sont décrites de manière homogène par

des réactions régulées qui impliquent des substrats, des produits, des activateurs, des inhibiteurs et des modulateurs comme éléments. Dans le graphe de réactions régulées, les molécules et les réactions sont représentées par des sommets étiquetés (typés).

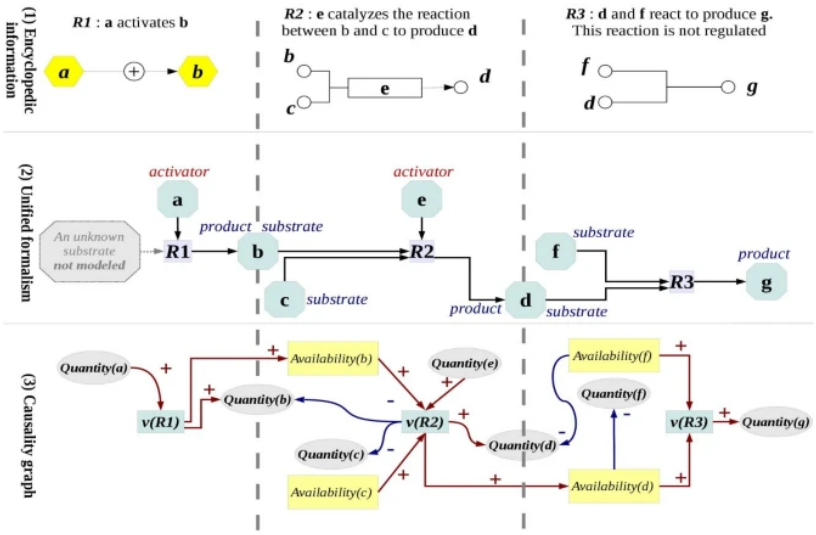


FIGURE 4.7: Deux niveaux d'abstraction utilisées pour interpréter des données de fichiers `BioPAX` sous forme de graphe : (1) graphe de réactions régulées sur la ligne (2), et graphe d'influence sur la ligne (3). Trois stypes de réactions types sont modélisées : une activation (colonne 1), une réaction contrôlée par une enzyme (colonne 2), et une réaction non contrôlée (colonne 3). Figure extraite de (Blavy et al., 2014)

Afin de faciliter l'interprétation de la causalité au sein du graphe de réactions régulées (Chindelevitch et al., 2012), un niveau d'abstraction supplémentaire est proposé, afin de retirer du graphe les sommets correspondants au réactions, pour ne conserver que des sommets qui seraient des substrats, produits et régulateurs au sein d'un graphe dirigé où les interactions sont étiquetées avec des influences positives et négatives. Le graphe d'influence calculé utilise deux classes de sommets et chaque composé impliqué dans une réaction est décrit par deux sommets : l'un des sommets représente la disponibilité (ie. la contribution du composé à la vitesse de réaction) et l'autre la quantité (ie. la concentration du composé). Le modèle générique de transformation des réactions régulées en graphe d'influence suit les principes décrits et illustré par la figure 4.7, extraite de Blavy et al., 2014.

4.4 Le défi de la reproductibilité et réplicabilité

Le dernier défi associé à l'extraction de réseaux de haute qualité concerne leur **reproductibilité / réplicabilité**. Ce défi englobe une composante liée à la reproductibilité ou réplicabilité des analyses biologiques qui mettent en évidence les relations entre

entités, et englobe également une composante informatique qui est liée à la reproductibilité ou réplicabilité du processus d'extraction des réseaux, ce qui suppose une utilisation pertinente des formats et structures de données sous-jacentes.

Vers la quantification d'un interactome reproductible

Les interactions entre protéines peuvent être étudiées à l'aide de nombreuses méthodes de détection des interactions, selon des approches biophysiques (par exemple, la cristallographie), biochimiques (par exemple, la purification par affinité) ou génétiques (par exemple, par double hybride). Il est important de noter que, puisque les différentes méthodes de détection sondent les interactions protéine-protéine de manière différente, elles produisent des résultats complémentaires qui, souvent, ne se recoupent pas entièrement. Par exemple, certaines méthodes de détection sont conçues pour détecter des interactions binaires (par exemple, le double hybride), tandis que d'autres sondent les interactions de groupes de protéines assemblées en complexes (par exemple, la purification par affinité). De plus, du fait de contraintes expérimentales, chaque méthode de détection génère des interactions faussement positives ou inversement ne détecte pas certaines interactions. De ce fait, des observations multiples d'une interaction avec différentes techniques renforcent la confiance dans cette interaction.

La construction d'un interactome fiable exige donc de combiner les données d'interaction produites par plusieurs preuves expérimentales indépendantes afin de réduire les faux positifs. Une solution alternative consiste à utiliser des scores de confiance des interactions, mais ce score est dépendant entre autre du nombre de preuves expérimentales indépendantes et de la nature des méthodes de détection d'interaction (IDM) (Villaveces et al, 2015). Pour être pertinents, ces méthodes de score nécessitent des jeux de données expérimentales fiables et non redondants.

Dans les travaux de Camille Juigné et Marc Melkonian, nous avons pu détecter les redondances explicites et implicites dans les réseaux d'interactions entre protéines, et nous avons ainsi pu éliminer toutes les redondances artificielles contenues dans l'agrégation des bases de données primaire d'APID. De cet ensemble de données d'interactions non-redondant, nous pouvons déduire un interactome reproductible ou réplicable, c'est à dire dans lequel chaque interaction entre protéine est supportée par au moins deux preuves expérimentales distinctes.

Nous avons montré que la détection de la redondance entre les curation events a un impact important sur la taille de l'interactome reproductible [30]. Pour l'Homme, l'interactome reproductible passe de 159192 à 70554 interactions reproductibles, ce qui correspond à une baisse de 55.7% (dont -30.3% due à l'élimination des redondances explicites et -25.4% due à l'élimination des redondances implicites). Pour la levure, l'impact des redondances est encore plus important, avec une chute de l'interactome reproductible de 52313 interactions reproductibles à 21311 après suppression des redondances explicites et implicites, ce qui correspond à une diminution de -59.3% de la taille de l'interactome reproductible (dont $-23,1\%$ dus aux redondances explicites et $-36,2\%$ dus aux implicites). Les résultats complets sont donnés dans la table 4.5 issue de [30]. Au final, nos travaux soulignent qu'un grand nombre d'interactions considérées comme reproductibles repose en fait sur des artefacts d'intégration.

TABLE 4.5: Impact de la suppression des redondances explicites et implicites sur le nombre d'événements de curation et sur la taille de l'interactome reproductible, pour l'Homme et la levure. La taille de l'interactome reproductible compte le nombre de couples de protéines (A, B) dont l'interaction est soutenue par au moins 2 curation events.

	Human	**Yeast**
Curation events		
Nombre de curation events initial	665,484	293,293
... après élimination des red. explicites	534,140	229,630
... après élimination des red. explicites + implicites	460,149	189,364
Taille apparente de l'interactome reproductible		
Nombre de curation events initial	159,192	52,313
... après élimination des red. explicites	111,009	40,235
... après élimination des red. explicites + implicites	70,554	21,311

4.5 Ce que nous avons appris dans ce chapitre

L'extraction de réseaux biologiques à partir des données multi-échelles soulève différents problèmes :

1. la recherche d'exhaustivité induit une problématique liée à l'agrégation des données, dont nous avons montré qu'elle nécessite une prise en compte très fine des ontologies riches décrivant les données, notamment afin de supprimer les redondances.

2. la recherche de l'abstraction permettant au mieux de répondre à une question donnée, dont nous avons montré qu'elle nécessite de maîtriser finement les spécifications des ontologies décrivant les données, et le développement d'outils *ad-hoc* pour extraire une information ciblée.

3. le recherche d'information fiable et reproductible, pour laquelle nous avons montré l'intérêt de prendre en compte toute la richesse des ontologies développées. Mes contributions ont permis d'identifier finement les redondances entre les entrées des différentes bases de données, et ainsi de quantifier la reproductibilité des interactions définissant un interactome.

⟶ **Méthodes :** Web Sémantique, ontologies, requêtes SPARQL, méthodes de parcours de graphe.

⟶ **Domaines d'application :** agro-écologie, biologie fondamentale.

⟶ **Collaborations :** Chrsitine Brun et Alain Guénoche (Marseille) ; Florence Gondret et François Moreews (UMR Pegase), Olivier Dameron et Anne Siegel (équipe Dyliss)

⟶ **Travaux en lien :** [33], [30], [29], [18], [08], [07], [05].

⟶ **Publications jointes en support en Annexe C** :

[33] Juigné C, Dameron O, Moreews F, Gondret F and **Becker E**.
Detection and correction of non-conformities and redundancies in complexes of molecules in `BioPAX`.
Actes JOBIM 2022, Rennes, 5-8 juillet 2022

[30] Melkonian M, Juigné C, Dameron O, Rabut G* and **Becker E***.
Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases.
Bioinformatics, 2022.

[29] Moreews F, Simon H, Siegel A, Gondret F, **Becker E.**
PAX2GRAPHML : a python library for large-scale regulation network analysis using `BioPAX`.
Bioinformatics, 2021.

Chapter 5

Projets et perspectives

5.1 Résumé des travaux présentés

Dans les travaux présentés au chapitre 2 sur la détection de signatures dans des données unimodales, nous avons vu que les analyses classiques de profilage n'adressent pas le problème de la spécificité des biomarqueurs, ni celui du potentiel prédictif et de l'interprétabilité biologique. Les stratégies de recherche actuelles privilégient en effet l'acquisition rapide de nouvelles données, puis leur analyse par des méthodes statistiques qui, du fait de la dimensionnalité défavorable des données -omiques, génèrent des faux positifs et des faux négatifs. Il existe donc une réelle nécessité scientifique à développer des méthodes permettant de consolider ces résultats de profilage pour en faire des biomarqueurs fiables. Les travaux présentés proposent des variations méthodologiques *a priori* pour remettre en question une signature afin de s'assurer de sa solidité. Ils soulignent également la nécessité de répliquer *a posteriori* les résultats avec des cohortes de validation indépendantes et un design expérimental pré-enregistré.

Ces travaux se poursuivent dans le chapitre 3 en passant de l'unimodal au multimodal, en intégrant dans notre cas des données -omiques de type miARN et des données d'imagerie médicale de type IRM-T1. Tout comme dans le chapitre 2, nous avons développé une méthode adaptée à une dimensionnalité défavorable des données, en nous intéressant aux cas des cohortes de petite taille sans données longitudinales. De plus, nous avons là encore souhaité intégrer les données avec une exigence d'interprétabilité biologique, ce qui nous a conduit à développer une méthode de prédiction de score de progression de maladie. Avec l'essor des données médicales et de la santé numérique, il est en effet important de développer des méthodes cohérentes avec la réalité des données disponibles, c'est à dire intégrant leurs problèmes de temporalité et de dimensionnalité. Les travaux présentés proposent une nouvelle méthode basée sur des autoencodeurs variationnels supervisés pour réduire la dimensionnalité des données et ainsi calculer des scores de progression de maladies à partir de données multimodales, sur des cohortes de petite taille, et en l'absence de données longitudinales. Les résultats obtenus sont prometteurs.

Le chapitre 4 s'intéresse également à l'intégration de données hétérogènes, mais avec une approche différente. Contrairement au chapitre 3 où les variables sont considérées comme indépendantes les unes des autres, le chapitre 4 est basé sur une approche systémique, c'est à dire en prenant compte et modélisant des relations entre les différentes variables. Or, le volume de données et de connaissances générées à différentes échelles au cours de ces 20 dernières années est tel que la représentation unifiée de ces connaissances, leur stockage, puis leur extraction est en soi un sujet de recherche. Mes travaux présentés au chapitre 4 proposent des approches pour

extraire des réseaux de meilleure qualité à partir des formats complexes de modélisation des connaissances biologiques.

5.2 Vers une recherche de biomarqueurs consolidée

Les travaux présentés soulignent que le *profilage* de données -omique reste largement exploratoire et les ensembles d'éléments -omiques identifiés comme différentiels doivent être consolidés en vue d'être utiles pour le diagnostic et de constituer une *signature* intéressante.

À court terme, il me semble important de développer un pipeline d'analyse unifié, qui intègre une analyse différentielle basée sur les meilleurs outils disponibles (Love et al., 2014a; Robinson et al., 2010a), mais qui intègre également les méthodes permettant d'investiguer le potentiel prédictif des éléments du profil, seuls ou en combinaison, en prenant garde aux biais potentiels si l'évaluation du potentiel prédictif est faite sur le même ensemble que celui ayant identifié les éléments du profil. Ce pipeline devrait donc intégrer des méthodes de re-échantillonnage, comme dans l'analyse de généralisabilité présentée dans 2.2.3, ainsi que des méthodes de classification simples, comme présentées dans l'analyse du potentiel prédictif dans 2.2.1. Ce pipeline pourra être utilisé par dans de nombreux projets par la suite.

À moyen terme, je pense qu'il est important d'approfondir la question de la spécificité des biomarqueurs. Dans les travaux présentés dans ce mémoire, la spécificité des biomarqueurs est abordée par la prisme des analyses des courbes ROC des comparaisons 2 à 2 (2.2.1). La spécificité de la signature est donc évaluée uniquement à l'échelle du schéma expérimental actuel, et rien ne permet de savoir si les éléments de la signature ne font pas partie d'autres signatures, mises en évidence par d'autres études (antérieures ou futures). On pourrait ainsi imaginer qu'un miARN soit identifié comme faisant partie de plusieurs signatures différentes, tout en ayant un potentiel prédictif au sein de chaque analyse pour discriminer les groupes témoins et traités. Afin d'aborder cette question, nous disposons des données du projet européen Phenomir (auquel j'ai participé, en collaboration avec Julien Bobe du LPGP INRAe), au sein duquel 12 analyses différentes ont été conduites afin d'identifier des biomarqueurs qui concernent les conditions de santé et d'élevage de truites arc-en-ciel parmi les miARNs circulants. L'extraction et le séquençage ont été réalisés conjointement, et l'identification des profils a été effectuée selon le même schéma d'analyse, afin que les résultats soient comparables. La comparaison des signatures obtenues pourra nous permettre au sein de ce projet d'aborder par un autre prisme la notion de spécificité des biomarqueurs.

Enfin, à plus long terme, la question de la reproductibilité des profils et signatures m'intéresse particulièrement. Dans les travaux de validation de Virgilio Kmetzsch présentés dans la section 2.2.4, nous notons qu'une part importante des miARNs identifiés comme faisant partie d'une signature sont bien retrouvés comme différentiellement exprimés dans une analyse de validation indépendante (35/65 pour la cohorte homogène C9orf72). Ces miARNs proviennent d'études différentes, et aucune étude n'identifie de signature parfaitement reproductible. Au vu de ces résultats, on peut se poser la question de savoir si certaines signatures sont plus reproductibles que d'autres. Si tel est le cas, est-ce uniquement lié au nombre et à

la taille des cohortes, ou peut-on lier cela (i) au type d'objets biologiques étudiés (transcrits de gènes, miARNs...), (ii) à la méthode quantitative utilisée (méthodes haut-débit, bas débit), (iii) à la prise en compte de connaissances *a priori*, ou (iv) au choix de la méthode d'analyse bioinformatique et au soin accordé à la consolidation de la signature. Cette question est essentielle afin de mieux calibrer les méthodes de recherche de biomarqueurs, et de diminuer la part de faux-positifs présente dans la littérature scientifique (Ioannidis, 2005) et dénoncée par la communauté scientifique (Baker, 2016).

5.3 Vers de meilleurs scores de progression de maladie exploitant les données disponibles

Il existe un besoin de définir des scores de progression de maladie qui soient adaptés aux nombreuses maladies de faible prévalence et d'évolution lente, et nous avons été confrontés à ce problème dans le cadre de l'étude de deux pathologies neurodégénératives. La faible prévalence limite la taille des cohortes disponibles, tandis que l'évolution lente rend le recueil de données longitudinales pour alimenter les cohortes irréaliste. Dans les travaux présentés, nous avons notamment proposé une nouvelle approche pour cela, basée sur l'utilisation d'auto-encodeurs variationnels supervisés.

À court terme, il est nécessaire de consolider et pérenniser la nouvelle approche développée et de la traduire en un package python. Cette solution n'a pas été retenue de prime abord car les données du projets sur lequel la méthode était appliquée ne sont pas publiques et ne peuvent pas être publiées sur GitHub, pour des raisons de confidentialité. Cependant, au vu des résultats prometteurs obtenus par l'approche et au vu de sa généricité, il est important de pouvoir la mettre en œuvre et la distribuer facilement. Ce package pourra servir à évaluer la méthode de façon plus robuste sur d'autres maladies. De plus, d'autres projet, en collaboration avec le CHU de Rennes, pourrait bénéficier de la même approche ou d'une adaptation de celle-ci, comme le projet PhenoCYP porté par Camille Tron et cherchant à caractériser la capacité hépatique résiduelle de patients à partir de données multimodales de grande dimension. Une demande a été déposée dans le cadre du réseau des ingénieurs CNRS du Programme National de Recherche en Intelligence Artificielle (PNRIA) pour poursuivre ce travail.

À moyen terme, la consolidation des résultats obtenus passe par une application à d'autres pathologies. Afin d'analyser l'impact de la faible taille des cohortes sur les résultats obtenus, il faudrait étudier une pathologie de prévalence plus importante afin de disposer de cohortes de plus grande taille, pour ensuite simuler l'impact de cohorte de faible taille. De plus, afin d'évaluer la pertinence du score obtenu, il serait intéressant de d'avoir une vérité de terrain à laquelle se comparer. Dans le cadre de la maladie d'Alzheimer, des cohortes de grande tailles sont disponibles, et certaines incluent des données d'imagerie médicale et de transcriptomique. C'est une piste à explorer afin de consolider les premiers résultats obtenus.

À plus long terme, j'aimerais poser la question de l'intégration de connaissances *a priori* pour améliorer le modèle proposé. L'hypothèse sous-jacente est de compenser la faible taille des cohortes par l'intégration de connaissances. Différentes informations peuvent être intégrées, comme par exemple pour les miARNs leur

localisation chromosomique, leurs gènes ou séquences cibles... Le principal défi sera d'identifier clairement quelles connaissances prendre en compte et comment les intégrer dans le modèle. Enfin, la question des modalités prises en compte se pose également. Dans le schéma proposé, les modalités prises en compte sont de l'imagerie médicale et des données de séquençage de miARNs circulant dans le plasma sanguin, et celles-ci semblent pertinentes car nous étudions une maladie neuro-dégénérative (d'où l'importance des données d'imagerie cérébrale) et dont les formes sporadiques et familiales peuvent être liées à des mutations de la protéine TDP-43, qui joue un rôle important dans la biogénèse des miARNs (Sreedharan et al., 2008) (d'où l'étude des miARNs). Les modalités prises en compte sont donc liées à la pathologie étudiée, et il sera nécessaire d'intégrer d'autres modalités pour répondre à d'autres problèmes ou pour améliorer les résultats en prenant plus de deux modalités en entrée. Ces travaux pourraient être menés dans le cadre du PEPR Santé Numérique.

5.4 Vers une amélioration de la qualité et de la pertinence des interactomes issus de connaissance

À court terme, les travaux réalisés pour identifier un interactome reproductible peuvent être raffinés afin de différencier reproductibilité et réplicabilité, de façon analogue aux définitions de Claerbout et Karrenbach pour l'informatique (Claerbout J, 1992). Plus précisément, il s'agit de différencier ce qui est reproductible de ce qui est réplicable, selon les techniques de mise en évidence utilisées (technique identique ou différentes) et selon le nombre d'équipes ayant mis en évidence l'interaction (même équipe ou deux équipes différentes).

À moyen terme, les travaux réalisés afin de mettre en évidence les interactions reproductibles chez la levure et chez l'homme nous permettent d'aborder la question de la complémentarité des méthodes de détection des interactions protéine-protéine. Plus précisément, je souhaite explorer la question suivante : lorsqu'une interaction est mise en évidence par deux techniques très différentes l'une de l'autre, peut-on déduire des propriétés telle qu'un niveau de confiance ou une affinité pour cette interaction ? Il s'agit en quelque sorte de savoir si l'on peut généraliser les résultats décrits dans Braun et al., 2009, dans lesquels un score de confiance standardisé est déduit des résultats de 4 méthodes de détection haut débit complémentaires.
Je propose de répondre à cette question en analysant toutes les interactions reproductibles, une fois les redondances explicites et implicites éliminées. Pour ces interactions, nous pourrons extraire les couples de méthodes de détection, et mesurer (i) la (di)similarité sémantique entre ces termes de l'ontologie Molecular Interaction, et (ii) des mesures d'information mutuelles. Ces valeurs pourront ensuite être comparées à (i) des métriques de fiabilité des interactions protéine-protéines utilisées par certaines banques comme `STRING` (Szklarczyk et al., 2021) ; (ii) la présence de d'une interaction entre orthologues des deux partenaires chez des espèces proches ou plus distantes, témoignant ainsi d'une stabilité évolutive de cette interaction ; et enfin (iii) à l'abondance de chacun des deux partenaires de l'interaction *in-vivo*, car on peut supposer pour certaines méthodes un biais technique favorisant la détection des interactions les plus abondantes.

Enfin, à plus long terme, se pose la question de l'utilisation de ces informations dans les scores de confiance des interactions. En effet, certaines bases de données

comme STRING (Szklarczyk et al., 2021), IntAct (Kerrien et al., 2012) ou HIPPIE (Alanis-Lobato et al., 2017) fournissent un score de confiance associé à chaque interaction. Ces scores sont souvent une combinaison de différents sous-scores, chacun étant calculé d'une manière spécifique. La plupart de ces scores prennent en compte le nombre de publications soutenant une interaction comme élément de leur score. Cependant, cette mesure du nombre de publications pourrait être remplacée par le nombre de preuves expérimentales non redondantes.

À long terme également se pose la question de l'impact des résultats spectaculaires d'AplhaFold2-multimer sur la topologie et la définition des interactomes (Evans et al., 2022). À l'heure actuelle, il est possible de prédire la conformation d'un complexe entre protéines à partir de la séquence des deux partenaires, avec une fiabilité largement supérieure à ce qui était fait auparavant. Deux questions se posent :

1. AlphaFold-multimer pourrait-il être utilisé pour prédire des interactions, par exemple en interprétant le score de confiance par résidu (*per-residue confidence score*, pLDDT) associé aux prédictions ? Si oui, quelle est la fiabilité de ces prédictions et comment les intégrer avec les informations existantes ? Si les prédictions s'avéraient fiables, cela pourrait nous permettre de prédire des interactions fiables à grande échelle, et ainsi mieux appréhender la topologie globale de l'interactome en réduisant drastiquement le nombre de faux négatifs.

2. Lorsqu'une protéine interagit avec plusieurs partenaires, la modélisation structurale faite par AlphaFold2-multimer devrait permettre de déterminer si les interactions sont mutuellement exclusives entre elles (XOR) ou peuvent avoir lieu simultanément (OR/AND). De ce fait, on peut envisager une contextualisation ou un enrichissement des interactomes avec ces informations, tout comme on peut envisager des adaptations des méthodes d'analyse des interactomes qui tiendraient compte de ces informations.

5.5 Vers une analyse systémique intégrative

L'approche systémique permet de prendre en compte les relations connues entre entités dans le schéma d'analyse globale. Du fait de cette approche « entités » et « relations entre entités », le formalisme des graphes est souvent utilisé pour établir des modèles. Pour autant, dans le cas de données hétérogènes, l'extraction de la connaissance sous-jacente, tout comme la recherche de l'abstraction permettant au mieux de répondre à une question donnée, restent des sujets délicats.

À court terme, nos projets de recherche en cours nous conduisent à nous intéresser à la bonne méthode pour extraire de sources de données au format BioPAX les informations concernant un sous-ensemble d'entités hétérogènes d'intérêt. Dans le cadre de la thèse de Camille Juigné, ces informations hétérogènes sont de type transcriptomique et de type métabolomique. Pour identifier nos données transcriptomiques dans les données au format BioPAX, nous pouvons nous baser sur les identifiants HGNC, lesquels sont liés aux entités ProteinReference de BioPAX, à partir desquelles on peut identifier les différences instances présentes *via* les relations entityRef. Pour les données de type métabolomique, le schéma d'intégration est plus complexe car nos données d'entrée ne possèdent pas d'identifiants non-ambigu.

Il sera donc nécessaire de rechercher les chaînes de caractères correspondant à nos différents métabolites dans la base de données ChEBI. Les identifiants ChEBI pourront ensuite nous permettre de retrouver les entités SmallMoleculeReference de BioPAX correspondantes, à partir desquelles on pourra identifier les instances présentes *via* les relations entityReference. Ce travail est en cours de réalisation.

À moyen terme, une fois que nous serons capables d'identifier les sommets d'intérêt dans les données BioPAX, nous souhaitons poser la question de la meilleure stratégie pour identifier les entités qui contrôleraient simultanément ces différentes entités. Autrement dit, il s'agira d'identifier au sein des réseaux extraits de la connaissance BioPAX les régulateurs ou ensembles minimaux couvrants de régulateurs contrôlant notre ensemble de cibles. Un premier verrou méthodologique est le choix de l'abstraction représentant les connaissances, et nous envisageons à ce sujet différentes approches : (i) ne pas abstraire les informations et utiliser des requêtes SPARQL pour calculer itérativement des métriques et indices ; (ii) ne pas abstraire les informations et utiliser une base de donnée orientée graphe pour représenter le sous-réseau extrait et l'interroger ; et enfin (iii) utiliser des abstractions de plus haut niveau, comme les graphes de réactions régulées ou les graphes d'influences présentés dans 4.3. Les méthodes d'analyse de ces différents réseaux sont variées : marches aléatoire avec redémarrage à l'origine, avec pondération ou non des arêtes, avec redémarrage totalement aléatoire ou guidé, arbres de Steiner, parcours en largeur... Celles-ci nécessiteront d'être adaptées à notre objectif de recherche de contrôleurs et évaluées.

À long terme, cette méthode pourrait conduire au développement d'un outil permettant de coupler des données de transcriptomique et de métabolomique avec une approche complémentaire de celle d'outils actuels tels que Diablo au sein de la suite mixOmics (Singh et al., 2019), dont l'objectif principal est l'exploration et l'intégration d'ensembles de données biologiques, avec un accent particulier sur la sélection des variables. La recherche et l'identification de contrôleurs ou d'ensembles minimaux de contrôleurs pour ces différentes variables permettrait en effet une compréhension et une interprétabilité plus importante.
Enfin, j'aimerai beaucoup à plus long terme travailler à la question plus générale des nouveaux schémas d'intégration de données, adaptés aux données multi-omiques ou aux données multi-modales. Le cadre de travail de l'équipe Dyliss est particulièrement propice à ces développements, de par la présence de collaborateurs spécialistes des schémas d'intégration de données, et de collègues spécialistes de la biologie des systèmes à différents niveaux.

Appendix A

Annexe A : Curriculum Vitae

A.1 État civil

Emmanuelle BECKER, née le 22/12/1980 à Saint-Avold (57), mariée, 3 enfants.

Coordonnées professionnelles
Equipe Dyliss
Irisa / Inria Rennes-Bretagne Atlantique
Campus de Beaulieu, 35042 Rennes cedex
emmanuelle.becker@univ-rennes1.fr

Page personnelle : https://www-dyliss.irisa.fr/team-members/emmanuelle-becker/

A.2 Études, diplômes et parcours professionnel

2018–2021	**Maîtresse de conférences** Équipe de recherche : Dyliss, dirigée par O. Dameron (IRISA) Équipe pédagogique : Bioinformatique et Biostatistiques (SVE)	UNIV. RENNES 1
2009–2017	**Maîtresse de conférences** Équipe de recherche : Remede, dirigée par M. Primig (IRSET) Équipe pédagogique : Bioinformatique et Biostatistiques (SVE)	UNIV. RENNES 1
2007–2009	**Chercheuse post-doctorante** Sujet : *Réseaux d'interactions protéine-protéine*, encadré par C. Brun, chargée de recherche CNRS. Laboratoire : TAGC, Inserm UMR 1090 Marseille.	UNIV. DE LA MÉDITERRANNÉE
2008	**Qualification aux fonctions de maître·sse de conférences** en sections **27**, **64** et **65**.	
2003–2007	**Doctorat de bioinformatique** Titre : *Prédictions bioinformatiques des propriétés des domaines de reconnaissance peptidique.* Directeur : Raphaël Guérois.	UNIV. PIERRE & MARIE CURIE
2002–2003	**DEA de bioinformatique** Spécialité : *Analyse de Génomes et Modélisation Moléculaire*	UNIV. PIERRE & MARIE CURIE
2000–2002	**Licence et maîtrise d'informatique** Spécialité : *Magistère d'Informatique et Modélisation*.	ENS DE LYON

A.3 Publications

En bioinformatique comme en biologie, l'ordre des auteurs est codifié. En particulier, les auteurs en dernières positions sont les personnes qui ont supervisé le travail. En bioinformatique, les publications se font majoritairement par des soumissions dans des journaux plutôt que des participations à des conférences, notamment pour toucher une audience plus large. La liste complète est disponible en Annexe B.

- Depuis ma thèse, j'ai publié **35 articles**, dont :
  - 31 articles de journaux
  - 3 articles de conférences nationales ou internationales
  - 1 article de revue
  - 8 en première autrice et 5 en dernière autrice.
- Les doctorant·e·s que j'ai encadré ont publié 4 articles en premier·e aut·eur·rice (+ 2 articles de conférences nationale/internationale).

A.4 Implication dans des projets de recherche financés

Je participe ou j'ai participé aux projets de recherche financés suivants :

- 2023-2027 : Encadrement prévu d'un ou une doctorant.e dans le cadre de l'ANR `Endovire` (porteur : Nathalie Volkoff)
- 2020-2023 : Encadrement de la thèse de Camille Juigné dans le cadre d'un financement INRAe `MP DigitBio` et Région Bretagne (porteurs : Florence Gondret et Emmanuelle Becker)
- 2020-2021 : Encadrement d'un ingénieur dans le cadre du projet européen `PhenoMir` (porteur : Julien Bobe)
- 2019-2022 : Encadrement de la thèse de Virgilio Kmetzsch dans le cadre du programme `Inria Project Lab Neuromarker` (porteur : Olivier Colliot)
- 2018 : Fiancement d'un stage de M2 dans le cadre du projet `Ubiq` (porteurs : Gwenaël Rabut et Emmanuelle Becker)

A.5 Encadrement

Je co-encadre actuellement une doctorante (début de thèse 2020) et j'ai co-encadré deux doctorant·e·s ayant soutenu en 2020 et 2022. J'ai précédemment encadré un ingénieur, et 14 étudiant·e·s de licence (2), master 1 (5), master 2 (6) ou erasmus traineeship (1).

- **Encadrements en cours (1 doctorante):**
  - Camille Juigné, **doctorante**, 3 ans (2020-2023)
    Titre : *Analyse de données biologiques hétérogènes modélisées avec des graphes multiplex pour comprendre l'efficacité alimentaire.*
    Co-encadrement avec Florence Gondret, taux réel 50 %.

Production : 1 publication (Bioinformatics), 1 article soumis, 2 articles en cours de rédaction, 1 exposé oral lors d'une conférence (`JOBIM 2022`).

- **Encadrements passés (2 doctorant·e·s, 1 ingénieur) :**

  – Virgilio Kmetzsch, **doctorant**, 3 ans (2019-2022)
  Titre : *Analyse multimodale de données génomiques et de neuroimagerie dans la démence fronto-temporale*.
  Co-encadrement avec Olivier Colliot, taux réel 50 %.
  Production : 3 publications de journaux (*JNNP*, *JHBI* et *ACTN*), 1 publication de conférence (`SPIE Medical Imaging 2022`), 2 exposés oraux lors de conférences (`SPIE Medical Imaging 2022` et `JOBIM 2022`).
  Soutenue le 26 septembre 2022.

  – Léo Milhade, **ingénieur**, 10 mois (2020-2021)
  Titre : *Mise en place de workflow pour la recherche de micro-ARNs caractéristiques de conditions phénotypiques chez la truite.*
  Co-encadrement avec Julien Bobe, taux réel 80 %.
  Production : package `prostPostProcess` + résultats en cours de valorisation

  – Méline Wery, **doctorante**, 3 ans (2017-2020)
  Titre : *intégration de données et analyse basée sur le raisonnement à chaque étape du développement de médicament.*
  Co-encadrement avec Anne Siegel, Olivier Dameron et Charles Bettembourg, taux réel 25 %.

- **Encadrements de stagiaires (14) :**

  – Nancy D'Arminio, erasmus traineeship de 4 mois - pré-doctorat (2021) : *extraction automatique de relations entre les ubiquitines ligases de la levure et leurs protéines subtrats d'après la littérature* (taux 50%).

  – Emmanuel Clostres, stage de 6 mois niveau M2 (2021) : *implication des miARNs dans les interactions plantes-rhizomicrobiome : automatisation de la recherche et l'identification de gènes cibles bactériens* (taux 25%)

  – Marc Melkonian, stage de 6 mois niveau M2 (2020) : *detection of redundancies in protein-protein interaction databases encoded by different PSI-MI identifiers using semantic web technologies* (taux 75%)
  Production : 1 publication (Bioinformatics)

  – Quentin Delhon, stage de 6 mois niveau M2 (2020) : *Discovery of key genes in multi-layered biological graphs (taux 75%).*
  Production : développements au sein de l'outil `pax2graphml`

  – Camille Juigné, stage de 6 mois niveau M2 (2019) : *intégration de données et analyse formelle du réseau d'interactions entre enzymes d'ubiquitination* (taux 75%).

- Fanny Casse, stage de 6 mois niveau M2 (2019) : *régulation par les miARNs des gènes régulant la fécondité et le développement embryonnaire précoce chez le poisson médaka* (taux 33%)
- Hugo Simon, stage de 2 mois niveau L3 (2019) : *analyses topologiques des réseaux de régulation et métaboliques* (taux 75%).
  Production : une partie de l'outil `pax2graphml` + associé à 1 publication (Bioinformatics)
- Anaëlle Caillarec-Joly, stage de 2 mois niveau M1 (2017) : *Modélisation quantitative de la stéroïdogénèse chez la lignée H295R : étude de l'impact de perturbations du système* (taux 75%)
  Production : associée à 1 publication (Bioinformatics)
- Thomas Chaussepied, stage de 2 mois niveau M1 (2017) : *Impact des antalgiques sur le développement des organes de la sphère uro-génitale chez le fœtus* (taux 50%)
- Victoria Potdevin, stage de 2 mois niveau M1 Agronome (2016) : *classification de signatures toxico-génomique* (taux 75%).
- Estelle Lecluze, stage de 6 mois niveau M2 (2015) : *Implémentation de modèles prédictifs pour l'identification et la classification de nouveaux perturbateurs endocriniens.* (taux 50%)
- Marine Brenet, stage de 2 mois niveau L3 (2015) : *Identification des gènes impliqués dans la sporulation chez la levure* (taux 50%)
- Paul Guéguen, stage de 2 mois niveau M1 (2014) : *Classification de données massives de toxico-génomique par des méthodes de clustering chevauchant et d'optimisation du critère de modularité de Newman* (taux 75%)
- Sofiane Omari, stage de 2 mois niveau M1 (2014) : *Etude de la sporulation chez la levure par intégration de données de transcriptomique et de protéomique quantitative* (taux 50%)

A.6 Tâches collectives

Je suis co-responsable du comité programme d'une conférence nationale, et j'ai été membre du comité de programme de 3 conférences dont 2 internationales. J'ai également participé à 5 commissions de sélection (postes CR ou MCU) et 1 jury de thèse.

- Conférences (4) :
  - co-responsable du comité de programme de la conférence Jobim 2022 (Rennes)
  - membre du comité de programme de la conférence Jobim 2021 (Paris)
  - membre du comité de programme de la conférence BBCC20 (en ligne)
  - membre du comité de programme de la conférence NETTAB/BBCC 2019 (joint meeting, Salerno)

- Comités de sélection (4) :
  - membre du comité de sélection du poste de chargé de recherche CR24 INRAe profils « IA et signalisation » et « Génomique comparée des poissons » (2022)
  - membre du comité de sélection du poste de maître·sse de conférences de l'Université d'Aix-Marseille profil « Bioinformatique et Génomique » (2021)
  - membre du comité de sélection du poste de maître·sse de conférences de l'Université de Lyon profil « Biochimie et interactions biomoléculaires » (2021)
  - membre du comité de sélection du poste de maître·sse de conférences de l'Université de Rennes profil « Modélisation des systèmes biologiques et biostatistiques » (2019)

- Jurys de thèse (1) :
  - Examinatrice de la thèse de Saran PANKAEW : *Study of impact of Pten-loss in leukemic development of T cell: Boolean modelling and transcriptomic analysis at single-cell resolution* (Aix-Marseille Université, 2022)

- Au sein du laboratoire (2) :
  - participation au jury de sélection des doctorant·e·s pour les bourses Moyens Incitatifs (2020)
  - membre du conseil scientifique de Biogenouest (réseau des 37 plates-formes technologiques du Grand Ouest en sciences du vivant et de l'environnement).

A.7 Enseignements et responsabilités pédagogiques

Ces cinq dernières années, je suis ou j'ai été responsable de 8 UE et je suis co-responsable du master de bioinformatique.

- **Responsabilités actuelles (6 UEs, 1 formation):**
  - **co-responsable du master mention Bioinformatique depuis 2017**
    Après 8 années en tant qu'intervenante dans le master de bioinformatique, j'en ai pris la co-responsabilité en 2017 en défendant une mention de bioinformatique. Cette responsabilité a notamment impliqué de monter de nouvelles maquettes (passant de 1 à 3 parcours), de monter un co-portage entre les UFRs SVE et Médecine de l'Université, et de mettre en place un accord Erasmus+. Le nombre d'étudiant·e·s a fortement augmenté dans cette période (de 20-25 à 40-50).
  - **responsable de 6 UEs de licence ou master** (détails ci-dessous) pour lesquelles j'ai crée les supports actuels.

- **Détails des unités d'enseignement en responsabilité (5 dernières années):**

  - **Atelier de biostatistiques 1**
    Années en responsabilité de l'UE : depuis 2012 - en cours
    Public visé : L2 Sciences de la Vie (SVE)

  - **Advanced R for data Analysis** (cours en langue anglaise)
    Années en responsabilité de l'UE : depuis 2017 - en cours
    Public visé : M1 Bioinformatique (SVE) + M1 Ecologues (OSUR)

  - **Programmation orientée objet en bioinformatique**
    Années en responsabilité de l'UE : depuis 2012 - en cours
    Public visé : M1 Bioinformatique (SVE)

  - **Réseaux biologiques**
    Années en co-responsabilité de l'UE : depuis 2012 - en cours (responsabilité partagée avec Anne Siegel)
    Public visé : M2 Bioinformatique (SVE)

  - **Introduction to computational ecology** (cours en langue anglaise)
    Années en responsabilité : depuis 2017 - en cours (responsabilité partagée avec Frédéric Hamelin)
    Public visé : M2 Ecologues parcours EFCE et MODE (OSUR)

  - **Method** (cours en langue anglaise)
    Années en responsabilité : depuis 2019 - en cours
    Public visé : M2 Informatique parcours SIF (ISTIC)

  - **Maths 2 : statistiques appliquées**
    Années en responsabilité de l'UE : 2016 - 2020
    Public visé : L3 Informatique (ENS Rennes)
  - **Outils bioinformatiques**
    Années en responsabilité de l'UE : 2012 - 2018
    Public visé : L2 Sciences de la Vie (SVE)

- Autres activités :

  - participation aux portes ouvertes et salon de l'étudiant ;
  - travail autour d'un référentiel de compétences en bio-statistiques ;
  - cours à l'Université de Salerne en 2018 et 2019 dans le cadre de l'accord Erasmus+ (12h par semaine).

Appendix B

Annexe B : Liste des publications

En bioinformatique comme en biologie, l'ordre des auteurs est codifié. En particulier, les auteurs en dernières positions sont les personnes qui ont supervisé le travail. En bioinformatique, les publications se font majoritairement par des soumissions dans des journaux plutôt que des participations à des conférences, notamment pour toucher une audience plus large.

Depuis ma thèse, j'ai publié **35 articles**, dont 8 en première autrice et 5 en dernière autrice. Les doctorant·e·s que j'encadre ou que j'ai encadré ont publié 4 articles en premier·e aut·eur·rice (+ 2 articles de conférences nationale/internationale).

Afin de faciliter la lecture de la liste de publication, j'utilise un code couleur permettant de différencier les contributions en lien avec le chapitre 2, le chapitre 3 ou le chapitre 4. Les contributions de type journaux, conférences ou articles de revue sont présentées successivement.

Articles de journaux (31)

[35] Kmetzsch V, Latouche M, Saracino D, Rinaldi D, Camuzat A, Garreau T, Le Ber I, Colliot O and **Becker E**.
Validation of circulating microRNA signatures as biomarkers in genetic frontotemporal dementia and amyotrophic lateral sclerosis.
accepted in Annals of Clinical and Translational Neurology

[34] Kmetzsch V, **Becker E**, Saracino D, Rinaldi D, Camuzat A, Le Ber I, and Colliot O.
Disease progression score estimation from multimodal imaging and microRNA data using supervised variational autoencoders.
accepted in IEEE J Biomed Health Inform

[32] Petit F*, Jamin S, Kernanec P-Y, **Becker E**, Halet G, and Primig M*.
EXOSC10/Rrp6 is essential for the eight-cell embryo/morula transition.
Dev. Biol, 2022

[30] Melkonian M, Juigné C, Dameron O, Rabut G* and **Becker E***.
Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases.
Bioinformatics, 2022.

[29] Moreews F, Simon H, Siegel A, Gondret F, **Becker E.**
PAX2GRAPHML : a python library for large-scale regulation network analysis using BIOPAX.
Bioinformatics, 2021.

[28] Kmetzsch V, Anquetil V, Saracino D, Rinaldi D, Camuzat A, Gareau T, Jornea L, Forlani S, Couratier P, Wallon D, Pasquier F, Robil N, PREV-DEMALS study group; de la Grange P, Moszer I, Le Ber I, Colliot O, **Becker E.**
Plasma microRNA signature in presymptomatic and symptomatic subjects with C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis.
J Neurol Neurosurg Psychiatry, 2021.

[27] Leverrier-Penna S, Michel A, Lecante LL, Costet N, Suglia A, Desdoits-Lethimonier C, Boulay H, Viel R, Chemouny JM, **Becker E**, Lavoué V, Rolland AD, Dejucq-Rainsford N, Vigneau C, Mazaud-Guittot S.
Exposure of human fetal kidneys to mild analgesics interferes with early nephrogenesis.
FASEB J, 2021

[26] Xie B, **Becker E**, Stuparevic I, Wery M , Descrimes M, Morillon A, and Primig M.
The anti-cancer drug 5-fluorouracil affects cell cycle regulators and long non-coding RNAs with regulatory potential in yeast.
RNA Biology 2019.

[25] Darde T, Gaudriault P, Beranger R, Lancien C, Caillarec-Joly A, Sallou O, Bonvallot N, Costet N, Chevrier C, Mazaud-Guittot S, Jégou B, Collin O, **Becker E**, Rolland A-D and Chalmel F.
TOXsIgN : A public repository for toxicological signatures at the IRSET.
Bioinformatics, 2018

[24] Leverrier-Penna S, Mitchell RT, **Becker E**, Lecante L, Ben Maamar M, Lavoué V, ristensen DM, Jégou B, Dejucq-Rainsford N, Mazaud-Gittot S.
Ibuprofen is deleterious for the development of first trimester human fetal ovary ex vivo.
Human Reprod, 2018

[23] Houzet L , Pérez-Losada M, Matusali G, Deleage C, Dereuddre-Bosquet N, Satie AP, Aubry F, **Becker E**, Jégou B, Le Grand R, Keele BF, Crandall KA and Dejucq-Rainsford N.
Seminal SIV in cynomolgus macaques is dominated by virus originating from multiple genital organs.
Journal of Virology, 2018

[22] Dugay F, Llamas-Gutierrez F, Gournay M, Medane S, Mazet F, Christian-Chiforeanu D, **Becker E**, Lamy R, Léna H, Rioux-Leclercq N, Belaud-Rotureau MA and Cabilic F.
Clinicopathological characteristics of ROS1- and RET-rearranged NSCLC in caucasian patients. Data from a cohort of 713 non-squamous NSCLC lacking KRAS/EGFR/HER2/BRAF/PIK3CA/ALK alterations.
Oncotarget, 2017

[21] **Becker E**, Com E, Lavigne R, Guilleux M-H, Evard B, Pineau C and Primig M.
The protein expression landscape of mitosis and meiosis in diploid budding yeast.
J. Proteomics, 2017

[20] Hao* C, Gely-Pernot* A, Kervarrec C, Boudjema M, **Becker E**, Khil P, Tevosian S, Jégou B and Smagulova F.
Exposure to the widely used herbicide atrazine results in deregulation of global tissue-specific RNA transcription in the third generation and is associated with a global decrease of histone trimethylation in mice.
Nucleic Acids Research, 2016

[19] Xie B, Horecka J, Chu A, Davis RW, **Becker E** and Primig M.
Ndt80 activates the meiotic ORC1 transcript isoform and SMA2 via a bidirectional middle sporulation element in Saccharomyces cerevisiae
RNA Biology, 2016.

[18] Chapple C E, Robisson B, Spinelli L, Guien C, **Becker E**, Brun C.
Extreme multifunctional proteins identified from a human protein interaction network.
Nature Communications, 2015

[17] Darde T, Sallou O, **Becker E**, Evrard B, Monjeau C, Le Bras Y, Jégou B, Collin O, Rolland A and Chalmel F.
The ReproGenomics Viewer: an integrative cross-species toolbox for the reproductive science community.
Nucleic Acids Research, Web-Server issue, 2015

[16] Liu* Y, Stuparevic* I, Xie B, **Becker E**, Law M and Primig M.
The conserved histone deacetylase Rpd3 and the DNA binding regulator Ume6 repress BOI1's meiotic transcript isoform during vegetative growth in Saccharomyces cerevisiae.
Mol. Microbiology, 2015.
Corrected in Mol. Microbiology, 2015, Volume 99, Issue 1, 217.

[15] Gely-Pernot* A, Hao* C, **Becker E**, Stuparevic I, Kervarrec C, Chalmel F, Primig M, Jégou B and Smagulova F.
Worldwide used herbicide atrazine affects meiosis in male mice.
BMC Genomics, 2015.

[14] **Becker E**, Liu Y, Lardenois A, Walther T, Horecka J, Stuparevic I, Law M, Lavigne R, Evrard E, Demougin P, Riffle M, Strich R, Davis R W, Pineau C and Primig M.
Integrated RNA- and protein profiling of fermentation and respiration in diploid budding yeast provides insight into nutrient control of cell growth and development.
J. of Proteomics, 2015

[13] Petit* F G, Kervarrec* C, Jamin* S P, Smagulova F, Hao C, **Becker E**, Jégou B, Chalmel F and Primig M.
Combining RNA- and Protein Profiling Data With Network Interactions Yields Clues About Gene Function in Mouse and Human Spermatogenesis
Biol. of Reproduction, 2015

[12] Stuparevic I, **Becker E**, Law M and Primig M.
The histone deacetylase Rpd3/Sin3/Ume6 complex represses an acetate inducible isoform of VTH2 in fermenting budding yeast cells.
FEBS Letters, 2015

[11] Lardenois A*, **Becker E***, Walther T*, Law M, Demougin P, Strich R and Primig M.
Global alterations of the transcriptional landscape during budding yeast growth and development in the absence of Ume6-dependant chromatin modification.
Mol. Genet. Genomics, 2015

[10] Lardenois A*, Stuparevic I*, Liu Y*, Law M J, **Becker E**, Smagulova F, Waern K, Guilleux M-H, Horecka J, Chu A, Kervarrec C, Strich R, Snyder M, Davis R W, Steinmetz L M, and Primig M.
The conserved histone deacetylase Rpd3 and its DNA binding subunit Ume6 control dynamic transcript architecture during mitotic growth and meiotic development.
Nucleic Acids Research, 2014

[09] Lavigne R, **Becker E**, Liu Y, Evrard B, Lardenois A, Primig M and Pineau C.
Direct iterative protein profiling (DIPP) - an innovative method for large-scale protein detection applied to budding yeast mitosis.
Mol Cell Proteomics, 2012

[08] **Becker E***, Robisson B*, Chapple CE, Guénoche A and Brun C.
Multifunctional proteins revealed by overlapping clustering in protein interaction network.
Bioinformatics, 2012

[07] Souiai O*, **Becker E***, Prieto C, Benkahla A, De las Rivas J and Brun C.
Functional integrative levels in the human interactome recapitulate organ organization.
PLoS One, 2011

[06] Aucher W*, **Becker E***, Ma E, Miron S, Martel A, Ochsenbein F, Marsolier-Kergoat M-C and Guerois R.
A strategy for interaction site prediction between phospho-binding modules and their partners identified from proteomic data.
Mol Cell Proteomics, 2010

[04] **Becker E**, Cotillard A, Meyer V, Madaoui H, Guerois R.
HmmKalign : a tool for generating sub-optimal HMM alignments.
Bioinformatics, 2007

[03] **Becker E***, Meyer V*, Madaoui H, Guerois R.
Detection of a tandem BRCT in Xrs2 and Nbs1 with functional implications in the DNA damage response.
Bioinformatics, 2006

[01] Mousson F, Lautrette A, Thuret JY, Agez M, Courbeyrette R, Amigues B, **Becker E**, Neumann JM, Guerois R, Mann C, Ochsenbein F.
Structural basis for the interaction of Asf1 with histone H3 and its functional implications.
Proc Natl Acad Sci U S A, 2005.

Articles de conférences nationales ou internationales avec comité de lecture (3) .

[33] Juigné C, Dameron O, Moreews F, Gondret F and **Becker E**.
Detection and correction of non-conformities and redundancies in complexes of molecules in BioPAX.
Actes JOBIM 2022, Rennes, 5-8 juillet 2022

[31] Kmetzsch V, **Becker E**, Saracino D, Anquetil V, Rinaldi D, Camuzat A, Gareau T, Le Ber I, Colliot O and the PREV-DEMALS study group.
A multimodal variational autoencoder for estimating progression scores from imaging and microRNA data in neurodegenerative diseases.
SPIE Medical Imaging 2022, 2022

[05] **Becker E.**, Guénoche A. and Brun C.
Systèmes de classes chevauchantes pour la recherche de protéines multifonctionnelles.
Actes JOBIM 2009, p.49-54, Nantes, 9-11 juin 2009

Articles de revue (1) .

[02] Madaoui H, **Becker E**, Guérois R
Sequence search methods and scoring functions for the design of protein structures.
Methods in Molecular Biology, 2006

Appendix C

Annexe C : Publications jointes

C.1 Contributions jointes en support au chapitre 2

[28] Kmetzsch V, Anquetil V, Saracino D, Rinaldi D, Camuzat A, Gareau T, Jornea L, Forlani S, Couratier P, Wallon D, Pasquier F, Robil N, PREV-DEMALS study group; de la Grange P, Moszer I, Le Ber I, Colliot O, **Becker E.**
Plasma microRNA signature in presymptomatic and symptomatic subjects with C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis.
J Neurol Neurosurg Psychiatry, 2021.

[35] Kmetzsch V, Latouche M, Saracino D, Rinaldi D, Camuzat A, Garreau T, Le Ber I, Colliot O and **Becker E**.
Validation of circulating microRNA signatures as biomarkers in genetic frontotemporal dementia and amyotrophic lateral sclerosis.
Annals of Clinical and Translational Neurology, 2022.

Original research

Plasma microRNA signature in presymptomatic and symptomatic subjects with *C9orf72*-associated frontotemporal dementia and amyotrophic lateral sclerosis

Virgilio Kmetzsch ,[1,2] Vincent Anquetil,[2] Dario Saracino,[1,2,3,4] Daisy Rinaldi,[2,3,4] Agnès Camuzat,[2,5] Thomas Gareau,[2] Ludmila Jornea,[2] Sylvie Forlani,[2] Philippe Couratier,[6] David Wallon,[7] Florence Pasquier,[8] Noémie Robil,[9] Pierre de la Grange,[9] Ivan Moszer,[2] Isabelle Le Ber,[2,3,4,10] Olivier Colliot,[1,2] Emmanuelle Becker,[11] The PREV-DEMALS study group

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jnnp-2020-324647).

For numbered affiliations see end of article.

Correspondence to
Dr Emmanuelle Becker, Dyliss team, Irisa / Inria Rennes-Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France; emmanuelle.becker@univ-rennes1.fr

Received 20 July 2020
Revised 30 September 2020
Accepted 27 October 2020
Published Online First 25 November 2020

To cite: Kmetzsch V, Anquetil V, Saracino D, *et al*. *J Neurol Neurosurg Psychiatry* 2021;**92**:485–493.

ABSTRACT

Objective To identify potential biomarkers of preclinical and clinical progression in chromosome 9 open reading frame 72 gene (*C9orf72)*-associated disease by assessing the expression levels of plasma microRNAs (miRNAs) in *C9orf72* patients and presymptomatic carriers.

Methods The PREV-DEMALS study is a prospective study including 22 *C9orf72* patients, 45 presymptomatic *C9orf72* mutation carriers and 43 controls. We assessed the expression levels of 2576 miRNAs, among which 589 were above noise level, in plasma samples of all participants using RNA sequencing. The expression levels of the differentially expressed miRNAs between patients, presymptomatic carriers and controls were further used to build logistic regression classifiers.

Results Four miRNAs were differentially expressed between patients and controls: miR-34a-5p and miR-345-5p were overexpressed, while miR-200c-3p and miR-10a-3p were underexpressed in patients. MiR-34a-5p was also overexpressed in presymptomatic carriers compared with healthy controls, suggesting that miR-34a-5p expression is deregulated in cases with *C9orf72* mutation. Moreover, miR-345-5p was also overexpressed in patients compared with presymptomatic carriers, which supports the correlation of miR-345-5p expression with the progression of *C9orf72*-associated disease. Together, miR-200c-3p and miR-10a-3p underexpression might be associated with full-blown disease. Four presymptomatic subjects in transitional/prodromal stage, close to the disease conversion, exhibited a stronger similarity with the expression levels of patients.

Conclusions We identified a signature of four miRNAs differentially expressed in plasma between clinical conditions that have potential to represent progression biomarkers for *C9orf72*-associated frontotemporal dementia and amyotrophic lateral sclerosis. This study suggests that dysregulation of miRNAs is dynamically altered throughout neurodegenerative diseases progression, and can be detectable even long before clinical onset.

Trial registration number NCT02590276.

INTRODUCTION

Frontotemporal dementia (FTD) designates neurodegenerative dementias characterised by progressive behavioural, executive and language impairments.[1] Amyotrophic lateral sclerosis (ALS) is a degenerative disease of motor neurons that leads to progressive muscle atrophy and motor deficit. FTD and ALS form a clinical continuum, as these two diseases may be associated in the same patients (FTD-ALS) or within families. They also share common pathophysiological mechanisms and genetic causes.[2] The most frequent genetic cause of familial FTD and ALS is a hexanucleotide (GGGGCC) repeat expansion in the chromosome 9 open reading frame 72 (*C9orf72*) gene.[3 4] This autosomal dominant mutation may cause neurodegeneration through *C9orf72* loss of function, aggregation of mutant RNA in nuclear foci and of dipeptide repeats generated by repeat-associated non-AUG translation, ultimately leading to pathological inclusions of TAR-DNA binding protein 43 (TDP-43).[5]

There are no effective treatments available in *C9orf72* disease to date, but several promising trials including antisense therapies are being developed. Presymptomatic *C9orf72* carriers represent an optimal target population for the development of new therapeutic interventions for FTD and ALS.[6 7] Therefore, it is of paramount importance to identify biomarkers of preclinical progression for FTD and ALS, which could be used to initiate and monitor potential disease-modifying treatments before any irreversible brain damage has occurred.

There is increasing evidence that microRNA (miRNA) expression in body fluids, such as plasma/serum[8] or cerebrospinal fluid (CSF),[9] correlates with the diagnosis and progression of many neurodegenerative diseases, including FTD[10] and ALS.[11] MicroRNAs are a class of small non-coding RNAs that negatively regulate gene expression by promoting translational repression and messenger RNA degradation.[12] Since TDP-43 promotes miRNA biogenesis,[13] the dysregulation of TDP-43 activity associated with FTD and ALS pathogenesis could impact miRNA expression levels.[14] Notably,

miRNAs originating from neurons and glial cells are released through extracellular vesicles, especially exosomes, and can be measured in different body fluids, including CSF and plasma.[15] Aberrant expression of miRNAs can be thus non-invasively detected in easily accessible body compartments, such as blood plasma, and potentially serve as biomarkers.[16]

Previous studies have explored selected plasma miRNAs as biomarkers for FTD/ALS[17] or FTD[18 19] using quantitative real-time PCR. Two of them have analysed the expression of a limited number of candidate miRNAs: nine miRNAs linked with apoptosis[18] or 37 brain-enriched miRNAs.[17] A wider miRNA profiling study[19] analysed 752 miRNAs, as a first attempt to perform an unbiased assessment of circulating miRNAs in patients with FTD. In addition, a more recent study[20] assessed the expression levels of 2313 miRNAs in a merged cohort of patients with FTD with different genetic forms (*C9orf72*, *MAPT*, *GRN*, *TBK1*) or with sporadic forms, by next generation RNA sequencing (RNA-seq). However, results among different studies have been conflicting so far, probably due to the heterogeneity of cohorts with respect to the underlying pathology (genetic or sporadic). Besides, these studies only compared healthy controls and symptomatic patients, focussing on evaluating potential diagnostic biomarkers. To date, no studies have evaluated plasma miRNAs as progression biomarkers for FTD or ALS in presymptomatic individuals.

The present work aims at investigating expression levels of plasma miRNAs in a large homogeneous genetic cohort of *C9orf72* mutation carriers, both in the presymptomatic and in the clinical phases, to identify potential non-invasive biomarkers of preclinical and clinical progression in *C9orf72*-associated FTD and ALS. We hypothesise that performing large scale RNA-seq analyses in plasma samples, without a priori assumptions, will reveal significant differences in miRNA expression levels between healthy controls, presymptomatic and symptomatic mutation carriers.

MATERIAL AND METHODS

Participants

PREV-DEMALS (https://clinicaltrials.gov/ Identifier: NCT02590276) is a national multicentric study focussed on *C9orf72* mutation carriers. Between 2015 and 2017, 111 individuals were investigated with the same protocol in four French university hospitals (Paris, Limoges, Lille and Rouen), as previously described.[6 21] Written informed consents were obtained from all participants.

This cohort included 22 patients (15 FTD, 4 FTD/ALS and 3 ALS) carrying a *C9orf72* expansion and 89 asymptomatic first-degree relatives of *C9orf72* patients (who have 50% risk to carry the mutation), out of 64 families. A pathogenic expansion was detected in 46 of them, denoted as the 'presymptomatic group'. The control group was formed by the 43 asymptomatic individuals that did not carry an expansion.

At inclusion, each participant's cognitive and behavioural clinical status was assessed based on standardised interview with relatives, comprehensive neurological examination, an extensive neuropsychological battery assessing all cognitive domains (including, notably, mini-mental state examination, Frontal Assessment Battery, Mattis Dementia Rating Scale and Ekman faces test) and behavioural scales (including Frontal Behavioural Inventory and Apathy Evaluation Scale) (table 1). The cognitive and behavioural evaluations and their scores have been described in more detail elsewhere[6 21] and in online supplemental appendix A1. Neuromuscular function was thoroughly evaluated by means of quantitative motor testing according to Medical Research Council muscle scale, assessment of upper and lower motor neuron signs and administration of ALS-FRS (ALS-Functional Rating Scale), evaluating the degree of functional impairment. All participants underwent a systematic standardised interview to investigate the presence of cramps, fatigue, muscle pain, muscle weakness, muscle stiffness or fasciculations. Electromyography was proposed to the participants with even subtle motor signs or complaints.

One participant was excluded because mild cerebellar syndrome was detected at a neurological examination, after inclusion. Thus, the present study comprises 110 individuals (22 patients, 45 presymptomatic carriers and 43 healthy controls), all of which underwent plasma sampling at their inclusion. The demographic and clinical characteristics of the studied population are shown in table 1.

The participants have then been clinically followed after their inclusion during a 3-year period, from 2017 to 2020. Four out of the 45 *C9orf72* presymptomatic carriers have developed subtle frontal cognitive and/or behavioural changes and/or motor signs/symptoms during this period, without fitting diagnostic criteria for FTD or ALS, suggesting they were in the transitional 'prodromal' phase at the moment of or just after their inclusion visit. These cases are described in online supplemental appendix A2. All analyses in the presymptomatic group were performed with (n=45) and without (n=41) the four prodromal subjects. We also analysed these cases separately in an additional complementary approach.

Plasma collection and preparation

Blood samples were collected on EDTA using the same standardised collection and handling procedures for all participants across the centres. The mean disease duration at sampling was 6.2±4.0 years in the patients' group. All were in fasted state. All samples were centralised at the ICM DNA and cell bank, and processed using the same protocol. Plasma was extracted at room temperature after centrifugation of blood samples at 2500 rpm for 10 min. Aliquots of 1 mL were stored in polypropylene tubes at −80°C.

MiRNA extraction and sequencing

MiRNA extraction was performed with a miRNeasy Serum/Plasma Kit (Qiagen) following the manufacturer's instructions. We used 200 µL of plasma quickly melted and directly added to 1 mL of QIAzol solution. MiRNAs were eluted in 14 µL of water; 5 µL were used for miRNA sequencing library preparation with QIAseq miRNA Library Kit (Qiagen) according to the manufacturer's protocol.

MiRNA sequencing was performed on Illumina NovaSeq 6000 in three independent batches, targeting a minimum of 10 million mapped reads per sample. Since batch effects may have a critical impact in high-throughput experiments, we randomly assigned each individual to one batch, equally distributing clinical status (control, presymptomatic and patient) and centres (Paris, Limoges, Lille and Rouen), to allow adjusting for batch effects during data analysis. Online supplemental table A1 describes the distribution of subjects across batches.

Raw reads to miRNA counts computation pipeline

Quality control of raw reads was performed with FastQC (Andrews S. 2010, http://www.bioinformatics.babraham.ac.uk/projects/fastqc). UMI-tools[22] and Cutadapt[23] were used respectively to extract UMIs and suppress adapting sequences as well as

Table 1 Demographic and clinical characteristics of the studied population

	Control (n=43)	Presymptomatic (n=45)	Patient (n=22)			
				χ^2 P value		
Female gender	23 (53.5%)	28 (62.2%)	10 (45.4%)	0.408		
				Kruskal-Wallis P value	Comparison	Dunn's test P value
Age at inclusion (years)	46.4±13.5	41.8±11.8	62.7±10.5	**<0.001**	Control vs presymptomatic	0.118
					Control vs patient	**<0.001**
					Presymptomatic vs patient	**<0.001**
ALS-FRS	39.5±1.3	39.5±1.9	33.4±7.7	**<0.001**	Control vs presymptomatic	0.827
					Control vs patient	**<0.001**
					Presymptomatic vs patient	**<0.001**
MMSE	29±1.2	28.5±1.4	17.8±8.4	**<0.001**	Control vs presymptomatic	0.183
					Control vs patient	**<0.001**
					Presymptomatic vs patient	**<0.001**
MDRS	142.1±1.8	141.2±3.0	97.3±36.7	**<0.001**	Control vs presymptomatic	0.431
					Control vs patient	**<0.001**
					Presymptomatic vs patient	**<0.001**
FAB	17±1.2	17.2±0.9	9.7±5.3	**<0.001**	Control vs presymptomatic	0.583
					Control vs patient	**<0.001**
					Presymptomatic vs patient	**<0.001**
Ekman faces test	30.1±2.6	30.1±2.3	18±9.1	**0.001**	Control vs presymptomatic	0.694
					Control vs patient	**<0.001**
					Presymptomatic vs patient	**0.001**
FBI	0.9±1.8	1.5±2.7	28.5±15.2	**<0.001**	Control vs presymptomatic	0.387
					Control vs patient	**<0.001**
					Presymptomatic vs patient	**<0.001**
AES	4.8±3.9	6.5±3.6	23.5±13.1	**<0.001**	Control vs presymptomatic	0.095
					Control vs patient	**<0.001**
					Presymptomatic vs patient	**0.004**

Values are expressed as mean±SD, or as number (%). Demographic characteristics were compared between groups using the χ^2 test for gender and Kruskal-Wallis with Dunn's test for numerical variables.
Statistically significant p values are in bold.
AES, Apathy Evaluation Scale; ALS-FRS, ALS Functional Rating Scale; FAB, Frontal Assessment Battery; FBI, Frontal Behavioral Inventory; MDRS, Mattis Dementia Rating Scale; MMSE, mini-mental state examination.

polyA tails. The resulting sequences were aligned with Bowtie[24] and sorted by genomic location with Samtools sort.[25] PCR bias was corrected with UMI-tools, its efficacy was assessed per chromosome with Samtools idxstats. After controlling for the overlap/ambiguity between miRNAs enrichment and Gencode annotation with FeatureCounts,[26] miRNAs were counted with miRDeep2.[27]

Statistical analysis

Statistical analyses were performed using R V.3.6.1 (R Foundation for Statistical Computing, Vienna, Austria). The differential expression of miRNAs between clinical groups was assessed with the R package EdgeR.[28] The analysis began with a count matrix with 2576 rows (one per miRNA i) and 110 columns (one per individual j). Only miRNAs considered above noise level (minimum count of 50 reads for at least one sample and a minimum total count of 1000) were retained for statistical analyses, reducing the count matrix to 589 rows. We assumed that miRNA counts followed a negative binomial distribution with mean μ_{ij} and dispersion ϕ_i and used generalised linear models to fit a log-linear model

$$log_2\mu_{ij} = x_j^T\beta_i$$

for each miRNA, where x_j is the vector of covariates that describes sample j and β_i is the vector of coefficients to be fitted for miRNA i. To control for possible batch, centre, age and gender effects, we added these variables as covariates in the model, in addition of clinical status. Raw counts were normalised using a trimmed mean of M-values.[29] Once the models were fitted, quasi-likelihood F-test was employed to determine the subset of miRNAs differentially expressed between clinical conditions (miRNA signature). Statistical significance was set at level $\alpha = 0.05$ and p values were adjusted for multiple testing using the Benjamini-Hochberg method.

Machine learning for binary classification

After the differentially expressed miRNAs were identified, we implemented logistic regression classifiers with L2 regularisation in Python 3.8.0 using scikit-learn[30] V.0.22.1. We used the expression levels of the miRNA signature as features to train binary classification models for each pairwise comparison between clinical status: controls versus presymptomatic individuals, controls versus patients and presymptomatic individuals versus patients. A stratified nested cross-validation strategy (online supplemental figure A1) was chosen to find the optimal hyperparameter (L2 regularisation coefficient) and to assess model performance using

the area under the receiver operating characteristic curve (ROC AUC). We computed 90% CIs for the ROC AUC scores from 2000 bootstrap samples, by taking the 5th and 95th percentiles of the bootstrap distribution. Stratification with respect to clinical status was performed to preserve the proportion of healthy controls, presymptomatic subjects and patients in each fold.

Generalisation analysis

Since the differentially expressed miRNAs were computed with the entire data set, the test folds of the cross-validation were also used in the feature selection for our classification models, which could inflate prediction performance. To estimate this possible bias, we then incorporated feature selection in the nested 5-fold cross-validation process: differentially expressed miRNAs were computed using only the outer cross-validation loop training data (four out of five folds) at each iteration. The nested cross-validation was repeated 100 times with different fold splits to assess the generalisation performance of our classifiers.

Analysis of the transitional stage to clinical FTD/ALS disease

Since we hypothesised that the expression levels of differentially expressed miRNAs might provide information relevant to *C9orf72* disease progression, we designed an experiment to evaluate prediction performance of clinical conversion to FTD/ALS in presymptomatic carriers. A logistic regression classifier was fitted with the expression levels of differentially expressed miRNAs from controls and patients. We used a regular 5-fold cross-validation to determine the optimal hyperparameter (L2 regularisation coefficient). Subsequently, this model was tested with the expression levels from the four known presymptomatic carriers who were in their transitional stage to the clinical disease. Scores from 0 to 1 were provided for each subject, indicating proximity with the expression levels of controls (scores near 0) or patients (scores closer to 1).

Target prediction and pathway analysis

A target-gene based miRNA enrichment analysis was performed, to discover potential biological functions regulated by the differentially expressed miRNAs. We used the publicly available tool DIANA-miRPath V.3,[31] which implements an in silico miRNA target prediction algorithm (DIANA-microT-CDS) as well as an experimentally validated miRNA:gene interaction dataset (DIANA-TarBase V.7.0). Both approaches were carried out to identify target genes and the associated Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, using the set of differentially expressed miRNAs as input. The enrichment analysis method consisted of Fisher's exact test (hypergeometric distribution) with Benjamini-Hochberg adjusted p value threshold of 0.05, giving as output a union set of associated KEGG pathways.

RESULTS

Differentially expressed miRNAs computed with the entire data set

Table 2 displays all miRNAs identified as differentially expressed, for each pairwise comparison between clinical status, after correction for multiple comparisons. Four miRNAs were computed as differentially expressed between healthy controls and patients: miR-34a-5p and miR-345-5p were overexpressed, while miR-200c-3p and miR-10a-3p were underexpressed in symptomatic mutation carriers. Interestingly, miR-34a-5p was identified as significantly overexpressed also in presymptomatic mutation carriers compared with healthy controls, suggesting that miR-34a-5p expression is associated with *C9orf72* mutation status. Additionally, miR-345-5p was also significantly overexpressed in patients when compared with presymptomatic carriers. When removing the four prodromal subjects from the presymptomatic group, the same miRNAs were identified as differentially expressed, indicating that the differences between the presymptomatic and other groups were not mainly driven by the four prodromal subjects.

Table 2 Differentially expressed miRNAs identified by EdgeR, after correction for multiple comparisons, for each pairwise comparison between clinical status: Control (n=43), presymptomatic (n=45) and patient (n=22)

miRNA	Log-fold change	P value	Adjusted p value
Control vs presymptomatic			
miR-34a-5p	−1.433	5.251e-16	3.093e-13
Control vs patient			
miR-34a-5p	−1.239	1.650e-8	9.720e-6
miR-345-5p	−0.540	1.131e-5	3.330e-3
miR-200c-3p	0.333	3.109e-5	6.104e-3
miR-10a-3p	0.697	7.141e-5	1.051e-2
Presymptomatic vs patient			
miR-345-5p	−0.528	3.610e-5	2.126e-2

miRNA, microRNA.

We considered these four miRNAs (miR-34a-5p, miR-345-5p, miR-200c-3p and miR-10a-3p) as our miRNA signature for further analyses. The complete output from EdgeR is available in online supplemental table A2.

Figure 1 displays boxplots with the expression levels, for each clinical group, of the four miRNAs identified as differentially expressed. There is a clear difference in miR-34a-5p expression levels between controls and *C9orf72* expansion carriers (presymptomatic and symptomatic). Moreover, the other three identified miRNAs differentiate the mutation carriers at different stages of the pathology: miR-345-5p showed increased expression in patients, while miR-200c-3p and miR-10a-3p exhibited decreased expression. An expression heatmap of the miRNA signature is displayed in online supplemental figure A2.

MiRNA signature to classify between clinical groups

To assess whether the identified miRNA signature could distinguish between clinical groups, we implemented logistic regression models using as features the expression levels of the four differentially expressed miRNAs (miR-34a-5p, miR-345-5p, miR-200c-3p and miR-10a-3p). The ROC AUC for the classification of healthy controls and presymptomatic mutation carriers was 0.90 (90% CI 0.83 to 0.95), for controls and patients was 0.90 (90% CI 0.82 to 0.97) and to distinguish presymptomatic carriers and patients was 0.80 (90% CI 0.67 to 0.90) (figure 2). The distributions of the bootstrapped ROC AUC scores are displayed in online supplemental figure A3.

Generalisation analysis

Since we used the entire data set to identify the miRNA signature, including test data, classification performance could be inflated. In order to assess the generality of our classification scores, we then incorporated feature selection in the nested cross-validation scheme (online supplemental figure A1), by using only the training data from the outer cross-validation loop to compute differentially expressed miRNAs. Figure 3 shows the distribution of miRNAs identified as differentially expressed after performing nested 5-fold cross-validation with 100 different

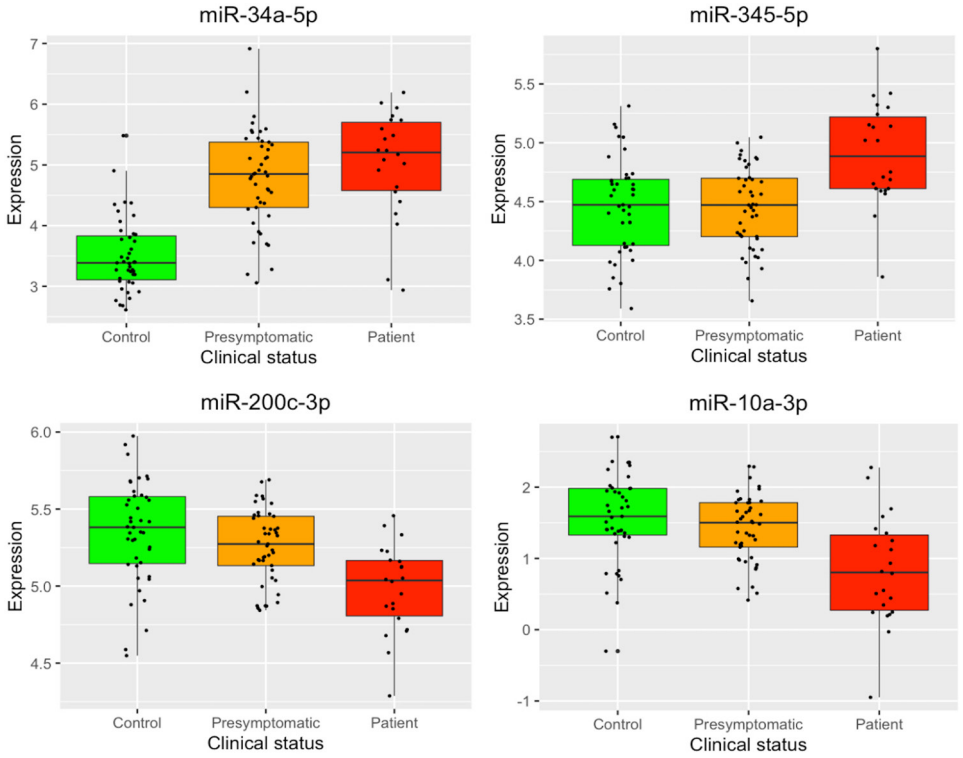


Figure 1 Boxplots depicting the normalised log2 expression levels of the four microRNAs identified as differentially expressed. Box boundaries represent the first and third quartiles and the median is indicated by the line dividing the IQR. The upper whiskers extend to the values that are within 1.5×IQR over the third quartiles. The lower whiskers extend to the values that are within 1.5×IQR under the first quartiles.

fold splits. Notably, the most frequent miRNAs (highlighted in blue) correspond to the ones computed using the entire data set: miR-34a-5p (500 occurrences) when comparing healthy controls and presymptomatic mutation carriers; miR-34a-5p, miR-345-5p, miR-200c-3p and miR-10a-3p (respectively 497, 335, 259 and 196 occurrences) for controls and patients; miR-345-5p (157 occurrences) when analysing presymptomatic subjects and patients.

Regarding prediction performance, the average ROC AUC when classifying controls versus presymptomatic subjects was 0.88 (90% CI 0.83 to 0.91), for controls versus patients was 0.89 (90% CI 0.83 to 0.94) and for presymptomatic individuals versus patients was 0.67 (90% CI 0.52 to 0.77). The distributions of the ROC AUC scores computed with 100 different fold splits are displayed in online supplemental figure A4.

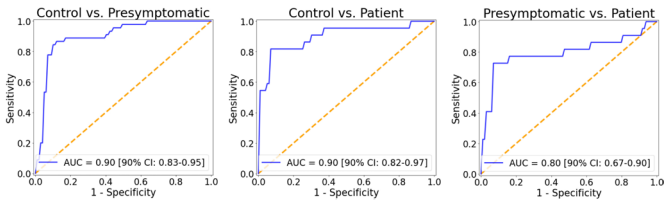


Figure 2 ROC (receiver operating characteristic) curves for each pairwise classification (control vs presymptomatic, control vs patient and presymptomatic vs patient) obtained with logistic regression using as features the expression levels of the microRNAs signature (miR-34a-5p, miR-345-5p, miR-200c-3p and miR-10a-3p). Bootstrapped 90% CIs are reported in brackets. AUC, area under the ROC curve.

Analysis of the transitional stage to clinical FTD/ALS disease

We evaluated the performance to predict the transitional stage to FTD/ALS disease by training a logistic regression classifier with the expression levels from patients and controls and testing with the expression levels of presymptomatic individuals. The probability scores computed for the four subjects in their transitional stage were all above 0.50, indicating a stronger similarity with patients: 0.54, 0.75, 0.80 and 0.82. The distribution of probability scores for all presymptomatic subjects is displayed in online supplemental figure A5.

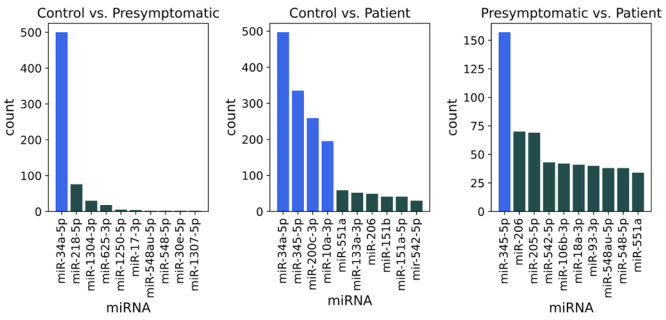


Figure 3 Number of times each miRNA was found differentially expressed, when performing a repeated 5-fold nested cross-validation for 100 times with different fold splits. In each step of the outer cross-validation loop, four of the five folds were used to identify differentially expressed miRNAs. Since one outer loop consists of five steps, and we performed 100 repetitions, 500 sets of miRNAs were computed for each pairwise comparison between groups, respectively: control vs presymptomatic, control vs patient and presymptomatic vs patient. MiRNAs from the signature computed with the entire data set are highlighted. miRNA, microRNA.

Table 3 Results from pathway analysis using the four differentially expressed microRNAs as input

Category	KEGG pathway	P value microT-CDS	P value TarBase
Cancer	Proteoglycans in cancer	**7.941e-4**	**4.259e-8**
	MicroRNAs in cancer	**1.386e-3**	**3.356e-8**
	Glioma	6.554e-2	**1.423e-2**
	Renal cell carcinoma	**1.098e-2**	9.254e-2
	Small cell lung cancer	3.220e-1	**3.341e-2**
Cell signalling/apoptosis	Hippo signalling pathway	**4.556e-2**	**5.622e-4**
	TGF-beta signalling pathway	5.008e-2	**9.288e-4**
	Thyroid hormone signalling pathway	**2.132e-3**	**1.502e-2**
	FoxO signalling pathway	2.368e-1	**1.449e-2**
	Neurotrophin signalling pathway	**9.801e-3**	3.113e-1
Intermediary metabolism	Lysine degradation	**1.606e-2**	**7.882e-4**
	Glycosphingolipid biosynthesis - lacto and neolacto series	**3.885e-10**	**4.423e-2**
Meiosis	Oocyte meiosis	2.487e-1	**2.446e-3**

Only significant pathways for at least one approach are shown. Statistically significant p values are in bold.
KEGG, Kyoto Encyclopedia of Genes and Genomes.

Target prediction and pathway analysis

Using the four differentially expressed miRNAs (miR-34a-5p, miR-345-5p, miR-200c-3p and miR-10a-3p) as input, we performed target prediction and pathway analysis with two methods available in DIANA-miRPath V.3. The in silico miRNA target prediction algorithm (microT-CDS) identified 31 influenced pathways (14 significant after Benjamini-Hochberg correction), while the experimentally supported approach (TarBase) resulted in 54 associated pathways (38 significant after Benjamini-Hochberg correction). Complete outputs concerning the list of the putative target genes and their related pathways are given in online supplemental tables A3 and A4. Table 3 reports the 13 pathways that were identified by both methods and have significant adjusted p values in at least one of them.

Online supplemental figure A6 shows miRNA versus KEGG pathways heatmaps, which depict the level of enrichment in significant KEGG pathways for the four differentially expressed miRNAs as computed by the two approaches.

DISCUSSION

The present study aimed to identify fluid biomarkers by analysing expression levels of plasma miRNAs without a priori knowledge in a large cohort of healthy controls, presymptomatic and symptomatic *C9orf72* carriers. We identified four miRNAs differentially expressed between clinical conditions: miR-34a-5p, miR-345-5p, miR-200c-3p and miR-10a-3p. Significantly higher expression of miR-34a-5p was found in mutation carriers when compared with healthy controls, which suggests that miR-34a-5p expression is deregulated in cases with *C9orf72* mutation. Additionally, we observed miR-345-5p expression to be significantly increased in patients when compared with presymptomatic carriers, which supports the correlation of miR-345-5p expression with the progression of *C9orf72*-associated disease. Finally, our results also suggest that miR-200c-3p and miR-10a-3p underexpression might be associated with full-blown disease as decreased expression levels were significant only between patients and healthy controls.

We used the expression levels of the miRNA signature to train logistic regression classifiers, which were able to differentiate individuals from different clinical groups with good predictive performance (figure 2). Notably, presymptomatic and symptomatic *C9orf72* carriers were distinguished with ROC AUC of 0.80 (90% CI 0.67 to 0.90), which suggests the suitability of plasma miRNAs for following preclinical progression and determining disease onset. We believe that this score was lower in our generalisation analysis (0.67, 90% CI 0.52 to 0.77) because the limited number of patients (22) led to a higher variability in the differentially expressed miRNAs in each step of the cross-validation loop (figure 3). Furthermore, we have obtained promising results regarding prediction performance of conversion from the presymptomatic to the clinical stage of FTD/ALS. The four presymptomatic subjects in transitional stage exhibited scores above 0.50, denoting a stronger similarity with the expression levels of patients. Although preliminary, these results suggest that the expression levels of our miRNA signature might be used as early predictors of the *C9orf72* disease conversion.

Previous studies have shown the potential of miRNAs in serum, plasma or CSF as diagnostic biomarkers for FTD and ALS,[9 17–20] focussing on comparing healthy controls and patients. However, our findings differ from preceding results: only two miRNAs from our signature (miR-345-5p and miR-200c-3p) were identified as differentially expressed in one of these studies,[20] none in the others.[17–19] Results are conflicting probably due to restricted choices for the analysed miRNAs[17 18] and heterogeneous cohorts, either with sporadic forms[18 19] or a mixture of sporadic and familial forms with different mutations.[20] To the best of our knowledge, the present work is the first to compare the expression levels of plasma miRNAs between presymptomatic and symptomatic carriers focussing on *C9orf72* mutation, in addition to providing a plasma miRNA signature that may contribute to the assessment of preclinical progression for *C9orf72*-associated FTD and ALS. Table 4 displays a comparison among studies evaluating miRNAs from blood samples (serum or plasma) of patients with FTD and/or ALS.

Overall, our work suggests that miR-34a-5p, miR-345-5p, miR-200c-3p and miR-10a-3p are likely involved in neuronal degeneration and *C9orf72*-associated pathogenesis. Among the KEGG pathways identified in this study, some involved in neurodevelopment (Hippo signalling and FoxO signalling), inflammation (TGF-beta signalling), intracellular transduction (neurotrophin signalling) and apoptosis (TGF-beta and FoxO signalling) were relevant as previously shown to be involved in *C9orf72*-disease.[32–34] Accordingly, these four miRNAs have been previously linked with a range of neurodevelopmental processes, neuropsychiatric and neurodegenerative conditions.[35–38] For instance, miR-200c and miR-34a family members are implicated

Table 4 Comparison of studies investigating miRNAs from blood samples (serum or plasma) of patients with FTD and/or ALS

	Freischmidt *et al* 2014*[11]	Sheinerman *et al* 2017†[17]	Piscopo *et al* 2018†[18]	Grasso *et al* 2019†[19]	Magen *et al* 2020†[20]	This study†
Disease	ALS	FTD, ALS	FTD	FTD	FTD, ALS	FTD, ALS
Cohort	Separate sporadic/genetic‡	Not mentioned	Sporadic	Sporadic	Mixed sporadic/genetic§	*C9orf72*
Patients, n= Discovery/replication	9/13 genetic 14 sporadic	50 FTD 50 ALS	54	10/48	52/117 FTD 115 ALS	22
Presymptomatic carriers, n=	18	–	–	–	–	45
Methods of analysis	Microarrays	37 selected miRNAs (qRT-PCR)	9 selected miRNAs (qRT-PCR)	752 selected miRNAs (qRT-PCR)	Large scale sequencing (RNA-seq)	Large scale sequencing (RNA-seq)
Major deregulated miRNAs	miR-4745-5p miR-3665 miR-1915-3p miR-4530 (validated from panel of 30 miRNAs)	miR-9/let-7e, miR-7/miR-451, miR335-5p/let-5e (FTD) miR-206/miR-338-3p, miR-9/miR-129-3p, miR-335-5p/miR-338-3p (ALS)	miR-127-3p	miR-663a miR-502-3p miR-206	Panels of 20, 147, 121 miRNAs for each cohort	miR-34a-5p miR-345-5p miR-200c-3p miR-10a-3p

*in serum.
†in plasma.
‡*SOD1*, *FUS*, *C9orf72*, *PFN1*.
§*C9orf72*, *MAPT*, *GRN*, *TBK1*.
ALS, amyotrophic lateral sclerosis; FTD, frontotemporal dementia; miRNA, microRNA; qRT-PCR, quantitative real-time PCR; RNA-seq, RNA sequencing.

in synaptic function, neuronal maturation, differentiation and survival.[39 40] Aberrant expression of miR-34a and miR-345 are also associated with neuronal apoptosis,[41] whereas members of miR-10a family were found to be differentially expressed in the muscle tissue of patients with ALS.[42]

How these four miRNAs are implicated in *C9orf72*-associated pathogenesis, and their relevance in brain pathology are important questions to go further. So far, only few studies addressing miRNA dysregulation in brain tissues of patients with FTD/ALS have been performed, and are summarised in online supplemental table A5. They specifically addressed *GRN*-associated,[43 44] sporadic FTD,[45 46] sporadic[47] or mixed genetic-sporadic ALS patients.[48] Notably, there was no miRNA dysregulation in common between the aforementioned studies, nor between any of those studies on the brain and ours on plasma. Those discrepancies may stem from the heterogeneity of the previous autoptic cohorts and the differences in the methods of miRNA expression analysis. Noteworthy, and differently from our investigation, none of the patient cohorts mentioned in online supplemental table A5 were exclusively made up of *C9orf72* carriers. Additionally, the observed differences between brain tissue and plasma miRNA profiles may be due to the tissue-specific expression of miRNA on the one hand, and to the time-dependent variations of detectable miRNAs all along the disease course on the other. Due to the disease process itself and other potential confounding factors, significant changes in miRNA expression are likely to occur between a relatively early phase of the disease, in which plasma miRNAs may be used as biomarkers, and the ultimate disease stage, at the moment of brain sampling. At this point, further miRNA profiling studies on *C9orf72* brain tissue are needed to better understand whether tissue miRNAs correlate with plasma expression profiles and their contribution to the disease pathogenesis.

Regardless, it is noteworthy that some studies pointed towards a direct relationship between these miRNAs and *C9orf72* pathogenesis. *C9orf72* stands as a putative target of miR-34a-5p, likely acting as a negative regulator of *C9orf72* mRNA expression.[49] Additionally, miR-200c-3p and miR-345-5p are down-regulated and up-regulated, respectively, in the extracellular vesicles secreted by induced astrocytes obtained from *C9orf72* patients.[50] Even if not completely explained so far, these important results parallel our study showing a comparable upregulation of miR-34a-5p and miR-345-5p and downregulation of miR-200c-3p in carriers, and provide converging evidence for a link between our set of miRNAs and *C9orf72*-pathogenesis, which will need further investigations.

Previous studies have provided the proof-of-concept that specific sets of miRNAs have the potential to serve as biomarkers of the preclinical/premanifest stages of other neurodegenerative diseases, such as ALS,[11] Huntington[39] and Prion diseases.[51] Our study supports the usefulness of our four miRNAs as biomarkers of disease progression from the presymptomatic to the symptomatic phase of *C9orf72* disease. Nevertheless, some of them may be dysregulated in a broader range of neurodegenerative conditions. For instance, miR-345 and miR-200c-3p were also dysregulated during the presymptomatic stage of Prion[51] and Huntington's diseases,[39] respectively. This would not prevent, however, their use in longitudinal monitoring of specific genetic neurodegenerative disorders, possibly in combination with other biomarkers. Together, all these studies and ours suggest that dysregulation of such miRNAs is dynamically altered throughout neurodegenerative diseases progression, and can be detectable even long before clinical onset.

The current study has limitations. First, the significant age difference between patients and the other clinical groups may have introduced a confounding factor, which we considered by including age as a covariate. Second, the absence of validation in other tissues or of a replication cohort means that further studies in independent cohorts are required to confirm our results, even though our generalisation analysis confirmed the identified miRNA signature. Finally, the limited number of patients does not allow any conclusions about the correlation of plasma miRNAs and different disease phenotypes. Future work will explore longitudinal analyses of plasma miRNAs to assess their use as biomarkers of FTD and ALS progression.

In summary, the current work revealed significant differences in miRNA expression levels in plasma when comparing healthy controls, presymptomatic and symptomatic *C9orf72* mutation carriers. Specifically, we highlighted the potential of miR-34a-5p, miR-345-5p, miR-200c-3p and miR-10a-3p expression levels in plasma as biomarkers of preclinical progression for *C9orf72*-associated FTD and ALS. Our results encourage the use of plasma miRNAs, possibly in combination with other markers, to improve the design of clinical trials for these neurodegenerative disorders.

Author affiliations
[1]Inria, Aramis project-team, F-75013, Paris, France
[2]Sorbonne Université, Paris Brain Institute – Institut du Cerveau – ICM, Inserm U1127, CNRS UMR 7225, AP-HP - Hôpital Pitié-Salpêtrière, Paris, France
[3]Centre de référence des démences rares ou précoces, IM2A, Département de Neurologie, AP-HP - Hôpital Pitié-Salpêtrière, Paris, France
[4]Département de Neurologie, AP-HP - Hôpital Pitié-Salpêtrière, Paris, France
[5]EPHE, PSL Research University, Paris, France
[6]CMRR Service de Neurologie, CHU de Limoges, Limoges, France
[7]Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Department of Neurology and CNR-MAJ, Normandy Center for Genomic and Personalized Medicine, Rouen, France
[8]Univ Lille, CHU, Inserm U1172, DISTALZ, LiCEND, Lille, France
[9]GenoSplice, Paris, France
[10]Paris Brain Institute – Institut du Cerveau – ICM, FrontLab, Paris, France
[11]Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France

Twitter Emmanuelle Becker @EmmanuelleBeck4

Acknowledgements We thank Justine Guégan, from iCONICS (ICM bioinformatics facility) for microRNA raw reads to counts pipeline handling. We also thank Yannick Marie and Delphine Bouteiller, from iGenSeq (ICM sequencing facility), for library preparation and sequencing. We thank the DNA and cell bank of the ICM for the technical assistance, notably Philippe Martin-Hardy (DNA and cell bank, ICM). The study was conducted with the support of the Centre d'Investigation Clinique Neuroscience (CIC 1422), GH Pitié Salpêtrière, Paris, and the Centre pour l'Acquisition et le Traitement des Images platform.

Collaborators The PREV-DEMALS study group includes: Eve Benchetrit (Hôpital de la Salpêtrière, Paris), Anne Bertrand (Hôpital de la Salpêtrière, Paris), Anne Bissery (Hôpital de la Salpêtrière, Paris), Marie-Paule Boncoeur (CHU Dypuytren, Limoges), Stéphanie Bombois (CHU Roger Salengro, Lille), Agnès Camuzat (ICM, Paris), Mathieu Chastan (CHU Charles Nicolle, Rouen), Yaohua Chen (CHU Roger Salengro, Lille), Marie Chupin (ICM, Paris), Olivier Colliot (ICM, Paris), Philippe Couratier (CHU Dypuytren, Limoges), Xavier Delbeuck (CHU Roger Salengro, Lille), Vincent Deramecourt (CHU Roger Salengro, Lille), Christine Delmaire (CHU Roger Salengro, Lille), Emmanuel Gerardin (CHU Charles Nicolle, Rouen), Claude Hossein-Foucher (CHU Roger Salengro, Lille), Bruno Dubois (Hôpital de la Salpêtrière, Paris), Marie-Odile Habert (Hôpital de la Salpêtrière, Paris), Didier Hannequin (CHU Charles Nicolle, Rouen), Géraldine Lautrete (CHU Dypuytren, Limoges), Thibaud Lebouvier (CHU Roger Salengro, Lille), Isabelle Le Ber (Hôpital de la Salpêtrière, Paris), Benjamin Le Toullec (ICM, Paris), Richard Levy (Hôpital de la Salpêtrière, Paris), Olivier Martinaud (CHU Charles Nicolle, Rouen), Kelly Martineau (ICM, Paris), Marie-Anne Mackowiak (CHU Roger Salengro, Lille), Jacques Monteil (CHU Dypuytren, Limoges), Florence Pasquier (CHU Roger Salengro, Lille), Gregory Petyt (CHU Roger Salengro, Lille), Pierre-François Pradat (Hôpital de la Salpêtrière, Paris), Assi-Hervé Oya (Hôpital de la Salpêtrière, Paris), Armelle Rametti-Lacroux (Hôpital de la Salpêtrière, Paris), Daisy Rinaldi (Hôpital de la Salpêtrière, Paris), Adeline Rollin-Sillaire (CHU Roger Salengro, Lille), François Salachas (Hôpital de la Salpêtrière, Paris), Sabrina Sayah (Hôpital de la Salpêtrière, Paris), David Wallon (CHU Charles Nicolle, Rouen).

Contributors VK had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Study concepts and study design: VK, VA, DS, ILB, OC and EB. Acquisition, analysis or interpretation of data: All authors. Manuscript drafting or manuscript revision for important intellectual content: All authors. Approval of final version of submitted manuscript: All authors. Literature research: VK, VA, DS and ILB. Statistical analysis: VK. Obtained funding: ILB and OC. Administrative, technical or material support: ILB, OC and EB. Study supervision: ILB, OC and EB.

Funding The research leading to these results has received funding from the programme 'Investissements d'avenir' ANR-10-IAIHU-06, from the French government under management of Agence Nationale de la Recherche as part of the 'Investissements d'avenir' programme, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), from Agence Nationale de la Recherche/DGOS (project ANR-PRTS PREV-DEMALS, grant number ANR-14-CE15-0016-07, promotion Assistance Publique-Hopitaux de Paris), from the Inria Project Lab Program (project Neuromarkers), from Fondation Vaincre Alzheimer FR-17035 and from the Institut Français de Bioinformatique (ANR-11-INSB-0013).

Competing interests OC reports having received consulting fees from AskBio (2020), having received fees for writing a lay audience short paper from Expression Santé (2019), having received speaker fees for a lay audience presentation from Palais de la découverte (2017) and that his laboratory has received grants from Qynapse (2017-present). Members from his laboratory have co-supervised a PhD thesis with myBrainTechnologies (2016-present). OC's spouse is an employee of myBrainTechnologies (2015-present). OC has submitted a patent to the International Bureau of the World Intellectual Property Organization (PCT/IB2016/0526993, Schiratti J-B, Allassonniere S, Colliot O, Durrleman S, A method for determining the temporal progression of a biological phenomenon and associated methods and devices) (2016). ILB served as a member of advisory boards for Prevail Therapeutic and received research grants from ANR, DGOS, PHRC, Vaincre Alzheimer Association, ARSla Association, Fondation Plan Alzheimer outside of the present work. PG is co-founder and director of GenoSplice. NR is employee at GenoSplice.

Patient consent for publication Not required.

Ethics approval This study was approved by the Comité de Protection des Personnes CPP Ile-De-France VI (CPP 68-15 and ID RCB 2015-A00856-43).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request (isabelle.leber@upmc.fr).

ORCID iD
Virgilio Kmetzsch http://orcid.org/0000-0003-3691-0180

REFERENCES

1 Rascovsky K, Hodges JR, Knopman D, *et al*. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain* 2011;134:2456–77.
2 Mackenzie IR, Rademakers R, Neumann M. Tdp-43 and FUS in amyotrophic lateral sclerosis and frontotemporal dementia. *Lancet Neurol* 2010;9:995–1007.
3 DeJesus-Hernandez M, Mackenzie IR, Boeve BF, *et al*. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9orf72 causes chromosome 9p-linked FTD and ALS. *Neuron* 2011;72:245–56.
4 Renton AE, Majounie E, Waite A, *et al*. A hexanucleotide repeat expansion in C9orf72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* 2011;72:257–68.
5 Mackenzie IRA, Frick P, Neumann M. The neuropathology associated with repeat expansions in the C9orf72 gene. *Acta Neuropathol* 2014;127:347–57.
6 Bertrand A, Wen J, Rinaldi D, *et al*. Early Cognitive, Structural, and Microstructural Changes in Presymptomatic *C9orf72* Carriers Younger Than 40 Years. *JAMA Neurol* 2018;75:236–45.
7 Eisen A, Kiernan M, Mitsumoto H, *et al*. Amyotrophic lateral sclerosis: a long preclinical period? *J Neurol Neurosurg Psychiatry* 2014;85:1232–8.
8 Grasso M, Piscopo P, Confaloni A, *et al*. Circulating miRNAs as biomarkers for neurodegenerative disorders. *Molecules* 2014;19:6891–910.
9 Schneider R, McKeever P, Kim T, *et al*. Downregulation of exosomal miR-204-5p and miR-632 as a biomarker for FTD: a GENFI study. *J Neurol Neurosurg Psychiatry* 2018;89:851–8.
10 Denk J, Oberhauser F, Kornhuber J, *et al*. Specific serum and CSF microRNA profiles distinguish sporadic behavioural variant of frontotemporal dementia compared with Alzheimer patients and cognitively healthy controls. *PLoS One* 2018;13:e0197329.
11 Freischmidt A, Müller K, Zondler L, *et al*. Serum microRNAs in patients with genetic amyotrophic lateral sclerosis and pre-manifest mutation carriers. *Brain* 2014;137:2938–50.
12 Huntzinger E, Izaurralde E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet* 2011;12:99–110.
13 Buratti E, Baralle FE. The multiple roles of TDP-43 in pre-mRNA processing and gene expression regulation. *RNA Biol* 2010;7:420–9.

14 Gascon E, Gao F-B. The emerging roles of microRNAs in the pathogenesis of frontotemporal dementia-amyotrophic lateral sclerosis (FTD-ALS) spectrum disorders. *J Neurogenet* 2014;28:30–40.
15 Li L, Wang J. Roles of extracellular microRNAs in central nervous system. *ExRNA* 2019;1:13.
16 Sohel MH. Extracellular/Circulating microRNAs: release mechanisms, functions and challenges. *Achiev Life Sci* 2016;10:175–86.
17 Sheinerman KS, Toledo JB, Tsivinsky VG, *et al*. Circulating brain-enriched microRNAs as novel biomarkers for detection and differentiation of neurodegenerative diseases. *Alzheimers Res Ther* 2017;9:89.
18 Piscopo P, Grasso M, Puopolo M, *et al*. Circulating miR-127-3p as a potential biomarker for differential diagnosis in frontotemporal dementia. *J Alzheimers Dis* 2018;65:455–64.
19 Grasso M, Piscopo P, Talarico G, *et al*. Plasma microRNA profiling distinguishes patients with frontotemporal dementia from healthy subjects. *Neurobiol Aging* 2019;84:240.e1–240.e12.
20 Magen I, Yacovzada N, Warren JD, *et al*. Classification and prediction of frontotemporal dementia based on plasma microRNAs. *medRxiv* 2020:2020.01.22.20018408.
21 Montembeault M, Sayah S, Rinaldi D, *et al*. Cognitive inhibition impairments in presymptomatic *C9orf72* carriers. *J Neurol Neurosurg Psychiatry* 2020;91:366–72.
22 Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res* 2017;27:491–9.
23 Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;17:10–12.
24 Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;10:R25.
25 Li H, Handsaker B, Wysoker A, *et al*. The sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–9.
26 Liao Y, Smyth GK, Shi W. featureCounts: an efficient General purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;30:923–30.
27 Friedländer MR, Mackowiak SD, Li N, *et al*. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* 2012;40:37–52.
28 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139–40.
29 Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-Seq data. *Genome Biol* 2010;11:R25.
30 Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12:2825–30.
31 Vlachos IS, Zagganas K, Paraskevopoulou MD, *et al*. DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res* 2015;43:W460–6.
32 Atkinson RAK, Fernandez-Martos CM, Atkin JD, *et al*. C9Orf72 expression and cellular localization over mouse development. *Acta Neuropathol Commun* 2015;3:59.
33 Farg MA, Konopka A, Soo KY, *et al*. The DNA damage response (DDR) is induced by the C9orf72 repeat expansion in amyotrophic lateral sclerosis. *Hum Mol Genet* 2017;26:2882–96.
34 Burberry A, Wells MF, Limone F, *et al*. C9Orf72 suppresses systemic and neural inflammation induced by gut bacteria. *Nature* 2020;582:89–94.
35 Fu J, Peng L, Tao T, *et al*. Regulatory roles of the miR-200 family in neurodegenerative diseases. *Biomed Pharmacother* 2019;119:109409.
36 Chua CEL, Tang BL. miR-34A in neurophysiology and neuropathology. *J Mol Neurosci* 2019;67:235–46.
37 Cosín-Tomás M, Antonell A, Lladó A, *et al*. Plasma miR-34a-5p and miR-545-3p as Early Biomarkers of Alzheimer's Disease: Potential and Limitations. *Mol Neurobiol* 2017;54:5550–62.
38 van den Berg MMJ, Krauskopf J, Ramaekers JG, *et al*. Circulating microRNAs as potential biomarkers for psychiatric and neurodegenerative disorders. *Prog Neurobiol* 2020;185:101732.
39 Jin J, Cheng Y, Zhang Y, *et al*. Interrogation of brain miRNA and mRNA expression profiles reveals a molecular regulatory network that is perturbed by mutant huntingtin. *J Neurochem* 2012;123:477–90.
40 Jauhari A, Singh T, Singh P, *et al*. Regulation of miR-34 family in neuronal development. *Mol Neurobiol* 2018;55:936–45.
41 Modi PK, Jaiswal S, Sharma P. Regulation of neuronal cell cycle and apoptosis by microRNA 34a. *Mol Cell Biol* 2016;36:84–94.
42 Kovanda A, Leonardis L, Zidar J, *et al*. Differential expression of microRNAs and other small RNAs in muscle tissue of patients with ALS and healthy age-matched controls. *Sci Rep* 2018;8:5609.
43 Kocerha J, Kouri N, Baker M, *et al*. Altered microRNA expression in frontotemporal lobar degeneration with TDP-43 pathology caused by progranulin mutations. *BMC Genomics* 2011;12:527.
44 Chen-Plotkin AS, Unger TL, Gallagher MD, *et al*. Tmem106B, the risk gene for frontotemporal dementia, is regulated by the microRNA-132/212 cluster and affects progranulin pathways. *J Neurosci* 2012;32:11213–27.
45 Hébert SS, Wang W-X, Zhu Q, *et al*. A study of small RNAs from cerebral neocortex of pathology-verified Alzheimer's disease, dementia with Lewy bodies, hippocampal sclerosis, frontotemporal lobar dementia, and non-demented human controls. *J Alzheimers Dis* 2013;35:335–48.
46 Gascon E, Lynch K, Ruan H, *et al*. Alterations in microRNA-124 and AMPA receptors contribute to social behavioral deficits in frontotemporal dementia. *Nat Med* 2014;20:1444–51.
47 Jawaid A, Woldemichael BT, Kremer EA, *et al*. Memory decline and its reversal in aging and neurodegeneration involve miR-183/96/182 biogenesis. *Mol Neurobiol* 2019;56:3451–62.
48 Helferich AM, Brockmann SJ, Reinders J, *et al*. Dysregulation of a novel miR-1825/TBCB/TUBA4A pathway in sporadic and familial ALS. *Cell Mol Life Sci* 2018;75:4301–19.
49 Lal A, Thomas MP, Altschuler G, *et al*. Capture of microRNA-bound mRNAs identifies the tumor suppressor miR-34a as a regulator of growth factor signaling. *PLoS Genet* 2011;7:e1002363.
50 Varcianna A, Myszczynska MA, Castelli LM, *et al*. Micro-Rnas secreted through astrocyte-derived extracellular vesicles cause neuronal network degeneration in C9orf72 ALS. *EBioMedicine* 2019;40:626–35.
51 Boese AS, Saba R, Campbell K, *et al*. Microrna abundance is altered in synaptoneurosomes during prion disease. *Mol Cell Neurosci* 2016;71:13–24.

ANNALS of Clinical and Translational Neurology
Open Access
AMERICAN NEUROLOGICAL ASSOCIATION

RESEARCH ARTICLE

MicroRNA signatures in genetic frontotemporal dementia and amyotrophic lateral sclerosis

Virgilio Kmetzsch[1,2], Morwena Latouche[3], Dario Saracino[1,4,5], Daisy Rinaldi[3,4,5], Agnès Camuzat[3,6], Thomas Gareau[3], the French Research Network on FTD/ALS, Isabelle Le Ber[3,4,5,7], Olivier Colliot[1] & Emmanuelle Becker[2]

[1]Sorbonne Université, Institut du Cerveau – Paris Brain Institute – ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France
[2]Univ Rennes, Inria, CNRS, IRISA, F-35000, Rennes, France
[3]Sorbonne Université, Institut du Cerveau – Paris Brain Institute – ICM, CNRS, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France
[4]Centre de référence des démences rares ou précoces, IM2A, Département de Neurologie, AP-HP - Hôpital Pitié-Salpêtrière, Paris, France
[5]Département de Neurologie, AP-HP - Hôpital Pitié-Salpêtrière, Paris, France
[6]EPHE, PSL Research University, Paris, France
[7]Paris Brain Institute – Institut du Cerveau – ICM, FrontLab, Paris, France

Correspondence
Emmanuelle Becker, Université de Rennes 1 – Campus Beaulieu, 263 Avenue Général Leclerc, 35042 Rennes, France. Tel: 0033 2 9984 2567; Fax: +33 2 99 84 71 71; E-mail: emmanuelle.becker@univ-rennes1.fr

Funding Information
The research leading to these results has received funding from the program "Investissements d'avenir" ANR-10-IAIHU-06, from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), from Agence Nationale de la Recherche/DGOS (project ANR-PRTS PREV-DEMALS, grant number ANR-14-CE15-0016-07, promotion Assistance Publique – Hôpitaux de Paris), from the Programme Hospitalier de Recherche Clinique (PHRC) Predict-PGRN (to ILB, promotion by Assistance Publique – Hôpitaux de Paris), from Fondation Plan Alzheimer (to ILB), from the Inria Project Lab Program (project Neuromarkers), from Fondation Vaincre Alzheimer FR-17035 and from the Institut Français de Bioinformatique (ANR-11-INSB-0013).

Received: 15 July 2022; Revised: 12 September 2022; Accepted: 20 September 2022

Annals of Clinical and Translational Neurology **2022; 9(11): 1778–1791**

doi: 10.1002/acn3.51674

Abstract

Objective: MicroRNAs are promising biomarkers of frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS), but discrepant results between studies have so far hampered their use in clinical trials. We aim to assess all previously identified circulating microRNA signatures as potential biomarkers of genetic FTD and/or ALS, using homogeneous, independent validation cohorts of *C9orf72* and *GRN* mutation carriers. **Methods**: 104 individuals carrying a *C9orf72* or a *GRN* mutation, along with 31 controls, were recruited through the French research network on FTD/ALS. All subjects underwent blood sampling, from which circulating microRNAs were extracted. We measured differences in the expression levels of 65 microRNAs, selected from 15 published studies about FTD or ALS, between 31 controls, 17 *C9orf72* presymptomatic subjects, and 29 *C9orf72* patients. We also assessed differences in the expression levels of 30 microRNAs, selected from five studies about FTD, between 31 controls, 30 *GRN* presymptomatic subjects, and 28 *GRN* patients. **Results**: More than half (35/65) of the selected microRNAs were differentially expressed in the *C9orf72* cohort, while only a small proportion (5/30) of microRNAs were differentially expressed in the *GRN* cohort. In multivariate analyses, only individuals in the *C9orf72* cohort could be adequately classified (ROC AUC up to 0.98 for controls versus presymptomatic subjects, 0.94 for controls versus patients, and 0.77 for presymptomatic subjects versus patients) with some of the signatures. **Interpretation**: Our results suggest that previously identified microRNAs using sporadic or mixed cohorts of FTD and ALS patients could potentially serve as biomarkers of *C9orf72*-associated disease, but not *GRN*-associated disease.

Introduction

Frontotemporal dementia (FTD) is a neurodegenerative disease characterized by brain atrophy in the frontal and temporal lobes, causing severe changes in personality and social behavior.[1] The most prevalent genetic causes of FTD are GGGGCC repeat expansions in the *C9orf72* gene and mutations in the *GRN* gene.[2,3] FTD shares disease pathways with amyotrophic lateral sclerosis (ALS), a debilitating motor neuron disease that causes progressive motor deficit and muscle wasting.[4] The *C9orf72* hexanucleotide repeat expansion has been identified as the most common genetic cause of both familial FTD and ALS, as well as of their sporadic counterparts.[2]

There are currently no disease-modifying treatments that can stop the course of FTD or ALS. New therapeutic trials depend on robust progression biomarkers to assess treatment outcomes. The study of FTD/ALS genetic forms is particularly important, since asymptomatic mutation carriers may provide insights about the early disease stages, before any irreversible neuronal damage.[5]

Among the potential noninvasive biomarkers of neurodegenerative diseases, circulating microRNAs (miRNAs) constitute a promising approach.[6] MiRNAs are short noncoding RNAs that negatively regulate gene expression.[7] There is increasing evidence of a link between miRNA expression levels and the diagnosis of FTD[8–12] and ALS.[11–22] However, there are strong inconsistencies between the identified miRNA signatures in different studies. The examined cohorts are highly heterogeneous, most of them being sporadic or mixed cohorts of sporadic and genetic forms. Importantly, it is unclear which miRNAs are specific to a particular genetic mutation or might serve as biomarkers for several genetic forms. It is also uncertain whether miRNAs found in sporadic forms are differentially expressed in genetic forms. Furthermore, several of the published articles lacked an independent validation cohort, which also might have caused disparity between results. Finally, differences in lifestyle factors (e.g. diet, exercise, cognitive training) across the studied cohorts also influence the levels of blood-based biomarkers, and thus may contribute to non-reproducible results.[23,24] This absence of convergence among different studies so far hinders the use of miRNAs in clinical trials.

The present work aims at testing circulating miRNA signatures identified in the literature, using two independent homogeneous cohorts of patients and presymptomatic carriers: one focused on *C9orf72* expansion carriers and another comprising *GRN* mutation carriers. For that purpose, we selected all published studies that identified specific miRNAs isolated from plasma or serum as potential biomarkers of FTD and/or ALS. With a preregistered study design, we investigated whether (1) miRNAs revealed in cohorts of sporadic patients (or in mixed cohorts with sporadic and genetic forms) may be biomarkers in *C9orf72* and/or *GRN* genetic forms, (2) miRNAs identified in a *C9orf72* cohort are validated in an independent *C9orf72* cohort, and (3) miRNAs discovered in a *C9orf72* cohort may be relevant in a *GRN* cohort.

We hypothesize that if a miRNA is a potential biomarker in a particular genetic form, it will be differentially expressed (adjusted p value below 0.05) between controls and presymptomatic subjects, controls and patients, or presymptomatic subjects and patients in an independent cohort of subjects carrying that mutation. Moreover, we consider that a miRNA signature will constitute an acceptable biomarker if a logistic regression model (using these miRNAs as features) classifies subjects between clinical groups with an area under the ROC curve greater than 0.70.[25]

Materials and Methods

This research was conducted according to the preregistration available in https://osf.io/4pw8f.

Participants of the validation cohorts

Between 2011 and 2021, 135 individuals were recruited through the French research network on FTD/ALS (Inserm RBM02-59) and investigated with the same protocol, as previously described in detail.[26] All participants signed written informed consents. This study was approved by the Comité de Protection des Personnes CPP Ile-De-France VI (CPP 36–09/ID RCB 2008-A01376-49 and CPP 68–15/ID RCB 2015-A00856-43).

Two cohorts were studied. One cohort was focused on *C9orf72* mutation carriers, including 29 patients (20 FTD, 6 FTD/ALS and 3 ALS) and 17 carriers in the presymptomatic phase. Another cohort was focused on *GRN* mutation carriers, comprising 28 FTD patients and 30 presymptomatic carriers. The control group, shared between the two cohorts, was made up of 31 neurologically healthy individuals that did not carry any of these mutations. Table 1 displays the demographic characteristics of the studied cohorts.

Standardized interviews with family members, full neurological examinations, quantitative motor testing, and extensive neuropsychological tests measuring all cognitive domains were used to assess each participant's cognitive and clinical conditions. All subjects underwent blood tests, and collected samples were stored in the Paris Brain Institute (ICM) DNA and cell bank.

Table 1. Demographic characteristics of the studied cohorts.

	C9orf72 patients	*C9orf72* presympto-matic carriers	*GRN* patients	*GRN* presymptomatic carriers	Controls
No.	29	17	28	30	31
Female, No. (%)	14 (48.3)	10 (58.8)	10 (35.7)	17 (56.7)	18 (58.1)
Age at inclusion, mean (SD)	66.2 (8.8)	51.7 (12.1)	62.9 (11.2)	42.5 (11)	47.1 (14.6)
Disease duration at inclusion, mean (SD)	4.9 (3.8)	–	3.2 (1.4)	–	–
GRN mutation (No.)	–	–	c.1157G > A (1), c.1231_1232dup (2), c.1252C > T (1), c.138 + 1G > A (1), c.1492G > T (1), c.1494_1498del (1), c.19 T > C (1), c.234_235del (1), c.25del (1), c.328C > T (2), c.380_381del (2), c.421_422del (2), c.559del (1), c.559delC (1), c.708 + 6_708 + 9del (1), c.745C > T (1), c.759_760del (1), c.768_769dup (1), c.87_90dup (1), c.900_901dup (1), c.907del (1), c.942C > A (1), Exon 1 del (2)	c.1201C > T (3), c.1231_1232dup (3), c.138 + 1G > A (1), c.1494_1498del (2), c.19 T > C (1), c.328C > T (1), c.361delG (1), c.380_381del (1), c.443_444del (1), c.675_676del (1), c.745C > T (3), c.768_769dup (2), c.813_816del (6), c.907del (2), c.988_989del (1), GRN del (1)	–

Plasma preparation, miRNA sequencing and computation pipeline

Blood samples from all participants were collected on EDTA following standardized collection and handling procedures. The mean disease duration at sampling was 4.9 (SD 3.8) years in the *C9orf72* patients' group and 3.2 (SD 1.4) years in the *GRN* patients' group. All participants were in fasted state. At the ICM DNA and cell bank, all samples were centralized and processed in conformity with the same procedure. Blood samples were centrifuged at 2500 rpm for 10 min before plasma was extracted at room temperature. At a temperature of −80°C, 1 ml aliquots were stored in polypropylene tubes.

Following the instructions provided by the manufacturer, miRNA extraction was carried out using a miRNeasy Serum/Plasma Kit (Qiagen). We used 200 µl of plasma that was progressively melted at 4°C and added directly to 1 ml of QIAzol solution. Using the QIAseq miRNA Library Kit (Qiagen) in accordance with the manufacturer's protocol, miRNAs were eluted in 14 µl of water; 5 µl were utilized to prepare the miRNA sequencing library. Targeting a minimum of 10 million mapped reads per sample, two independent batches of miRNA sequencing were performed on the Illumina NovaSeq 6000.

The quantification of miRNAs was carried out according to recommendations by Potla et al.[27] The quality of reads was assessed with FastQC (Andrews S. 2010, http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Next, UMI-tools[28] and Cutadapt[29] were used to clean sequences and extract unique molecular identifiers (UMIs). Then, the resulting reads were aligned to the mature miRNA sequences from the miRBase (https://www.mirbase.org) database version 22.1, using Bowtie.[30] After that, the PCR duplicates were removed with UMI-tools. Finally, miRNA count tables were created with Samtools idxstats.[31]

Selected studies

We aimed to find all papers that identified specific miRNAs extracted from human plasma or serum as potential biomarkers of FTD and/or ALS, excluding reviews and meta-analyses. We thus conducted the following search in PubMed (https://pubmed.ncbi.nlm.nih.gov/), on March 10, 2022:

(microRNA[Title] OR microRNAs[Title] OR miR[Title] OR miRNA[Title]) AND (serum[Title] OR circulating[Title] OR plasma[Title]) AND (ALS[Title] OR FTD[Title] OR amyotrophic[Title] OR frontotemporal[Title] OR (neurodegenerative[Title] AND (frontotemporal[Title/Abstract] OR amyotrophic[Title/Abstract]))) NOT mice[Title/Abstract] NOT mouse[Title/Abstract] NOT extracellular vesicles[Title] NOT review[PT] NOT meta-analysis[PT] NOT (comment[PT])

Table 2. Selected studies investigating circulating microRNA expression (from serum or plasma) of patients with FTD or ALS. Columns indicate each reference, studied disease (FTD, ALS, or both), type of the analyzed cohort (sporadic, genetic, or mixed), number of patients in the discovery and replication (if available) cohorts, number of presymptomatic carriers included in the study, method of analysis (qRT-PCR, microarrays, RNA-sequencing, or a combination), and the identified miRNA signature.

Article	Disease	Cohort	Patients, No. (discovery/replication)	Presymto-matic carriers, No.	Method of analysis	MiRNA signature
Grasso et al., 2019[8]	FTD	Sporadic	10/48 split of same cohort	–	qRT-PCR of 752 miRNAs	miR-663a, miR-502-3p, miR-206
Piscopo et al., 2018[9]	FTD	Sporadic	54	–	qRT-PCR of 9 miRNAs linked with apoptosis	miR-127-3p
Denk et al., 2018[10]	FTD	Sporadic	48	–	qRT-PCR of 96 miRNAs identified in preliminary study	let-7b-5p, let-7 g-5p, miR-106a-5p, miR-106b-5p, miR-18b-5p, miR-223-3p, miR-26a-5p, miR-26b-5p, miR-301a-3p, miR-30b-5p, miR-146a-5p, miR-15a-5p, miR-22-3p, miR-320a, miR-320b, miR-92a-3p, miR-1246
Kmetzsch et al., 2021[11]	FTD, ALS	Genetic (*C9orf72*)	22	45	RNA-sequencing of 2576 miRNAs	miR-34a-5p, miR-345-5p, miR-200c-3p, miR-10a-3p
Sheinerman et al., 2017[12]	FTD, ALS	Unspecified	For each disease, 25/25 split of same cohort	–	qRT-PCR of 37 brain-enriched miRNAs	miR-9/let-7e, miR-7/miR-451, miR-335-5p/let-7e (FTD) and miR-206/miR-338-3p, miR-9/miR-129-3p, miR-335-5p/miR-338-3p (ALS)
Magen et al., 2021[13]	ALS	Mixed sporadic and genetic (*C9orf72*)	126/122 split of same cohort	–	RNA-sequencing of 125 miRNAs identified in longitudinal study	miR-181a-5p, miR-181b-5p
Soliman et al., 2021[14]	ALS	Mixed sporadic and genetic (unspecified mutation)	30	–	qRT-PCR of 7 miRNAs involved in ALS	miR-206, miR-143-3p, miR-142-3p
Dobrowolny et al., 2021[15]	ALS	Mixed sporadic and genetic (unspecified mutation)	13/23	–	RNA-sequencing followed by qRT-PCR	miR-151a-5p, miR-199a-5p, miR-423–3p
Raheja et al., 2018[16]	ALS	Mixed sporadic and genetic (*C9orf72*, *SOD1*)	23	–	qRT-PCR of 191 miRNAs identified on prior study	miR-29b-3p, miR-320c, miR-34a-5p, miR-29c-3p, miR-320a, miR-22-3p, miR-1, miR-133a-3p, miR-191-5p, miR-144-5p, miR-320b, miR-423-3p, miR-192-5p, miR-133b, miR-194-5p, miR-7-1-3p, miR-19a-3p, miR-425-5p, miR-145-5p, miR-144-3p
Waller et al., 2017[17]	ALS	Sporadic	27/23	–	qRT-PCR of 750 miRNAs	miR-206, miR-143-3p, miR-374b-5p
Tasca et al., 2016[18]	ALS	Sporadic	14	-	qRT-PCR of 9 muscle-specific,	miR-206, miR-133a, miR-133b, miR-27a

(*Continued*)

Table 2 Continued.

Article	Disease	Cohort	Patients, No. (discovery/replication)	Presymto-matic carriers, No.	Method of analysis	MiRNA signature
					inflammatory, or angiogenic miRNAs	
Takahashi et al., 2015[19]	ALS	Sporadic	16/48 split of same cohort	–	Microarrays, followed by qRT-PCR of 9 miRNAs	miR-4649-5p, miR-4299
Freischmidt et al., 2015[20]	ALS	Sporadic	18/20	–	Microarrays of 1733 miRNAs, followed by qRT-PCR of 2 miRNAs	miR-1234-3p, miR-1825
Freischmidt et al., 2014[21]	ALS	Separate sporadic and genetic (*SOD1*, FUS, *C9orf72*)	9/13 (genetic), 14 (sporadic)	18	Microarrays of 1733 miRNAs and qRT-PCR of 4 miRNAs	miR-4745-5p, miR-3665, miR-1915-3p, miR-4530
De Felice et al., 2014[22]	ALS	Sporadic	10	-	qRT-PCR of 1 miRNA	miR-338-3p

This search resulted in 19 journal articles. Two papers[6,32] were excluded because they were review studies, one[33] was discarded because it was focused on protein levels, and one[34] was excluded because it was focused on one microRNA from serum exosomes.

Our final selection therefore contained 15 articles. These selected papers, along with the studied diseases (FTD, ALS, or both), cohort types (sporadic, genetic, or mixed), cohort sizes, methods of analyses (qRT-PCR, microarrays, RNA-sequencing, or a combination), and the identified miRNA signatures are displayed in Table 2. We note that three of these studies[8–10] identified three different miRNA signatures associated with FTD, 10 articles[13–22] pointed out 10 distinct miRNA signatures related to ALS, one investigation[11] revealed another miRNA signature for both FTD and ALS, and one study[12] found two separate signatures for FTD and ALS. In all, considering both FTD and ALS, we thus analyzed 16 miRNA signatures previously identified in the literature. Since *C9orf72* expansions can cause both diseases, all miRNA signatures were tested with our *C9orf72* cohort. However, since *GRN* mutations cause exclusively FTD, only five miRNA signatures (the ones associated with FTD) were tested with our *GRN* cohort.

Regarding the sizes of the identified miRNA signatures, most of the studies[8,9,11,13–15,17–22] pointed out four or less miRNAs, one article identified six miRNAs per disease,[12] and two investigations[10,16] proposed larger signatures containing respectively 17 or 20 miRNAs. When analyzing the intersections between the different signatures, we note that some miRNAs were identified by multiple studies, for instance miR-206,[8,12,14,17,18] but most miRNAs were found by a single study. When considering the union of the miRNA signatures from the selected articles, the set of miRNAs associated with either FTD or ALS is composed of **65** miRNAs, and the set of miRNAs associated only with FTD is composed of **30** miRNAs.

As previously mentioned in the description of our computation pipeline, we used the miRNA nomenclature from the miRBase database version 22.1. However, most of the selected articles were based on previous versions of the miRBase. We thus performed the following conversions for compatibility:

- miR-320a: miR-320a-5p plus miR-320a-3p
- miR-9: miR-9-5p plus miR-9-3p
- let-7e: let-7e-5p plus let-7e-3p
- miR-1: miR-1-5p plus miR-1-3p
- miR-133-a: miR-133-a-5p plus miR-133a-3p
- miR-27a: miR-27a-5p plus miR-27a-3p
- miR-7: miR-7-5p
- miR-451: miR-451a
- miR-129-3p: miR-129-1-3p

Differential expression

Differential expression analyses were performed using the R package EdgeR.[35] After microRNA extraction, sequencing and quality control steps, our dataset contained the expression levels of 2656 miRNAs (denoted by *i*, corresponding to all miRNA sequences mapped in miRBase version 22.1) for each of the 135 subjects (represented by *j*). First, we created two count matrices: one containing the miRNA counts from the *C9orf72* patients, presymptomatic subjects and controls, and another containing the miRNA counts from the *GRN* patients, presymptomatic individuals, and controls. Second, for each count matrix, we used generalized linear models (GLM) to fit a log-

linear model to each miRNA, following a negative binomial distribution with mean μ_{ij} and dispersion ϕ_i:

$$\log_2 \mu_{ij} = x_j^T \beta_i$$

where x_j denotes the covariates describing sample j and β_i denotes the coefficients to be fitted for miRNA i. To control for possible age, sex, center and batch effects, we included them as covariates in the models, in addition to the clinical group (control, presymptomatic or patient). A trimmed mean of M-values[36] was used to normalize raw counts. Finally, after the log-linear models were fitted, quasi-likelihood (QL) F-tests were used to identify the differentially expressed miRNAs.

Concretely, we tested each of the 65 miRNAs associated with either FTD or ALS in the literature, to identify which were differentially expressed between (a) controls versus *C9orf72* presymptomatic subjects, (b) controls versus *C9orf72* patients, and (c) *C9orf72* presymptomatic subjects versus *C9orf72* patients. Additionally, we tested the 30 miRNAs associated with only FTD in the literature, to highlight which were differentially expressed between (d) controls versus *GRN* presymptomatic subjects, (e) controls versus *GRN* patients, and (f) *GRN* presymptomatic subjects versus *GRN* patients.

All p values were 2-tailed, and the level of statistical significance was set at 0.05. The Benjamini-Hochberg[37] procedure was used to adjust p values for multiple testing. Additionally, we considered as suggestive, but not statistically significant, miRNAs with adjusted p values between 0.05 and 0.1. Our differential expression analyses resulted in two meta-signatures for each cohort: one meta-signature including miRNAs with adjusted p values below 0.05 (thus including only statistically significant miRNAs), and one containing miRNAs with adjusted p values below 0.1 (thus including both statistically significant and suggestive miRNAs).

Binary classification

To test if the miRNA signatures described in the literature could discriminate between clinical groups, we trained L2-regularized logistic regression classifiers, using Python 3.8.5 with scikit-learn[38] 0.23.2. We first organized the miRNA expression data into six datasets, one for each relevant pairwise comparison: (a) controls versus *C9orf72* presymptomatic subjects, (b) controls versus *C9orf72* patients, (c) *C9orf72* presymptomatic subjects versus *C9orf72* patients, (d) controls versus *GRN* presymptomatic subjects, (e) controls versus *GRN* patients, and (f) *GRN* presymptomatic subjects versus *GRN* patients. A total of 18 classifiers were trained for each of the comparisons (a), (b) and (c): 16 classifiers used as features each of the miRNA signatures identified in the literature, and two were trained with meta-signatures containing the differentially expressed miRNAs identified in the univariate analyses (a), (b) and (c), respectively with adjusted p values lower than 0.05 and 0.1. In addition, seven classifiers were built for each of the comparisons (d), (e) and (f): five of them used as features each of the miRNA signatures associated with FTD in the literature, and two were trained with meta-signatures containing the differentially expressed miRNAs identified in the univariate analyses (d), (e) and (f), respectively with adjusted p values lower than 0.05 and 0.1.

Each logistic regression model was trained with a stratified nested 5-fold cross-validation strategy followed by bootstrapping to compute confidence intervals (CIs), as previously detailed.[11] The inner cross-validation loop was used for hyperparameter (L2 regularization coefficient) search. The outer cross-validation loop was used to assess each classifier's performance by computing the area under the ROC curve. Since a k-fold cross-validation does not provide an unbiased estimator of the variance,[39] it cannot be used to compute CIs. Therefore, we used 2000 bootstrap samples to compute empirical 90% CIs for the ROC AUC scores, by considering the 5th and 95th percentiles of the bootstrap distribution. A miRNA signature was considered an acceptable biomarker for a given comparison if the corresponding ROC AUC was above 0.70.

Results

Differential expression in the *C9orf72* cohort

The first analysis consisted of testing which of the 65 miRNAs identified in the literature as potential biomarkers of FTD and/or ALS were differentially expressed in our *C9orf72* cohort. For this analysis, we considered the miRNA counts obtained from sequencing plasma samples from the *C9orf72* patients, *C9orf72* presymptomatic subjects, and controls. After negative binomial generalized linear models were fitted for each miRNA, we thus performed 65 quasi-likelihood F-tests per pairwise comparison (controls versus *C9orf72* presymptomatic subjects, controls versus *C9orf72* patients, and *C9orf72* presymptomatic subjects versus *C9orf72* patients), adjusting for multiple comparisons using the Benjamini-Hochberg[37] method.

All differentially expressed miRNAs identified in this analysis are displayed in Table 3 (second to fourth columns). We can see that a considerable amount of miRNAs (**35** of the 65 miRNAs identified in the literature) were significantly differentially expressed (adjusted p

Table 3. Differentially expressed miRNAs for at least one pairwise comparison between clinical groups, considering both cohorts. The * indicates in which comparisons each miRNA was significantly differentially expressed (adjusted *p* values below 0.05), while the (+) denotes adjusted *p* values between 0.05 and 0.1.

miRNA	Controls vs. *C9orf72* presymptomatic subjects	Controls vs. *C9orf72* patients	*C9orf72* presymptomatic subjects vs. *C9orf72* patients	Controls vs. *GRN* presymptomatic subjects	Controls vs. *GRN* patients	*GRN* presymptomatic subjects vs. *GRN* patients
miR-34a-5p	*	*				
miR-338-3p	*		*			
miR-142-3p	*		*			
miR-320a	*		*			
miR-145-5p	*		*			
miR-92a-3p	*				(+)	
let-7g-5p	*					
miR-199a-5p	*		*			
miR-206	*	*	*			
miR-30b-5p	*		*			
miR-191-5p	*					
miR-27a	*		*			
miR-320b	*		*			
miR-143-3p	*		*			
miR-1246	*					
miR-223-3p	*		*			
miR-144-3p	*					
miR-451		*	(+)		*	(+)
miR-194-5p		*	*			
miR-144-5p		*	(+)			
miR-29b-3p		*	*			
miR-29c-3p		*	*			
miR-192-5p		*				
miR-19a-3p		*	*			
miR-502-3p		*			*	
miR-15a-5p		*			*	
miR-374b-5p	(+)		*			
miR-7-1-3p			*			
miR-320c	(+)		*			
miR-106b-5p			*		(+)	
miR-146a-5p	(+)		*			
miR-133b		(+)	*			
let-7b-5p	(+)		*		(+)	
miR-345-5p			*			
miR-22-3p			*			
miR-7					*	(+)
miR-18b-5p		(+)			*	
miR-151a-5p	(+)					
miR-1234-3p	(+)					
miR-26a-5p	(+)					
miR-301a-3p	(+)		(+)			
let-7e		(+)				
miR-106a-5p		(+)			(+)	
miR-1915-3p		(+)				
miR-9			(+)			

values smaller than 0.05) in at least one comparison: miR-34a-5p, miR-338-3p, miR-142-3p, miR-320a, miR-145-5p, miR-92a-3p, let-7 g-5p, miR-199a-5p, miR-206, miR-30b-5p, miR-191-5p, miR-27a, miR-320b, miR-143-3p, miR-1246, miR-223-3p, miR-144-3p, miR-451, miR-194-5p, miR-144-5p, miR-29b-3p, miR-29c-3p, miR-192-

5p, miR-19a-3p, miR-502-3p, miR-15a-5p, miR-374b-5p, miR-7-1-3p, miR-320c, miR-106b-5p, miR-146a-5p, miR-133b, let-7b-5p, miR-345-5p, and miR-22-3p. Moreover, the following **9** miRNAs had a p value between 0.05 and 0.1, close to significance value: miR-151a-5p, miR-1234-3p, miR-26a-5p, miR-301a-3p, let-7e, miR-18b-5p, miR-106a-5p, miR-1915-3p, and miR-9.

The complete output from the differential expression analyses in the *C9orf72* cohort, including log-fold changes indicating the intensity of underexpression or overexpression, as well as computed p values, are displayed in Table S1. Expression heatmaps of the differentially expressed miRNAs are shown in Figure S1.

Differential expression in the *GRN* cohort

The second analysis focused on identifying which of the 30 miRNAs linked with FTD in the literature were differentially expressed in our *GRN* cohort. For this experiment, we used the miRNA counts acquired by sequencing plasma samples from the *GRN* patients, *GRN* presymptomatic individuals, and controls. Once negative binomial generalized linear models were fitted for each miRNA, we conducted 30 quasi-likelihood F-tests per pairwise comparison (controls versus *GRN* presymptomatic subjects, controls versus *GRN* patients, and *GRN* presymptomatic subjects versus *GRN* patients), adjusting for multiple comparisons using the Benjamini-Hochberg[37] method.

Table 3 (fifth to seventh columns) shows all differentially expressed miRNAs identified in this experiment. Contrary to what was observed with the *C9orf72* cohort, we note that only a small proportion of miRNAs (**5** of the 30 miRNAs identified in the literature) were significantly differentially expressed (adjusted p values lower than 0.05), all of them when comparing controls and *GRN* patients: miR-451, miR-15a-5p, miR-502-3p, miR-7, and miR-18b-5p. Additionally, **4** miRNAs had a p value close to significance value, between 0.05 and 0.1: miR-106a-5p, miR-92a-3p, miR-106b-5p, and let-7b-5p.

Table S2 summarizes the complete results of the differential expression experiments in the *GRN* cohort, including log-fold changes reflecting the degree of underexpression or overexpression of each miRNA in each pairwise comparison, and the calculated p values. Figure S1 displays expression heatmaps of the differentially expressed miRNAs.

Finally, Table 3 also allows comparing the results obtained with the *C9orf72* and the *GRN* cohorts. Remarkably, the three comparisons involving the *C9orf72* cohort revealed significantly differentially expressed miRNAs, but that was the case for only one comparison involving the *GRN* cohort (controls versus *GRN* patients). We also note that only a small minority of miRNAs (3) was significantly differentially expressed in both cohorts.

Taken together, these results offer evidence for the potential contribution of miRNAs identified in previous studies as biomarkers of *C9orf72*-associated disease, but not *GRN*-associated disease.

Binary classification in the *C9orf72* cohort

The first set of binary classification experiments focused on the *C9orf72* cohort. We trained 18 logistic regression classifiers for each pairwise comparison between clinical groups (controls versus *C9orf72* presymptomatic subjects, controls versus *C9orf72* patients, and *C9orf72* presymptomatic subjects versus *C9orf72* patients). For each of the 18 classifiers, we used as features the expression levels of distinct sets of miRNAs: the 16 miRNA signatures identified in the previously published studies about FTD and ALS, and two meta-signatures obtained from our differential expression analyses with the *C9orf72* cohort (one containing miRNAs with adjusted p value <0.05, and another including miRNAs with adjusted p value <0.1).

Figure 1 displays the areas under the ROC curves obtained by each of the 18 logistic regression classifiers in the three pairwise comparisons, as well as the empirical 90% confidence intervals. We observe that more than half of the classifiers (10 of the 18) achieved a ROC AUC greater than 0.70 in at least one comparison: those using the miRNA signatures from Kmetzsch et al., 2021,[11] Sheinerman et al., 2017[12] (FTD), Sheinerman et al., 2017[12] (ALS), Soliman et al., 2021,[14] Dobrowolny et al., 2021,[15] Raheja et al., 2018,[16] Waller et al., 2017,[17] Tasca et al., 2016,[18] and the two meta-signatures from our differential expression analyses. The miRNA signatures with the largest ROC AUC were from Kmetzsch et al., 2021[11] (0.98 for controls versus presymptomatic subjects), Raheja et al., 2018[16] (0.94 for controls versus patients), and the meta-signature with $p < 0.1$ (0.77 for presymptomatic subjects versus patients).

As well as in the differential expression analyses, these findings provide support for the potential use of several miRNAs identified in previous studies as biomarkers of *C9orf72*-associated FTD and ALS.

Binary classification in the *GRN* cohort

The second set of binary classification experiments consisted of training seven logistic regression classifiers for each pairwise comparison in the *GRN* cohort (controls versus *GRN* presymptomatic subjects, controls versus *GRN* patients, and *GRN* presymptomatic subjects versus *GRN* patients). Each of the seven classifiers was trained with a different set of features: the expression levels of the

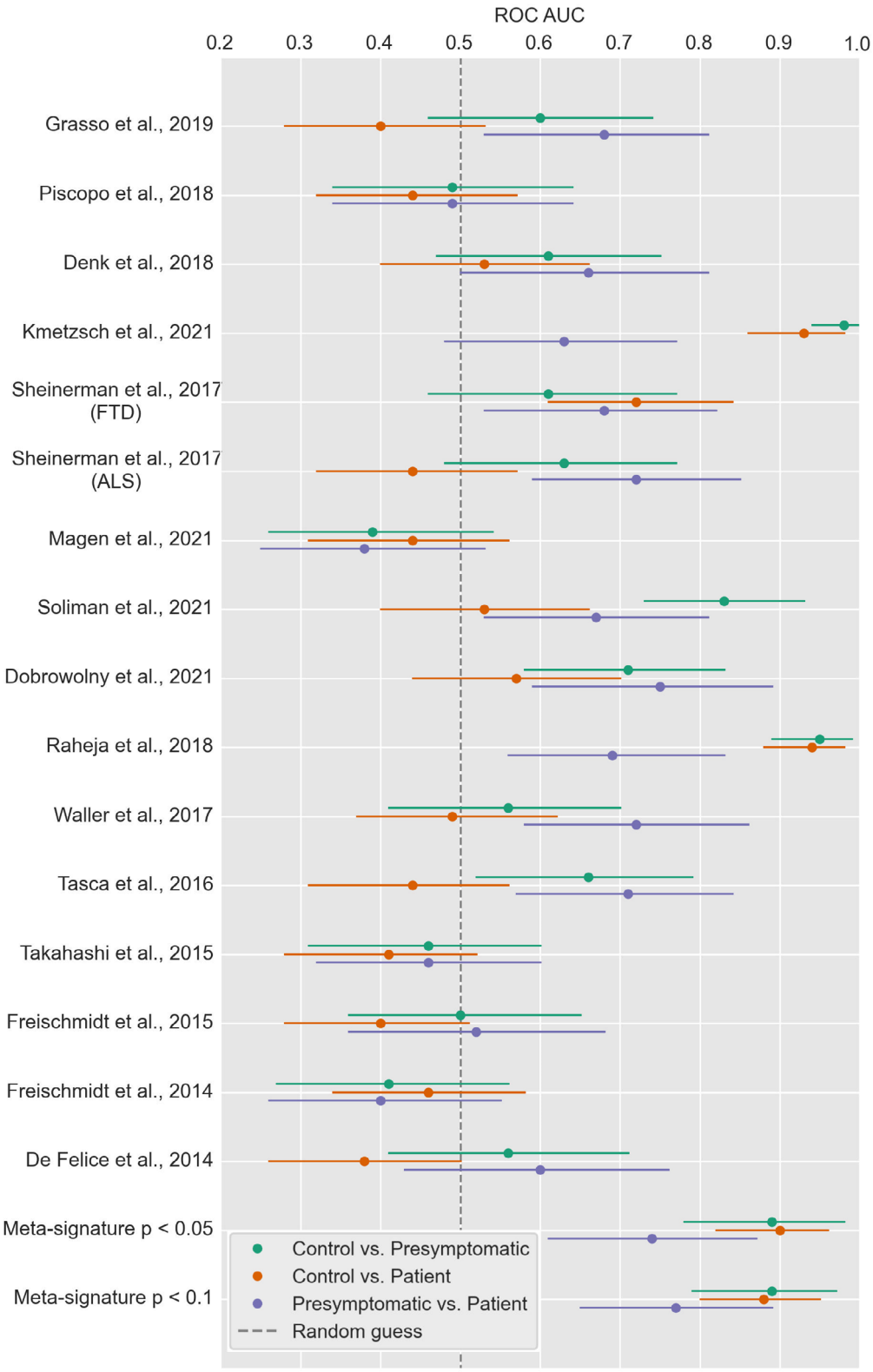


Figure 1. Area under the ROC curve results when classifying groups from the *C9orf72* cohort. The solid circles indicate the areas under the ROC curves obtained for each pairwise comparison using 18 different miRNA signatures. The whiskers denote empirical 90% confidence intervals obtained with 2000 bootstrap samples.

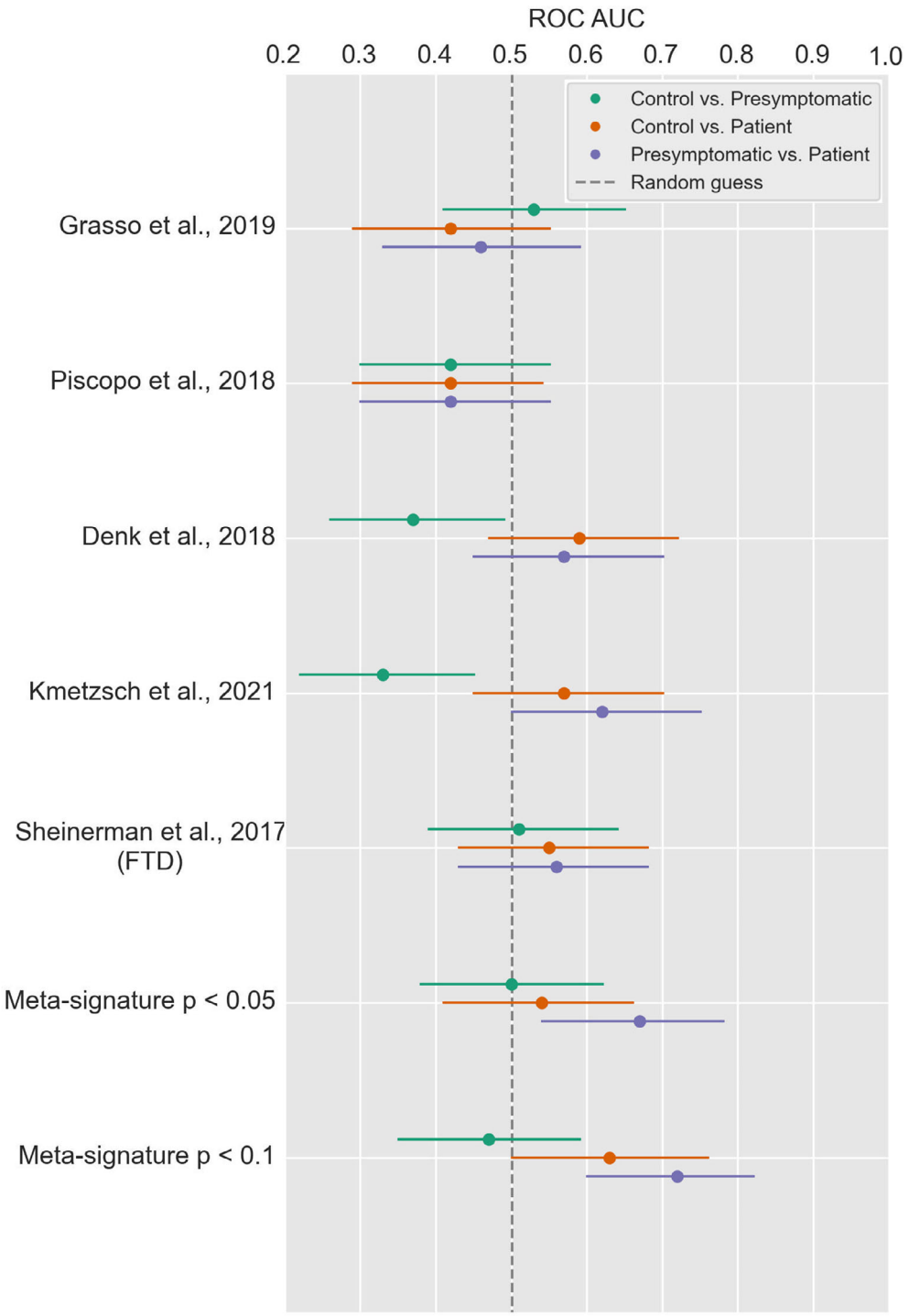


Figure 2. Area under the ROC curve results when classifying groups from the *GRN* cohort. The solid circles indicate the areas under the ROC curves obtained for each pairwise comparison using seven different miRNA signatures. The whiskers denote empirical 90% confidence intervals obtained with 2000 bootstrap samples.

five miRNA signatures linked with FTD in previously published studies, and two meta-signatures obtained in our differential expression analyses with the *GRN* cohort (one consisting of miRNAs with an adjusted p value smaller than 0.05, and another comprising miRNAs with an adjusted p value smaller than 0.1).

The areas under the ROC curves and empirical 90% confidence intervals for each of the seven logistic regression classifiers, in the three pairwise comparison involving the *GRN* cohort, are shown in Figure 2. Strikingly, only one ROC AUC was (slightly) greater than 0.70: when classifying *GRN* presymptomatic subjects and *GRN* patients, using the meta-signature from our differential expression analysis comprising miRNAs with p value lower than 0.1. The miRNA signatures with the largest ROC AUC were from Grasso et al., 2019[8] (0.53 for controls versus presymptomatic subjects), and the meta-signature with $p < 0.1$ (0.63 for controls versus patients, and 0.72 for presymptomatic subjects versus patients).

It is noteworthy that these results are consistent with our differential expression analyses with the *GRN* cohort, offering evidence that most miRNAs identified in previous studies about FTD are not useful biomarkers of *GRN*-associated disease.

Discussion

The goal of this study was to assess all circulating miRNA signatures previously published in the literature as possible biomarkers of FTD and/or ALS, by testing them in two separate homogenous cohorts of *C9orf72* and *GRN* mutation carriers, comprising patients and presymptomatic subjects. The results of this work demonstrate that (1) several miRNAs identified in sporadic or mixed FTD/ALS cohorts could potentially be used as biomarkers of *C9orf72* disease; (2) some miRNAs revealed in a *C9orf72* cohort are validated in an independent *C9orf72* cohort; and (3) most miRNAs associated with FTD in sporadic or mixed cohorts, or in a cohort of *C9orf72* mutation carriers, are not relevant biomarkers of *GRN* disease.

First, differential expression results (Table 3) showed that more than half (35/65) of the miRNAs linked with FTD and/or ALS in the literature were significantly differentially expressed in the *C9orf72* cohort. Remarkably, only four of the 15 selected studies included *C9orf72* mutation carriers,[11,13,16,21] three of which focused exclusively on ALS.[13,16,21] Therefore, these outcomes reveal similarities in miRNA dysregulation between individuals with sporadic forms of FTD/ALS and *C9orf72*-associated disease. Classification results with the *C9orf72* cohort (Fig. 1) also corroborate these findings, since half of the examined miRNA signatures (8/16) yielded at least one pairwise comparison with acceptable performance, and all comparisons employing the meta-signatures had acceptable performance. We considered as an acceptable biomarker any signature with a ROC AUC above 0.70, in accordance with the recommendation by Mandrekar.[25] This means that such a signature has some discriminatory power and is worthy of further exploration. It does not mean that the signature is suitable for clinical use.

Next, we observed that a miRNA signature previously identified in a homogeneous *C9orf72* cohort[11] and another one revealed in a mixed cohort of sporadic and familial ALS[16] displayed an outstanding result (ROC AUC above 0.90) when classifying controls versus *C9orf72* presymptomatic subjects and controls versus *C9orf72* patients (Fig. 1). These two signatures have in common the presence of miR-34a-5p, which has the smallest adjusted p value in the differential expression analyses regarding these comparisons (Table S1, respectively $p = 2.42E\text{-}08$ and $p = 5.06E\text{-}06$). In contrast, the performance of both of these signatures classifying *C9orf72* presymptomatic individuals from patients was unsatisfactory. Indeed, neither of them contained miR-206, which is the most differentially expressed miRNA in this comparison (Table S1, $p = 9.04E\text{-}05$). The overexpression of miR-206 in ALS patients had already been evidenced,[40] and the results of the present work extend this association also to *C9orf72*-disease. Nevertheless, even using the expression levels of miR-206, the classification of *C9orf72* presymptomatic subjects versus patients led to lower performances than comparisons involving the control group: the highest ROC AUC was 0.77, using the meta-signature with $p < 0.1$.

Finally, our results with the *GRN* cohort suggest that previously identified miRNAs have a weaker correlation with disease diagnosis and progression in this genetic form. Only a small proportion (5/30) of the miRNAs associated with FTD in previous studies was significantly differentially expressed in the *GRN* cohort (Table 3), and not a single miRNA was differentially expressed between controls and presymptomatic *GRN* carriers. Regarding the classification experiments, none of the studied miRNA signatures in the *GRN* cohort exhibited an acceptable performance (Fig. 2), and the only ROC AUC slightly above 0.70 was obtained when classifying *GRN* presymptomatic carriers and patients using the largest meta-signature (miRNAs with p value <0.1). One should note that none of the previous studies included *GRN* participants. Thus, our results demonstrate that miRNAs associated with sporadic FTD or genetic FTD due to *C9orf72* are not relevant for *GRN*-associated disease. Further studies are needed to determine if other miRNAs, not analyzed in the present paper, are useful in *GRN*-associated disease.

Validation studies using independent datasets, such as this one, are crucial to assess the utility of biomarker candidates, fostering research rigor and reproducibility. Notably, we carefully defined our research questions and analysis plan before data analysis, and preregistered our

study. Preregistration has the strong benefit of leaving no flexibility for changes in analytical decisions after observing the data, which has been highlighted as a major source of false discoveries and replication failure.[41]

The main limitation of this work is the size of the studied cohorts, particularly the small group of *C9orf72* presymptomatic carriers (17) in comparison with the other groups, due to the rarity of genetic FTD. Additionally, due to the low number of *C9orf72* patients with different phenotypes (20 FTD, 6 FTD/ALS and 3 ALS), no conclusions can be drawn concerning the relationship between miRNAs and different disease manifestations. Moreover, this study did not investigate the influence of lifestyle factors in microRNA levels. Future work will explore the combination of circulating microRNAs with other biomarkers, such as gray matter volume,[5] white matter integrity,[42] and neurofilament light chain level.[43] Multimodality will be crucial to accurately assess progression in *GRN*-associated FTD, and will likely improve the understanding of *C9orf72*-associated disease.

In summary, the present work revealed that most miRNAs previously identified in sporadic or mixed FTD/ALS cohorts are potential biomarkers of *C9orf72*-associated FTD/ALS, but not of *GRN*-associated FTD. Longitudinal studies are needed to confirm our findings and to determine miRNAs expression levels changes throughout disease progression, before circulating miRNAs can be used to assess treatment outcomes in *C9orf72*-associated disease clinical trials.

Acknowledgements

We thank Justine Guégan for microRNA raw reads to counts pipeline handling. The study was conducted with the support of the Centre d'Investigation Clinique Neuroscience (CIC 1422), GH Pitié Salpêtrère, Paris, and the Centre pour l'Acquisition et le Traitement des Images platform. The research leading to these results has received funding from the program "Investissements d'avenir" ANR-10-IAIHU-06, from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), from Agence Nationale de la Recherche/DGOS (project ANR-PRTS PREV-DEMALS, grant number ANR-14-CE15-0016-07, promotion Assistance Publique – Hôpitaux de Paris), from the Programme Hospitalier de Recherche Clinique (PHRC) Predict-PGRN (to ILB, promotion by Assistance Publique – Hôpitaux de Paris), from Fondation Plan Alzheimer (to ILB), from the Inria Project Lab Program (project Neuromarkers), from Fondation Vaincre Alzheimer FR-17035 and from the Institut Français de Bioinformatique (ANR-11-INSB-0013).

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Virgilio Kmetzsch had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Conception and design of the study: all authors. Data acquisition, analysis and interpretation: all authors. Statistical analysis: VK. Drafting of the manuscript: VK. Revising the manuscript: all authors. The French research network on FTD/ALS contributed to data acquisition: Sophie Auriacombe (CHU Pellegrin, Bordeaux), Serge Belliard (CHU Rennes), Frédéric Blanc (Hôpitaux Civils, Strasbourg), Claire Boutoleau-Bretonnière (CHU Laennec, Nantes), Alexis Brice (Hôpital Pitié-Salpêtrière, Paris), Agnès Camuzat (ICM, Paris), Mathieu Ceccaldi (CHU La Timone, Marseille), Philippe Couratier (CHU Limoges), Vincent Deramecourt (CHU Roger Salengro, Lille), Mira Didic (CHU La Timone, Marseille), Charles Duyckaerts (Hôpital Pitié-Salpêtrière, Paris), Frédérique Etcharry-Bouyx (CHU Angers), Maïté Formaglio (CHU Lyon), Véronique Golfier (CHU Rennes), Didier Hannequin (CHU Charles Nicolle, Rouen), Lucette Lacomblez (Hôpital Pitié-Salpêtrière, Paris), Isabelle Le Ber (Hôpital Pitié-Salpêtrière, Paris), Bernard-François Michel (CH Sainte-Marguerite, Marseille), Jérémie Pariente (CHU Rangueil, Toulouse), Florence Pasquier (CHU Lille), Daisy Rinaldi (CHU Pitié-Salpêtrière, Paris), Adeline Rollin-Sillaire (CHU Roger Salengro, Lille), Mathilde Sauvée (CHU Grenoble Alpes), François Sellal (CH Colmar), Christel Thauvin-Robinet (CHU Dijon), Catherine Thomas-Anterion (CH Plein-Ciel, Lyon), Martine Vercelletto (CHU Laennec, Nantes) and David Wallon (CHU Charles Nicolle Rouen).

REFERENCES

1. Neary D, Snowden J, Mann D. Frontotemporal dementia. Lancet Neurol. 2005;4(11):771-780. doi:10.1016/S1474-4422(05)70223-4
2. DeJesus HM, Mackenzie IR, Boeve BF, et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron. 2011;72(2):245-256. doi:10.1016/j.neuron.2011.09.011
3. Renton AE, Majounie E, Waite A, et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron. 2011;72(2):257-268. doi:10.1016/j.neuron.2011.09.010
4. Pasinelli P, Brown RH. Molecular biology of amyotrophic lateral sclerosis: insights from genetics. Nat Rev Neurosci. 2006;7(9):710-723. doi:10.1038/nrn1971

5. Rohrer JD, Nicholas JM, Cash DM, et al. Presymptomatic cognitive and neuroanatomical changes in genetic frontotemporal dementia in the genetic frontotemporal dementia initiative (GENFI) study: a cross-sectional analysis. Lancet Neurol. 2015;14(3):253-262. doi:10.1016/S1474-4422(14)70324-2
6. Grasso M, Piscopo P, Crestini A, Confaloni A, Denti MA. Circulating microRNAs in neurodegenerative diseases. In: Igaz P, ed. Circulating MicroRNAs in Disease Diagnostics and their Potential Biological Relevance. Experientia Supplementum. Springer; 2015:151-169. doi:10.1007/978-3-0348-0955-9_7
7. Huntzinger E, Izaurralde E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat Rev Genet. 2011;12(2):99-110. doi:10.1038/nrg2936
8. Grasso M, Piscopo P, Talarico G, et al. Plasma microRNA profiling distinguishes patients with frontotemporal dementia from healthy subjects. Neurobiol Aging. 2019;84:240.e1-240.e12. doi:10.1016/j.neurobiolaging.2019.01.024
9. Piscopo P, Grasso M, Puopolo M, et al. Circulating miR-127-3p as a potential biomarker for differential diagnosis in frontotemporal dementia. J Alzheimers Dis. 2018;65(2):455-464. doi:10.3233/JAD-180364
10. Denk J, Oberhauser F, Kornhuber J, et al. Specific serum and CSF microRNA profiles distinguish sporadic behavioural variant of frontotemporal dementia compared with Alzheimer patients and cognitively healthy controls. PloS One. 2018;13(5):e0197329. doi:10.1371/journal.pone.0197329
11. Kmetzsch V, Anquetil V, Saracino D, et al. Plasma microRNA signature in presymptomatic and symptomatic subjects with C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis. J Neurol Neurosurg Psychiatry. 2021;92(5):485-493. doi:10.1136/jnnp-2020-324647
12. Sheinerman KS, Toledo JB, Tsivinsky VG, et al. Circulating brain-enriched microRNAs as novel biomarkers for detection and differentiation of neurodegenerative diseases. Alzheimers Res Ther. 2017;9(1):89. doi:10.1186/s13195-017-0316-0
13. Magen I, Yacovzada NS, Yanowski E, et al. Circulating miR-181 is a prognostic biomarker for amyotrophic lateral sclerosis. Nat Neurosci. 2021;24(11):1534-1541. doi:10.1038/s41593-021-00936-z
14. Soliman R, Mousa NO, Rashed HR, et al. Assessment of diagnostic potential of some circulating microRNAs in amyotrophic lateral sclerosis patients, an Egyptian study. Clin Neurol Neurosurg. 2021;208:106883. doi:10.1016/j.clineuro.2021.106883
15. Dobrowolny G, Martone J, Lepore E, et al. A longitudinal study defined circulating microRNAs as reliable biomarkers for disease prognosis and progression in ALS human patients. Cell Death Discov. 2021;7:4. doi:10.1038/s41420-020-00397-6
16. Raheja R, Regev K, Healy BC, et al. Correlating serum microRNAs and clinical parameters in amyotrophic lateral sclerosis. Muscle Nerve. 2018;58(2):261-269. doi:10.1002/mus.26106
17. Waller R, Goodall EF, Milo M, et al. Serum miRNAs miR-206, 143-3p and 374b-5p as potential biomarkers for amyotrophic lateral sclerosis (ALS). Neurobiol Aging. 2017;55:123-131. doi:10.1016/j.neurobiolaging.2017.03.027
18. Tasca E, Pegoraro V, Merico A, Angelini C. Circulating microRNAs as biomarkers of muscle differentiation and atrophy in ALS. Clin Neuropathol. 2016;35(1):22-30. doi:10.5414/NP300889
19. Takahashi I, Hama Y, Matsushima M, et al. Identification of plasma microRNAs as a biomarker of sporadic amyotrophic lateral sclerosis. Mol Brain. 2015;8(1):67. doi:10.1186/s13041-015-0161-7
20. Freischmidt A, Müller K, Zondler L, et al. Serum microRNAs in sporadic amyotrophic lateral sclerosis. Neurobiol Aging. 2015;36(9):2660.e15-2660.e20. doi:10.1016/j.neurobiolaging.2015.06.003
21. Freischmidt A, Müller K, Zondler L, et al. Serum microRNAs in patients with genetic amyotrophic lateral sclerosis and pre-manifest mutation carriers. Brain J Neurol. 2014;137(Pt 11):2938-2950. doi:10.1093/brain/awu249
22. De Felice B, Annunziata A, Fiorentino G, et al. miR-338-3p is over-expressed in blood, CFS, serum and spinal cord from sporadic amyotrophic lateral sclerosis patients. Neurogenetics. 2014;15(4):243-253. doi:10.1007/s10048-014-0420-2
23. Simrén J, Gustafson DR. A blood-based biomarker of cognitive decline and interaction with lifestyle. JAMA Netw Open. 2022;5(3):e223602. doi:10.1001/jamanetworkopen.2022.3602
24. Henriksen K, O'Bryant SE, Hampel H, et al. The future of blood-based biomarkers for Alzheimer's disease. Alzheimers Dement J Alzheimers Assoc. 2014;10(1):115-131. doi:10.1016/j.jalz.2013.01.013
25. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol. 2010;5(9):1315-1316. doi:10.1097/JTO.0b013e3181ec173d
26. Le Ber I, Guedj E, Gabelle A, et al. Demographic, neurological and behavioural characteristics and brain perfusion SPECT in frontal variant of frontotemporal dementia. Brain J Neurol. 2006;129(Pt 11):3051-3065. doi:10.1093/brain/awl288
27. Potla P, Ali SA, Kapoor M. A bioinformatics approach to microRNA-sequencing analysis. Osteoarthr Cartil Open. 2021;3(1):100131. doi:10.1016/j.ocarto.2020.100131
28. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. Genome Res. 2017;27(3):491-499. doi:10.1101/gr.209601.116

29. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17(1):10-12. doi:10.14806/ej.17.1.200
30. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25. doi:10.1186/gb-2009-10-3-r25
31. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinforma Oxf Engl. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352
32. Brennan S, Keon M, Liu B, Su Z, Saksena NK. Panoramic visualization of circulating MicroRNAs across neurodegenerative diseases in humans. Mol Neurobiol. 2019;56(11):7380-7407. doi:10.1007/s12035-019-1615-1
33. Freischmidt A, Goswami A, Limm K, et al. A serum microRNA sequence reveals fragile X protein pathology in amyotrophic lateral sclerosis. Brain. 2021;144(4):1214-1229. doi:10.1093/brain/awab018
34. Xu Q, Zhao Y, Zhou X, Luan J, Cui Y, Han J. Comparison of the extraction and determination of serum exosome and miRNA in serum and the detection of miR-27a-3p in serum exosome of ALS patients. Intractable Rare Dis Res. 2018;7(1):13-18. doi:10.5582/irdr.2017.01091
35. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinforma Oxf Engl. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616
36. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25. doi:10.1186/gb-2010-11-3-r25
37. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289-300.
38. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12 (Oct):2825-2830.
39. Bengio Y, Grandvalet Y. No unbiased estimator of the variance of K-fold cross-validation. J Mach Learn Res. 2004;5(Sep):1089-1105.
40. Toivonen JM, Manzano R, Oliván S, Zaragoza P, García-Redondo A, Osta R. MicroRNA-206: a potential circulating biomarker candidate for amyotrophic lateral sclerosis. PLoS One. 2014;9(2):e89065. doi:10.1371/journal.pone.0089065
41. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. Proc Natl Acad Sci USA. 2018;115(11):2600-2606. doi:10.1073/pnas.1708274114
42. Bertrand A, Wen J, Rinaldi D, et al. Early cognitive, structural, and microstructural changes in Presymptomatic C9orf72 carriers younger than 40 years. JAMA Neurol. 2018;75(2):236-245. doi:10.1001/jamaneurol.2017.4266
43. Saracino D, Dorgham K, Camuzat A, et al. Plasma NfL levels and longitudinal change rates in C9orf72 and GRN-associated diseases: from tailored references to clinical applications. J Neurol Neurosurg Psychiatry. 2021;92 (12):1278-1288. doi:10.1136/jnnp-2021-326914

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Expression heatmaps of differentially expressed microRNAs. Rows represent microRNAs and columns represent individuals ordered by clinical status (control, presymptomatic, and patient). The log2 expression values of microRNAs are standardized (mean of 0 and standard deviation of 1), and z-scores are indicated by colors (blue indicates underexpression and red indicates overexpression). (A) All 35 differentially expressed microRNAs identified in the *C9orf72* cohort. (B) Zoom over five of the most differentially expressed microRNAs identified in the *C9orf72* cohort. (C) All five differentially expressed microRNAs identified in the *GRN* cohort.

Table S1 Complete output from differential expression analyses in the *C9orf72* cohort, for each pairwise comparison between the clinical groups. The columns show the 30 studied miRNAs, the log-fold change when comparing the clinical groups, the unadjusted *p* values, and finally the adjusted *p* values after Benjamini-Hochberg. For each pairwise comparison, a positive log-fold change means that the miRNA is overexpressed in the first group. Controls ($n = 31$), *C9orf72* presymptomatic subjects ($n = 17$), and *C9orf72* patients ($n = 29$). Adjusted *p* values lower than 0.05 are shown in bold.

Table S2 Complete output from differential expression analyses in the *GRN* cohort, for each pairwise comparison between clinical groups. The columns show the 30 studied miRNAs, the log-fold change when comparing the clinical groups, the unadjusted *p* values, and finally the adjusted *p* values after Benjamini-Hochberg. For each pairwise comparison, a positive log-fold change means that the miRNA is overexpressed in the first group. Controls ($n = 31$), *GRN* presymptomatic subjects ($n = 30$), *GRN* patients ($n = 28$). Adjusted *p* values lower than 0.05 are shown in bold.

C.2 Contributions jointes en support au chapitre 3

[31] Kmetzsch V, **Becker E**, Saracino D, Anquetil V, Rinaldi D, Camuzat A, Gareau T, Le Ber I, Colliot O and the PREV-DEMALS study group.
A multimodal variational autoencoder for estimating progression scores from imaging and microRNA data in neurodegenerative diseases.
SPIE Medical Imaging 2022, 2022

[34] Kmetzsch V, **Becker E**, Saracino D, Rinaldi D, Camuzat A, Le Ber I, and Colliot O.
Disease progression score estimation from multimodal imaging and microRNA data using supervised variational autoencoders.
accepted in IEEE J Biomed Health Inform

A multimodal variational autoencoder for estimating progression scores from imaging and microRNA data in rare neurodegenerative diseases

Virgilio Kmetzsch, Emmanuelle Becker, Dario Saracino, Vincent Anquetil, Daisy Rinaldi, Agnès Camuzat, Thomas Gareau, Isabelle Le Ber, Olivier Colliot

A multimodal variational autoencoder for estimating progression scores from imaging and microRNA data in rare neurodegenerative diseases

Virgilio Kmetzsch[a, b], Emmanuelle Becker[c], Dario Saracino[a, b, d, e], Vincent Anquetil[b], Daisy Rinaldi[b, d, e], Agnès Camuzat[b, f], Thomas Gareau[b], Isabelle Le Ber[b, d, e, g], Olivier Colliot[b, a], and The PREV-DEMALS study group

[a]Inria, Aramis project-team, F-75013, Paris, France
[b]Sorbonne Université, Paris Brain Institute – Institut du Cerveau – ICM, Inserm U1127, CNRS UMR 7225, AP-HP - Hôpital Pitié-Salpêtrière, Paris, France
[c]Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France
[d]Centre de référence des démences rares ou précoces, IM2A, AP-HP, Paris, France
[e]DMU Neurosciences, AP-HP - Hôpital Pitié-Salpêtrière, Paris, France
[f]EPHE, PSL Research University, Paris, France
[g]Paris Brain Institute – Institut du Cerveau – ICM, FrontLab, Paris, France

ABSTRACT

Frontotemporal dementia (FTD) is a rare neurodegenerative disease, often of genetic origin, with no effective treatment. There is a substantial pathophysiological overlap with amyotrophic lateral sclerosis (ALS), mutations in the *C9orf72* gene being their most common genetic cause. In these disorders, no single biomarker can accurately measure progression, thus it is crucial to combine complementary information from multiple modalities to evaluate new therapeutic interventions. In particular, neuroimaging and transcriptomic (microRNA) data have been shown to have value to track FTD and ALS progression. As these conditions are rare, large samples are not available, hence the need for methods to fuse multimodal data from small samples. In this paper, we propose a method for computing a disease progression score (DPS) from cross-sectional multimodal data, based on variational autoencoders (VAE). We show that unsupervised training leads to the estimation of meaningful latent spaces, where subjects with similar disease states are clustered together and from which a DPS may be inferred. Models were evaluated on 14 patients, 40 presymptomatic mutation carriers and 37 healthy controls from the PREV-DEMALS study. Since there is no ground truth for the DPS, we used the inferred scores to perform pairwise classification as a proxy metric. Presymptomatic subjects and patients were classified with an average area under the ROC curve of 0.83 and 0.94, respectively without and with feature selection. The proposed approach has the potential to leverage cross-sectional multimodal datasets with small sample sizes in order to objectively measure disease progression.

Keywords: Multimodal, neuroimaging, transcriptomics, microRNA, variational autoencoder, deep learning, disease progression score, neurodegenerative disease

1. INTRODUCTION

Frontotemporal dementia (FTD) is a rare heterogeneous neurodegenerative disease characterized by progressive behavioral changes, executive dysfunction and language impairments.[1] A large proportion of FTD cases are due to genetic mutations, the most frequent being an expansion in the *C9orf72* gene.[2,3] *C9orf72* expansions are also an important genetic cause of amyotrophic lateral sclerosis (ALS), a motor neuron disease leading to muscle atrophy, progressive weakness and eventual paralysis.[4] These fatal disorders, which may occasionally co-occur in *C9orf72*-mutated individuals, have no effective treatment to date.

Further author information: (Send correspondence to Olivier Colliot.)
Olivier Colliot: E-mail: olivier.colliot@sorbonne-universite.fr, Telephone: +33 1 57 27 43 65

Presymptomatic carriers of the *C9orf72* mutation, with no current clinical symptoms, are an ideal population for the evaluation of new disease-modifying treatments, before any irreversible brain damage has occurred. Previous work demonstrated the importance of neuroimaging[5] and transcriptomics (microRNA)[6] biomarkers to better understand the *C9orf72* disease progression. However, when analysed independently, neuroimaging and microRNA data provide incomplete views of FTD and ALS. Therefore, in order to monitor the effect of experimental therapies, it is critical to leverage the complementary information provided by these modalities.

Since different biomarkers characterize a disease in different stages, several biomarkers could be combined to represent the entire progression with a single disease progression score (DPS). Many approaches have been developed for data-driven disease progression modeling, including event-based models (EBM),[7,8] a vertex-wise model of brain pathology fitted with expectation-maximisation,[9] non-linear mixed-effects models,[10,11] alternating least squares to fit sigmoid functions,[12] Gaussian processes,[13] Recurrent Neural Networks[14] and M-estimation.[15] Most of these approaches require a large amount of longitudinal data, which is not available for FTD/ALS. The only published methods that infer a disease progression score from cross-sectional data are event-based models,[7,8] but these approaches do not scale well for hundreds of biomarkers, such as microRNA data.

In the present work, we proposed a method for inferring a disease progression score (a latent trait) based on a multimodal variational autoencoder (VAE).[16] VAEs are powerful generative models that project data in a regularized latent low dimensional space and have been shown to be effective in high dimensional low sample size settings.[17] These models have already been used with multimodal data,[18] although not with the goal of inferring a DPS. We hypothesized that the inferred score, based on cross-sectional neuroimaging and microRNA data, could represent the distance traveled along the underlying FTD/ALS pathophysiological pathway, and thus be used to monitor disease progression and evaluate novel treatments.

2. MATERIALS AND METHODS

2.1 Studied population

Participants were recruited through the PREV-DEMALS study (https://clinicaltrials.gov, ID NCT02590276), a cohort focused on *C9orf72* expansion carriers, comprising neuroimaging and microRNA sequencing data. MicroRNAs (miRNAs) are a class of noncoding RNAs that negatively regulate gene expression,[19] being detected in blood plasma and correlating with the progression of many neurodegenerative diseases,[20] including FTD and ALS.

Our study comprised 110 individuals, divided into three groups: 22 symptomatic carriers of a pathogenic expansion (patient group), 45 asymptomatic carriers (presymptomatic group) and 43 asymptomatic non-carriers (control group). Written informed consents were obtained from all participants and the study was approved by the ethics committee (Comité de Protection des Personnes CPP Ile-De-France VI, CPP 68-15 and ID RCB 2015-A00856-43).

2.2 Data acquisition and preprocessing

All individuals had transcriptomic data available, consisting of the expression levels of 589 miRNAs. However, only 91 (14 patients, 40 presymptomatic carriers and 37 controls) had also neuroimaging data available, consisting in grey matter volumes extracted from anatomical MRI (T1) including 68 cortical regions of interest (ROIs) (Desikan atlas) and 18 subcortical ROIs (Aseg atlas) as well as the estimated total intracranial volume, thus resulting in 87 imaging features. Details regarding features and population can be found in Ref. 6 and Ref. 5. Subjects were divided into two datasets: 19 subjects with only microRNA data, used as a discovery set for feature selection, and 91 subjects with multimodal neuroimaging and microRNA data, used as input to our models. Features were rescaled from 0 to 1 and ordered via principal component analysis in the transposed data matrix: we projected features into the first principal component and used the coordinate values to sort them.

We also conducted experiments with two simulated datasets, based on the real one. To build the simulated data matrices, we simply increased (or decreased) each feature value by 5% or 15% for all patients and healthy controls, to accentuate their means' difference. The presymptomatic participants remained unchanged.

2.3 Multimodal variational autoencoder

In order to build disease progression scores, we propose a multimodal variational autoencoder for estimating a latent space representation. Let $x \in \mathcal{X}$ represent a set of multimodal data, where each point is a vector with concatenated neuroimaging and microRNA data. A variational autoencoder (VAE)[16] is a generative model which aims to learn the training data distribution using a latent representation model:

$$p(x) = \int p(x|z)p(z)dz, \tag{1}$$

where $z \in \mathcal{Z}$ is a lower dimensional latent variable and $p(z)$ is its prior distribution (commonly a multivariate unit Gaussian). VAEs learn two mappings in the form of neural networks: an encoder $q_\phi(z|x)$ which maps data x to its latent representation z, and a decoder $p_\theta(x|z)$ which maps from the latent representation z back to the input space. Since the marginal log-likelihood of the data is intractable, VAEs are trained to maximize the variational lower bound of the marginal log-likelihood, known as ELBO (Evidence Lower Bound):

$$\log p(x) \geq \mathbb{E}_{z\sim q_\phi}[\log p_\theta(x|z)] - D_{KL}[q_\phi(z|x)||p(z)], \tag{2}$$

where $D_{KL}[q_\phi(z|x)||p(z)]$ is the Kullback-Leibler divergence between the approximated posterior $q_\phi(z|x)$ and the prior distribution $p(z)$ and acts as a regularization term.

Our encoder consisted of a 1-dimensional convolution layer, followed by two fully-connected layers, while the decoder was implemented with two fully-connected layers followed by a 1-dimensional transposed convolutional layer. After each layer, batch normalization[21] was applied for its regularization properties and to avoid vanishing or exploding gradients. The nonlinear activation function was the rectified linear unit (ReLU) $f(x) = max(0, x)$ in all layers except the decoder's last one, which used a sigmoid function $f(x) = \frac{1}{1+e^{-x}}$ in order to have the output normalized between 0 and 1. The loss function was optimized using Adam.[22]

Two slightly different networks were used as our final models. For the experiments with no feature selection (589 miRNAs + 87 neuroimaging features), we identified that 64 channels and a kernel of dimension 80 with a stride of 10 was a good parametrization, along with a hidden layer of 400 units and a latent space of dimension 5. For the experiments with the discovery set and feature selection (68 miRNAs + 87 neuroimaging features), we chose 32 channels, a kernel of dimension 20 with a stride of 5, along with a hidden layer of 50 units and 2-dimensional latent space. The VAEs were implemented in Python 3.8.5 using PyTorch 1.8.1, and trained with batches of 32 subjects for 250 epochs using a learning rate of 10^{-3}.

2.4 Computing disease progression scores in the latent space

We used a stratified 5-fold cross-validation strategy, training the VAE with four folds and testing with the remaining fold in each iteration. Training was unsupervised: no clinical labels were used. Our hypothesis was that the VAE would identify a meaningful latent space, placing subjects with the same clinical status (and similar disease stage among presymptomatic individuals and patients) closer together.

Once each model was trained, we projected the training data in the latent space and used the clinical labels (patient, presymptomatic subject or control) to compute the centroid of each group. We then defined the trajectory to traverse the latent space as the line passing through the centroids of the presymptomatic and the patient groups. Finally, we encoded the test fold in the latent space and computed the DPS for each subject as the coordinate of their projection in this line.

Since there is no ground truth for the DPS, we applied a proxy metric to assess model performance: the inferred scores were used to classify subjects according to their clinical status. Therefore, labels were used during test time to compute the area under the receiver operating characteristic curve (ROC AUC) averaged over the five folds.

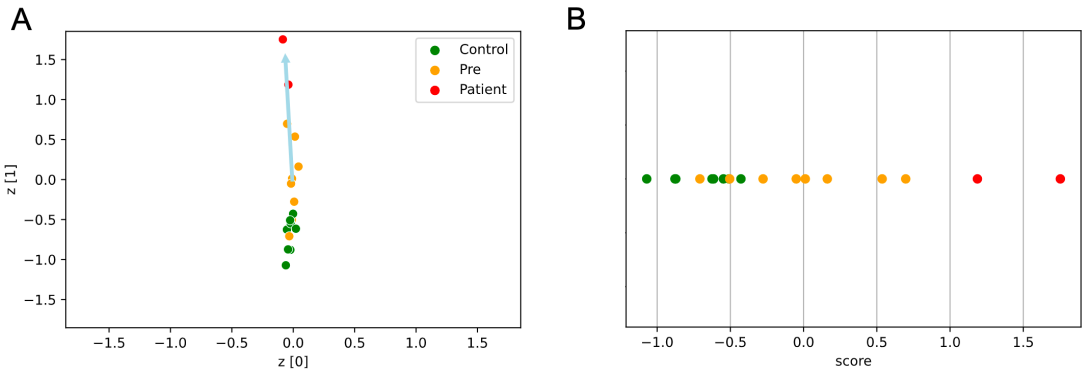


Figure 1. Example of (A) a 2-dimensional latent space encoding a test fold comprising 2 patients, 8 presymptomatic individuals and 8 controls and the trajectory to traverse this latent space (blue arrow); (B) the corresponding disease progression scores.

3. RESULTS

Figure 1 depicts an example of a two-dimensional latent space obtained after training the VAE with four folds and using the trained model to encode the remaining test fold. In this particular fold, we observe a perfect separation between patients and the other two groups, and a clear (although not perfect) distinction between presymptomatic individuals and controls.

Table 1 displays the mean and standard deviation of the area under the ROC curve obtained after a 5-fold cross-validation, for each pairwise comparison between clinical groups. The models were initially trained without any feature selection, with the expression levels of 589 miRNAs and the grey matter volumes of 87 ROIs. Then, we used 19 subjects as a discovery set to identify the most differentially expressed miRNAs between clinical groups, reducing the dimension of the microRNA data to 68. Classification performance improved when feature selection was applied. Table 1 also shows the results with the two simulated datasets. As expected, performance increases when a dataset with more discriminating features is used as input.

Table 1. Area under the ROC curve (mean ± SD) for each pairwise classification, obtained using the inferred disease progression scores respectively without feature selection, with feature selection and with two simulated datasets.

Comparison | Features	589 miRNAs, 87 T1 ROIs	68 miRNAs, 87 T1 ROIs	Simulated dataset (5%)	Simulated dataset (15%)
Control vs Presymptomatic	0.57 ± 0.15	0.62 ± 0.20	0.76 ± 0.19	0.89 ± 0.12
Control vs Patient	0.88 ± 0.11	0.95 ± 0.07	0.99 ± 0.02	1.00 ± 0.00
Presymptomatic vs Patient	0.83 ± 0.19	0.94 ± 0.12	0.95 ± 0.10	0.98 ± 0.03

Finally, Fig. 2 presents the visualization of the inferred scores for all 91 subjects after a 5-fold cross validation. The scores were computed for each subject when included in a test fold, without or with feature selection in the microRNA data. There is a superior performance (better separation between groups) when miRNAs are selected using the discovery set.

4. DISCUSSION

We proposed a multimodal variational autoencoder for combining imaging and transcriptomic (microRNA) data. It allowed inferring a single score to represent disease progression, using only cross-sectional neuroimaging and microRNA data from less than a hundred subjects. We showed that variational autoencoders built with shallow 1-dimensional convolutional neural networks were able to infer meaningful latent spaces, putting closer together subjects from the same clinical groups (patients, presymptomatic individuals and controls) without using any

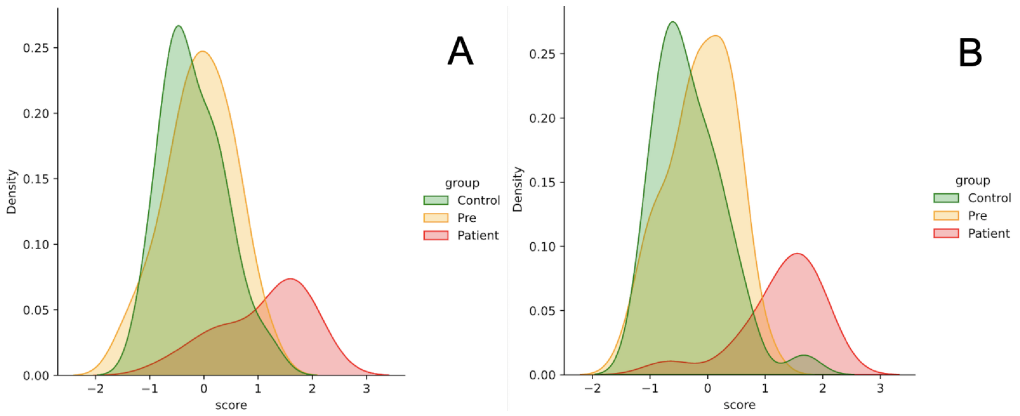


Figure 2. Disease progression score densities conditioned on clinical status, obtained from 5-fold cross-validation with 91 individuals, when using (A) all 589 miRNAs and all 87 T1 MRI ROIs and (B) selected 68 miRNAs and all 87 T1 MRI ROIs.

labels during training. We were able to encode individuals from the test sets into the latent spaces and compute their corresponding DPS. Then using only the computed scores, presymptomatic subjects and patients were distinguished with an average ROC AUC of 0.83 and 0.94, respectively without and with feature selection.

Our experiments with the simulated datasets showed that more informative features will lead to better results. In addition, the presented approach is generic enough to be used with datasets from other neurodegenerative diseases, even though our experiments focused only on *C9orf72*-associated FTD and ALS. So our results motivate further experiments with other neurodegenerative diseases with well established biomarkers.

The current study has a limitation: the absence of ground truth for the progression scores, which led us to use classification performance as a proxy metric. Long-term longitudinal data would be needed to confirm the accuracy of the inferred DPS. For instance, we hypothesize that, for presymptomatic subjects, a higher DPS implies an earlier disease onset, and we would need long-term follow-up data to confirm this hypothesis. Future work could explore different network architectures (e.g. 2-dimensional inputs, different number of layers, feature maps and kernel sizes), investigate the integration of different prior information to order the input data and analyze different methods to traverse the latent space.

In summary, our results encourage the use of the proposed approach as a tool to measure disease progression in rare neurodegenerative diseases and evaluate potential treatments.

ACKNOWLEDGMENTS

The PREV-DEMALS study group includes: Eve Benchetrit (Hôpital de la Salpêtrière, Paris), Anne Bertrand (Hôpital de la Salpêtrière, Paris), Anne Bissery (Hôpital de la Salpêtrière, Paris), Marie-Paule Boncoeur (CHU Dypuytren, Limoges), Stéphanie Bombois (CHU Roger Salengro, Lille), Agnès Camuzat (ICM, Paris), Mathieu Chastan (CHU Charles Nicolle, Rouen), Yaohua Chen (CHU Roger Salengro, Lille), Marie Chupin (ICM, Paris), Olivier Colliot (ICM, Paris), Philippe Couratier (CHU Dypuytren, Limoges), Xavier Delbeuck (CHU Roger Salengro, Lille), Vincent Deramecourt (CHU Roger Salengro, Lille), Christine Delmaire (CHU Roger Salengro, Lille), Emmanuel Gerardin (CHU Charles Nicolle, Rouen), Claude Hossein-Foucher (CHU Roger Salengro, Lille), Bruno Dubois (Hôpital de la Salpêtrière, Paris), Marie-Odile Habert (Hôpital de la Salpêtrière, Paris), Didier Hannequin (CHU Charles Nicolle, Rouen), Géraldine Lautrete (CHU Dypuytren, Limoges), Thibaud Lebouvier (CHU Roger Salengro, Lille), Isabelle Le Ber (Hôpital de la Salpêtrière, Paris), Benjamin Le Toullec (ICM, Paris), Richard Levy (Hôpital de la Salpêtrière, Paris), Olivier Martinaud (CHU Charles Nicolle, Rouen), Kelly Martineau (ICM, Paris), Marie-Anne Mackowiak (CHU Roger Salengro, Lille), Jacques Monteil (CHU Dypuytren, Limoges), Florence Pasquier (CHU Roger Salengro, Lille), Gregory Petyt (CHU Roger Salengro, Lille), Pierre-François Pradat (Hôpital de la Salpêtrière, Paris), Assi-Hervé Oya (Hôpital de la Salpêtrière, Paris), Armelle Rametti-Lacroux (Hôpital de la Salpêtrière, Paris), Daisy Rinaldi (Hôpital de la Salpêtrière, Paris), Adeline

Rollin-Sillaire (CHU Roger Salengro, Lille), François Salachas (Hôpital de la Salpêtrière, Paris), Sabrina Sayah (Hôpital de la Salpêtrière, Paris), David Wallon (CHU Charles Nicolle, Rouen).

The research leading to these results has received funding from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6), from Agence Nationale de la Recherche (project PREV-DEMALS, grant number ANR-14-CE15-0016-07), and from the Inria Project Lab Program (project Neuromarkers).

We thank Justine Guégan, from iCONICS (ICM bioinformatics facility) for microRNA raw reads to counts pipeline handling. We also thank Yannick Marie and Delphine Bouteiller, from iGenSeq (ICM sequencing facility), for library preparation and sequencing. Part of this work was carried out on the DNA and cell bank of ICM. We gratefully acknowledge Philippe Martin-Hardy and Ludmila Jornea for technical assistance. The study was conducted with the support of the Centre d'Investigation Clinique Neuroscience (CIC 1422), GH Pitié Salpêtrière, Paris, and the Centre pour l'Acquisition et le Traitement des Images platform.

REFERENCES

[1] Rascovsky, K., Hodges, J. R., Knopman, D., Mendez, M. F., Kramer, J. H., Neuhaus, J., van Swieten, J. C., Seelaar, H., Dopper, E. G. P., Onyike, C. U., Hillis, A. E., Josephs, K. A., Boeve, B. F., Kertesz, A., Seeley, W. W., Rankin, K. P., Johnson, J. K., Gorno-Tempini, M.-L., Rosen, H., Prioleau-Latham, C. E., Lee, A., Kipps, C. M., Lillo, P., Piguet, O., Rohrer, J. D., Rossor, M. N., Warren, J. D., Fox, N. C., Galasko, D., Salmon, D. P., Black, S. E., Mesulam, M., Weintraub, S., Dickerson, B. C., Diehl-Schmid, J., Pasquier, F., Deramecourt, V., Lebert, F., Pijnenburg, Y., Chow, T. W., Manes, F., Grafman, J., Cappa, S. F., Freedman, M., Grossman, M., and Miller, B. L., "Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia," *Brain* **134**, 2456–2477 (Sept. 2011).

[2] DeJesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., Nicholson, A. M., Finch, N. A., Flynn, H., Adamson, J., Kouri, N., Wojtas, A., Sengdy, P., Hsiung, G.-Y. R., Karydas, A., Seeley, W. W., Josephs, K. A., Coppola, G., Geschwind, D. H., Wszolek, Z. K., Feldman, H., Knopman, D. S., Petersen, R. C., Miller, B. L., Dickson, D. W., Boylan, K. B., Graff-Radford, N. R., and Rademakers, R., "Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS," *Neuron* **72**, 245–256 (Oct. 2011).

[3] Renton, A. E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J. R., Schymick, J. C., Laaksovirta, H., van Swieten, J. C., Myllykangas, L., Kalimo, H., Paetau, A., Abramzon, Y., Remes, A. M., Kaganovich, A., Scholz, S. W., Duckworth, J., Ding, J., Harmer, D. W., Hernandez, D. G., Johnson, J. O., Mok, K., Ryten, M., Trabzuni, D., Guerreiro, R. J., Orrell, R. W., Neal, J., Murray, A., Pearson, J., Jansen, I. E., Sondervan, D., Seelaar, H., Blake, D., Young, K., Halliwell, N., Callister, J. B., Toulson, G., Richardson, A., Gerhard, A., Snowden, J., Mann, D., Neary, D., Nalls, M. A., Peuralinna, T., Jansson, L., Isoviita, V.-M., Kaivorinne, A.-L., Hölttä-Vuori, M., Ikonen, E., Sulkava, R., Benatar, M., Wuu, J., Chiò, A., Restagno, G., Borghero, G., Sabatelli, M., ITALSGEN Consortium, Heckerman, D., Rogaeva, E., Zinman, L., Rothstein, J. D., Sendtner, M., Drepper, C., Eichler, E. E., Alkan, C., Abdullaev, Z., Pack, S. D., Dutra, A., Pak, E., Hardy, J., Singleton, A., Williams, N. M., Heutink, P., Pickering-Brown, S., Morris, H. R., Tienari, P. J., and Traynor, B. J., "A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD," *Neuron* **72**, 257–268 (Oct. 2011).

[4] Pasinelli, P. and Brown, R. H., "Molecular biology of amyotrophic lateral sclerosis: insights from genetics," *Nature Reviews Neuroscience* **7**, 710–723 (Sept. 2006). Number: 9 Publisher: Nature Publishing Group.

[5] Bertrand, A., Wen, J., Rinaldi, D., Houot, M., Sayah, S., Camuzat, A., Fournier, C., Fontanella, S., Routier, A., Couratier, P., Pasquier, F., Habert, M.-O., Hannequin, D., Martinaud, O., Caroppo, P., Levy, R., Dubois, B., Brice, A., Durrleman, S., Colliot, O., Le Ber, I., and Predict to Prevent Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis (PREV-DEMALS) Study Group, "Early Cognitive, Structural, and Microstructural Changes in Presymptomatic C9orf72 Carriers Younger Than 40 Years," *JAMA neurology* **75**(2), 236–245 (2018).

[6] Kmetzsch, V., Anquetil, V., Saracino, D., Rinaldi, D., Camuzat, A., Gareau, T., Jornea, L., Forlani, S., Couratier, P., Wallon, D., Pasquier, F., Robil, N., Grange, P. d. l., Moszer, I., Ber, I. L., Colliot, O., and Becker, E., "Plasma microRNA signature in presymptomatic and symptomatic subjects with C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis," *Journal of Neurology, Neurosurgery & Psychiatry* (Nov. 2020). Publisher: BMJ Publishing Group Ltd Section: Neurodegeneration.

[7] Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., Scahill, R. I., Tabrizi, S. J., Ourselin, S., Fox, N. C., and Alexander, D. C., "An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease," *NeuroImage* **60**, 1880–1889 (Apr. 2012).

[8] Venkatraghavan, V., Bron, E. E., Niessen, W. J., and Klein, S., "Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling," *NeuroImage* **186**, 518–532 (Feb. 2019).

[9] Marinescu, R. V., Eshaghi, A., Lorenzi, M., Young, A. L., Oxtoby, N. P., Garbarino, S., Crutch, S. J., Alexander, D. C., and Alzheimer's Disease Neuroimaging Initiative, "DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders," *NeuroImage* **192**, 166–177 (May 2019).

[10] Schiratti, J.-B., Allassonnière, S., Colliot, O., and Durrleman, S., "A Bayesian Mixed-Effects Model to Learn Trajectories of Changes from Repeated Manifold-Valued Observations," *Journal of Machine Learning Research* **18**(133), 1–33 (2017).

[11] Koval, I., Bône, A., Louis, M., Lartigue, T., Bottani, S., Marcoux, A., Samper-González, J., Burgos, N., Charlier, B., Bertrand, A., Epelbaum, S., Colliot, O., Allassonnière, S., and Durrleman, S., "AD Course Map charts Alzheimer's disease progression," *Scientific Reports* **11**, 8020 (Apr. 2021). 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 science;Computer science Subject_term_id: computational-science;computer-science.

[12] Jedynak, B. M., Lang, A., Liu, B., Katz, E., Zhang, Y., Wyman, B. T., Raunig, D., Jedynak, C. P., Caffo, B., and Prince, J. L., "A computational neurodegenerative disease progression score: Method and results with the Alzheimer's disease neuroimaging initiative cohort," *NeuroImage* **63**, 1478–1486 (Nov. 2012).

[13] Lorenzi, M., Filippone, M., Frisoni, G. B., Alexander, D. C., and Ourselin, S., "Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease," *NeuroImage* **190**, 56–68 (Apr. 2019).

[14] Mehdipour Ghazi, M., Nielsen, M., Pai, A., Cardoso, M. J., Modat, M., Ourselin, S., and Sørensen, L., "Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling," *Medical Image Analysis* **53**, 39–46 (Apr. 2019).

[15] Mehdipour Ghazi, M., Nielsen, M., Pai, A., Modat, M., Jorge Cardoso, M., Ourselin, S., and Sørensen, L., "Robust parametric modeling of Alzheimer's disease progression," *NeuroImage* **225**, 117460 (Jan. 2021).

[16] Kingma, D. P. and Welling, M., "Auto-Encoding Variational Bayes," *arXiv:1312.6114 [cs, stat]* (May 2014). arXiv: 1312.6114.

[17] Chadebec, C., Thibeau-Sutre, E., Burgos, N., and Allassonnière, S., "Data Augmentation in High Dimensional Low Sample Size Setting Using a Geometry-Based Variational Autoencoder," *arXiv:2105.00026 [cs, stat]* (Apr. 2021). arXiv: 2105.00026.

[18] Antelmi, L., Ayache, N., Robert, P., and Lorenzi, M., "Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data," in [*International Conference on Machine Learning*], 302–311, PMLR (May 2019). ISSN: 2640-3498.

[19] Huntzinger, E. and Izaurralde, E., "Gene silencing by microRNAs: contributions of translational repression and mRNA decay," *Nature Reviews Genetics* **12**, 99–110 (Feb. 2011). Number: 2 Publisher: Nature Publishing Group.

[20] Grasso, M., Piscopo, P., Confaloni, A., and Denti, M. A., "Circulating miRNAs as biomarkers for neurodegenerative disorders," *Molecules (Basel, Switzerland)* **19**, 6891–6910 (May 2014).

[21] Ioffe, S. and Szegedy, C., "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv:1502.03167 [cs]* (Mar. 2015). arXiv: 1502.03167.

[22] Kingma, D. P. and Ba, J., "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]* (Jan. 2017). arXiv: 1412.6980.

Disease progression score estimation from multimodal imaging and microRNA data using supervised variational autoencoders

Virgilio Kmetzsch, Emmanuelle Becker, Dario Saracino, Daisy Rinaldi, Agnès Camuzat, Isabelle Le Ber, and Olivier Colliot, *Member, IEEE*, for the PREV-DEMALS study group

***Abstract*— Frontotemporal dementia and amyotrophic lateral sclerosis are rare neurodegenerative diseases with no effective treatment. The development of biomarkers allowing an accurate assessment of disease progression is crucial for evaluating new therapies. Concretely, neuroimaging and transcriptomic (microRNA) data have been shown useful in tracking their progression. However, no single biomarker can accurately measure progression in these complex diseases. Additionally, large samples are not available for such rare disorders. It is thus essential to develop methods that can model disease progression by combining multiple biomarkers from small samples. In this paper, we propose a new framework for computing a disease progression score (DPS) from cross-sectional multimodal data. Specifically, we introduce a supervised multimodal variational autoencoder that can infer a meaningful latent space, where latent representations are placed along a disease trajectory. A score is computed by orthogonal projections onto this path. We evaluate our framework with multiple synthetic datasets and with a real dataset containing 14 patients, 40 presymptomatic genetic mutation carriers and 37 controls from the PREV-DEMALS study. There is no ground truth for the DPS in real-world scenarios, therefore we use the area under the ROC curve (AUC) as a proxy metric. Results with the synthetic datasets support this choice, since the higher the AUC, the more accurate the predicted simulated DPS. Experiments with the real dataset demonstrate better performance in comparison with a state-of-the-art approach. The proposed framework thus leverages cross-sectional multimodal datasets with small sample sizes to objectively measure disease progression, with potential application in clinical trials.**

***Index Terms*— Disease progression score, Deep learning, MicroRNA, Multimodal data, Neurodegenerative disease, Neuroimaging, Variational autoencoder**

The research leading to these results has received funding from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6), from Agence Nationale de la Recherche (project PREV-DEMALS, grant number ANR-14-CE15-0016-07), and from the Inria Project Lab Program (project Neuromarkers).

V. Kmetzsch, D. Saracino and O. Colliot are with Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France (e-mails: virgilio.kmetzsch@inria.fr, dario.saracino@icm-institute.org, olivier.colliot@cnrs.fr).

E. Becker is with Univ Rennes, Inria, CNRS, IRISA, F-35000, Rennes, France (e-mail: emanuelle.becker@univ-rennes1.fr).

D. Rinaldi and A. Camuzat are with Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France (e-mails: daisy.rinaldi@icm-institute.org, agnes.camuzat@icm-institute.org).

I. Le Ber is with Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Institute of Memory and Alzheimer's Disease (IM2A), Centre of excellence of neurodegenerative disease (CoEN), Department of Neurology, DMU Neurosciences, F-75013, Paris, France (e-mail: isabelle.leber@icm-institute.org).

I. Introduction

FRONTOTEMPORAL dementia (FTD) and amyotrophic lateral sclerosis (ALS) are rare neurodegenerative disorders that have devastating personal and social consequences. Progressive cognitive and behavioural changes, emotional instability, and language impairment are the main symptoms of FTD [1]. ALS is a motor neuron disease characterized by gradual muscle wasting, ultimately leading to disability [2]. FTD and ALS may be sporadic (no previous family history) or genetically inherited. The most common genetic cause of FTD and ALS is a hexanucleotide repeat expansion in the *C9orf72* gene [3], [4]. These fatal conditions can sometimes coexist in *C9orf72*-mutated individuals, and have no cure or standard treatment to date.

Carriers of the *C9orf72* mutation that do not present clinical symptoms are considered presymptomatic, since they have a very high probability of manifesting FTD and/or ALS later in life. Clinical trials for potential therapies are likely to be most effective at this presymptomatic stage, before any irreversible brain damage has occurred. However, the evaluation of new treatments depends on an accurate measure of disease progression, which is not evident without observable symptoms. Therefore, it is crucial to identify biomarkers to assess disease progression in presymptomatic subjects. Indeed, previous work has shown the relevance of neuroimaging [5], [6] and transcriptomic (microRNA) [7] biomarkers for a better understanding of *C9orf72*-disease in presymptomatic carriers. Nevertheless, when these modalities are analysed separately, they provide only an incomplete picture of these complex neurodegenerative diseases. It is thus essential to develop methods that leverage the complementary information available from different modalities to accurately measure disease progression. As different biomarkers characterise distinct disease stages, various biomarkers can be combined to represent the entire

disease course with a single measure, commonly referred in the literature as the *disease progression score* (DPS).

The idea of computing disease progression scores falls within the larger topic of modeling disease progression. In the past years, many approaches have been developed for data-driven modeling of disease progression, such as event-based models (EBM) [8], [9], different algorithms fitting logistic functions to biomarker trajectories [10], [11], non-linear mixed-effects models [12], [13], a vertex-wise model of brain diseases fitted with expectation-maximisation [14], Gaussian processes [15], topological profiles reflecting brain connectivity [16], Bayesian multi-task learning [17], and recurrent neural networks [18].

Most of these approaches require longitudinal data. For instance, the authors of [10] assume that the longitudinal dynamic of each biomarker can be represented as a sigmoidal function of the DPS. They propose a joint optimization algorithm to compute the DPS, fit one sigmoid function per biomarker using alternating least squares, and apply their work to hundreds of patients with Alzheimer's disease (AD). Similarly, a more recent method [11], also applied to AD, uses M-estimation to map each subject's age to a DPS, jointly fitting generalized logistic functions to the longitudinal dynamics of biomarkers as functions of the DPS. Schiratti et al [12] proposed a general non-linear mixed-effects model for longitudinal data based on concepts from Riemannian geometry. The application of this framework to AD, called AD Course Map [13], allowed to map each subject to their corresponding disease stage. The authors of [15] proposed a probabilistic approach based on Gaussian process regression from time-series of biomarker measurements. Yet another framework, named Data-driven Inference of Vertexwise Evolution (DIVE) [14] consists in identifying clusters of vertex-wise biomarker measurements in the brain, and estimating representative trajectories for these clusters. Finally, [18] uses recurrent neural networks to predict biomarker values without parametric assumptions about trajectories, with application to AD. To the best of our knowledge, the only disease modeling approaches that infer a DPS from cross-sectional data are EBM [8], [9]. These models explore the temporal sequence in which biomarkers become abnormal in the course of a disease. They have been successfully applied to a variety of diseases including AD [8], [9], [19]–[22], multiple sclerosis [23], [24], Parkinson's disease [25], Huntington's disease [26] as well as FTD [27], [28] and ALS [29]. However, in these works, EBMs were applied to a relatively small number of features (typically 10-50) and it is unkwown if they would perform well in higher dimensions.

Despite the recognized importance of estimating neurodegenerative diseases progression, research has tended to focus mostly on higher prevalence conditions. Existing solutions are thus inadequate to model rare diseases with high-dimensional cross-sectional data, for three main reasons. First, we observe that longitudinal data is needed for the vast majority of approaches. However, *C9orf72*-associated FTD and ALS are slowly progressive conditions in the presymptomatic phase, which hinders the collection of meaningful longitudinal data. Second, most published methods benefit from large samples, which are not available for very low prevalence disorders such as genetic FTD and ALS. Finally, it is unclear if event-based models, the only methods suitable for cross-sectional data, can be robustly applied to high-dimensional microRNA expression data, which comprise hundreds of biomarkers.

In this paper, we thus present a novel framework to estimate disease progression scores for rare neurodegenerative disorders using only cross-sectional data. To that purpose, we introduce a new supervised multimodal variational autoencoder (VAE) trained with neuroimaging and microRNA data. Our working hypothesis is that disease progression scores may be modelled as underlying latent traits. Concretely, we aim to learn a meaningful latent space, where the relative positions of latent representations indicate the distance travelled along the disease pathophysiological pathway.

VAEs are powerful generative models that project data into a low-dimensional regularized latent space [30]. These models have been previously used with multimodal data [31]–[33], but not for the purpose of inferring a DPS. Usually VAEs are trained in an unsupervised manner. However, extensions have been proposed for semi-supervised [34]–[36] or supervised [37] tasks. These studies demonstrate that providing supervision to the model imposes specific semantics on the latent space, resulting in more meaningful and robust representations. In our context, explicit labels (control, presymptomatic, patient) are already available for all subjects. We thus add supervision during training, leveraging this information to improve the separation of the groups in the latent space. Additionally, we propose to split high-dimensional (neuroimaging and microRNA data) and low-dimensional (demographic information) modalities. Our model thus couples two neural networks with different inputs: (1) an encoder/decoder that learns a latent space from the high-dimensional features, and (2) a classifier having as input the latent variables concatenated with the low dimensional features, useful for the classification task. As no ground truth is available for the DPS in real-world scenarios, we evaluate our models with a proxy metric: the area under the ROC curve (AUC) for each pairwise classification between clinical groups, computed using only the inferred DPS.

A preliminary version of this work has been published at the SPIE Medical Imaging 2022 conference [38]. Compared to the conference version, the present paper introduces the following novelties: (1) a supervised instead of a standard unsupervised VAE approach, (2) data split between low-dimensional and high-dimensional modalities, (3) disease trajectory computation in the latent space using principal curves instead of straight lines, (4) additional experiments with multiple synthetic datasets, (5) a comparison with event-based models, and (6) an ablation study.

The manuscript is organized as follows. Section II explains our proposed framework, section III describes the analyzed datasets, section IV details our experiments and corresponding outcomes, and finally section V examines the meaning of our results and highlights the broader implications of our study.

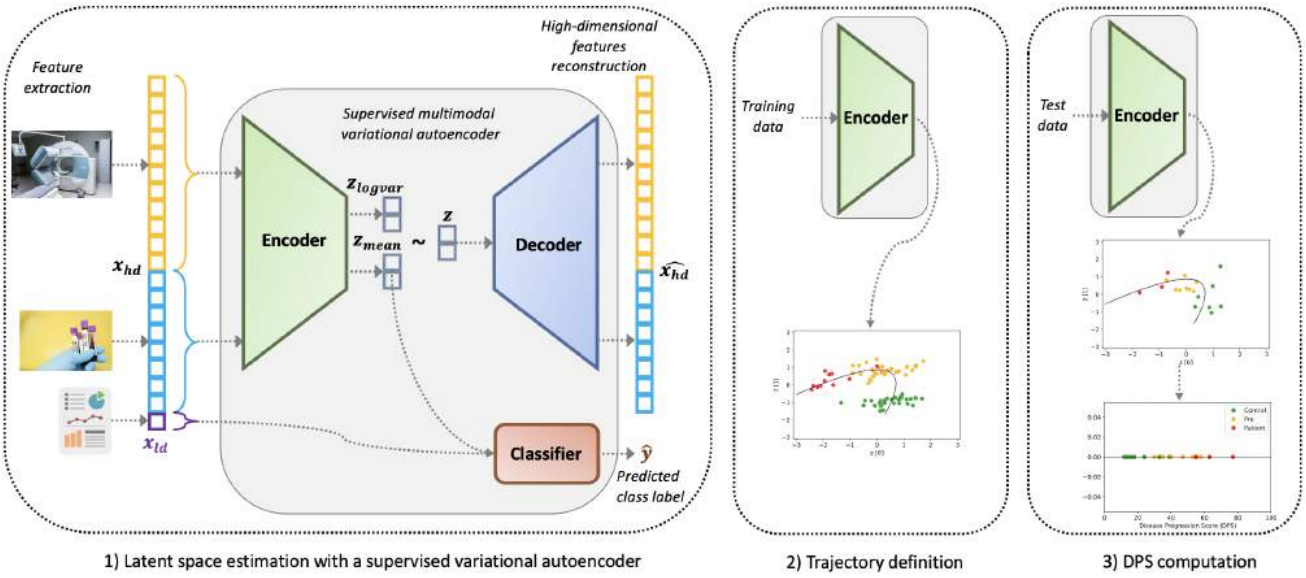


Fig. 1: Illustration of the proposed framework for disease progression scores (DPS) computation. 1) High-dimensional (neuroimaging and microRNAs expression data) and low-dimensional (demographic information) features are extracted; the former are fed to the encoder, the latter are concatenated with latent codes and fed to the classifier. 2) Once the model is trained, all training examples are encoded in the latent space and a principal curve is calculated to define the disease trajectory. 3) Test examples are encoded in the latent space and the latent representations are orthogonally projected onto the previously computed curve; the DPS correspond to their coordinates along the curve.

II. Methodology

We consider a dataset $(\mathcal{X}, \mathcal{Y}) = \{(x_1, y_1), ..., (x_n, y_n)\}$. The i-th subject is characterized by a feature vector $x_i \in \mathbb{R}^m$ and a label $y_i \in \{0, 1, 2\}$ denoting the clinical group (control, presymptomatic, patient). Our aim is to estimate a DPS, denoted as $v_i \in [1, 100]$ (the interval for the scores is arbitrary), where a greater score corresponds to a higher disease severity. To that purpose, we assume that the observations have corresponding latent variables $z_i \in \mathcal{R}^\ell$. We will thus aim to estimate a latent representation and the DPS will be computed from a trajectory in the latent space.

Our framework is composed of three main steps, as illustrated in Fig. 1. First, we propose a supervised multimodal variational autoencoder to estimate the latent space. We leverage the fact that participants belong to different groups to introduce some supervision in order to improve the VAE training. The model aims at simultaneously reconstructing the data and classifying the participants. We propose to split low-dimensional sociodemographic data (denoted $\mathcal{X}_{ld}$, used only for the classification) from high-dimensional multimodal neuroimaging and transcriptomic data (denoted $\mathcal{X}_{hd}$, used both for reconstruction and classification). Second, we build a curve representing disease trajectory in the latent space. Finally, data from new subjects, not included in the training set, are encoded in the latent space and projected onto this trajectory, in order to obtain their DPS.

In this section, we first explain the three main steps of our framework, then we describe implementation details.

A. Supervised multimodal VAE

A variational autoencoder (VAE) [30] is a generative model that learns the training data distribution $p(x)$ using a latent representation model:

$$p(x) = \int p(x|z)p(z)dz,$$

where z is a continuous latent variable living in a lower dimensional space and $p(z)$ is its prior distribution, commonly a Gaussian with zero mean and identity covariance matrix. The solution of the inference problem to describe the latent space is given by deriving the posterior $p(z|x)$. However, there is no closed-form solution for complex real-world datasets. Therefore, VAEs introduce the idea of learning a variational approximation $q_\phi(z|x)$ of the true posterior, in the form of a neural network referred to as the *encoder*. The encoder maps data x to a mean vector z_{mean} and a log-variance vector z_{logvar}, that parametrize a Gaussian distribution from which we obtain the latent representation z. VAEs are also equipped with a generative function $p_\theta(x|z)$, parametrized by a neural network referred to as the *decoder*. The decoder transforms the latent representation z back to the original input space.

During training, the vanilla VAE aims at maximizing the variational lower bound of the marginal log-likelihood, known as the evidence lower bound (ELBO). This is equivalent to minimizing a loss function with two terms:

$$\mathcal{L} = \mathcal{L}_r(x, \hat{x}) + \mathcal{L}_{KL}(q_\phi(z|x), p(z)).$$

The first term is the reconstruction error between the input data x and the reconstructed data $\hat{x}$, typically a mean squared error (MSE). The second term is the Kullback-Leibler divergence

between the approximated posterior $q_\phi(z|x)$ and the prior distribution $p(z)$, acting as a regularization term.

We propose to insert a supervised branch in the vanilla VAE architecture in order to exploit the fact that our samples have different diagnostic labels, even though their DPS is unknown. Denoting y as the true class label and $\hat{y}$ as the predicted class label, we define our training objective as:

$$\mathcal{L} = \alpha_1 \cdot \mathcal{L}_r(x, \hat{x}) + \alpha_2 \cdot \mathcal{L}_{KL}(q_\phi(z|x), p(z)) + \alpha_3 \cdot \mathcal{L}_c(y, \hat{y}),$$

where $\mathcal{L}_r$ and $\mathcal{L}_{KL}$ correspond to the ELBO in vanilla VAEs and $\mathcal{L}_c$ is a cross-entropy term that penalizes the classification error. The hyperparameters α_k control the relative weights between the different loss terms ($\sum_{k=1}^{3} \alpha_k = 1$).

Before training, we split the high-dimensional modalities (miRNA expression and neuroimaging) from the low-dimensional (demographic information). As it will be mentioned later in the datasets description, we consider one low-dimensional feature and $m-1$ high-dimensional features, although the same concepts can be applied to more low-dimensional features. So we use $m-1$ features to feed the encoder and one feature concatenated to the latent code to feed the classifier. Features are rescaled from 0 to 1. Our encoder consists of fully-connected layers of sizes $(m-1) \to 50 \to 2$, meaning our latent space is 2-dimensional. The decoder is implemented with fully-connected layers of sizes $2 \to 50 \to (m-1)$. The nonlinear activation function is the leaky rectified linear unit (ReLU) in all layers except the decoder's last layer which uses a sigmoid function to constrain the output between 0 and 1. The classifier network has one fully connected layer of $3 \to 3$ units, with a softmax function to normalize the output to probabilities over the predicted classes. We use the mean squared error as the reconstruction loss $\mathcal{L}_r$ and the cross-entropy as the classification loss $\mathcal{L}_c$.

B. Trajectory definition

Once the model is trained, the next step is to encode the training data in the latent space. We then compute the straight line passing through the centroids of the control and patient clusters. This straight line could be used in downstream analyses as a rudimentary disease trajectory in the latent space. Instead, we obtain an improved nonlinear trajectory by using this line as initialization for the principal curve algorithm [39]. A principal curve is a smooth one-dimensional curve passing through the *middle* of given data points. The algorithm detailed in [39] finds a nonparametric curve by iteratively minimizing the orthogonal distances to the points until convergence.

C. DPS computation

Once the disease trajectory curve is computed in the latent space, we can encode the test data. The next step is to orthogonally project the latent codes onto the computed curve. The DPS $v_i \in [1, 100]$ for each subject is the coordinate of their projection along this curve, 1 corresponding to the beginning and 100 to the end of the curve. The pseudo-code from model training to DPS computation is shown in Algorithm 1.

Algorithm 1 DPS computation from latent representation

Input: features $\mathcal{X} = \{x_i\}_{i=1}^n \in \mathbb{R}^m$, labels $\mathcal{Y} = \{y_i\}_{i=1}^n \in \{0, 1, 2\}$, training set indices I_{tr} and test set indices I_{te} for one data split into training and test set.
Output: DPS $\{v_i\}_{i=I_{te}}$ of the subjects in the test set.
/* first step: supervised VAE training */
for epoch in [1,250] **do**
  Sample batches $(\mathcal{X}_j, \mathcal{Y}_j)$ from $(\mathcal{X}_{I_{tr}}, \mathcal{Y}_{I_{tr}})$
  for each batch $(\mathcal{X}_j, \mathcal{Y}_j)$ **do**
    $\mathcal{X}_{hd}, \mathcal{X}_{ld} \leftarrow$ split_high_low_dimension$(\mathcal{X}_j)$
    $\mathcal{Z}_{mean}, \mathcal{Z}_{logvar} \leftarrow$ encoder$(\mathcal{X}_{hd})$
    Draw latent codes $\mathcal{Z} \sim \mathcal{N}(\mathcal{Z}_{mean}, e^{\mathcal{Z}_{logvar}})$
    $\hat{\mathcal{Y}}_y \leftarrow$ classifier(concatenate$(\mathcal{X}_{ld}, \mathcal{Z}_{mean})$)
    $\hat{\mathcal{X}}_{hd} \leftarrow$ decoder$(\mathcal{Z})$
    $\mathcal{L}_r \leftarrow$ mean_squared_error$(\mathcal{X}_{hd}, \hat{\mathcal{X}}_{hd})$
    $\mathcal{L}_{KL} \leftarrow$ kl_divergence$(\mathcal{N}(\mathcal{Z}_{mean}, e^{\mathcal{Z}_{logvar}}), \mathcal{N}(0, I))$
    $\mathcal{L}_c \leftarrow$ cross_entropy$(\mathcal{Y}_y, \hat{\mathcal{Y}}_y)$
    $\mathcal{L} \leftarrow \alpha_1 \cdot \mathcal{L}_r + \alpha_2 \cdot \mathcal{L}_{KL} + \alpha_3 \cdot \mathcal{L}_c$
    Compute gradients, update network to minimize $\mathcal{L}$
  end for
end for
/* second step: trajectory definition */
$\mathcal{Z}, \_ \_ \leftarrow encoder(\mathcal{X}_{I_{te}})$
$c_{control} \leftarrow$ mean$(\{\mathcal{Z}_j : y_j == 0\})$
$c_{patient} \leftarrow$ mean$(\{\mathcal{Z}_j : y_j == 2\})$
$pc \leftarrow$ principal_curve$(c_{control}, c_{patient}, degree = 2)$
/* third step: DPS computation */
for i in I_{te} **do**
  $z_{pc} \leftarrow$ projection of z_i into pc
  $v_i \leftarrow$ coordinate of $z_{pc} \in [0, 100]$
end for
return $\{v_i\}_{i=I_{te}}$

D. Implementation details

The hyperparameters of the training objective were set as $\alpha_1 = 0.2$, $\alpha_2 = 0.2$, and $\alpha_3 = 0.6$. The loss function was optimized using Adam [40], with a learning rate of 10^{-3}, batches of 32 observations and 250 epochs.

We carried out the experiments on a computer equipped with a 2.4 GHz Intel Quad-Core Core i5 processor and 16 GB of RAM. Models were implemented in Python 3.8.5 using PyTorch 1.8.1 and Scikit-learn 0.23.2 [41]. For the principal curves computation, we used the implementation provided in the Python package pcurvepy 0.0.10 (`https://pypi.org/project/pcurvepy/`), specifying 2 as the degree of the smoothing spline.

III. Datasets

A. Synthetic datasets

Since ground truth disease progression scores are not available in real-world scenarios, we created synthetic datasets to better evaluate the proposed framework. Multiple datasets were generated, with different noise levels and distinct proportions of features correlating with the DPS.

Let $Y \in \{0, 1, 2\}$ indicate the class labels (respectively control, presymptomatic and patient). We created $n = 111$

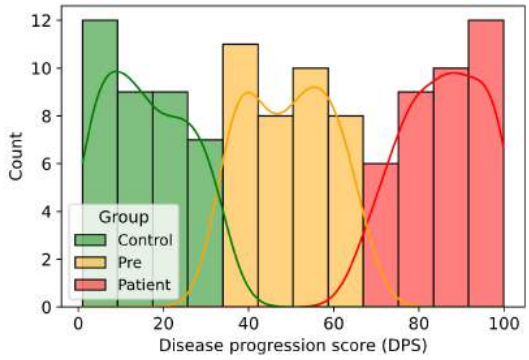


Fig. 2: Synthetic ground truth disease progression scores $\{v_i\}_{i=1}^{n} \in [0, 100)$ for $n = 111$ subjects (37 subjects per group).

synthetic participants (a number close to that of our real dataset) with class labels denoted by y_i $(i = 1, ..., 111)$,

$$y_{i=1,...,37} = 0$$
$$y_{i=38,...,74} = 1$$
$$y_{i=75,...,111} = 2.$$

Next, we modeled the disease progression scores as continuous random variables following uniform distributions. Let $V \in [1, 100)$ represent the DPS values. We defined the conditional distribution of the DPS given the class labels as follows:

$$V|Y = 0 \sim \mathrm{U}[1, 34)$$
$$V|Y = 1 \sim \mathrm{U}[34, 67)$$
$$V|Y = 2 \sim \mathrm{U}[67, 100)$$

We then sampled the corresponding DPS v_i from the conditional distributions defined above. The obtained disease progression scores are displayed in Fig. 2.

Once the synthetic ground truth DPS were created, we generated multiple datasets $\mathcal{D} \in \mathbb{R}^{n \times m}$ containing $n = 111$ participants and $m = 160$ features. In order to simulate two modalities, features were initially sampled from two distributions: half from a negative binomial distribution (typical of miRNA expression data) and half from a normal distribution (representative of various real-world datasets). We denote the columns of $\mathcal{D}$ by $C_1, ..., C_m$. The format of the synthetic datasets is illustrated in Fig. 3.

Each created dataset had a distinct proportion of features correlating with the DPS and different noise levels. The number of features from each modality to positively and negatively correlate with the DPS is denoted as f, and the standard deviation of the added zero-mean Gaussian noise as s. We used $f = \{0, 2, 5, 10, 15, 20, 25, 30, 35, 40\}$ and $s = \{0.001, 0.2, 0.5, 0.8, 1, 5\}$ and thus obtained a total of 60 synthetic datasets. The algorithm describing the set of operations performed for their generation is in Appendix I Algorithm A.1.

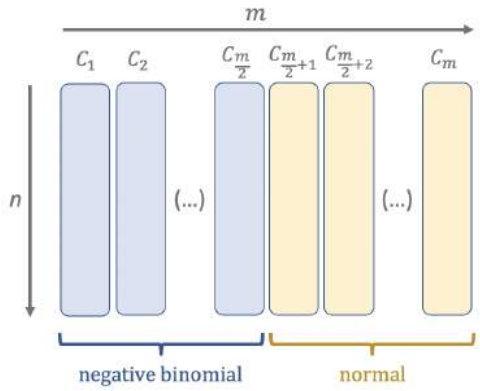


Fig. 3: Format of the synthetic datasets $\mathcal{D} \in \mathbb{R}^{n \times m}$ containing m features from n individuals. Half of the features are initially sampled from a negative binomial distribution and half from a normal distribution.

B. Real dataset

Participants were recruited through the PREV-DEMALS study (https://clinicaltrials.gov, ID NCT02590276), a French multicentric prospective cohort focused on *C9orf72* expansion carriers. Written informed consents were obtained from all participants. The study was approved by the ethics committee (Comité de Protection des Personnes CPP Ile-De-France VI, CPP 68-15 and ID RCB 2015-A00856-43). A detailed description of this cohort and its demographic profile can be found in [7].

We included 110 individuals in our analyses, divided into three groups, according to their clinical status:

- Patient group: 22 symptomatic (15 FTD, 4 FTD/ALS and 3 ALS) carriers of a pathogenic *C9orf72* expansion;
- Presymptomatic group: 45 asymptomatic carriers;
- Control group: 43 asymptomatic non-carriers.

The dataset comprised multimodal data including microRNA (miRNA) sequencing data and neuroimaging data. These two modalities are described below.

1) MicroRNA data: MicroRNAs are a class of small non-coding RNAs that negatively regulate gene expression [42]. MicroRNAs expression in blood plasma has been shown to correlate with the diagnosis and progression of many neurodegenerative diseases [43], including FTD and ALS. All individuals included in this cohort underwent plasma sampling, from which miRNA sequencing was performed. Plasma collection and preparation, miRNA extraction and sequencing, quality control and the computational pipeline to obtain the miRNA counts are detailed in [7]. The initial miRNA dataset contained expression levels for all miRNAs mapped in the human genome (2576 miRNAs). We retained the 589 miRNAs with expression profiles above noise level (minimum total count of 1000 reads and at least 50 reads for one sample). A trimmed mean of M-values [44] implemented in the R package EdgeR [45] was used to normalize the raw counts.

2) Neuroimaging data: Neuroimaging data consisted of grey matter volumes extracted from T1-weighted anatomical magnetic resonance imaging (MRI), including the estimated total

intracranial volume (TIV), 68 cortical regions of interest (ROIs) using the Desikan atlas and 18 subcortical ROIs using the Aseg nomenclature, thus resulting in 87 neuroimaging features. The TIV was used to normalize the volume of each ROI,

$$NV_{ROI} = \frac{TIV_m \times V_{ROI}}{TIV},$$

where V_{ROI} is the original volume of the ROI, NV_{ROI} is the corresponding normalized volume and TIV_m is the average TIV computed across all subjects. The MRI acquisition parameters, quality check and processing pipeline are thoroughly described in [6].

Only 91 subjects (14 patients, 40 presymptomatic carriers and 37 controls) had MRI scans collected. Hence, we divided our dataset into two subsets: 19 subjects that only had miRNA data available, and 91 subjects with multimodal neuroimaging and miRNA data. The former subset was used as a discovery set for miRNA feature selection: we used these 19 individuals to perform differential expression analysis (as described in [7]). The 68 miRNAs with the lowest p-values were selected for all downstream analyses.

Lastly, we also included age as demographic information for all subjects. So the total dimension of each feature vector was $m = 87 + 68 + 1 = 156$.

IV. Experiments and Results

A. Synthetic datasets

We applied our framework to 60 synthetic datasets (described in Section III-A) with different noise levels and distinct number of features correlating with the ground truth DPS. Each synthetic dataset was divided into a training set of 90 subjects (30 per clinical group) and a test set of 21 individuals (7 per group). We trained one model per dataset, using the same hyperparameters as the experiments with the real dataset. After training each model, we computed the DPS for the subjects from the test set. We then calculated the Spearman correlations between the simulated ground-truth scores and the predicted scores. Finally, we evaluated the ROC AUC for each pairwise comparison between the three simulated clinical groups.

Fig. 4 presents the computed trajectories and the DPS obtained when 50% of the features are correlated (25% positively and 25% negatively correlated) with the disease progression, for different noise levels. The correlation matrices illustrate the strength of the relationships between the simulated features, for all investigated noise levels.

The results of the Spearman correlation between the estimated DPS and the ground truth data, as well as the average ROC AUC scores for the three pairwise comparison between groups, are showed in Fig. 5. As expected, we can observe that the DPS is very well estimated for lower noise levels and higher proportion of relevant features, while the performances decrease when the noise level becomes very high and when only few features are correlated with the DPS. Importantly, we observe that the Spearman correlation of the DPS and the ROC AUC have similar behaviors, indicating that the ROC AUC of pairwise comparisons is a reasonable proxy to evaluate the DPS, as will be done with the real dataset.

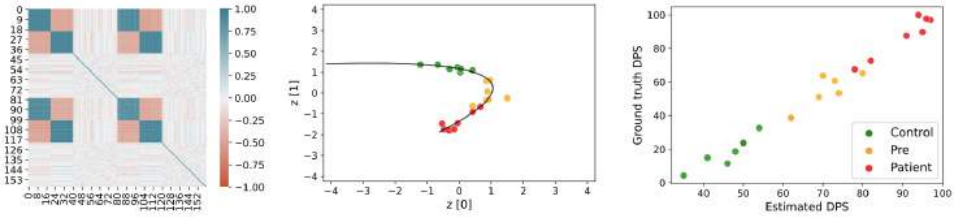
B. Real dataset

Experiments with the real dataset (described in Section III-B) were carried out with a cross-validation of 100 stratified randomized folds. For each fold split, we trained a model using 73 training subjects, and then computed the DPS for the 18 individuals in the test set. Fig. 6 displays an example of the latent space trajectory computed with one representative training data split, the corresponding test set projected in the latent space, and the obtained disease progression scores.

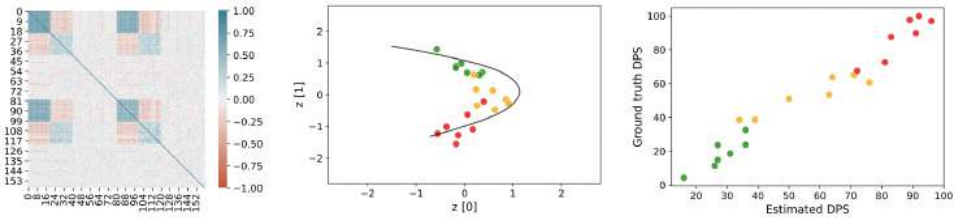
Unlike for the synthetic dataset, there is no ground truth for the DPS in the real dataset. We thus applied a proxy metric to assess model performance: using only the inferred DPS, we did pairwise comparisons between the clinical groups and computed the corresponding areas under the ROC curves. Specifically, we present the following experiments: (1) evaluation of the proposed method, (2) comparison with a state-of-the-art method for modeling disease progression, the discriminative event-based model, (3) ablation study, and (4) variation of hyperparameters.

1) Evaluation of the proposed method: First, we used the DPS computed in each fold to build ROC curves for the three pairwise comparisons between clinical groups. The average ROC curves are shown in Fig. 7. The ROC AUC for the classification of controls and presymptomatic subjects was 0.74 ± 0.13, for controls and patients was 0.98 ± 0.05 and to distinguish presymptomatic carriers and patients was 0.96 ± 0.07. These results reveal that it is harder to differentiate controls from presymptomatic individuals than it is to distinguish between patients and the other two groups. The histogram displayed in Fig. 8 illustrates the disease progression scores computed over all 100 test folds (18 subjects per test fold, corresponding to 1800 DPS). The distribution shapes highlight a clear separation between the patient group and the other groups. The distribution of the DPS for the presymptomatic group is more spread, which was expected as this group is the most heterogeneous. Some presymptomatic subjects are very far from onset and the neurodegenerative process has barely begun, they are thus closer to controls. Other presymptomatic subjects are closer to disease onset and thus their DPS is closer to that of patients.

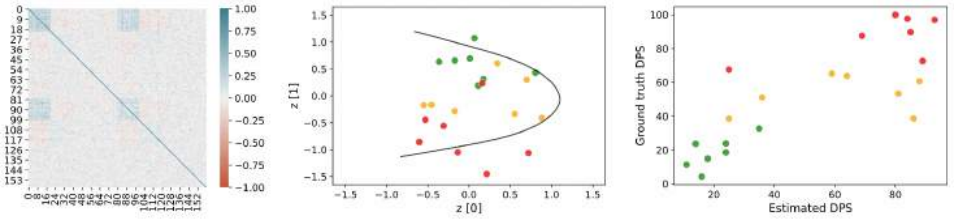
2) Comparison with DEBM: Next, we compared our results to a discriminative event-based model (DEBM) [9], a method that also infers a DPS from cross-sectional data. For that experiment, the same cross-validation strategy of 100 stratified folds was applied. We built the DEBM models and computed the DPS using the Python package pyebm 2.0.3 (`https://pypi.org/project/pyebm/`). Table I displays the corresponding ROC AUC results for each pairwise comparison. We can observe that our model achieves a substantially better classification performance for all pairwise comparisons. Additionally, our approach used less computing time: our framework took 2 seconds per fold for training and DPS computation, while the DEBM algorithm took on average 180 seconds per fold.



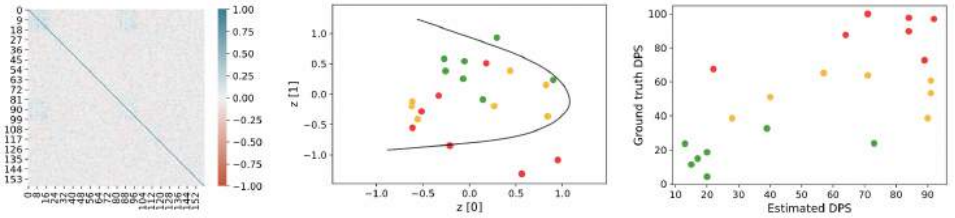
(a) Noise standard deviation = 0.001



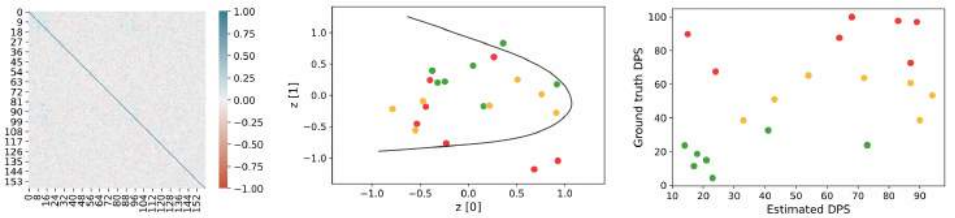
(b) Noise standard deviation = 0.2



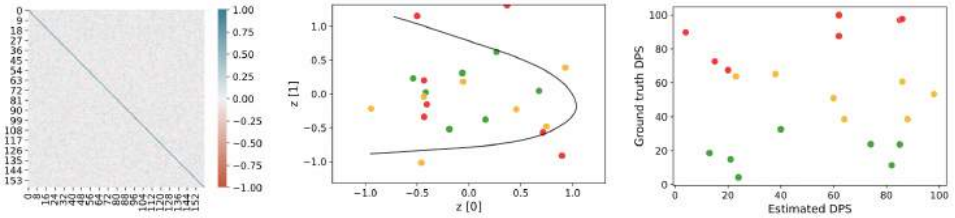
(c) Noise standard deviation = 0.5



(d) Noise standard deviation = 0.8



(e) Noise standard deviation = 1



(f) Noise standard deviation = 5

Fig. 4: Results on synthetic data when 50% of the features are correlated with the disease progression score. The rows indicate different noise levels (zero-mean Gaussian noise with different standard deviations). Each column displays, respectively: (1) correlation matrices showing the strength of the relationships between the simulated features, (2) inferred trajectories and test sets projected in the latent space, and (3) estimated DPS vs. ground truth DPS.

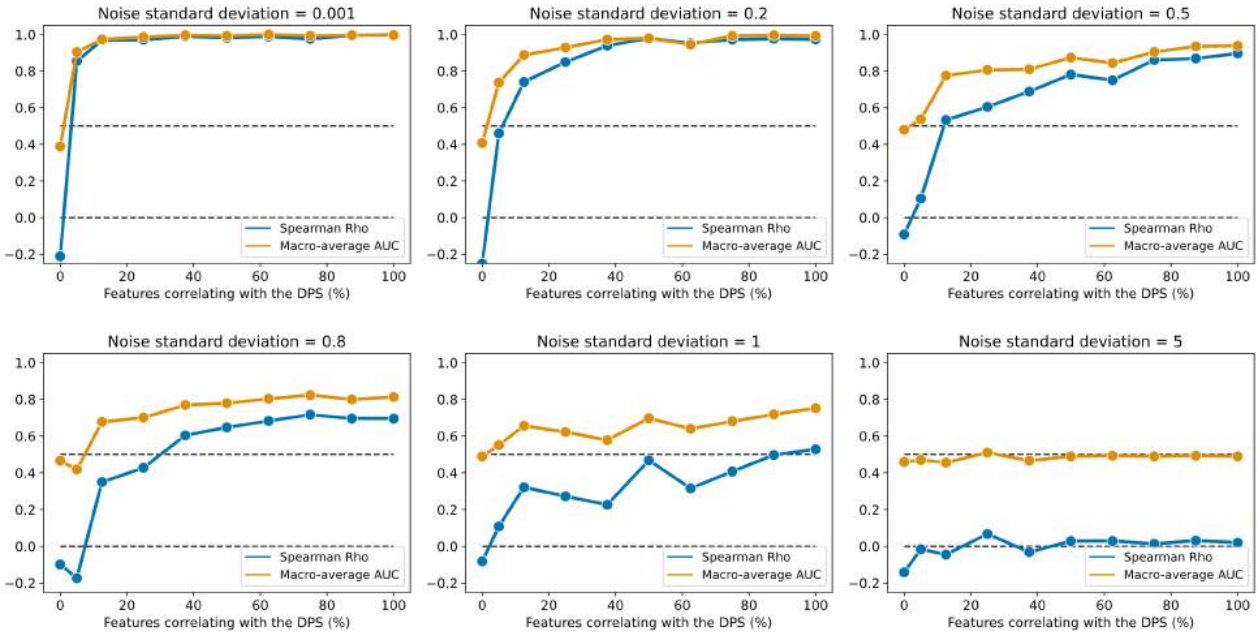
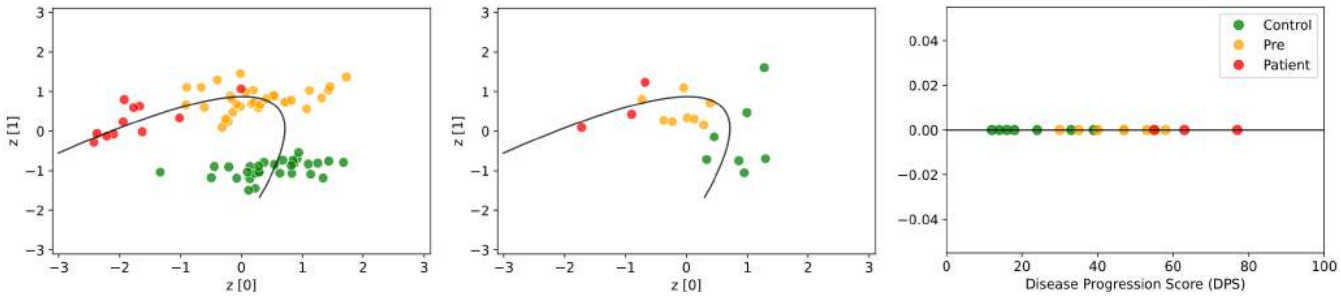


Fig. 5: Results on synthetic data. Macro-average ROC AUC and Spearman correlation between ground truth and estimated DPS, for different noise levels (zero-mean Gaussian with 0.001, 0.2, 0.5, 0.8, 1, and 5 as standard deviation) and several proportions (0% to 100%) of features correlating with the disease progression score. Random chance is denoted by the dashed lines (ROC AUC = 0.5 and Spearman Rho = 0).



(a) Training data projected in the latent space (b) Test data projected in the latent space (c) Computed scores for the test data

Fig. 6: Results on real data. (a) Training data projected in the latent space and the corresponding computed trajectory for one of the 100 fold splits. (b) Test data projected in the latent space, along with the previously computed trajectory. (c) Scores computed after the projection of the latent representation of the test data onto the trajectory.

TABLE I: Results on real data: comparison between our approach and a discriminative event-based model (DEBM) [9]. ROC AUC (mean ± standard deviation) over 100 stratified splits.

Comparison	Our model	DEBM
Control vs. Pre	**0.74 ± 0.13**	0.67 ± 0.14
Control vs. Patient	**0.98 ± 0.05**	0.76 ± 0.17
Pre vs. Patient	**0.96 ± 0.07**	0.65 ± 0.17

3) Ablation study: Afterwards, to investigate the impact of certain components of our framework, we conducted an ablation study. We changed some elements of the proposed approach to obtain three alternative models:

- Linear instead of curved trajectory: rather than computing the trajectory in the latent space using principal curves, we simply used a straight line.
- No supervised branch: we removed the classification component of the loss function, thus performing unsupervised training.
- Joint low-dimensional modality: we concatenated the low-dimensional modality (demographic information) with the high-dimensional modalities (neuroimaging and miRNA expression) in the encoder input, and used only the latent codes as input for the classifier.

For each alternative model, we conducted the same cross-validation strategy of 100 stratified folds, computing the DPS for the test sets and the corresponding areas under the ROC curves. The results, displayed in Table II, show that the proposed model has a better and more stable performance in all comparisons, with the highest average ROC AUC and lowest standard deviation among the splits.

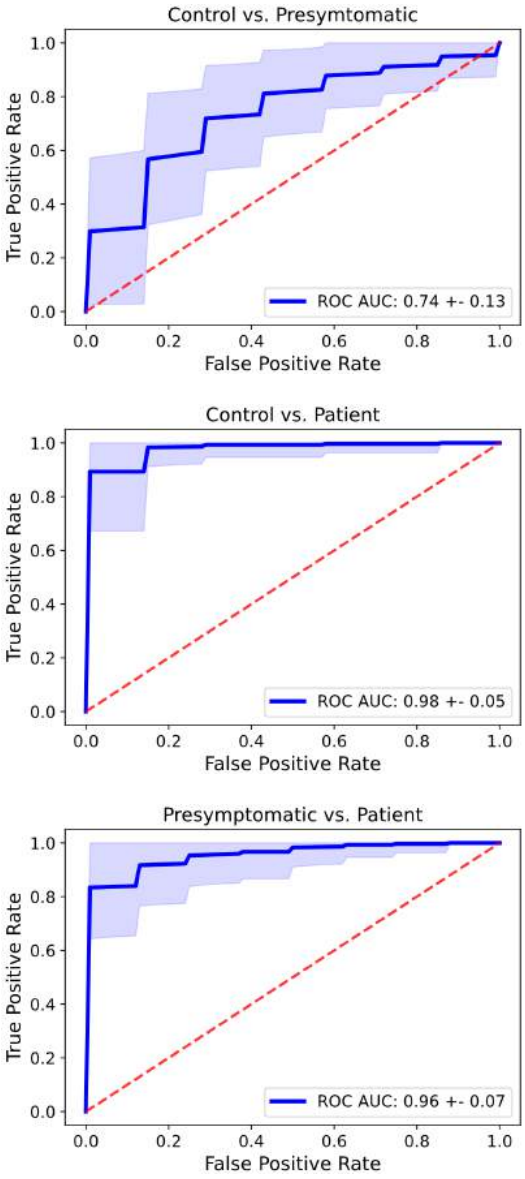


Fig. 7: Results on real data. Average ROC (receiver operating characteristic) curves for each pairwise comparison between clinical groups, over 100 stratified splits. The shaded areas correspond to one standard deviation. The areas under the ROC curves (ROC AUC) are shown as mean ± standard deviation. Random chance is indicated by the dashed line.

4) Variation of hyperparameters: Finally, we checked whether our results were robust to reasonable changes in the hyperparameters. Notably, we tested different numbers of hidden units in the fully-connected layers, and different combinations of the relative weights between the loss terms. These results are summarized in Appendix II Table B.1 and Table B.2. The slightly different but overall similar results demonstrate that our hyperparameter choice is not overfitting the data.

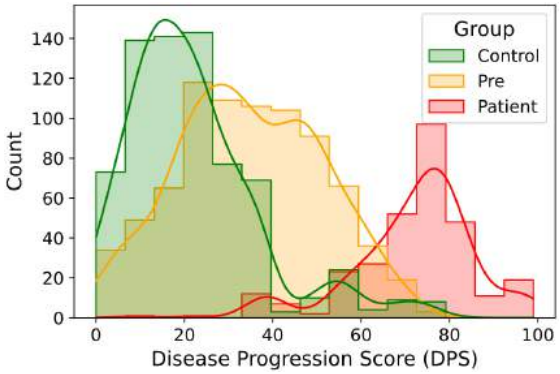


Fig. 8: Results on real data. Histogram of the disease progression scores (DPS) inferred for 18 test subjects over 100 stratified splits. The distribution shapes are approximated with kernel density estimates.

TABLE II: Results on real data: ablation study. ROC AUC results (mean ± standard deviation) for the proposed model and three alternative models from the ablation study, respectively using a linear instead of a curved trajectory, removing the classification branch, and concatenating the low-dimensional modality with the high-dimensional ones.

Comparison	Proposed model	Linear trajectory	No supervision	Joint low-dim.
Control vs. Pre	**0.74 ± 0.13**	0.62 ± 0.15	0.67 ± 0.15	0.72 ± 0.15
Control vs. Patient	**0.98 ± 0.05**	0.93 ± 0.12	0.96 ± 0.06	0.95 ± 0.17
Pre vs. Patient	**0.96 ± 0.07**	0.93 ± 0.11	0.94 ± 0.10	0.91 ± 0.18

V. Discussion

In this paper, we proposed a new approach for estimating disease progression scores from cross-sectional neuroimaging and transcriptomic data that is applicable in small samples, which are typically found in rare diseases. The approach was designed and evaluated on data from *C9orf72*-associated FTD and ALS, but is potentially applicable to other diseases. Results on synthetic data demonstrated the ability of the method to accurately estimate the DPS, and experiments on real data, in the absence of ground truth DPS, showed the separation of different diagnostic classes. The findings of this study validated the usefulness of supervised variational autoencoders to infer disease trajectories from cross-sectional multimodal data, indicating that a single disease progression score may be used to represent progression of neurodegenerative diseases. Remarkably, our results revealed that the DPS may be inferred using only cross-sectional data from a small sample of subjects.

Experiments with a cohort of *C9orf72*-mutation carriers demonstrate that subjects from the same clinical groups (patients, presymptomatic individuals and controls) are clustered together in the latent space (Fig. 6), allowing the inference of a disease trajectory. After training the model, data from new individuals is encoded in the latent space and orthogonally projected onto this trajectory to compute the DPS. Notably, using only the computed DPS, we are able to classify presymptomatic subjects and patients with an average ROC AUC of 0.96 over 100 stratified fold splits (Fig. 7). Of the

three possible pairwise comparisons between clinical groups, this is the most relevant. It illustrates how much the DPS reflects the degree of disease progression in mutation carriers. Unsurprisingly, it is harder to differentiate between controls and presymptomatic individuals, as indicated by the average ROC AUC of 0.74 and displayed in Fig. 8. This stems from the fact that, during earlier disease stages, most biomarker levels are closer to normal ranges, so the presymptomatic class is more heterogeneous.

To the best our of knowledge, event-based models are the only published methods to compute disease progression scores from cross-sectional data, other approaches requiring longitudinal data. The comparison summarized in Table I reveals that our approach resulted in considerably higher ROC AUC than DEBM for all pairwise classifications. This suggests that the proposed approach is more suitable than event-based models for DPS computation with high-dimensional features, such as microRNA data. Indeed, published studies using event-based models explored a substantially lower number of features. For instance, in Alzheimer's disease, EBM experiments were carried our with 13 to 50 [8], [9], [19]–[22] biomarkers. Studies focusing on FTD analyzed 21 [27] or 7 [28] biomarkers, while multiple sclerosis was investigated with 25 [23] or 24 [24] biomarkers. Other conditions such as Parkinson's disease [25], ALS [29] and Huntington's disease [26] were modeled with respectively 42, 19 and 8 biomarkers. Nevertheless, the EBM model presents useful additional features, beyond the computation of DPS. In particular, it can provide a temporal ordering of when the different biomarkers become abnormal, which is useful for understanding disease progression. Moreover, a balance has to be found between the number of features and the number of subjects in each dataset. Indeed, we also had to perform feature selection to decrease the number of microRNAs in our study. It should be noted that this feature selection was unbiased, since it was performed using a completely separate set of participants that was not used in the rest of the study. The proposed framework was able to achieve a good performance with 156 features and less than a hundred subjects, thus demonstrating its potential for dealing with higher dimensional datasets.

An ablation study evaluated the impact of different components of our approach (Table II). We observed that each component positively impacted the framework's performance. First, it can be seen that a curved trajectory better fits the disease pathway in the latent space when compared to a straight line. The use of principal curves has been inspired from their application in a similar task: pseudotime inference for single-cell transcriptomics, as shown in [46]. In that context, pseudotime represents an underlying temporal variable driving a smooth transition between cellular states, and principal curves are used to infer a trajectory in a low-dimensional space. Second, it is clear that the addition of supervision with a classifier branch improves the separation between clinical groups in the latent space. Rather than discrete clusters, our experiments demonstrate that latent representations are placed along a continuous path. Specifically, supervision adds meaning to the relative positions between points in the latent space. Finally, results show the contribution of splitting high and low-dimensional features. When using the low-dimensional features concatenated with the latent codes as inputs to the classifier, the model's performance is enhanced. The same pattern is observed in [37], although in a totally different context (failure detection in robotics). Concretely, a low-dimensional feature can directly contribute to the classifier, without the need for encoding.

Regarding the experiments with simulated datasets, it is crucial to highlight the relationship of the average ROC AUC with the Spearman correlation between ground truth and estimated DPS (Fig. 5). The simulation supports that the higher the ROC AUC, the more accurate the predicted DPS. Therefore, for real-world scenarios without ground truth DPS, our choice of the ROC AUC as proxy metric is corroborated. Furthermore, evidence was found that the models do not overfit the data, since it is clear that larger noise levels lead to poorer results, eventually equivalent to random chance. The effect of noise is further illustrated in Fig. 4. We observe that lower noise levels induce more evident clusters and more meaningful trajectories in the latent space. Consequently, the estimated DPS are closer to the ground truth. These simulations also confirm one intuition behind our model: the more features correlate with disease progression, the closer the estimated DPS are to the ground truth.

Our study has the following limitations. First, there is no ground truth for the progression scores in real datasets. Although the experiments with synthetic data showed that the ROC AUC is an adequate proxy metric, long-term follow-up of patients will be necessary to assess the accuracy of the computed DPS. For instance, we need follow-up data to confirm the hypothesis that a higher DPS implies an earlier disease onset for a presymptomatic subject. Another limitation was the lack of a replication cohort. This will be necessary to further support the clinical relevance of our findings. Future work will concentrate on the integration of more data sources, such as positron emission tomography (PET) scans and neurofilament light chain (NfL) levels in blood.

In conclusion, we proposed a new approach to measure disease progression from multimodal imaging and microRNA data in rare neurodegenerative disorders using only cross-sectional data. Even though we focused on *C9orf72*-associated FTD and ALS, our framework is generic. It has the potential to be useful for a variety of other diseases, enabling the evaluation of novel treatments even when only cross-sectional data from small cohorts are available.

Acknowledgment

The PREV-DEMALS study group includes: Eve Benchetrit (Hôpital de la Salpêtrière, Paris), Anne Bertrand (Hôpital de la Salpêtrière, Paris), Anne Bissery (Hôpital de la Salpêtrière, Paris), Marie-Paule Boncoeur (CHU Dypuytren, Limoges), Stéphanie Bombois (CHU Roger Salengro, Lille), Agnès Camuzat (ICM, Paris), Mathieu Chastan (CHU Charles Nicolle, Rouen), Yaohua Chen (CHU Roger Salengro, Lille), Marie Chupin (ICM, Paris), Olivier Colliot (ICM, Paris), Philippe Couratier (CHU Dypuytren, Limoges), Xavier Delbeuck (CHU Roger Salengro, Lille), Vincent Deramecourt (CHU Roger

Salengro, Lille), Christine Delmaire (CHU Roger Salengro, Lille), Emmanuel Gerardin (CHU Charles Nicolle, Rouen), Claude Hossein-Foucher (CHU Roger Salengro, Lille), Bruno Dubois (Hôpital de la Salpêtrière, Paris), Marie-Odile Habert (Hôpital de la Salpêtrière, Paris), Didier Hannequin (CHU Charles Nicolle, Rouen), Géraldine Lautrete (CHU Dypuytren, Limoges), Thibaud Lebouvier (CHU Roger Salengro, Lille), Isabelle Le Ber (Hôpital de la Salpêtrière, Paris), Benjamin Le Toullec (ICM, Paris), Richard Levy (Hôpital de la Salpêtrière, Paris), Olivier Martinaud (CHU Charles Nicolle, Rouen), Kelly Martineau (ICM, Paris), Marie-Anne Mackowiak (CHU Roger Salengro, Lille), Jacques Monteil (CHU Dypuytren, Limoges), Florence Pasquier (CHU Roger Salengro, Lille), Gregory Petyt (CHU Roger Salengro, Lille), Pierre-François Pradat (Hôpital de la Salpêtrière, Paris), Assi-Hervé Oya (Hôpital de la Salpêtrière, Paris), Armelle Rametti-Lacroux (Hôpital de la Salpêtrière, Paris), Daisy Rinaldi (Hôpital de la Salpêtrière, Paris), Adeline Rollin-Sillaire (CHU Roger Salengro, Lille), François Salachas (Hôpital de la Salpêtrière, Paris), Sabrina Sayah (Hôpital de la Salpêtrière, Paris), David Wallon (CHU Charles Nicolle, Rouen).

We thank Vincent Anquetil for plasma collection and preparation, Justine Guégan and Thomas Gareau for microRNA raw reads to counts pipeline handling, Yannick Marie and Delphine Bouteiller for library preparation and sequencing, and Philippe Martin-Hardy and Ludmila Jornea for technical assistance.

APPENDIX I
ALGORITHM FOR SYNTHETIC DATASETS GENERATION

Algorithm A.1 Synthetic datasets generation

Input: number of subjects n, number of features m, disease progression scores values v_i (i=1,...,n).
Output: set L containing the datasets $\mathcal{D} \in \mathbb{R}^{n \times m}$
$L = \{\}$
for f in $\{0, 2, 5, 10, 15, 20, 25, 30, 35, 40\}$ **do**
  for s in $\{0.001, 0.2, 0.5, 0.8, 1, 5\}$ **do**
    $C_{1,...,\frac{m}{2}} \leftarrow$ NB($r = 3000$, $p = 0.75$, $size = (n, \frac{m}{2})$)
    $C_{\frac{m}{2}+1,...,m} \leftarrow \mathcal{N}$($\mu$=1000, σ=200, $size$=$(n, \frac{m}{2})$)
    /* f features from each modality positively correlate with disease progression */
    for j in $\{1, ..., f\} \cup \{\frac{m}{2} + 1, ..., \frac{m}{2} + 1 + f\}$ **do**
      $C_j \leftarrow v \odot C_j$
    end for
    /* the next f features are negatively correlated */
    for j in $\{f, ..., 2f\} \cup \{\frac{m}{2} + 1 + f, ..., \frac{m}{2} + 1 + 2f\}$ **do**
      $C_j \leftarrow \frac{1}{v} \odot C_j$
    end for
    /* normalize and add zero-mean Gaussian noise */
    for j in {1,...,m} **do**
      $C_j \leftarrow \frac{C_j - min(C_j)}{max(C_j) - min(C_j)}$
    end for
    $\mathcal{D} \leftarrow \mathcal{D} + \mathcal{N}$($\mu$=0, σ=s, $size$=(n, m))
    $L = L \cup \{D\}$
  end for
end for
return L

APPENDIX II
RESULTS VARYING HYPERPARAMETERS

TABLE B.1: Results on real data. ROC AUC results (mean ± standard deviation) over 100 stratified splits when changing the number of units of the hidden layers. Original results, with 50 units, are shown in bold.

Hidden units	**50**	100	80	25
Control vs. Pre	**0.74 ± 0.13**	0.73 ± 0.13	0.71 ± 0.12	0.71 ± 0.13
Control vs. Patient	**0.98 ± 0.05**	0.98 ± 0.04	0.97 ± 0.05	0.98 ± 0.05
Pre vs. Patient	**0.96 ± 0.07**	0.96 ± 0.06	0.96 ± 0.06	0.96 ± 0.06

TABLE B.2: Results on real data. ROC AUC results (mean ± standard deviation) over 100 stratified splits when changing the weights of the loss function terms. Original results, with α_1=0.2, α_2=0.2, α_3=0.6, are shown in bold.

Weights α_k	**0.2, 0.2, 0.6**	0.1, 0.1, 0.8	0.1, 0.2, 0.7	0.3, 0.2, 0.5
Control vs. Pre	**0.74 ± 0.13**	0.72 ± 0.12	0.73 ± 0.12	0.72 ± 0.14
Control vs. Patient	**0.98 ± 0.05**	0.97 ± 0.08	0.97 ± 0.06	0.98 ± 0.05
Pre vs. Patient	**0.96 ± 0.07**	0.94 ± 0.10	0.95 ± 0.09	0.96 ± 0.07

REFERENCES

[1] K. Rascovsky, J. R. Hodges, D. Knopman, M. F. Mendez, J. H. Kramer, J. Neuhaus, J. C. van Swieten, *et al.*, "Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia," *Brain*, vol. 134, pp. 2456–2477, Sept. 2011.
[2] P. Pasinelli and R. H. Brown, "Molecular biology of amyotrophic lateral sclerosis: insights from genetics," *Nature Reviews Neuroscience*, vol. 7, pp. 710–723, Sept. 2006.
[3] M. DeJesus-Hernandez, I. R. Mackenzie, B. F. Boeve, A. L. Boxer, M. Baker, N. J. Rutherford, *et al.*, "Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS," *Neuron*, vol. 72, pp. 245–256, Oct. 2011.
[4] A. E. Renton, E. Majounie, A. Waite, J. Simón-Sánchez, S. Rollinson, J. R. Gibbs, *et al.*, "A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD," *Neuron*, vol. 72, pp. 257–268, Oct. 2011.
[5] J. D. Rohrer, J. M. Nicholas, D. M. Cash, J. van Swieten, E. Dopper, L. Jiskoot, R. van Minkelen, S. A. Rombouts, M. J. Cardoso, S. Clegg, M. Espak, S. Mead, D. L. Thomas, E. De Vita, M. Masellis, *et al.*, "Presymptomatic cognitive and neuroanatomical changes in genetic frontotemporal dementia in the Genetic Frontotemporal dementia Initiative (GENFI) study: a cross-sectional analysis," *The Lancet. Neurology*, vol. 14, pp. 253–262, Mar. 2015.
[6] A. Bertrand, J. Wen, D. Rinaldi, M. Houot, S. Sayah, A. Camuzat, *et al.*, "Early Cognitive, Structural, and Microstructural Changes in Presymptomatic C9orf72 Carriers Younger Than 40 Years," *JAMA neurology*, vol. 75, no. 2, pp. 236–245, 2018.
[7] V. Kmetzsch, V. Anquetil, D. Saracino, D. Rinaldi, A. Camuzat, T. Gareau, L. Jornea, S. Forlani, P. Couratier, D. Wallon, F. Pasquier, N. Robil, P. d. l. Grange, I. Moszer, I. L. Ber, O. Colliot, and E. Becker, "Plasma microRNA signature in presymptomatic and symptomatic subjects with C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 92, pp. 485–493, May 2021.
[8] H. M. Fonteijn, M. Modat, M. J. Clarkson, J. Barnes, M. Lehmann, N. Z. Hobbs, R. I. Scahill, S. J. Tabrizi, S. Ourselin, N. C. Fox, and D. C. Alexander, "An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease," *NeuroImage*, vol. 60, pp. 1880–1889, Apr. 2012.
[9] V. Venkatraghavan, E. E. Bron, W. J. Niessen, and S. Klein, "Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling," *NeuroImage*, vol. 186, pp. 518–532, Feb. 2019.
[10] B. M. Jedynak, A. Lang, B. Liu, E. Katz, Y. Zhang, B. T. Wyman, D. Raunig, C. P. Jedynak, B. Caffo, and J. L. Prince, "A computational neurodegenerative disease progression score: Method and results with the Alzheimer's disease neuroimaging initiative cohort," *NeuroImage*, vol. 63, pp. 1478–1486, Nov. 2012.

[11] M. Mehdipour Ghazi, M. Nielsen, A. Pai, M. Modat, M. Jorge Cardoso, S. Ourselin, and L. Sørensen, "Robust parametric modeling of Alzheimer's disease progression," *NeuroImage*, vol. 225, p. 117460, Jan. 2021.

[12] J.-B. Schiratti, S. Allassonnière, O. Colliot, and S. Durrleman, "A Bayesian Mixed-Effects Model to Learn Trajectories of Changes from Repeated Manifold-Valued Observations," *Journal of Machine Learning Research*, vol. 18, no. 133, pp. 1–33, 2017.

[13] I. Koval, A. Bône, M. Louis, T. Lartigue, S. Bottani, A. Marcoux, J. Samper-González, N. Burgos, B. Charlier, A. Bertrand, S. Epelbaum, O. Colliot, S. Allassonnière, and S. Durrleman, "AD Course Map charts Alzheimer's disease progression," *Scientific Reports*, vol. 11, p. 8020, Apr. 2021.

[14] R. V. Marinescu, A. Eshaghi, M. Lorenzi, A. L. Young, N. P. Oxtoby, S. Garbarino, S. J. Crutch, D. C. Alexander, and Alzheimer's Disease Neuroimaging Initiative, "DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders," *NeuroImage*, vol. 192, pp. 166–177, May 2019.

[15] M. Lorenzi, M. Filippone, D. C. Alexander, and S. Ourselin, "Disease Progression Modeling and Prediction through Random Effect Gaussian Processes and Time Transformation," *NeuroImage*, vol. 190, pp. 56–68, Apr. 2019. arXiv: 1701.01668.

[16] S. Garbarino, M. Lorenzi, N. P. Oxtoby, E. J. Vinke, R. V. Marinescu, A. Eshaghi, M. A. Ikram, W. J. Niessen, O. Ciccarelli, F. Barkhof, J. M. Schott, M. W. Vernooij, and D. C. Alexander, "Differences in topological progression profile among neurodegenerative diseases from imaging data," *eLife*, vol. 8, p. e49298, 2019.

[17] L. M. Aksman, M. A. Scelsi, A. F. Marquand, D. C. Alexander, S. Ourselin, A. Altmann, and for ADNI, "Modeling longitudinal imaging biomarkers with parametric Bayesian multi-task learning," *Human Brain Mapping*, vol. 40, pp. 3982–4000, Sept. 2019.

[18] M. Mehdipour Ghazi, M. Nielsen, A. Pai, M. J. Cardoso, M. Modat, S. Ourselin, and L. Sørensen, "Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling," *Medical Image Analysis*, vol. 53, pp. 39–46, Apr. 2019.

[19] A. L. Young, N. P. Oxtoby, P. Daga, D. M. Cash, N. C. Fox, S. Ourselin, J. M. Schott, D. C. Alexander, and on behalf of the Alzheimer's Disease Neuroimaging Initiative, "A data-driven model of biomarker changes in sporadic Alzheimer's disease," *Brain*, vol. 137, pp. 2564–2577, Sept. 2014.

[20] N. P. Oxtoby, A. L. Young, D. M. Cash, T. L. S. Benzinger, A. M. Fagan, J. C. Morris, R. J. Bateman, N. C. Fox, J. M. Schott, and D. C. Alexander, "Data-driven models of dominantly-inherited Alzheimer's disease progression," *Brain*, vol. 141, pp. 1529–1544, May 2018.

[21] N. C. Firth, S. Primativo, E. Brotherhood, A. L. Young, K. X. Yong, S. J. Crutch, D. C. Alexander, and N. P. Oxtoby, "Sequences of cognitive decline in typical Alzheimer's disease and posterior cortical atrophy estimated using a novel event-based model of disease progression," *Alzheimer's & Dementia*, vol. 16, pp. 965–973, July 2020.

[22] D. Archetti, S. Ingala, V. Venkatraghavan, V. Wottschel, A. L. Young, M. Bellio, E. E. Bron, S. Klein, F. Barkhof, D. C. Alexander, N. P. Oxtoby, G. B. Frisoni, and A. Redolfi, "Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in Alzheimer's disease," *NeuroImage : Clinical*, vol. 24, p. 101954, July 2019.

[23] I. Dekker, M. M. Schoonheim, V. Venkatraghavan, A. J. Eijlers, I. Brouwer, E. E. Bron, S. Klein, M. P. Wattjes, A. M. Wink, J. J. Geurts, B. M. Uitdehaag, N. P. Oxtoby, D. C. Alexander, H. Vrenken, J. Killestein, F. Barkhof, and V. Wottschel, "The sequence of structural, functional and cognitive changes in multiple sclerosis," *NeuroImage : Clinical*, vol. 29, p. 102550, Dec. 2020.

[24] A. Eshaghi, R. V. Marinescu, A. L. Young, N. C. Firth, F. Prados, M. Jorge Cardoso, C. Tur, F. De Angelis, N. Cawley, W. J. Brownlee, N. De Stefano, M. Laura Stromillo, M. Battaglini, S. Ruggieri, C. Gasperini, *et al.*, "Progression of regional grey matter atrophy in multiple sclerosis," *Brain*, vol. 141, pp. 1665–1677, June 2018.

[25] N. P. Oxtoby, L.-A. Leyland, L. M. Aksman, G. E. C. Thomas, E. L. Bunting, P. A. Wijeratne, A. L. Young, A. Zarkali, M. M. X. Tan, F. D. Bremner, P. A. Keane, H. R. Morris, A. E. Schrag, D. C. Alexander, and R. S. Weil, "Sequence of clinical and neurodegeneration events in Parkinson's disease progression," *Brain*, vol. 144, pp. 975–988, Feb. 2021.

[26] P. A. Wijeratne, E. B. Johnson, S. Gregory, N. Georgiou-Karistianis, J. S. Paulsen, R. I. Scahill, S. J. Tabrizi, and D. C. Alexander, "A Multi-Study Model-Based Evaluation of the Sequence of Imaging and Clinical Biomarker Changes in Huntington's Disease," *Frontiers in Big Data*, vol. 4, p. 662200, Aug. 2021.

[27] J. L. Panman, V. Venkatraghavan, E. L. v. d. Ende, R. M. E. Steketee, L. C. Jiskoot, J. M. Poos, E. G. P. Dopper, *et al.*, "Modelling the cascade of biomarker changes in GRN-related frontotemporal dementia," *Journal of Neurology, Neurosurgery & Psychiatry*, Jan. 2021.

[28] E. L. van der Ende, E. E. Bron, J. M. Poos, L. C. Jiskoot, J. L. Panman, J. M. Papma, *et al.*, "A data-driven disease progression model of fluid biomarkers in genetic frontotemporal dementia," *Brain: A Journal of Neurology*, p. awab382, Oct. 2021.

[29] M. C. Gabel, R. J. Broad, A. L. Young, S. Abrahams, M. E. Bastin, R. A. L. Menke, A. Al-Chalabi, L. H. Goldstein, S. Tsermentseli, D. C. Alexander, M. R. Turner, P. N. Leigh, and M. Cercignani, "Evolution of white matter damage in amyotrophic lateral sclerosis," *Annals of Clinical and Translational Neurology*, vol. 7, pp. 722–732, May 2020.

[30] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114 [cs, stat]*, May 2014. arXiv: 1312.6114.

[31] L. Antelmi, N. Ayache, P. Robert, and M. Lorenzi, "Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data," in *International Conference on Machine Learning*, pp. 302–311, PMLR, May 2019. ISSN: 2640-3498.

[32] Y. Xu, X. Liu, L. Pan, X. Mao, H. Liang, G. Wang, and T. Chen, "Explainable Dynamic Multimodal Variational Autoencoder for the Prediction of Patients with Suspected Central Precocious Puberty," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2021.

[33] J. Cheng, M. Gao, J. Liu, H. Yue, H. Kuang, J. Liu, and J. Wang, "Multimodal Disentangled Variational Autoencoder With Game Theoretic Interpretability for Glioma Grading," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, pp. 673–684, Feb. 2022.

[34] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-Supervised Learning with Deep Generative Models," *arXiv:1406.5298 [cs, stat]*, Oct. 2014. arXiv: 1406.5298.

[35] S. N, B. Paige, J.-W. van de Meent, A. Desmaison, N. Goodman, P. Kohli, F. Wood, and P. Torr, "Learning Disentangled Representations with Semi-Supervised Deep Generative Models," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[36] F. Berkhahn, R. Keys, W. Ouertani, N. Shetty, and D. Geißler, "Augmenting Variational Autoencoders with Sparse Labels: A Unified Framework for Unsupervised, Semi-(un)supervised, and Supervised Learning," *arXiv:1908.03015 [cs, stat]*, Nov. 2019. arXiv: 1908.03015.

[37] T. Ji, S. T. Vuppala, G. Chowdhary, and K. Driggs-Campbell, "Multi-Modal Anomaly Detection for Unstructured and Uncertain Environments," in *Proceedings of the 2020 Conference on Robot Learning*, pp. 1443–1455, PMLR, Oct. 2021. ISSN: 2640-3498.

[38] V. Kmetzsch, E. Becker, D. Saracino, V. Anquetil, D. Rinaldi, A. Camuzat, T. Gareau, I. Le Ber, and O. Colliot, "A multimodal variational autoencoder for estimating progression scores from imaging and microRNA data in rare neurodegenerative diseases," in *SPIE Medical Imaging 2022*, (San Diego, California, United States), Feb. 2022. In Press.

[39] T. Hastie and W. Stuetzle, "Principal Curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989.

[40] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017. arXiv: 1412.6980.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[42] E. Huntzinger and E. Izaurralde, "Gene silencing by microRNAs: contributions of translational repression and mRNA decay," *Nature Reviews Genetics*, vol. 12, pp. 99–110, Feb. 2011.

[43] M. Grasso, P. Piscopo, A. Confaloni, and M. A. Denti, "Circulating miRNAs as biomarkers for neurodegenerative disorders," *Molecules (Basel, Switzerland)*, vol. 19, pp. 6891–6910, May 2014.

[44] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biology*, vol. 11, p. R25, Mar. 2010.

[45] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics (Oxford, England)*, vol. 26, pp. 139–140, Jan. 2010.

[46] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit, "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics," *BMC Genomics*, vol. 19, p. 477, June 2018.

C.3 Contributions jointes en support au chapitre 4

[29] Moreews F, Simon H, Siegel A, Gondret F, **Becker E.**
PAX2GRAPHML : a python library for large-scale regulation network analysis using BIOPAX.
Bioinformatics, 2021.

[30] Melkonian M, Juigné C, Dameron O, Rabut G* and **Becker E***.
Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases.
Bioinformatics, 2022.

[33] Juigné C, Dameron O, Moreews F, Gondret F and **Becker E**.
Detection and correction of non-conformities and redundancies in complexes of molecules in BioPAX.
Actes JOBIM 2022, Rennes, 5-8 juillet 2022

Bioinformatics, 37(24), 2021, 4889–4891
doi: 10.1093/bioinformatics/btab441
Advance Access Publication Date: 15 June 2021
Applications Note

Systems biology

PAX2GRAPHML: a python library for large-scale regulation network analysis using BioPAX

François Moreews [1,2,*], Hugo Simon[1], Anne Siegel[1], Florence Gondret[2] and Emmanuelle Becker[1]

[1]Univ Rennes, Inria, CNRS, IRISA, Rennes, France and [2]Pegase, Inrae, Institut Agro, 35590 Saint-Gilles, France

*To whom correspondence should be addressed.

Associate Editor: Lenore Cowen

Received on December 16, 2020; revised on May 18, 2021; editorial decision on June 9, 2021; accepted on June 17, 2021

Abstract

Summary: PAX2GRAPHML is an open-source Python library that allows to easily manipulate BioPAX source files as regulated reaction graphs described in.graphml format. The concept of regulated reactions, which allows connecting regulatory, signaling and metabolic levels, has been used. Biochemical reactions and regulatory interactions are homogeneously described by regulated reactions involving substrates, products, activators and inhibitors as elements. PAX2GRAPHML is highly flexible and allows generating graphs of regulated reactions from a single BioPAX source or by combining and filtering BioPAX sources. Supported by the graph exchange format .graphml, the large-scale graphs produced from one or more data sources can be further analyzed with PAX2GRAPHML or standard Python and R graph libraries.

Availability and implementation: https://pax2graphml.genouest.org.

Contact: francois.moreews@irisa.fr

1 Introduction

BioPAX is a standard format encoding biological processes like gene regulation, metabolic pathways or signaling events, that facilitates the inter-operability between data sources and network analysis tools. However, this rich knowledge-oriented data format that finely captures the complexity of biological networks cannot be easily handled without appropriated tools. Software have been recently proposed to design, visualize (Babur *et al.*, 2010; Shannon *et al.*, 2003), parse (Turei *et al.*, 2016), validate (Rodchenkov *et al.*, 2013), query (Babur *et al.*, 2014) and analyze BioPAX files. However, an important missing feature to analyze BioPAX data sources is the ability to interpret BioPAX files into graph structures including the role of physical entities as substrate, product or regulator in the reactions.

An accurate format for representing the variety and complexity of the biological reactions is the concept of regulated reactions connecting regulatory, signaling and metabolic levels (Blavy *et al.*, 2014). In this conceptual framework, both biochemical reactions and regulatory interactions are described homogeneously as regulated reactions involving substrates, products, activators, inhibitors and modulators as key elements. In the reaction graph generated from regulated reactions, the molecules and the reactions are represented as typed nodes, as shown in Figure 1.

Thus, we propose to extend the BioPAX toolbox with a Python library able to interpret BioPAX files as graphs of regulated reactions. With PAX2GRAPHML, the graphs are represented in the .graphml format, allowing the manipulation of nodes and edges properties. The PAX2GRAPHML tool also enables extracting subgraphs, by filtering the original files according to specific properties of the nodes (genes or proteins) or by merging different graphs. It also implements basic methods to explore the graphs. Thanks to the .graphml exchange format support, generated graphs can be further analyzed with already existing graph libraries in Python or R.

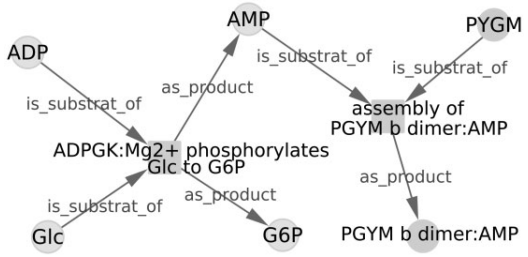


Fig. 1. Example of reaction graph manipulated by PAX2GRAPHML showing reactions and entities as nodes

2 Format and package description

PAX2GRAPHML is able to process all BioPAX files to generate regulated reaction graphs, which can be further interpreted into positive and negative oriented influences. It is available on pypi and as a docker image.

Table 1. BioPAX files transformation of datasources available in PC into regulated reaction graphs with PAX2GRAPHML

Data sources	Nodes	Reaction nodes	Entity nodes	Edges	Substrate of	Product of	Activator of	Inhibitor of
PC* all sources	175 262	85 750	89 512	639 945	84 496	95 743	52 009	407 697
CTD	44 639	19 814	24 825	98 993	18 538	19 073	35 077	26 305
HumanCyc	5733	1778	3955	11 890	3875	4455	3560	0
INOH	4315	2188	2127	7409	4247	3162	0	0
Intact complex	2187	563	1624	2869	2306	563	0	0
KEGG**	3133	1560	1573	7041	3488	3553	0	0
Mirtarbase	32 727	15 064	17 663	395 703	0	15 064	0	380 639
Panther	3766	1662	2104	5263	2914	2165	142	42
PID	9403	4495	4908	14 827	6613	4544	3233	437
Reactome	31 718	11 404	20 314	46 541	27 210	15 358	3717	256
Reconx	6956	2821	4135	20 485	7722	7689	5074	0
PC* all sources except CTD and Mirtarbase	119 285	55 184	64 101	167 515	83 782	66 056	16 939	7 38
PID and HumanCyc	12 561	4518	8043	16 085	3575	5859	6221	430
PID and HumanCyc and KEGG	15 564	6079	9485	19 651	5332	7652	6237	430
PID and HumanCyc and KEGG and Reactome	38 585	10 398	28 187	43 899	16 087	17 371	9756	685

Note: Single datasources transformations were performed with the sub-package pax_import. Combination of several datasources was performed by filtering the PC* all sources graphml file with the sub-package extract. Nodes are either reactions or entities (proteins, small molecules, etc.). The numbers of regulated reactions computed by PAX2GRAPHML, together with the number of substrates, products, activators, inhibitors, are indicated. All these graphs can be directly downloaded from the PAX2GRAPHML website.

*'PC' version 12, September 2019.

**KEGG, July 2011 (only human, hsa* files).

In PAX2GRAPHML, PaxTools (Babur *et al.*, 2014) is used internally to extract sub-classes of patterns and further interpret them as regulated reactions. These extracted patterns form the building elements of a *regulated reaction graph* (Blavy *et al.*, 2014). Each regulated reaction graph pattern is centered on a reaction node linked to one or several substrate nodes and product nodes. The reaction node can also be linked to modulator nodes (activators or inhibitors). Substrates and modulators are inputs of the reaction node, whereas products are outputs of the reaction node. All nodes (reaction, substrate, product or modulator) are associated with their own metadata in the graph.

PAX2GRAPHML is composed of four sub-packages. (i) The sub-package pax_import is dedicated to global or parametrized import of BioPAX files from Pathway Commons (PC) to be further interpreted as regulated reaction graph. (ii) The sub-package properties allow to manipulate nodes and edges properties of the generated graphs. All aliases contained in BioPAX have been incorporated in the .graphml format as node properties to represent genes, protein and compounds. Additional annotations can also be directly imported from specific files. (iii) The sub-package extract allows modifying either the generated reaction graph or the influence graph, including sub-graphs selection or graphs merging. (iv) The sub-package graph_explore includes IO functions and analysis of the generated graphs. It also includes classical graph metrics (degree, betweenness, closeness, connected components) as preliminary steps. More sophisticated analyses can be further performed with graph-tool or other advanced libraries (Csardi and Nepusz, 2006).

The PAX2GRAPHML website provides a complete documentation and the pre-processed database resources. Regulated reaction graphs and influence graphs produced from 16 data sources of PC can be downloaded as ready-to-use data for further analyses with PAX2GRAPHML. Files are automatically updated using databanks synchronization and a processing software (Filangi *et al.*, 2008).

3 Application

PAX2GRAPHML was first applied to the complete PC databank. The regulated reaction graph produced in.graphml format has a size of 363 MB (13% of the initial BioPAX file size). PAX2GRAPHML was also applied to each data source of PC considered independently. As shown in Table 1, the regulated reaction concept used to unify the different BioPAX reaction types facilitates the comparison of the content of each resource. Notably, this revealed that Mirtarbase and CTD are the main contributors of PC in terms of nodes, edges, and especially inhibition reactions.

Generating the regulated reaction graph from 16 BioPAX datasources with PAX2GRAPHML lasted 7 days on a virtual machine with 48 G RAM. Conveniently, the generated files can be downloaded on PAX2GRAPHML website as ready-to-use data resources, which is automatically updated.

Customized graphs can be produced for any subsets of the databases. To achieve this, users can either filter the overall regulated reaction graph, or can merge the regulated reaction graphs produced from two or more databases selected according to their specific interest. The two functionalities (filtering and merging) are available within the PAX2GRAPHML package. As an illustration, Table 1 shows that filtering out CTD and Mirtarbase from PC eliminates 32% of the nodes (36% of reaction nodes and 28% of entity nodes) and 74% of the edges. Table 1 also illustrates that the combination of PID with successively HumanCyc, KEGG and Reactome improves coverage of both reaction nodes (from 4495 to 10 398) and entities (from 4908 to 28 187).

By managing BioPAX data extraction into regulated graphs, PAX2GRAPHML simplifies the implementation of many methods for regulation network analysis and understanding of the controlling steps of the biological pathways.

Financial Support: none declared.

Conflict of Interest: none declared.

References

Babur,O. *et al.* (2010) ChiBE: interactive visualization and manipulation of BioPAX pathway models. *Bioinformatics*, **26**, 429–431.

Babur,O. *et al.* (2014) Pattern search in BioPAXmodels. *Bioinformatics*, **30**, 139–140.

Blavy,P. *et al.* (2014) Using a large-scale knowledge database on reactions and regulations to propose key upstream regulators of various sets of molecules participating in cell metabolism. *BMC Syst. Biol.*, **8**, 32.

Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJ. **Complex Syst.***, **1695**, 1–9.

Filangi,O. *et al.* (2008) BioMAJ: a flexible framework for databanks synchronization and processing. *Bioinformatics (Oxford, England)*, **24**, 1823–1825.

Rodchenkov,I. *et al.* (2013) The BioPAX validator. *Bioinformatics*, **29**, 2659–2660.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Turei,D. *et al.* (2016) OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods*, **13**, 966–967.

Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases

Marc Melkonian, Camille Juigné, Olivier Dameron, Gwenaël Rabut, Emmanuelle Becker

Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases

Marc Melkonian, Camille Juigné, Olivier Dameron, Gwenaël Rabut, Emmanuelle Becker

Databases and Ontologies

Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases

Marc Melkonian [1,2], Camille Juigné [1,3], Olivier Dameron[1], Gwenaël Rabut [2,*] and Emmanuelle Becker [1,*]

[1] Univ Rennes, Inria, CNRS, IRISA, F-35000, Rennes, France
[2] Univ Rennes, CNRS, IGDR - UMR 6290, F-35000, Rennes, France
[3] Pegase, Inrae, Institut Agro, 35590 Saint-Gilles, France.

*To whom correspondence should be addressed, equal contribution.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Information on protein-protein interactions is collected in numerous primary databases with their own curation process. Several meta-databases aggregate primary databases to provide more exhaustive datasets. In addition to exhaustivity, aggregation contributes to reliability by providing an overview of the various studies and detection methods supporting an interaction. However, interactions listed in different primary databases are partly redundant because some publications reporting protein-protein interactions have been curated by multiple primary databases. Mere aggregation can thus introduce a bias if these redundancies are not identified and eliminated. To overcome this bias, meta-databases rely on the Molecular Interaction ontology that describes interaction detection methods, but they do not fully take advantage of the ontology's rich semantics, which leads to systematically overestimating interaction reproducibility.

Results: We propose a precise definition of explicit and implicit redundancy, and show that both can be easily detected using Semantic Web technologies. We apply this process to a dataset from the APID meta-database and show that while explicit redundancies were detected by the APID aggregation process, about 15% of APID entries are implicitly redundant and should not be taken into account when presenting confidence-related metrics. More than 90% of implicit redundancies result from the aggregation of distinct primary databases, while the remaining occurs between entries of a single database. Finally, we build a "reproducible interactome" with interactions that have been reproduced by multiple methods or publications. The size of the reproducible interactome is drastically impacted by removing redundancies for both yeast (-59%) and human (-56%), and we show that this is largely due to implicit redundancies.

Availability: Software, data and results are available at `https://gitlab.com/nnet56/reproducible-interactome`, `https://reproducible-interactome.genouest.org/`, Zenodo (doi:10.5281/zenodo.5595037) and NDEx (doi:10.18119/N94302, doi:10.18119/N97S4D

Contact: emmanuelle.becker@irisa.fr, gwenael.rabut@inserm.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein-protein interactions (PPIs) play an ubiquitous and fundamental role in all biological processes. Description of PPIs is essential to understand how proteins operate at the molecular level and the construction of accurate

and comprehensive protein interaction networks (or interactomes) is an important aim of biological research (Bonetta, 2010; Cafarelli *et al.*, 2017; Luck *et al.*, 2020; Huttlin *et al.*, 2021).

PPIs can be probed using numerous interaction detection methods (IDMs), following biophysical (e.g. x-ray crystallography), biochemical (e.g. affinity purification) or genetic approaches (e.g. yeast two-hybrid). Importantly, since different IDMs probe PPIs in a different manner, they produce complementary results that often do not fully overlap. For instance, some IDMs are designed to detect binary interactions of proteins probed in pairs (e.g. yeast two-hybrid), while others probe interactions of protein groups assembled in complexes (e.g. affinity purification). Consequently, the biological interpretation of PPI networks depends on the underlying IDMs that have been used to produce them. Moreover, since IDMs can generate false positive and false negative interactions, multiple observations of a given PPI with different experimental techniques reinforce the confidence in this PPI. Accurate IDM annotation and interpretation is thus an important issue in interactome studies.

Information on published PPIs is collected in primary databases such as IntAct (Kerrien *et al.*, 2012), MINT (Calderone *et al.*, 2020), BioGRID (Oughtred *et al.*, 2019), DIP (Salwinski *et al.*, 2004) or HPRD (Keshava Prasad *et al.*, 2009). The major databases report IDMs using a controlled vocabulary defined by the Proteomics Standards Initiative-Molecular Interactions (PSI-MI) consortium (Sivade Dumousseau *et al.*, 2018). This vocabulary is structured in an ontology to represent the hierarchical relationships between IDM families by a directed acyclic graph.

Each primary database follows its own curation process with different literature mining, filtering, and reporting techniques. To address the resulting need for integration, several meta-databases aggregate information from multiple primary databases to provide more exhaustive PPI datasets. Some of these meta-databases, such as the Agile Protein Interactomes DataServer (APID) (Alonso-López *et al.*, 2016; Alonso-López *et al.*, 2019), HINT (Das and Yu, 2012) or mentha (Calderone *et al.*, 2013), focus exclusively on experimentally determined PPIs, while others, such as IID (Kotlyar *et al.*, 2019) or STRING (Szklarczyk *et al.*, 2019) also integrate predicted interactions, text mining results or other information.

The accurate aggregation of PPIs from multiple and partly redundant sources is not a trivial task (Turinsky *et al.*, 2010; Klapa *et al.*, 2013). Although the primary databases refer to the PSI-MI ontology, they do not necessarily select identical terms to annotate PPIs (Alonso-López *et al.*, 2019). Hence, a PPI observed in a single experiment reported in a given publication can be annotated with distinct IDM terms in different primary databases. Such annotation differences are usually not taken into account or corrected during the aggregation process.

APID, which unifies data from five of the largest PPI databases (Alonso-López *et al.*, 2016; Alonso-López *et al.*, 2019), implements an integration method that takes redundancy into account and enables to distinguish *'experimental evidences'* (i.e experimental observations reported in publications) from *'curation events'* (i.e. entries in PPI databases). For a given protein pair, multiple entries annotated with identical IDM and identical PubMed publication identifier (PMID) are considered as duplicates and counted as a single experimental evidence. In addition, IDMs are classified into *'binary'* and *'indirect'* methods and IDMs corresponding to related binary methods (e.g. 'two hybrid array' and 'two hybrid pooling approach') are assigned a common method type (e.g. 'two hybrid'). This common method type is then used instead of the original IDM to identify duplicate entries across multiple databases. This custom integration process is not fully satisfying since it is restricted to binary interactions and it does not take advantage of the PSI-MI ontology.

We propose a novel approach to integrate PPI information from primary databases. We define the conventional **explicit redundancy** and extend it with **implicit redundancy** based on parent-related terms in the PSI-MI ontology. We present a method relying on Semantic Web technologies that successfully detects and reconciles implicit redundancies in curation events compiled from multiple primary databases, opening the way to an improved automated curation process. Once curated for both explicit and implicit redundancies, the integrated set of experimental evidences can be used to determine the reproducible interactome supported by multiple experiments.

2 Approach

2.1 Explicit and implicit redundancy

Let us consider a pair of proteins (A, B) and count the number of non-redundant experiments reporting their interaction.

Primary databases such as BioGRID or IntAct can provide several entries corresponding to this protein pair. Usually, these entries differ in the IDM, the PMID, or both. An entry in these databases can thus be defined by a quadruplet

$$(A, B, M_i, P_x)$$

where A and B are the proteins, M_i is the IDM (such as 'affinity chromatography technology', 'anti-tag coimmunoprecipitation' or 'two hybrid', for the most frequent ones), and P_x is the PMID of the original article describing their interaction. When two entries only differ in the IDM, this should signify that the original article has observed the interaction using several experimental techniques. When two entries only differ in the PMID, this should signify that the interaction has been reproduced in two distinct studies using the same detection method.

For meta-databases such as APID, populated by aggregating curation events from other databases, an entry can be defined by a quintuplet

$$(A, B, M_i, P_x, D_a)$$

where D_a indicates the primary database indexing the interaction. Meta-databases can contain different types of redundancies:

- **Explicit redundancy** occurs when distinct entries referring to the same protein pair (A, B) and the same PMID P_x have an identical IDM M_i. This happens when two primary databases registered the same experimental evidence using the same IDM term. Explicit redundancies are detected and unified by APID and other meta-databases.
- **Implicit redundancy** occurs when distinct entries referring to the same protein pair and the same PMID have been annotated with different IDMs although they correspond to the same experimental evidence. In practice, this occurs when curators select IDM terms at different levels of the ontology, one being more general and the other more specific. For example, the interaction of the human proteins MDM2 and TP53 is listed in APID as (MDM2, TP53, 'anti tag Co-immunoprecipitation', PMID:17159902, INTACT:7156209) and also as (MDM2, TP53, 'affinity chromatography technology', PMID:17159902, BIOGRID:680279). Although biologists would naturally recognize one observation annotated twice at different granularities, the redundancy is not explicit. Implicit redundancy should not be confused with the common case where several experimental techniques are used in a single publication to validate a given PPI. Therefore, detecting implicit redundancies requires knowledge on IDMs.

Hereafter, we take advantage of the PSI-MI ontology to identify these two cases, as illustrated in Figure 1.

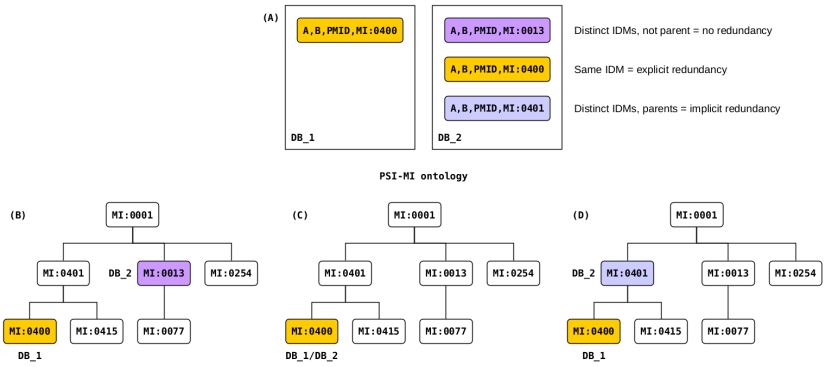


Fig. 1. Illustration of the different types of redundancy across primary databases. (A) Curation events from two databases (`DB_1` and `DB_2`). Depending on the IDM reported by `DB_2`, one can identify no redundancy (purple), explicit redundancy (yellow), or implicit redundancy (blue). Ontology representations of the different cases are presented in panels (B), (C) and (D).

2.2 Definitions

Following the notation introduced in 2.1, we consider two entries, E_i and E_j, of a meta-database, defined by their respective quintuplets of the form $(A, B, M_i, P_x, D_\alpha)$. Note that here we do not consider the experimental role of A and B, therefore all PPIs are symetric and the order of A and B is irrelevant.
E_i and E_j present explicit redundancy if and only if:

$$\begin{cases} E_i = & (A, B, M_i, P_x, D_a) \\ E_j = & (A, B, M_i, P_x, D_b) \end{cases}$$

E_i and E_j present implicit redundancy if and only if:

$$\begin{cases} E_i = & (A, B, M_i, P_x, D_a) \\ E_j = & (A, B, M_j, P_x, D_b) \\ M_i & \textit{is an ancestor of } M_j, \end{cases}$$

where an ancestor can be a direct or an indirect parent.

Ontologies such as PSI-MI (Sivade Dumousseau *et al.*, 2018) can be used to formalize the subsumption relations between IDMs. Note that with the notions provided in section 2.1, explicit and implicit redundancies might be observed among entries originating from different databases (inter-database redundancy, $D_a \neq D_b$) but also from the same database (intra-database redundancy, $D_a = D_b$). We will discuss later (Section 5.3) the meaning of intra-database redundancies, which can correspond either to multiple curation events, but also to variations of an IDM (for example, switching the experimental role (*'bait'* or *'prey'*) of the A and B proteins).

3 Methods

3.1 Source PPI datasets

PPI curation events integrated by APID were downloaded from the APID website on March 23, 2020, last update of APID in January, 2019) for two species (*Homo sapiens* and *Saccharomyces cerevisiae*) in the MITAB25 format (Kerrien *et al.*, 2007). These files aggregate the curated events from five primary databases in a standard format.

In MITAB25 formated data, each line represents a curation event. Interacting proteins are identified by their Uniprot accession numbers. The organism is identified with its NCBI taxonomy identifier. Various information on the experimental evidence is also provided, notably the PMID of the source publication and the PSI-MI code of the IDM used to detect the interaction. Some information such as the direction of the interaction (which protein was used as a *'bait'* and which as a *'prey'*) is not available in this format, but it is usually recorded in primary databases or in more recent MITAB formats (MITAB27). If necessary, missing information might be retrieved using the primary database interaction identifier which is provided and offers full tractability.

3.2 RDF schema and triplestore

The global RDF schema used to integrate all information is presented in Figure 2. It relies on the following ontologies:

- Biological Pathway Exchange (BioPAX) is an ontology developed as a standard for representing molecular interactions, including protein-protein interactions (Demir *et al.*, 2010). We followed the level 3 of the BioPAX specification.
- Proteomics Standards Initiative-Molecular Interactions (PSI-MI) is an ontology edited by the HUPO-PSI. It is dedicated to describe experimental IDMs (Sivade Dumousseau *et al.*, 2018). We used version 1.2.

Raw PPI curation events from the MITAB file were first imported into a MySQL database. A Perl script was used to connect to this database, to exclude curation events that are not considered by APID (see below), and to convert it into a RDF dataset following the BioPAX v3 standard. The resulting interaction data were merged with the PSI-MI ontology, available as an OWL file, into a triplestore powered by the Apache Foundation's JENA suite (v3.14.0). The complete workflow is described in Supplementary Figure S1.

In its integration process, the APID meta-database does not consider curation events annotated with IDMs that do not correspond to a specific experimental method (Alonso-López *et al.*, 2019). To be able to compare our results with APID, we also excluded from our analysis the very same curation events. These are the ones annotated with the IDMs 'molecular interaction', 'interaction detection method', 'biophysical', 'experimental interaction detection', 'inference', 'inferred by author', 'inferred by curator', 'in vitro', 'in vivo', 'unspecified method', or 'phenotype-based detection assay'.

3.3 SPARQL queries

Queries were run using SPARQL Protocol and RDF Query Language (SPARQL). The JENA suite was used to run the SPARQL queries. All queries used to detect redundancies are available in supplementary data

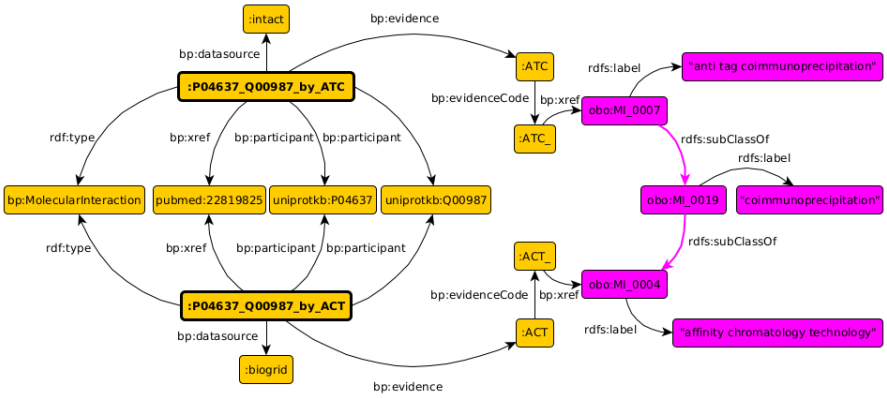


Fig. 2. Scheme representing two curation events reporting the interaction between the ubiquitin ligase MDM2 (UniprotKB: P04637) and the tumor protein 53 (UniprotKB: Q00987) in the BioPAX level 3 ontology (yellow nodes). These two curation events (highlighted in bold) were annotated by different databases (BioGRID and Intact). They refer to the same publication (PMID: 22819825), but different IDMs were used to annotate the interaction ('anti tag coimmunoprecipitation' and 'affinity chromatography technology'). The PSI-MI ontology (purple nodes) reveals that 'affinity chromatography technology' is an ancestor of 'anti tag coimmunoprecipitation', indicating an implicit redundancy between the two curation events.

(Figures S2, S3, S4, S5, S6, S7). As an example, Figure 3 presents the SPARQL query used to detect implicit redundancies in curation events, if one term is an ancestor of the other in the PSI-MI ontology. For each implicit redundancy detected, we conserved only the curation event with the most precise IDM.

```
SELECT DISTINCT ?p1 ?p2 ?pmid ?dm_name1
WHERE {
    ?ppi1 rdf:type bp:MolecularInteraction ;
          bp:participant ?p1, ?p2 ;
          bp:xref ?pmid ;
          bp:evidence ?dm_name1 .
    ?dm_name1 bp:evidenceCode ?m_vocab1 .
    ?m_vocab1 bp:xref ?dm_code1 .
    FILTER ( STR(?p1) < STR(?p2) )
    FILTER NOT EXISTS {
        ?ppi2 rdf:type bp:MolecularInteraction ;
              bp:participant ?p1, ?p2 ;
              bp:xref ?pmid ;
              bp:evidence ?dm_name2 .
        ?dm_name2 bp:evidenceCode ?m_vocab2 .
        ?m_vocab2 bp:xref ?dm_code2 .
        ?dm_code2 rdfs:subClassOf+ ?dm_code1 .
    }
}
```

Fig. 3. SPARQL query to select curation events without explicit nor implicit redundancies. (Note: prefixes are not shown)

3.4 Availability and implementation

The code is available at `https://gitlab.com/nnet56/reproducible-interactome`. The results are available at `https://reproducible-interactome.genouest.org/` and on the Zenodo open data repository (doi:10.5281/zenodo.5595037). The non-redundant interactomes are also accessible on the NDEx platform to facilitate their analysis and manipulation with classical algorithms (doi:10.18119/N94302 (human), doi:10.18119/N97S4D (yeast)).

4 Results

4.1 Overview of analyzed curation events

We analysed the same curation events as the APID database to assess the efficiency of redundancy detection methods. A summary of these curation events is presented in Table 1. The downloaded MITAB files contain $700,484$ curation events for *Homo sapiens* and $305,102$ for *Saccharomyces cerevisiae* (hereinafter referred to as human and yeast, respectively). Together, BioGRID and IntAct represent approximately 85% of all curation events in both species. The contribution of HPRD and BioPlex, restricted to human data, accounts for 13.9% of human curation events. For both species, most PPIs appear in only one or two curation events. PPIs reported by a single curation event represent 49.3% and 60.7% of interacting pairs in human and yeast, respectively.

4.2 Interaction detection methods (IDMs)

The most frequent IDMs in all curation events are listed in Table 1. Among them, 'affinity chromatography technology', 'tandem affinity purification', 'anti tag coimmunoprecipitation' and 'two hybrid' cover more than 58% of human and 76% of yeast curation events. Interestingly, these IDMs include terms with parent–child relationships in the PSI-MI ontology. For example, 'affinity chromatography technology' is a direct ancestor of 'anti tag coimmunoprecipitation'. The presence of such chains is suggestive of possible implicit redundancies between curation events, as defined in sections 2.1 and 2.2.

4.3 Quantification of implicit redundancies

Thanks to the expressiveness of the SPARQL language, we identified both explicit and implicit redundancies among curation events (example query in Figure 3). For constituting a non-redundant dataset, we selected the most precise curation events and discard the redundant and less precise ones since they do not add information.

The occurrence of redundancy among curation events is significant (Table 2). We detected and discarded $73,991$ (11.1%) and $40,266$ (13.7%) implicitly redundant curation events for human and yeast, respectively. Taking into account both explicit and implicit redundancies resulted in removing 30.9% of curation events for human and 35.4% for yeast.

Table 1. Human and yeast curation events (CEs) analysed in this study. Excluded Interaction Detection Methods (IDMs) concern 5.00% ($n = 35,000$) of all curation events in human and 3.87% ($n = 11,809$) in yeast. Only IDMs annotated with a frequency higher than 2% are shown.

Contributing Databases			**Most frequent Interaction Detection Methods**			**Currations events for** (P_a, P_b)		
Databases	**CEs**	**(%)**	**Interaction Detection Methods**	**Counts**	**(%)**	**Occurrences**	**Counts**	**(%)**
Human								
BioGRID	378,910	(54.1%)	Affinity chromatography technology	291,621	(41.63%)	One	161,031	(49.30%)
IntAct	215,577	(30.8%)	Two hybrid	71,969	(10.27%)	Two	91,742	(28.08%)
BIOPLEX!	55,151	(7.9%)	Anti tag coimmunoprecipitation	49,428	(7.06%)	[3-10[	69,015	(21.13%)
HPRD	42,327	(6.0%)	Pull down	42,423	(6.06%)	[10-50[	4,763	(1.46%)
DIP	8,519	(1.2%)	Biochemical	40,544	(5.79%)	≥ 50	113	(0.03%)
			Anti bait coimmunoprecipitation	27,745	(3.96%)			
			In vivo	21,118	(3.01%)			
			Two hybrid array	20,813	(2.97%)			
			Validated two hybrid	14,525	(2.07%)			
Yeast								
BioGRID	133,998	(43.9%)	Affinity chromatography technology	88,681	(29.07%)	One	83,799	(60.73%)
IntAct	130,025	(42.6%)	Tandem affinity purification	84,842	(27.81%)	Two	28,496	(20.65%)
DIP	41,079	(13.5%)	Anti tag coimmunoprecipitation	35,363	(11.59%)	[3-10[	21,792	(15.79%)
			Two hybrid	24,752	(8.11%)	[10-50[	3,799	(2.75%)
			Pull down	13,960	(4.58%)	≥ 50	99	(0.07%)
			Inferred by author	10,894	(3.57%)			
			Protein complementation assay	6,825	(2.24%)			
			Enzymatic study	6,817	(2.23%)			

Table 2. Impact of the removal of both explicit and implicit redundancies on the number of curation events and on the apparent size of the reproducible interactome, for human and yeast. (EEs: Experimental Evidences)

	Human	(%)	**Yeast**	(%)
Curation events				
Initial curation events	665,484	(100%)	293,293	(100%)
Curation events without explicit redundancies	534,140	(80.3%)	229,630	(78.3%)
Curation events without explicit and implicit redundancies	460,149	(69.1%)	189,364	(64.6%)
Apparent size of the reproducible interactome (PPIs supported by ≥ 2 EEs)				
Initial	159,192	(100%)	52,313	(100%)
Without explicit redundancies	111,009	(69,7%)	40,235	(76.9%)
Without explicit and implicit redundancies	70,554	(44,3%)	21,311	(40.7%)

Importantly, detection of redundancy between curation events has a strong impact on the apparent size of the reproducible interactome (i.e PPIs supported by at least two experimental evidences) (Table 2, Supplementary Figures S8 and S9). For human, the reproducible interactome drops from $159,192$ to $70,554$ PPIs (-55.7%: -30.3% due to explicit redundancies and -25.4% due to implicit ones). For yeast, the impact of redundancies is even worse, with a drop of the reproducible interactome from 52,313 PPIs to 21,311 after removal of both explicit and implicit redundancies (-59.3%: -23.1% due to explicit redundancies and -36.2% due to implicit ones). In other words, for human, discarding 11.1% of implicitly redundant curation events accounts for reducing by 25.4% the reproducible interactome. Similarly, for yeast, discarding 13.7% of implicitly redundant curation events accounts for reducing by 36.2% the reproducible interactome.

4.4 Implicit redundancies mostly result from the integration of the different primary databases

We then investigated whether implicit redundancy was already present in source databases (intra-database redundancy), or if it was a consequence of the integration of different source databases (inter-database redundancy). The vast majority originates from inter-database redundancies for both human (91.1%) and yeast (95.0%) (see Supplementary Tables S1 and S2). The couple of databases that generates the largest part of the implicit redundancies is BioGRID and IntAct. This is consistent with the fact that BioGRID and IntAct are the two most contributing source databases. Intra-database redundancies will be further discussed in section 5.3.

4.5 Frequently redundant identification methods

We computed the frequency of the pairs of detection methods involved in implicit redundancies. For human, the most frequent implicitly redundant couples of IDMs and their parent-child relationships in the PSI-MI ontology are displayed in Figure 4.

The most frequent couple is 'affinity chromatography technology' and 'anti tag coimmunoprecipition', which is responsible for $25,333$ redundancies. The term 'affinity chromatography technology' is also frequently observed with other descendants such as "pull down" ($n = 9,896$), 'anti bait coimmunoprecipitation' ($n = 6,617$), or "tandem affinity purification" ($n = 5,968$). Two-hydrid techniques are also introducing redundancies, for example with 'two hybrid', and its descendants 'two hybrid array' ($n = 16,113$), 'two hybrid prey polling approach' ($n = 11,238$), 'validated two hybrid' ($n = 11,123$), or 'two hybrid pooling approach' ($n = 10,713$). A similar situation is observed in yeast (the complete list of implicit redundancies for both human and yeast is available as Supplementary Tables S3 and S4). Implicit redundancies are thus widespread all along the PSI-MI ontology, and not limited to binary IDMs. This highlights the need for a general approach to reconcile

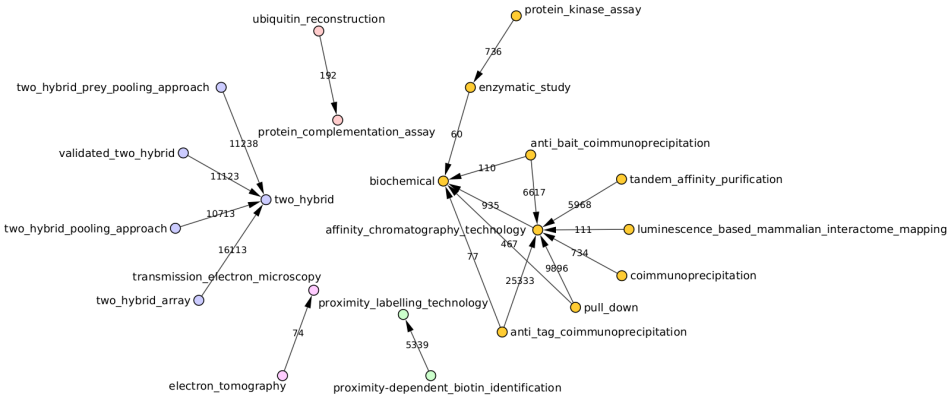


Fig. 4. Couples of related interaction detection methods (IDMs) from the PSI-MI ontology frequently identified in implicit redundancies of human PPIs. The arrows connect the most specific to the most general term according to the PSI-MI ontology. Only implicit redundancies with at least 50 occurrences are shown. Nodes connected to a common IDM are represented with the same color.

curation events during the integration of multiple primary databases.

The fact that implicit redundancies are observed between very different terms of the PSI-MI ontology suggests that different primary databases have different policies for annotating IDMs, as previously noted for IntAct and BioGRID (Alonso-López *et al.*, 2019). We therefore further analysed the IDMs used by each primary database.

We observed that IntAct and DIP use a wide range of IDMs for both human and yeast PPIs (165 for IntAct and 89 for DIP) while BioGRID, HPRD and BioPlex use much fewer (12, 3 and 1 IDMs, respectively) and more general IDMs. Hence, the strong discrepancies in database annotation policies are the source of inter-database implicit redundancies.

Overall, we observed that implicit redundancy (i) occurs between a wide range of the PSI-MI ontology terms, regardless of the species, (ii) mostly results from the integration of different primary databases with different annotation policies, and (iii) happens for all database combinations.

5 Discussion

The construction of a reliable interactome demands to combine interaction data produced by several independent experimental evidences and IDMs in order to reduce false positives. Since experimental evidences are curated and stored in several primary databases, a unification of these databases is required. The Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI) developed the PSICQUIC specification and web services that facilitate data retrieval from multiple databases and assist their integration but do not elaborate on redundancy detection (del Toro *et al.*, 2013). In several (meta-) databases, PPIs are annotated with a confidence score, which is calculated using the number of independent experimental evidences and the nature of IDMs (Villaveces *et al.*, 2015). To be relevant, these algorithmic require reliable, non-redundant, datasets of experimental evidences. Therefore, several primary databases have decided to coordinate their curation efforts in the frame of the IMEx consortium in order to provide a single non-redundant set of homogeneously annotated protein interaction data (Orchard *et al.*, 2012; Porras *et al.*, 2020).

Here, we propose a formalisation of both explicit and implicit redundancy between experimental evidence entries in order to integrate PPIs from any database that uses the PSI-MI ontology. Knowledge about IDMs is extracted from the PSI-MI ontology, while the method to identify redundancies is based on Semantic Web technologies.

5.1 The Semantic Web is adapted for identifying implicit redundancies

Alonso-López *et al.* (2019) pointed two problems related to redundancy identification: (i) there may be a parent-child relationship between IDM terms, and (ii) the path from a child term to its ancestors may not be unique due to multiple inheritance. We propose the notion of **implicit redundancy** to address the logical implications of two database entries describing the interaction of the same protein pair with IDMs that have a descendant-ancestor relationship. The Semantic Web is designed to perform integrated reasoning on data annotations and ontologies. In particular, it makes handling simple and multiple hierarchies straightforward. In the raw data of APID that aggregates BioGRID, IntAct, HPRD, BioPlex and DIP, we were able to identify both explicit and implicit redundancies. Our work reveals that implicit redundancies are a widespread phenomenon resulting from the different curation choices of the various databases and that it is of similar importance than explicit redundancies. Therefore, we demonstrated the relevance of both the notion of implicit redundancy and of the choice of the Semantic Web as a technical framework for addressing the redundancy identification problem. Moreover, new explicit and implicit redundancies will continue to occur over the natural updates of the various databases.

The PSI-MI ontology that describes the IDMs is evolving. For example, during the time of our project, we noticed that the term 'three hybrid', which was initially a child of the term 'two hybrid', is now a child of 'transcriptional complementation assay'. This modification is highly relevant since 'two hybrid' is a binary identification method, whereas 'three hybrid' is not, and having a non-binary identification method as a direct child of a binary one was not consistent. Therefore, just like the databases are regularly updated, the ontologies are also corrected and enriched, which also has an incidence on redundancies. By allowing to automate redundancy detection as the integration of databases scales up, the Semantic Web facilitates the reliable interpretation of the results in the perspective of the construction of a reproducible interactome.

5.2 Widespread inter-databases implicit redundancies

Implicit redundancies primarily arise from the integration of different databases (91.1% and 95.0% of inter-database redundancies for human

and yeast, respectively). In our study, we clearly highlight that this is due to the granularity of IDMs used in the primary databases. Indeed, while some databases like IntAct refer to numerous detailed terms from the PSI-MI ontology (165 and 89 terms used to annotate human and yeast PPIs, respectively), other databases like BioGRID merely use general and high level terms (only 12 terms used for both human and yeast).

Therefore, if the integration of different PPI databases is necessary to better cover the interactome, a particular attention has to be paid to detect the widespread inter-database implicit redundancies. A simple method could be to define priorities between databases depending on whether they use precise or general terms to annotate PPIs. In case of multiple curations events referring to the same proteins and the same PMID, the ones from the database with the highest priority would be selected. However, this would be an approximate approach whereas we propose an exact solution, robust to possible changes of annotation policy by primary databases.

Primary databases of the IMEx consortium coordinate and share their curation efforts to produce a non-redundant dataset of PPI experimental evidences (Orchard *et al.*, 2012). IMEx members use common curation rules to harmonize their annotation process. The unicity of the curation events is ensured by allowing PPIs from a given PMID to be annotated only once, and all data are centralized in IntAct. Both this work from the IMEx consortium and ours emphasize the need for a general approach to assemble non-redundant PPI datasets.

5.3 Intra-database redundancies

Our analysis also identified a significant number of apparently redundant curation events within primary databases (Supplementary Figures S8 and S9). Such intra-database redundancy may originate from multiple independent annotations of identical experimental evidences within primary databases, as noted by Alonso-López *et al.* (2019). Yet, further inspection of such curation events indicates that intra-database redundancy primarily occurs when independent experiments from the same publication have been annotated in a given database with identical or related IDMs, leading to apparent explicit or implicit intra-database redundancies. For instance, we observed that the vast majority of the explicit intra-database redundancies originating from BioGRID are due to PPIs probed with both partners as baits and preys (6229 out of 8696 explicit redundancies involving exactly two curation events for yeast and 12283 out of 15385 for human). Intra-database redundancy can also occur when a PPI has been identified with a high-throughput experiment and then validated using the same or a related method performed at low-throughput. Hence, this currently leads to the unification of curation events that actually report distinct experimental evidences. To correct this, our method could be extended by taking into account additional information, such as the experimental role of each protein.

5.4 Towards a reproducible interactome

The size of the reproducible interactome is drastically impacted by removing redundancies for both human (-55.7%) and yeast (-59.3%), and we show that this is largely due to implicit redundancies. Indeed, we observe that filtering the curation events involved in implicit redundancy (11 to 14 %) leads to a drastic (25 to 36 %) reduction of the apparently reproducible interactome. This implies that a large number of PPIs currently considered as reproducible actually relies on integration artefacts. Thus, more experimental data are still needed to further improve the size and confidence level of the reproducible interactome. Information on PPIs that have not yet been reproduced can help to prioritize such experiments. Knowledge-based methods as presented in this article will be necessary to support the integration of the continuously increasing experimental evidences and publications.

Acknowledgements

The GenOuest platform provided computational support and Web hosting.

Funding

This work has been supported by Univ Rennes with a Defi Emergent 2019 grant to EB and GR.

References

Alonso-López, D. *et al.* (2016). APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucleic Acids Res*, **44**(W1), W529–535.

Alonso-López, D. *et al.* (2019). APID database: redefining protein–protein interaction experimental evidences and binary interactomes. *Database*, **2019**. baz005.

Bonetta, L. (2010). Interactome under construction. *Nature*, **468**(7325), 851–852.

Cafarelli, T. *et al.* (2017). Mapping, modeling, and characterization of protein–protein interactions on a proteomic scale. *Current Opinion in Structural Biology*, **44**, 201–210.

Calderone, A. *et al.* (2013). mentha: a resource for browsing integrated protein-interaction networks. *Nat Methods*, **10**(8), 690–691.

Calderone, A. *et al.* (2020). Using the MINT Database to Search Protein Interactions. *Curr Protoc Bioinformatics*, **69**(1), e93.

Das, J. and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol*, **6**, 92.

del Toro, N. *et al.* (2013). A new reference implementation of the PSICQUIC web service. *Nucleic Acids Res*, **41**(Web Server issue), W601–606.

Demir, E. *et al.* (2010). The BioPAX community standard for pathway data sharing. *Nature biotechnology*, **28**(9), 935–942.

Huttlin, E. L. *et al.* (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, **184**(11), 3022–3040.

Kerrien, S. *et al.* (2007). Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol*, **5**, 44.

Kerrien, S. *et al.* (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Res*, **40**(Database issue), D841–846.

Keshava Prasad, T. S. *et al.* (2009). Human Protein Reference Database–2009 update. *Nucleic Acids Res*, **37**(Database issue), D767–772.

Klapa, M. I. *et al.* (2013). Reconstruction of the experimentally supported human protein interactome: what can we learn? *BMC systems biology*, **7**, 96.

Kotlyar, M. *et al.* (2019). IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res*, **47**(D1), D581–D589.

Luck, K. *et al.* (2020). A reference map of the human binary protein interactome. *Nature*, **580**(7803), 402–408.

Orchard, S. *et al.* (2012). Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods*, **9**(4), 345–350.

Oughtred, R. *et al.* (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res*, **47**(D1), D529–D541.

Porras, P. *et al.* (2020). Towards a unified open access dataset of molecular interactions. *Nature communications*, **11**(1), 6144.

Salwinski, L. *et al.* (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, **32**(Database issue), D449–451.

Sivade Dumousseau, M. *et al.* (2018). Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions. *BMC Bioinformatics*, **19**(1), 134.

Szklarczyk, D. *et al.* (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*, **47**(D1), D607–D613.

Turinsky, A. L. *et al.* (2010). Literature curation of protein interactions: measuring agreement across major public databases. *Database*, **2010**. baq026.

Villaveces, J. M. *et al.* (2015). Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database*, **2015**. bau131.

Detection and correction of non-conformities and redundancies in complexes of molecules in BioPAX

Camille Juigné[1,2], Olivier Dameron[1], François Moreews[1,2], Florence Gondret[2] and Emmanuelle Becker[1]
[1] Univ Rennes, Inria, CNRS, IRISA - UMR 6074, F-35000 Rennes, France
[2] PEGASE, INRAE, Institut Agro, 35590, Saint Gilles, France

Corresponding author: camille.juigne@irisa.fr

Abstract *Complexes play major roles in the interactions between biochemical entities and the regulation of biological pathways. The Biological Pathway Exchange format (BioPAX) facilitates the integration of data sources describing such interactions. In BioPAX, a black-box complex is a complex with no component, as its composition is unknown. The BioPAX specification explicitly prevents complexes to have any component that is another complex unless this component is a black-box complex. We observed that the Reactome pathway database contains such invalid recursive complexes of complexes. We propose a method to identify and fix them. For the Homo sapiens version of Reactome, 5,734 of the 14,840 complexes (39%) are invalid. They participate to 6,994 of the 21,813 interactions (32%). All the complexes identified as invalid could be fixed. In addition to improving BioPAX compliance, the benefits of fixing invalid recursive complexes are two-folds. First, it led to a decrease in false negative when identifying the interactions in which a physical entity can participate as a component of a complex. With the original Reactome, a set of proteins of interest participated in 554 interactions either directly or as direct parts of complexes. Fixing invalid recursive complexes allowed to identify 429 additional interactions (+77%). Second, it decreased artificial redundancies in Reactome by identifying sets of structurally equivalent complexes that share the same components with the same stoichiometric coefficients and have the same cellular location. The original Reactome contained 241 redundant complexes. Fixing invalid recursive complexes allowed to identify 92 additional redundant complexes (+38%).*

Keywords BioPAX, Curation, Molecular complex, SPARQL, Reactome

1 Introduction

Complexes and interactions in biology. Understanding how biological systems adapt to their environment needs to better capture, describe and model the interactions between their biochemical entities. A rigorous approach is required to support further analyses because these systems typically involve an intricate network of interactions between numerous participants. Among them, complexes are a major class of entities that results from the chemical assembly of several molecules (nucleic acids, proteins and other molecules) that bind each other at the same time and place, and form single multimolecular machines. Biologically, they play an important role in transcription, RNA splicing and polyadenylation machinery, protein export and transport etc [1,2]. From the data analysis perspective, they cause some indirection between molecules and interactions, i.e. this introduces an additional node (the complex) between two entities that are no longer directly connected by a link. Therefore, molecules can either participate directly to an interaction or also be a component of a complex that participates to an interaction. The latter should also be considered.

Complexes in BioPAX. The Biological Pathway Exchange format [1] (BioPAX) is a well established formalism to represent biological pathways at the molecular and cellular levels, including interactions [3]. In the BioPAX ontology, the four top level classes are `Pathway`, `Interaction`, `Physical Entity` and `Gene`. Interactions represent biological relationships between two or more entities. They encompass molecular interactions, controls and conversions. Physical entities encompass small molecules,

1. http://www.biopax.org/release/biopax-level3-documentation.pdf

proteins, DNA, RNA and complexes. Complexes are defined in BioPAX as "*physical entities whose structure is comprised of other physical entities bound to each other non-covalently, at least one of which is a macromolecule (e.g. protein, DNA, or RNA)*". If BioPAX requires that at least one of the components of a complex is a macromolecules, the specification also addresses the situation where another component would be a complex. It explicitely mentions: "*complexes should not be defined recursively [...] i.e. a complex should not be a component of another complex. [...] Exceptions are black-box complexes (i.e. complexes in which the component property is empty), which may be used as components of other complexes because their constituent parts are unknown*".

The Reactome use-case. BioPAX is based on Semantic Web technologies, with RDF facilitating integration, SPARQL facilitating querying and OWL facilitating knowledge-based reasoning. All the major reference pathway databases are available in BioPAX. Among them, Reactome [2] is a free, open-source, curated and peer-reviewed pathway database [4]. It is widely used in genome analysis, modeling, systems biology, clinical research and education. We noticed the presence of invalid recursive complexes, i.e. complexes composed of other complexes that are not black-box complexes. Interestingly, these invalid complexes were not detected by the BioPAX validator [3] [5]. Their presence in Reactome precludes further robust analyses.

2 Objectives

Our hypothesis is that recursive complexes can cause false negatives when identifying the interactions in which a physical entity can participate. For example, if A, B and C are physical entities, A can directly participate in several interactions, but also indirectly when associated to B as a complex, or to B and C as another complex. If (A,B) and (A,B,C) are valid complexes, the two situations can be correctly processed by identifying the interactions matching with the criterion "having a participant that is a complex composed of A". However, if (A,B,C) complex is composed of (A,B) complex and of C, all the interactions in which (A,B,C) participates would fail to meet the aforementioned criterion because their participants are not "a complex composed of A" but "a complex composed of a complex composed of A".

We propose a method for identifying and fixing these invalid recursive complexes. Using this method, we determine whether fixing invalid recursive complexes, thereby repairing the topology of the graph, is beneficial by allowing to identify new interactions in which a molecule participates.

In addition, fixing invalid recursive complexes could also result in the identification of redundant complexes. For example, ((A,B),C) or (A,(B,C)) or ((A,C),B) could be distinct invalid complexes with their own identifiers, but they would be all fixed as complexes having the same components A, B and C using our method. We also study whether some of the invalid complexes that were fixed at the previous step become equivalent to other fixed invalid complexes or to original valid complexes.

3 Materials and methods

We developed semantically-rich SPARQL queries for identifying and fixing invalid recursive complexes, and to detect the resulting redundancies. We applied this method on the Reactome pathways database as a use-case study. For the sake of reproducibility, we provide a jupyter notebook detailing the analysis [4].

3.1 Identify invalid complexes

We first analyzed all the configurations of BioPAX complexes according to the nature of their components. Figure 1 summarizes the BioPAX validity of the different categories. We wrote a SPARQL query for each category to compute their relative size out of the total number of complexes. The query in figure 2 identifies the invalid recursive complexes.

2. https://reactome.org/
3. https://biopax.baderlab.org/
4. https://github.com/cjuigne/non_conformities_detection_biopax

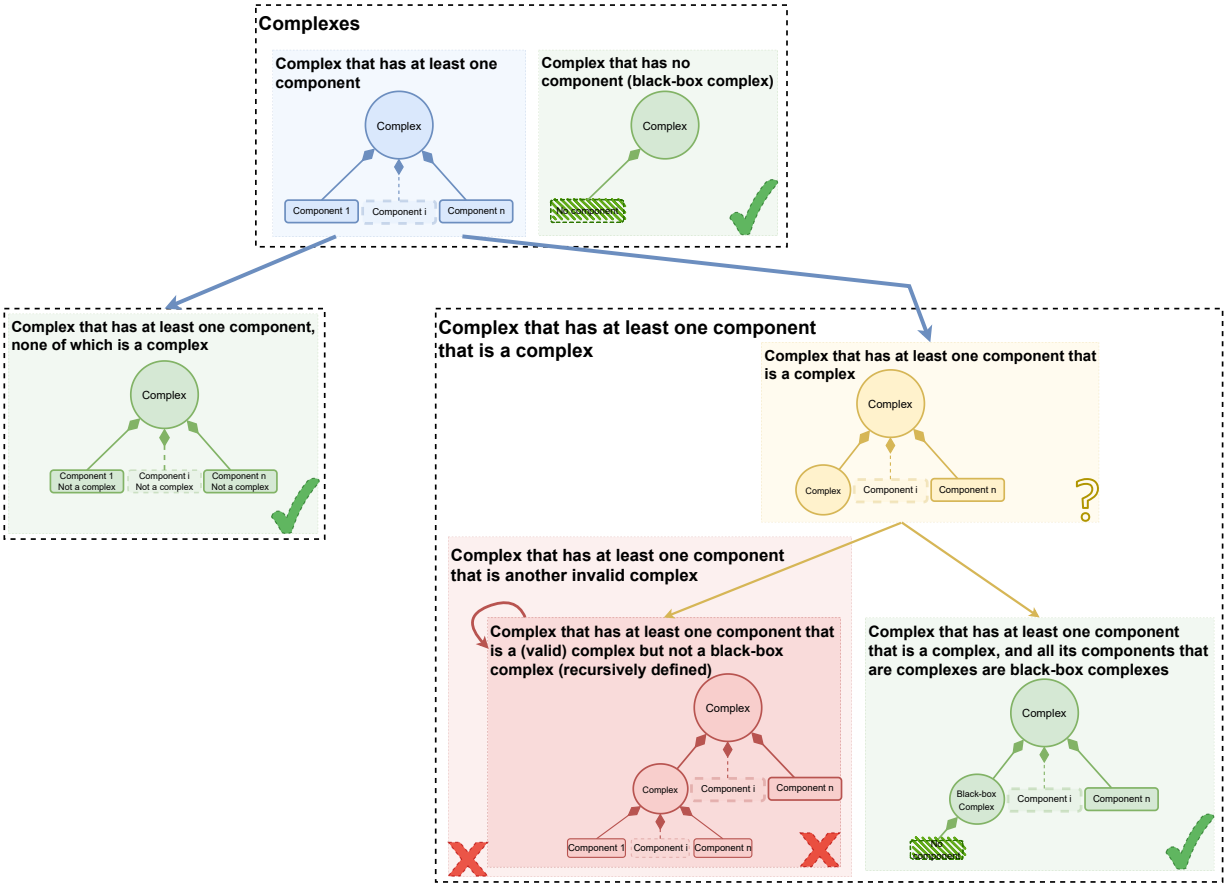


Fig. 1. Validity of the categories of BioPAX complexes. Complexes are represented by circles and the other physical entities by boxes with rounded corners. Composition is represented by a diamond head arrow with the diamond on the side of the whole. A valid complex can have components that are themselves complexes only if they are all black-box complexes, i.e. they do not have any component. Note that an invalid complex can itself be a component of another complex (which therefore becomes invalid as well).

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX bp3: <http://www.biopax.org/release/biopax-level3.owl#>

SELECT DISTINCT ?invalidComplex
WHERE {
  ?invalidComplex rdf:type bp3:Complex .
  ?invalidComplex bp3:component ?complexComponent .
  ?complexComponent rdf:type bp3:Complex .
  ?complexComponent bp3:component ?componentOfComplexComponent .
}
```

Fig. 2. SPARQL query for identifying invalid recursive complexes in BioPAX, i.e. complexes composed of at least another complex that has components.

3.2 Fix the invalid complexes

Fixing an invalid recursive complex consists in (1) collapsing as direct components all its direct or indirect atomic components (i.e. either not complexes or black-box complexes; they correspond to the leaves in the tree of components), (2) deleting all the other components, (3) setting the correct values for stoichiometric coefficients, and (4) preserving all its other attributes.

For computing the values of the stoichiometric coefficients in step (3), we need to trace all the stoichiometric coefficients from each leaf up to the root complex, and to take into account the fact that a given physical entity can be a component of several parts of the recursive complex. Tracing stoichiometry is illustrated by figure 3 where complex Y was composed of γ Z, and X was itself composed of α Y. This resulted in $\alpha \times \gamma$ Z in X (via Y). The second situation is illustrated by the

fact that, in addition to being a component of Y, Z was also a direct component of X with β as stoichiometric coefficient value. Overall, this resulted in $\alpha \times \gamma + \beta$ occurrences of Z in X.

We note $S_y(z)$ the global stoichiometric coefficient value of z at y, i.e. the number of occurrences of z in y, and $C(y)$ the set of the direct components of y. Formula 1 recursively computes the stoichiometric coefficient value of any physical entity z.

$$\begin{cases} S_z(z) = 1 & \\ S_y(z) = 0 & \text{if } (y \neq z) \wedge (C(y) = \emptyset) \\ S_y(z) = \sum_{p \in C(y)} S_y(p) \times S_p(z) & \text{otherwise} \end{cases} \tag{1}$$

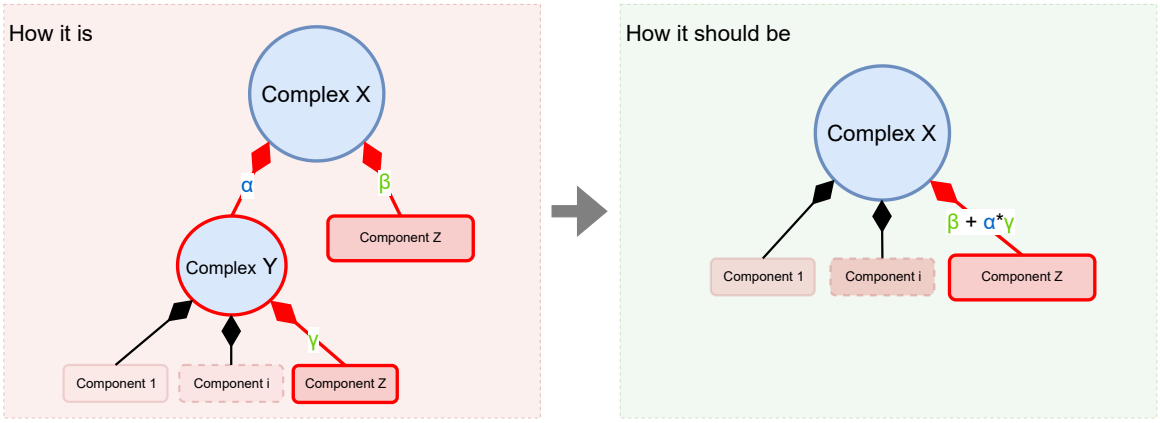


Fig. 3. Fixing an invalid recursive complex consists in collapsing as direct components all its direct and indirect components that are leaves in the composition tree of the complex and computing the correct stoichiometric coefficient values with equation 1.

3.3 Identify redundant complexes

We considered as redundant the complexes that have exactly the same components with the same stoichiometric values and the same cellular location. Figure 4 illustrates how invalid recursive complexes can be the cause of redundancy due to the order by which the components are nested in the complexes. Fixing the invalid complexes made possible the detection of redundancy. For that, we developed a SPARQL query that identifies the pairs of complexes that have the same components and properties but different identifiers.

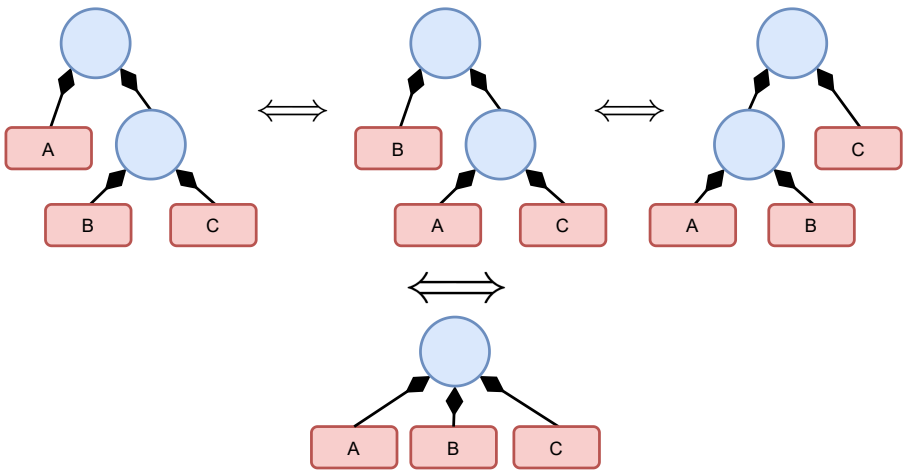


Fig. 4. Example of invalid recursive complexes leading to redundancy. Fixing them made possible the detection of redundancy.

3.4 Application to Reactome

We used the Reactome pathway database (Homo sapiens version 79 (2022-02-03)) as a use-case study. First, we identified the invalid recursive complexes, the molecules they were composed of, and the interactions they participated in. Second, we fixed the invalid recursive complexes and generated their valid representation. Third we generated a corrected version of Reactome with no invalid recursive complexes: we deleted the composition relations and the stoichiometric coefficients of the invalid complexes, and then imported their valid representation.

3.5 Evaluation for determining the reactions a molecule participates to

Complexes induce more than one level of indirection between biological entities and the reactions to which they participate (as products, substrates, activators or inhibitors).

1. From a list of molecules (ChEBI ID or UniProt ID) we located the corresponding entities
   — for each molecule we looked for the corresponding entity reference, which in BioPAX means a node where all the non-changing aspects of the entity are stored.
   — then, we identify all the physical entities associated to these entity references. Having several physical entities associated to a single entity reference allows to associated them with different contexts such as cellular location.
2. From these physical entities, we identified the interactions in which they participate, including as a component of a valid or invalid complex:
   — as direct participant of an interaction
   — before fixing invalid complexes, the interactions that had a complex in which the molecule was a direct component
   — after fixing invalid complexes, the interactions that have a complex in which the molecule is a *direct* component. Note that this corresponds to the interactions that have a participant that is an invalid complex in which the molecule is an *indirect* component.

4 Results

4.1 Invalid complexes identified in Reactome

The Human subset of Reactome v79 is composed of a total of 14,840 complexes. Among them, we identified 858 black-box complexes. Among the remaining 13,982 complexes (that had at least one known component), 8,248 complexes had no component that was itself a complex with known components. Together with the black-box complexes, they represented the 9,106 valid complexes. On the opposite, 5,734 complexes had at least one component that was a complex with components. Altogether, invalid recursive complexes represented 39% of the 14,840 complexes in Reactome. None of them have been detected by the BioPAX-validator tool.

Invalid recursive complex participated to 6,994 of the 21,813 interactions (32%).

4.2 Fixed invalid complexes

All the 5,734 complexes were fixed.

In the Reactome dataset, the average number of direct components in a complex was 2 ($\sigma = 2.6$), with `R-HSA-5626171` and `R-HSA-72069` having each a maximum of 65 components. After fixing the invalid complexes, the average number of direct components was 4 ($\sigma = 8.7$), with `R-HSA-156656` having the maximum of 174 components. Figure 5 illustrates the number of direct components counted before and after fixing invalid complexes.Indeed, fixing the invalid complexes allowed to repair the topology of the graph.

4.3 Identified redundant complexes

Among the 14,840 complexes from the original Reactome, the query identified 133 pairs of redundant complexes involving 241 distinct complexes. They constituted 117 maximal cliques of equivalent complexes (size ranging from 2 to 4), corresponding to 124 complexes in excess.

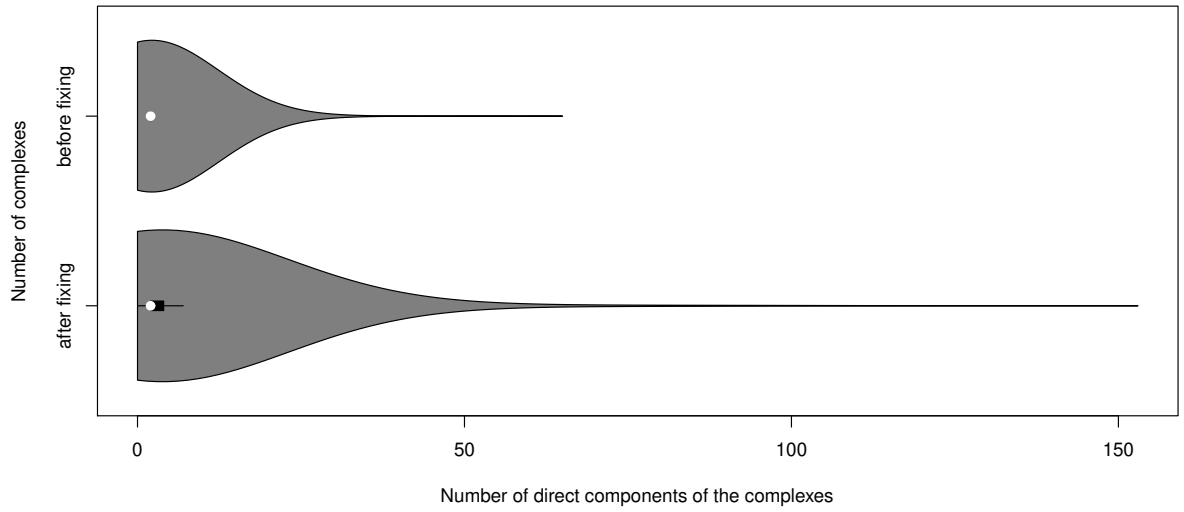


Fig. 5. Number of direct components of the complexes in the Human dataset of Reactome v79 before (top) and after (bottom) fixing invalid recursive complexes.

After fixing the invalid complexes, the query identified 210 pairs of redundant complexes involving 333 distinct complexes. They constituted 157 maximal cliques of equivalent complexes (size ranging from 2 to 6), corresponding to 176 complexes in excess. Figure 6 illustrates how fixing structurally different complexes revealed their redundancy.

The files `reactome-v79-complexes-orig-redundant.ttl` and `reactome-v79-complexes-valid-redundant.ttl` contain symmetric `dyliss:hasSameCompositionAs` relations between pairs of redundant complexes.

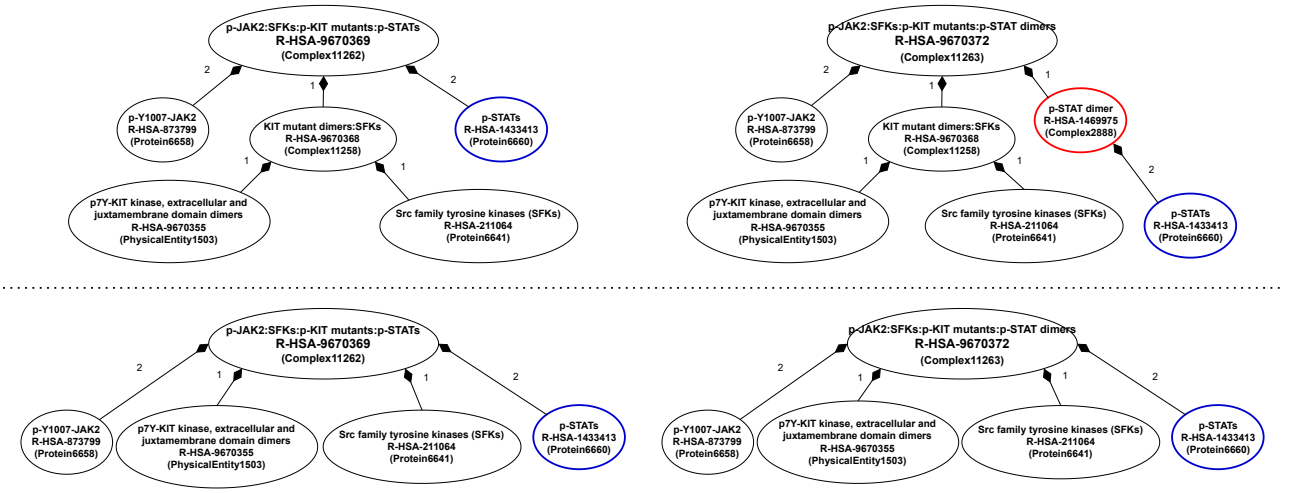


Fig. 6. Original compositions of complexes R-HSA-9670369 (top left) and `R-HSA-9670372` (top right) from Reactome. Both are invalid recursive complexes. The fixed versions (bottom left and right, respectively) have more direct components than the original. Both fixed versions have the same components with the same stoichiometric coefficients, which reveals their redundancy. The structure difference between the original versions is highlighted in red: `R-HSA-9670372` is composed of an intermediate dimer of the p-STATs protein whereas `R-HSA-9670369` is directly composed of p-STATs with a stoichiometric coefficient of 2.

4.4 Evaluation of the benefits of the complete procedure

Complexes play an important role in Reactome, and the components of invalid recursive complexes participate in many interactions. When retrieving the interactions molecules participate to, we quantified the false negatives caused by invalid complexes and that become true positives after fixing the

invalid complexes, both in general and in an experimental use-case. As shown in table 1, complexes are the most common interaction participants in Reactome. The 5,734 invalid complexes were composed of 7,119 molecules (proteins, small molecules and physical entities). These molecules were direct participants of 7,832 interactions. They were direct participants or *direct* component of complexes that were direct participants of 12,148 interactions. The 55% increase demonstrates the relevance of considering complexes. They were direct participants or direct or *indirect* (which amounts to fixing the invalid recursive complexes) component of complexes that were direct participants of 13,211 interactions. The additional 9% increase demonstrates the benefit of fixing invalid complexes. Therefore, these 7,119 molecules participate to 13,211 interactions, 1063 (9%) of which are false negative caused by invalid complexes.

Type of interaction participants	Total number of participants
Complex	**10560**
Protein	7558
Biochemical reaction	7128
Small molecule	2980
Physical entity	1157
DNA	661
RNA	163
Degradation	7
Template reaction	6

Tab. 1. BioPAX Types of interaction participants in Reactome Homo sapiens v79.

To evaluate the benefit of the procedure, we considered eight modules of 7 to 67 unique transcripts of genes (Table 2). These modules were initially identified from an experimental microarray thanks to weighted gene co-expression network analysis [6] describing the correlation patterns among genes across microarray samples, and related to a phenotype of interest in animal production. From these experimental modules, we wanted to retrieve all the interactions associated to one of these genes (via their transcripts or encoded proteins). We compared the number of interactions found when identifying proteins from one of the modules as a direct participant, as the direct component of a complex that was a direct participant (no correction) and as the direct component of a complex after fixing and validation (i.e. by considering invalid complexes). Table 2 shows that more interactions were extracted after fixing the invalid complexes than when considering only the atomic entities, and even more when considering invalid complexes, i.e., using complex SPARQL paths through the *component* property of complexes to obtain their final components.

Module	ID UniProt	Corresponding ProteinReference	Corresponding Protein	Interactions (Q1)	Interactions (Q2)	Interactions (Q3)	Gain
module (1)	96	67	129	58	164	238	180
module (2)	44	27	32	20	73	157	137
module (3)	22	12	59	6	8	15	9
module (4)	14	10	14	5	13	89	84
module (5)	11	9	39	6	12	24	18
module (6)	10	7	18	3	70	120	117
module (7)	9	9	22	7	77	79	72
module (8)	7	49	200	71	137	261	190

Tab. 2. Number of interactions found for each *module* composed of *ID UniProt* proteins by queries. Query1: Interactions these proteins are direct components. Query2: Interactions that have a participant of which the molecule is a direct component. Query3: Interactions that have a participant that is an invalid complex of which the molecule is an indirect component.

5 Discussion

Invalid recursive complexes affect a large portion of Reactome. They constitute 39% of the complexes and participate in 32% of the interactions. They are not detected by the BioPAX validator. Therefore, its is crucial to address this problem.

The procedure developed herein to fix invalid complexes and identify redundant complexes helped to correct non-compliance with BioPAX specifications. By doing so, it also allowed to simplify any further analyses of biological networks, as recursive complexes that should not have existed can now be processed without having to modify standard queries or scripts based on BioPAX libraries such as Paxtools [7] or PyBioPAX [8].

As a first consequence, fixing invalid complexes allowed to decrease the number of false negative when retrieving the interactions to which a molecule of interest participates.

As a second consequence, fixing invalid complexes revealed redundancies that would have been difficult to identify. As Reactome contains mappings to ComplexPortal [9], we could suppose that redundant complexes are associated to the same ComplexPortal identifier. Future work will assess (1) whether all the complexes that are associated to the same ComplexPortal identifier were recognized as redundant with our method, and (2) whether all our predicted redundancies are confirmed by ComplexPortal (and if they are not, we could propose new mappings). However, ComplexPortal seems to incompletely support stoichiometry, which may be a severe limitation.

After addressing the question of the mappings to ComplexPortal, we will study whether our approach is applicable to other resources. The first step will confirm that our queries are compatible with any BioPAX-compliant database. The second step will assess the importance of invalid recursive complexes in the other species of Reactome and in the other major BioPAX pathway databases (Kegg, MetaCYC, wikipathways). Eventually, a longer-term perspective will be to study whether our redundancy queries are relevant for generating mappings when combining several BioPAX databases.

References

[1] Javad Zahiri, Abbasali Emamjomeh, Samaneh Bagheri, Asma Ivazeh, Ghasem Mahdevar, Hessam Sepasi Tehrani, Mehdi Mirzaie, Barat Ali Fakheri, and Morteza Mohammad-Noori. Protein complex prediction: A survey. *Genomics*, 112(1):174–183, 2019.

[2] Victor Spirin and Leonid A Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21):12123–12128, 2003.

[3] Emek Demir, Michael P Cary, Suzanne Paley, Ken Fukuda, and et al. The BioPAX community standard for pathway data sharing. *Nature biotechnology*, 28(9):935–942, 2010.

[4] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, Chuan Deng, Thawfeek Varusai, Eliot Ragueneau, Yusra Haider, Bruce May, Veronica Shamovsky, Joel Weiser, Timothy Brunson, Nasim Sanati, Liam Beckman, Xiang Shao, Antonio Fabregat, Konstantinos Sidiropoulos, Julieth Murillo, Guilherme Viteri, Justin Cook, Solomon Shorser, Gary Bader, Emek Demir, Chris Sander, Robin Haw, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase 2022. *Nucleic acids research*, 2021. In press.

[5] Igor Rodchenkov, Emek Demir, Chris Sander, and Gary D. Bader. The BioPAX Validator. *Bioinformatics (Oxford, England)*, 29(20):2659–2660, October 2013.

[6] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, December 2008.

[7] Emek Demir, Ozgün Babur, Igor Rodchenkov, Bülent Arman Aksoy, Ken I Fukuda, Benjamin Gross, Onur Selçuk Sümer, Gary D Bader, and Chris Sander. Using biological pathway data with paxtools. *PLoS computational biology*, 9(9):e1003194, 2013.

[8] Benjamin M. Gyori and Charles Tapley Hoyt. PyBioPAX: biological pathway exchange in python. *Journal of Open Source Software*, 7(71):4136, 2022.

[9] Birgit H M Meldal, Livia Perfetto, Colin Combe, Tiago Lubiana, João Vitor Ferreira Cavalcante, Hema Bye-A-Jee, Andra Waagmeester, Noemi Del-Toro, Anjali Shrivastava, Elisabeth Barrera, Edith Wong, Bernhard Mlecnik, Gabriela Bindea, Kalpana Panneerselvam, Egon Willighagen, Juri Rappsilber, Pablo Porras, Henning Hermjakob, and Sandra Orchard. Complex portal 2022: new curation frontiers. 2021. In press.

Bibliography

Aksman, L. M., M. A. Scelsi, A. F. Marquand, D. C. Alexander, S. Ourselin, A. Altmann, and for ADNI (Sept. 2019). "Modeling longitudinal imaging biomarkers with parametric Bayesian multi-task learning". In: *Human Brain Mapping* 40.13, pp. 3982–4000. DOI: 10.1002/hbm.24682.

Alanis-Lobato, G., M. A. Andrade-Navarro, and M. H. Schaefer (Jan. 2017). "HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks". In: *Nucleic Acids Res.* 45.D1, pp. D408–D414.

Alonso-López, D., M. A. Gutiérrez, K. P. Lopes, C. Prieto, R. Santamaría, and J. De Las Rivas (2016). "APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks". In: *Nucleic Acids Res* 44.W1, W529–535.

Alonso-López, D., F. J. Campos-Laborie, M. A. Gutiérrez, L. Lambourne, M. A. Calderwood, M. Vidal, and J. De Las Rivas (Jan. 2019). "APID database: redefining protein–protein interaction experimental evidences and binary interactomes". In: *Database* 2019. DOI: 10.1093/database/baz005. eprint: https://academic.oup.com/database/article-pdf/doi/10.1093/database/baz005/27644182/baz005.pdf.

Anders, S. and W. Huber (Oct. 2010). "Differential expression analysis for sequence count data". In: *Genome Biol.* 11.10, R106.

Antelmi, L., N. Ayache, P. Robert, and M. Lorenzi (May 2019). "Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data". In: *International Conference on Machine Learning*. PMLR, pp. 302–311.

Aranda, B., H. Blankenburg, S. Kerrien, F. S. L. Brinkman, A. Ceol, E. Chautard, J. M. Dana, J. De Las Rivas, M. Dumousseau, E. Galeota, A. Gaulton, J. Goll, R. E. W. Hancock, R. Isserlin, R. C. Jimenez, J. Kerssemakers, J. Khadake, D. J. Lynn, M. Michaut, G. O'Kelly, K. Ono, S. Orchard, C. Prieto, S. Razick, O. Rigina, L. Salwinski, M. Simonovic, S. Velankar, A. Winter, G. Wu, G. D. Bader, G. Cesareni, I. M. Donaldson, D. Eisenberg, G. J. Kleywegt, J. Overington, S. Ricard-Blum, M. Tyers, M. Albrecht, and H. Hermjakob (June 2011). "PSICQUIC and PSISCORE: accessing and scoring molecular interactions". In: *Nat. Methods* 8.7, pp. 528–529.

Archetti, D., S. Ingala, V. Venkatraghavan, V. Wottschel, A. L. Young, M. Bellio, E. E. Bron, S. Klein, F. Barkhof, D. C. Alexander, N. P. Oxtoby, G. B. Frisoni, and A. Redolfi (July 2019). "Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in Alzheimer's disease". In: *NeuroImage : Clinical* 24, p. 101954. DOI: 10.1016/j.nicl.2019.101954.

Arroyo, J. D., J. R. Chevillet, E. M. Kroh, I. K. Ruf, C. C. Pritchard, D. F. Gibson, P. S. Mitchell, C. F. Bennett, E. L. Pogosova-Agadjanyan, D. L. Stirewalt, J. F. Tait, and M. Tewari (Mar. 2011). "Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma". In: *Proc. Natl. Acad. Sci. U. S. A.* 108.12, pp. 5003–5008.

Ashburner, M, C. A. Ball, J. A. Blake, D Botstein, H Butler, J. M. Cherry, A. P. Davis, K Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L Issel-Tarver, A Kasarskis, S Lewis, J. C. Matese, J. E. Richardson, M Ringwald, G. M. Rubin, and

G Sherlock (May 2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium". In: *Nat. Genet.* 25.1, pp. 25–29.

Baig, H., P. Fontanarrosa, V. Kulkarni, J. A. McLaughlin, P. Vaidyanathan, B. Bartley, J. Beal, M. Crowther, T. E. Gorochowski, R. Grünberg, G. Misirli, J. Scott-Brown, E. Oberortner, A. Wipat, and C. J. Myers (June 2020). "Synthetic biology open language (SBOL) version 3.0.0". In: *J. Integr. Bioinform.* 17.2-3.

Baker, M. (May 2016). "1,500 scientists lift the lid on reproducibility". In: *Nature* 533.7604, pp. 452–454.

Barba, L. A. (2018). *Terminologies for Reproducible Research*. DOI: 10.48550/ARXIV.1802.03311.

Bauer-Mehren, A., L. I. Furlong, and F. Sanz (2009). "Pathway databases and tools for their exploitation: benefits, current limitations and challenges". In: *Mol Syst Biol* 5, p. 290.

Belcour, A., J. Girard, M. Aite, L. Delage, C. Trottier, C. Marteau, C. Leroux, S. M. Dittami, P. Sauleau, E. Corre, J. Nicolas, C. Boyen, C. Leblanc, J. Collén, A. Siegel, and G. V. Markov (Feb. 2020). "Inferring biochemical reactions and metabolite structures to understand metabolic pathway drift". In: *iScience* 23.2, p. 100849.

Blake, J. A. and C. J. Bult (2006). "Beyond the data deluge: Data integration and bio-ontologies". In: *Journal of Biomedical Informatics* 39.3, pp. 314–320. DOI: https://doi.org/10.1016/j.jbi.2006.01.003.

Blavy, P., F. Gondret, S. Lagarrigue, J. van Milgen, and A. Siegel (Mar. 2014). "Using a large-scale knowledge database on reactions and regulations to propose key upstream regulators of various sets of molecules participating in cell metabolism". In: *BMC Syst. Biol.* 8.1, p. 32.

Bourgon, R., R. Gentleman, and W. Huber (May 2010). "Independent filtering increases detection power for high-throughput experiments". In: *Proc. Natl. Acad. Sci. U. S. A.* 107.21, pp. 9546–9551.

Braun, P., M. Tasan, M. Dreze, M. Barrios-Rodiles, I. Lemmens, H. Yu, J. M. Sahalie, R. R. Murray, L. Roncari, A.-S. de Smet, K. Venkatesan, J.-F. Rual, J. Vandenhaute, M. E. Cusick, T. Pawson, D. E. Hill, J. Tavernier, J. L. Wrana, F. P. Roth, and M. Vidal (Jan. 2009). "An experimentally derived confidence score for binary protein-protein interactions". In: *Nat. Methods* 6.1, pp. 91–97.

Brennan, S., M. Keon, B. Liu, Z. Su, and N. K. Saksena (Nov. 2019). "Panoramic Visualization of Circulating MicroRNAs Across Neurodegenerative Diseases in Humans". In: *Molecular Neurobiology* 56.11, pp. 7380–7407. DOI: 10.1007/s12035-019-1615-1.

Buchka, S., A. Hapfelmeier, P. P. Gardner, R. Wilson, and A. L. Boulesteix (May 2021). "On the optimistic performance evaluation of newly introduced bioinformatic methods". In: *Genome Biol* 22.1, p. 152.

Cannata, N., E. Merelli, and R. B. Altman (2005). "Time to organize the bioinformatics resourceome". In: *PLoS Comput Biol* 1.7, e76.

Chadebec, C., E. Thibeau-Sutre, N. Burgos, and S. Allassonnière (Apr. 2021). "Data Augmentation in High Dimensional Low Sample Size Setting Using a Geometry-Based Variational Autoencoder". In: *arXiv:2105.00026 [cs, stat]*.

Chibon, F. (May 2013). "Cancer gene expression signatures - the rise and fall?" In: *Eur. J. Cancer* 49.8, pp. 2000–2009.

Chindelevitch, L., P.-R. Loh, A. Enayetallah, B. Berger, and D. Ziemek (Feb. 2012). "Assessing statistical significance in causal graphs". In: *BMC Bioinformatics* 13.1, p. 35.

Claerbout J, K. M. (1992). "Electronic documents give reproducible research a new meaning". In: *Proceedings of the 62nd Annual International Meeting of the Society of Exploration Geophysics*, pp. 601 –604.

Clerx, M., M. T. Cooling, J. Cooper, A. Garny, K. Moyle, D. P. Nickerson, P. M. F. Nielsen, and H. Sorby (July 2020). "CellML 2.0". In: *J. Integr. Bioinform.* 17.2-3.

Community, T. T. W. (July 2022). *The Turing Way: A handbook for reproducible, ethical and collaborative research*. Version 1.0.2. DOI: 10.5281/zenodo.6909298.

Csardi, G. and T. Nepusz (2006). "The igraph software package for complex network research". In: *InterJournal* Complex Systems, p. 1695.

De Felice, B., A. Annunziata, G. Fiorentino, M. Borra, E. Biffali, C. Coppola, R. Cotrufo, J. Brettschneider, M. L. Giordana, T. Dalmay, G. Wheeler, and R. D'Alessandro (Oct. 2014). "miR-338-3p is over-expressed in blood, CFS, serum and spinal cord from sporadic amyotrophic lateral sclerosis patients". In: *neurogenetics* 15.4, pp. 243–253. DOI: 10.1007/s10048-014-0420-2.

Dekker, I., M. M. Schoonheim, V. Venkatraghavan, A. J. Eijlers, I. Brouwer, E. E. Bron, S. Klein, M. P. Wattjes, A. M. Wink, J. J. Geurts, B. M. Uitdehaag, N. P. Oxtoby, D. C. Alexander, H. Vrenken, J. Killestein, F. Barkhof, and V. Wottschel (Dec. 2020). "The sequence of structural, functional and cognitive changes in multiple sclerosis". In: *NeuroImage : Clinical* 29, p. 102550. DOI: 10.1016/j.nicl.2020.102550.

Delgado, F. M. and F. Gómez-Vela (Apr. 2019). "Computational methods for Gene Regulatory Networks reconstruction and analysis: A review". In: *Artif. Intell. Med.* 95, pp. 133–145.

Demir, E., O. Babur, I. Rodchenkov, B. A. Aksoy, K. I. Fukuda, B. Gross, O. S. Sümer, G. D. Bader, and C. Sander (2013). "Using biological pathway data with paxtools". In: *PLoS computational biology* 9.9, e1003194.

Demir, E., M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'Eustachio, C. Schaefer, J. Luciano, F. Schacherer, I. Martinez-Flores, Z. Hu, V. Jimenez-Jacinto, G. Joshi-Tope, K. Kandasamy, A. C. Lopez-Fuentes, H. Mi, E. Pichler, I. Rodchenkov, A. Splendiani, S. Tkachev, J. Zucker, G. Gopinath, H. Rajasimha, R. Ramakrishnan, I. Shah, M. Syed, N. Anwar, O. Babur, M. Blinov, E. Brauner, D. Corwin, S. Donaldson, F. Gibbons, R. Goldberg, P. Hornbeck, A. Luna, P. Murray-Rust, E. Neumann, O. Ruebenacker, M. Samwald, M. van Iersel, S. Wimalaratne, K. Allen, B. Braun, M. Whirl-Carrillo, K.-H. Cheung, K. Dahlquist, A. Finney, M. Gillespie, E. Glass, L. Gong, R. Haw, M. Honig, O. Hubaut, D. Kane, S. Krupa, M. Kutmon, J. Leonard, D. Marks, D. Merberg, V. Petri, A. Pico, D. Ravenscroft, L. Ren, N. Shah, M. Sunshine, R. Tang, R. Whaley, S. Letovksy, K. H. Buetow, A. Rzhetsky, V. Schachter, B. S. Sobral, U. Dogrusoz, S. McWeeney, M. Aladjem, E. Birney, J. Collado-Vides, S. Goto, M. Hucka, N. Le Novère, N. Maltsev, A. Pandey, P. Thomas, E. Wingender, P. D. Karp, C. Sander, and G. D. Bader (Sept. 2010a). "The BioPAX community standard for pathway data sharing". In: *Nat. Biotechnol.* 28.9, pp. 935–942.

Demir, E., M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'Eustachio, C. a. Schaefer, F. Schacherer, I. Martinez-Flores, Z. Hu, V. Jimenez-Jacinto, G. Joshi-Tope, K. Kandasamy, A. C. Lopez-Fuentes, E. Mi Huchler, I. Rodchenkov, A. Splendiani, S. Tkachev, J. Zucker, G. Gopinath, H. Rajasimha, R. Ramakrishnan, I. Shah, N. Syeand Anwar, O. Babur, M. Blinov, E. Brauner, D. Corwin, S. Donaldson, F. Gibbons, R. Goldberg, P. Hornbeck, P. Luna Augustiy-Rust, E. Neumann, O. Ruebenacker, O. Reubenacker, M. Samwald, M. van Iersel, S. Wimalaratne, K. Allen, M. Braun Burk andillo, K.-H. Cheung, K. Dahlquist, A. Finney, M. Gillespie, E. Glass, L. Gong, R. Haw, M. Honig, O. Hubaut, S. Kand

Krupa, M. Kutmon, J. Leonard, D. Marks, D. Merberg, V. Petri, A. Pico, D. Ravenscroft, L. Ren, N. Shah, Sugot, R. Tang, R. Whaley, S. Letovksy, K. H. Buetow, A. Rzhetsky, V. Schachter, B. S. Sobral, U. Dogrusoz, S. a. M. McWeeney, E. Birney, J. Collado-Vides, S. Goto, M. Hucka, N. Le Novère, N. Maltsev, A. Pandey, P. Thomas, E. P. D. Wingender, C. Sander, and G. D. Bader (2010b). "The BioPAX community standard for pathway data sharing". In: *Nature biotechnology* 28.9, pp. 935–942.

Denk, J., F. Oberhauser, J. Kornhuber, J. Wiltfang, K. Fassbender, M. L. Schroeter, A. E. Volk, J. Diehl-Schmid, J. Prudlo, A. Danek, B. Landwehrmeyer, M. Lauer, M. Otto, H. Jahn, and FTLDc study group (2018). "Specific serum and CSF microRNA profiles distinguish sporadic behavioural variant of frontotemporal dementia compared with Alzheimer patients and cognitively healthy controls". In: *PloS One* 13.5, e0197329. DOI: 10.1371/journal.pone.0197329.

Diehl, A. G. and A. P. Boyle (Apr. 2016). "Deciphering ENCODE". In: *Trends Genet.* 32.4, pp. 238–249.

Dobrowolny, G., J. Martone, E. Lepore, I. Casola, A. Petrucci, M. Inghilleri, M. Morlando, A. Colantoni, B. M. Scicchitano, A. Calvo, G. Bisogni, A. Chiò, M. Sabatelli, I. Bozzoni, and A. Musarò (Jan. 2021). "A longitudinal study defined circulating microRNAs as reliable biomarkers for disease prognosis and progression in ALS human patients". In: *Cell Death Discovery* 7, p. 4. DOI: 10.1038/s41420-020-00397-6.

Ende, E. L. van der, E. E. Bron, J. M. Poos, L. C. Jiskoot, J. L. Panman, J. M. Papma, et al. (Oct. 2021). "A data-driven disease progression model of fluid biomarkers in genetic frontotemporal dementia". In: *Brain: A Journal of Neurology*, awab382. DOI: 10.1093/brain/awab382.

Eshaghi, A., R. V. Marinescu, A. L. Young, N. C. Firth, F. Prados, M Jorge Cardoso, C. Tur, F. De Angelis, N. Cawley, W. J. Brownlee, N. De Stefano, M Laura Stromillo, M. Battaglini, S. Ruggieri, C. Gasperini, et al. (June 2018). "Progression of regional grey matter atrophy in multiple sclerosis". In: *Brain* 141.6, pp. 1665–1677. DOI: 10.1093/brain/awy088.

Evans, R., M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, and D. Hassabis (2022). "Protein complex prediction with AlphaFold-Multimer". In: *bioRxiv*. DOI: 10.1101/2021.10.04.463034. eprint: https://www.biorxiv.org/content/early/2022/03/10/2021.10.04.463034.full.pdf.

Fabregat, A., K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay, L. Matthews, B. May, M. Milacic, K. Rothfels, V. Shamovsky, M. Webber, J. Weiser, M. Williams, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio (Jan. 2016). "The reactome pathway knowledgebase". In: *Nucleic Acids Res.* 44.D1, pp. D481–7.

Filangi, O., Y. Beausse, A. Assi, L. Legrand, J.-M. Larré, V. Martin, O. Collin, C. Caron, H. Leroy, and D. Allouche (Aug. 2008). "BioMAJ: a flexible framework for databanks synchronization and processing". In: *Bioinformatics* 24.16, pp. 1823–1825.

Firth, N. C., S. Primativo, E. Brotherhood, A. L. Young, K. X. Yong, S. J. Crutch, D. C. Alexander, and N. P. Oxtoby (July 2020). "Sequences of cognitive decline in typical Alzheimer's disease and posterior cortical atrophy estimated using a novel event-based model of disease progression". In: *Alzheimer's & Dementia* 16.7, pp. 965–973. DOI: 10.1002/alz.12083.

Fonteijn, H. M., M. Modat, M. J. Clarkson, J. Barnes, M. Lehmann, N. Z. Hobbs, R. I. Scahill, S. J. Tabrizi, S. Ourselin, N. C. Fox, and D. C. Alexander (Apr. 2012).

"An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease". In: *NeuroImage* 60.3, pp. 1880–1889. DOI: 10.1016/j.neuroimage.2012.01.062.

Freischmidt, A., A. Goswami, K. Limm, V. L. Zimyanin, M. Demestre, H. Glaß, K. Holzmann, A. M. Helferich, S. J. Brockmann, P. Tripathi, A. Yamoah, I. Poser, P. J. Oefner, T. M. Böckers, E. Aronica, A. C. Ludolph, P. M. Andersen, A. Hermann, J. Weis, J. Reinders, K. M. Danzer, and J. H. Weishaupt (2021). "A serum microRNA sequence reveals fragile X protein pathology in amyotrophic lateral sclerosis". In: *Brain* 144.4, pp. 1214–1229. DOI: 10.1093/brain/awab018.

Freischmidt, A., K. Müller, L. Zondler, P. Weydt, B. Mayer, C. A. F. von Arnim, A. Hübers, J. Dorst, M. Otto, K. Holzmann, A. C. Ludolph, K. M. Danzer, and J. H. Weishaupt (Sept. 2015). "Serum microRNAs in sporadic amyotrophic lateral sclerosis". In: *Neurobiology of Aging* 36.9, 2660.e15–2660.e20. DOI: 10.1016/j.neurobiolaging.2015.06.003.

Freischmidt, A., K. Müller, L. Zondler, P. Weydt, A. E. Volk, A. L. Božič, M. Walter, M. Bonin, B. Mayer, C. A. F. von Arnim, M. Otto, C. Dieterich, K. Holzmann, P. M. Andersen, A. C. Ludolph, K. M. Danzer, and J. H. Weishaupt (Nov. 2014). "Serum microRNAs in patients with genetic amyotrophic lateral sclerosis and pre-manifest mutation carriers". In: *Brain: A Journal of Neurology* 137.Pt 11, pp. 2938–2950. DOI: 10.1093/brain/awu249.

Gabel, M. C., R. J. Broad, A. L. Young, S. Abrahams, M. E. Bastin, R. A. L. Menke, A. Al-Chalabi, L. H. Goldstein, S. Tsermentseli, D. C. Alexander, M. R. Turner, P. N. Leigh, and M. Cercignani (May 2020). "Evolution of white matter damage in amyotrophic lateral sclerosis". In: *Annals of Clinical and Translational Neurology* 7.5, pp. 722–732. DOI: 10.1002/acn3.51035.

Garbarino, S., M. Lorenzi, N. P. Oxtoby, E. J. Vinke, R. V. Marinescu, A. Eshaghi, M. A. Ikram, W. J. Niessen, O. Ciccarelli, F. Barkhof, J. M. Schott, M. W. Vernooij, and D. C. Alexander (2019). "Differences in topological progression profile among neurodegenerative diseases from imaging data". In: *eLife* 8. Ed. by F. P. de Lange, e49298. DOI: 10.7554/eLife.49298.

Gavin, A.-C., M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga (Jan. 2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes". In: *Nature* 415.6868, pp. 141–147.

Gillespie, M., B. Jassal, R. Stephan, M. Milacic, K. Rothfels, A. Senff-Ribeiro, J. Griss, C. Sevilla, L. Matthews, C. Gong, C. Deng, T. Varusai, E. Ragueneau, Y. Haider, B. May, V. Shamovsky, J. Weiser, T. Brunson, N. Sanati, L. Beckman, X. Shao, A. Fabregat, K. Sidiropoulos, J. Murillo, G. Viteri, J. Cook, S. Shorser, G. Bader, E. Demir, C. Sander, R. Haw, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio (2021). "The reactome pathway knowledgebase 2022". In: *Nucleic acids research*. DOI: https://doi.org/10.1093/nar/gkab1028.

Gleeson, P., S. Crook, R. C. Cannon, M. L. Hines, G. O. Billings, M. Farinella, T. M. Morse, A. P. Davison, S. Ray, U. S. Bhalla, S. R. Barnes, Y. D. Dimitrova, and R. A. Silver (June 2010). "NeuroML: a language for describing data driven models of neurons and networks with a high degree of biological detail". In: *PLoS Comput. Biol.* 6.6, e1000815.

Grasso, M., P. Piscopo, A. Crestini, A. Confaloni, and M. A. Denti (2015). "Circulating microRNAs in Neurodegenerative Diseases". In: *Circulating microRNAs in Disease Diagnostics and their Potential Biological Relevance*. Ed. by P. Igaz. Experientia Supplementum. Basel: Springer, pp. 151–169. DOI: 10.1007/978-3-0348-0955-9_7.

Grasso, M., P. Piscopo, G. Talarico, L. Ricci, A. Crestini, G. Tosto, M. Gasparini, G. Bruno, M. A. Denti, and A. Confaloni (Dec. 2019). "Plasma microRNA profiling distinguishes patients with frontotemporal dementia from healthy subjects". In: *Neurobiology of Aging* 84, 240.e1–240.e12. DOI: 10.1016/j.neurobiolaging.2019.01.024.

Hafner, M., P. Landgraf, J. Ludwig, A. Rice, T. Ojo, C. Lin, D. Holoch, C. Lim, and T. Tuschl (Jan. 2008). "Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing". In: *Methods* 44.1, pp. 3–12.

Hagberg, A. A., D. A. Schult, and P. J. Swart (2008). "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference*. Ed. by G. Varoquaux, T. Vaught, and J. Millman. Pasadena, CA USA, pp. 11 –15.

Hastie, T. and W. Stuetzle (1989). "Principal Curves". In: *Journal of the American Statistical Association* 84.406, pp. 502–516. DOI: 10.1080/01621459.1989.10478797.

Hermjakob, H., L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K Shanker, S. G. N. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, and R. Apweiler (Feb. 2004). "The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data". In: *Nat. Biotechnol.* 22.2, pp. 177–183.

Heroux, M. A., L. Barba, M. Parashar, V. Stodden, and M. Taufer (n.d.). "Toward a Compatible Reproducibility Taxonomy for Computational and Computing Sciences." In: (). DOI: 10.2172/1481626.

Ho, Y., A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sørensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, and M. Tyers (Jan. 2002). "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry". In: *Nature* 415.6868, pp. 180–183.

Ioannidis, J. P. A. (Aug. 2005). "Why most published research findings are false". In: *PLoS Med.* 2.8, e124.

Ito, T, T Chiba, R Ozawa, M Yoshida, M Hattori, and Y Sakaki (Apr. 2001). "A comprehensive two-hybrid analysis to explore the yeast protein interactome". In: *Proc. Natl. Acad. Sci. U. S. A.* 98.8, pp. 4569–4574.

Jedynak, B. M., A. Lang, B. Liu, E. Katz, Y. Zhang, B. T. Wyman, D. Raunig, C. P. Jedynak, B. Caffo, and J. L. Prince (Nov. 2012). "A computational neurodegenerative disease progression score: Method and results with the Alzheimer's disease neuroimaging initiative cohort". In: *NeuroImage* 63.3, pp. 1478–1486. DOI: 10.1016/j.neuroimage.2012.07.059.

Jelizarow, M., V. Guillemot, A. Tenenhaus, K. Strimmer, and A. L. Boulesteix (2010). "Over-optimism in bioinformatics: an illustration". In: *Bioinformatics* 26.16, pp. 1990–1998.

Jiang, H. and A. M. English (July 2002). "Quantitative analysis of the yeast proteome by incorporation of isotopically labeled leucine". In: *J. Proteome Res.* 1.4, pp. 345–350.

Karimi, E., E. Geslain, A. Belcour, C. Frioux, M. Aïte, A. Siegel, E. Corre, and S. M. Dittami (May 2021). "Robustness analysis of metabolic predictions in algal microbial communities based on different annotation pipelines". In: *PeerJ* 9, e11344.

Karp, P. D., C. A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahrén, S. Tsoka, N. Darzentas, V. Kunin, and N. López-Bigas (Oct. 2005). "Expansion of the BioCyc collection of pathway/genome databases to 160 genomes". In: *Nucleic Acids Res.* 33.19, pp. 6083–6089.

Keating, S. M., D. Waltemath, M. König, F. Zhang, A. Dräger, C. Chaouiya, F. T. Bergmann, A. Finney, C. S. Gillespie, T. Helikar, S. Hoops, R. S. Malik-Sheriff, S. L. Moodie, I. I. Moraru, C. J. Myers, A. Naldi, B. G. Olivier, S. Sahle, J. C. Schaff, L. P. Smith, M. J. Swat, D. Thieffry, L. Watanabe, D. J. Wilkinson, M. L. Blinov, K. Begley, J. R. Faeder, H. F. Gómez, T. M. Hamm, Y. Inagaki, W. Liebermeister, A. L. Lister, D. Lucio, E. Mjolsness, C. J. Proctor, K. Raman, N. Rodriguez, C. A. Shaffer, B. E. Shapiro, J. Stelling, N. Swainston, N. Tanimura, J. Wagner, M. Meier-Schellersheim, H. M. Sauro, B. Palsson, H. Bolouri, H. Kitano, A. Funahashi, H. Hermjakob, J. C. Doyle, M. Hucka, and SBML Level 3 Community members (Aug. 2020). "SBML Level 3: an extensible format for the exchange and reuse of biological models". In: *Mol. Syst. Biol.* 16.8, e9110.

Kerrien, S., B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob (2012). "The IntAct molecular interaction database in 2012". In: *Nucleic Acids Res* 40.Database issue, pp. D841–846.

Kerrien, S., S. Orchard, L. Montecchi-Palazzi, B. Aranda, A. F. Quinn, N. Vinod, G. D. Bader, I. Xenarios, J. Wojcik, D. Sherman, M. Tyers, J. J. Salama, S. Moore, A. Ceol, A. Chatr-Aryamontri, M. Oesterheld, V. Stümpflen, L. Salwinski, J. Nerothin, E. Cerami, M. E. Cusick, M. Vidal, M. Gilson, J. Armstrong, P. Woollard, C. Hogue, D. Eisenberg, G. Cesareni, R. Apweiler, and H. Hermjakob (2007). "Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions". In: *BMC Biol* 5, p. 44.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg (Apr. 2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome Biol.* 14.4, R36.

Klein, A. M., L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner (May 2015). "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells". In: *Cell* 161.5, pp. 1187–1201.

Kmetzsch, V., V. Anquetil, D. Saracino, D. Rinaldi, A. Camuzat, T. Gareau, L. Jornea, S. Forlani, P. Couratier, D. Wallon, F. Pasquier, N. Robil, P. d. l. Grange, I. Moszer, I. L. Ber, O. Colliot, and E. Becker (May 2021). "Plasma microRNA signature in presymptomatic and symptomatic subjects with C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis". In: *Journal of Neurology, Neurosurgery & Psychiatry* 92.5, pp. 485–493. DOI: 10.1136/jnnp-2020-324647.

Kolodziejczyk, A. A., J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann (May 2015). "The technology and biology of single-cell RNA sequencing". In: *Mol. Cell* 58.4, pp. 610–620.

Koval, I., A. Bône, M. Louis, T. Lartigue, S. Bottani, A. Marcoux, J. Samper-González, N. Burgos, B. Charlier, A. Bertrand, S. Epelbaum, O. Colliot, S. Allassonnière, and S. Durrleman (Apr. 2021). "AD Course Map charts Alzheimer's disease progression". In: *Scientific Reports* 11.1, p. 8020. DOI: 10.1038/s41598-021-87434-1.

Langfelder, P. and S. Horvath (Dec. 2008). "WGCNA: an R package for weighted correlation network analysis". In: *BMC Bioinformatics* 9.1, p. 559.

Le Novère, N., M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M. I. Aladjem, S. M. Wimalaratne, F. T. Bergman, R. Gauges, P. Ghazal, H. Kawaji, L. Li, Y. Matsuoka, A. Villéger, S. E. Boyd, L. Calzone, M. Courtot, U. Dogrusoz, T. C. Freeman, A. Funahashi, S. Ghosh, A. Jouraku, S. Kim, F. Kolpakov, A. Luna, S. Sahle, E. Schmidt, S. Watterson, G. Wu, I. Goryanin, D. B. Kell, C. Sander, H. Sauro, J. L. Snoep, K. Kohn, and H. Kitano (Aug. 2009). "The Systems Biology Graphical Notation". In: *Nat. Biotechnol.* 27.8, pp. 735–741.

Lemmens, K., T. Dhollander, T. De Bie, P. Monsieurs, K. Engelen, B. Smets, J. Winderickx, B. De Moor, and K. Marchal (May 2006). "Inferring transcriptional modules from ChIP-chip, motif and microarray data". In: *Genome Biol.* 7.5, R37.

Leng, N., J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. G. Smits, J. D. Haag, M. N. Gould, R. M. Stewart, and C. Kendziorski (Apr. 2013). "EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments". In: *Bioinformatics* 29.8, pp. 1035–1043.

Li, D., M. S. Zand, T. D. Dye, M. L. Goniewicz, I. Rahman, and Z. Xie (Sept. 2022). "An evaluation of RNA-seq differential analysis methods". In: *PLoS One* 17.9, e0264246.

Li, J. and R. Tibshirani (Oct. 2013). "Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data". In: *Stat. Methods Med. Res.* 22.5, pp. 519–536.

Liu, S., Z. Wang, R. Zhu, F. Wang, Y. Cheng, and Y. Liu (Sept. 2021). "Three differential expression analysis methods for RNA sequencing: Limma, EdgeR, DESeq2". In: *J. Vis. Exp.* 175.

Lorenzi, M., M. Filippone, D. C. Alexander, and S. Ourselin (Apr. 2019). "Disease Progression Modeling and Prediction through Random Effect Gaussian Processes and Time Transformation". In: *NeuroImage* 190, pp. 56–68. DOI: 10.1016/j.neuroimage.2017.08.059.

Love, M. I., W. Huber, and S. Anders (2014a). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biol.* 15.12, p. 550.

— (2014b). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biol.* 15.12, p. 550.

Macosko, E. Z., A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll (May 2015). "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets". In: *Cell* 161.5, pp. 1202–1214.

Magen, I., N. S. Yacovzada, E. Yanowski, A. Coenen-Stass, J. Grosskreutz, C.-H. Lu, L. Greensmith, A. Malaspina, P. Fratta, and E. Hornstein (Nov. 2021). "Circulating miR-181 is a prognostic biomarker for amyotrophic lateral sclerosis". In: *Nature Neuroscience* 24.11, pp. 1534–1541. DOI: 10.1038/s41593-021-00936-z.

Marinescu, R. V., A. Eshaghi, M. Lorenzi, A. L. Young, N. P. Oxtoby, S. Garbarino, S. J. Crutch, D. C. Alexander, and Alzheimer's Disease Neuroimaging Initiative

(May 2019). "DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders". In: *NeuroImage* 192, pp. 166–177. DOI: 10.1016/j.neuroimage.2019.02.053.

Marx, V. (June 2013). "Biology: The big challenges of big data". In: *Nature* 498.7453, pp. 255–260.

McCarthy, D. J., Y. Chen, and G. K. Smyth (May 2012). "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation". In: *Nucleic Acids Res.* 40.10, pp. 4288–4297.

Mehdipour Ghazi, M., M. Nielsen, A. Pai, M. J. Cardoso, M. Modat, S. Ourselin, and L. Sørensen (Apr. 2019). "Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling". In: *Medical Image Analysis* 53, pp. 39–46. DOI: 10.1016/j.media.2019.01.004.

Mehdipour Ghazi, M., M. Nielsen, A. Pai, M. Modat, M. Jorge Cardoso, S. Ourselin, and L. Sørensen (Jan. 2021). "Robust parametric modeling of Alzheimer's disease progression". In: *NeuroImage* 225, p. 117460. DOI: 10.1016/j.neuroimage.2020.117460.

Mi, H., A. Muruganujan, J. T. Casagrande, and P. D. Thomas (Aug. 2013). "Large-scale gene function analysis with the PANTHER classification system". In: *Nat. Protoc.* 8.8, pp. 1551–1566.

Mitchell, P. S., R. K. Parkin, E. M. Kroh, B. R. Fritz, S. K. Wyman, E. L. Pogosova-Agadjanyan, A. Peterson, J. Noteboom, K. C. O'Briant, A. Allen, D. W. Lin, N. Urban, C. W. Drescher, B. S. Knudsen, D. L. Stirewalt, R. Gentleman, R. L. Vessella, P. S. Nelson, D. B. Martin, and M. Tewari (July 2008). "Circulating microRNAs as stable blood-based markers for cancer detection". In: *Proc. Natl. Acad. Sci. U. S. A.* 105.30, pp. 10513–10518.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold (July 2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nat. Methods* 5.7, pp. 621–628.

Nurse, P. (Sept. 2021). "Biology must generate ideas as well as data". In: *Nature* 597.7876, p. 305.

Oda, Y, K Huang, F. R. Cross, D Cowburn, and B. T. Chait (June 1999). "Accurate quantitation of protein expression and site-specific phosphorylation". In: *Proc. Natl. Acad. Sci. U. S. A.* 96.12, pp. 6591–6596.

Ong, S.-E., B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann (May 2002). "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics". In: *Mol. Cell. Proteomics* 1.5, pp. 376–386.

Orchard, S., S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. S. Brinkman, F. Brinkman, G. Cesareni, A. Chatr-aryamontri, E. Chautard, C. Chen, M. Dumousseau, J. Goll, R. E. Hancock, R. Hancock, L. I. Hannick, I. Jurisica, J. Khadake, D. J. Lynn, U. Mahadevan, L. Perfetto, A. Raghunath, S. Ricard-Blum, B. Roechert, L. Salwinski, V. Stümpflen, M. Tyers, P. Uetz, I. Xenarios, and H. Hermjakob (2012). "Protein interaction data curation: the International Molecular Exchange (IMEx) consortium". In: *Nat Methods* 9.4, pp. 345–350.

Oxtoby, N. P., L.-A. Leyland, L. M. Aksman, G. E. C. Thomas, E. L. Bunting, P. A. Wijeratne, A. L. Young, A. Zarkali, M. M. X. Tan, F. D. Bremner, P. A. Keane, H. R. Morris, A. E. Schrag, D. C. Alexander, and R. S. Weil (Feb. 2021). "Sequence of clinical and neurodegeneration events in Parkinson's disease progression". In: *Brain* 144.3, pp. 975–988. DOI: 10.1093/brain/awaa461.

Oxtoby, N. P., A. L. Young, D. M. Cash, T. L. S. Benzinger, A. M. Fagan, J. C. Morris, R. J. Bateman, N. C. Fox, J. M. Schott, and D. C. Alexander (May 2018).

"Data-driven models of dominantly-inherited Alzheimer's disease progression". In: *Brain* 141.5, pp. 1529–1544. DOI: 10.1093/brain/awy050.

Panman, J. L., V. Venkatraghavan, E. L. v. d. Ende, R. M. E. Steketee, L. C. Jiskoot, J. M. Poos, E. G. P. Dopper, et al. (Jan. 2021). "Modelling the cascade of biomarker changes in GRN-related frontotemporal dementia". In: *Journal of Neurology, Neurosurgery & Psychiatry*. DOI: 10.1136/jnnp-2020-323541.

Peng, R. D., F. Dominici, and S. L. Zeger (May 2006). "Reproducible epidemiologic research". In: *Am. J. Epidemiol.* 163.9, pp. 783–789.

Peri, S., J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. K. B. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K Shanker, H. N. Shivashankar, B. P. Rashmi, M. A. Ramya, Z. Zhao, K. N. Chandrika, N Padma, H. C. Harsha, A. J. Yatish, M. P. Kavitha, M. Menezes, D. R. Choudhury, S. Suresh, N. Ghosh, R Saravana, S. Chandran, S. Krishna, M. Joy, S. K. Anand, V Madavan, A. Joseph, G. W. Wong, W. P. Schiemann, S. N. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G. C. Blobe, C. V. Dang, J. G. N. Garcia, J. Pevsner, O. N. Jensen, P. Roepstorff, K. S. Deshpande, A. M. Chinnaiyan, A. Hamosh, A. Chakravarti, and A. Pandey (Oct. 2003). "Development of human protein reference database as an initial platform for approaching systems biology in humans". In: *Genome Res.* 13.10, pp. 2363–2371.

Picelli, S., Å. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg (Nov. 2013). "Smart-seq2 for sensitive full-length transcriptome profiling in single cells". In: *Nat. Methods* 10.11, pp. 1096–1098.

Piscopo, P., M. Grasso, M. Puopolo, E. D'Acunto, G. Talarico, A. Crestini, M. Gasparini, R. Campopiano, S. Gambardella, A. E. Castellano, G. Bruno, M. A. Denti, and A. Confaloni (Jan. 2018). "Circulating miR-127-3p as a Potential Biomarker for Differential Diagnosis in Frontotemporal Dementia". In: *Journal of Alzheimer's Disease* 65.2, pp. 455–464. DOI: 10.3233/JAD-180364.

Poldrack, R. A., G. Huckins, and G. Varoquaux (May 2020). "Establishment of best practices for evidence for prediction: A review". In: *JAMA Psychiatry* 77.5, pp. 534–540.

Porras, P., E. Barrera, A. Bridge, N. Del-Toro, G. Cesareni, M. Duesbury, H. Hermjakob, M. Iannuccelli, I. Jurisica, M. Kotlyar, L. Licata, R. C. Lovering, D. J. Lynn, B. Meldal, B. Nanduri, K. Paneerselvam, S. Panni, C. Pastrello, M. Pellegrini, L. Perfetto, N. Rahimzadeh, P. Ratan, S. Ricard-Blum, L. Salwinski, G. Shirodkar, A. Shrivastava, and S. Orchard (2020). "Towards a unified open access dataset of molecular interactions". In: *Nature communications* 11.1, p. 6144. DOI: https://doi.org/10.1038/s41467-020-19942-z.

Raheja, R., K. Regev, B. C. Healy, M. A. Mazzola, V. Beynon, F. von Glehn, A. Paul, C. Diaz-Cruz, T. Gholipour, B. I. Glanz, P. Kivisakk, T. Chitnis, H. L. Weiner, J. D. Berry, and R. Gandhi (Aug. 2018). "Correlating serum microRNAs and clinical parameters in Amyotrophic lateral sclerosis". In: *Muscle & nerve* 58.2, pp. 261–269. DOI: 10.1002/mus.26106.

Rau, A., M. Gallopin, G. Celeux, and F. Jaffrézic (Sept. 2013). "Data-based filtering for replicated high-throughput transcriptome sequencing experiments". In: *Bioinformatics* 29.17, pp. 2146–2152.

Reverter, A. and E. K. F. Chan (Nov. 2008). "Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks". In: *Bioinformatics* 24.21, pp. 2491–2497.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth (Jan. 2010a). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1, pp. 139–140.

— (Jan. 2010b). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1, pp. 139–140.

Robinson, M. D. and A. Oshlack (Mar. 2010). "A scaling normalization method for differential expression analysis of RNA-seq data". In: *Genome Biol.* 11.3, R25.

Rodchenkov, I., E. Demir, C. Sander, and G. D. Bader (2013). "The BioPAX Validator". In: *Bioinformatics* 29.20, pp. 2659–2660.

Rodchenkov, I., O. Babur, A. Luna, B. A. Aksoy, J. V. Wong, D. Fong, M. Franz, M. C. Siper, M. Cheung, M. Wrana, H. Mistry, L. Mosier, J. Dlin, Q. Wen, C. O'Callaghan, W. Li, G. Elder, P. T. Smith, C. Dallago, E. Cerami, B. Gross, U. Dogrusoz, E. Demir, G. D. Bader, and C. Sander (Jan. 2020a). "Pathway Commons 2019 Update: integration, analysis and exploration of pathway data". In: *Nucleic Acids Res.* 48.D1, pp. D489–D497.

— (Jan. 2020b). "Pathway Commons 2019 Update: integration, analysis and exploration of pathway data". In: *Nucleic Acids Res.* 48.D1, pp. D489–D497.

Romero, P., J. Wagg, M. L. Green, D. Kaiser, M. Krummenacker, and P. D. Karp (2005). "Computational prediction of human metabolic pathways from the complete human genome". In: *Genome Biol.* 6.1, R2.

Schiratti, J.-B., S. Allassonnière, O. Colliot, and S. Durrleman (2017). "A Bayesian Mixed-Effects Model to Learn Trajectories of Changes from Repeated Manifold-Valued Observations". In: *Journal of Machine Learning Research* 18.133, pp. 1–33.

Schmidt, D., M. D. Wilson, C. Spyrou, G. D. Brown, J. Hadfield, and D. T. Odom (July 2009). "ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions". In: *Methods* 48.3, pp. 240–248.

Schreiber, F., B. Sommer, G. D. Bader, P. Gleeson, M. Golebiewski, M. Hucka, S. M. Keating, M. König, C. Myers, D. Nickerson, and D. Waltemath (July 2019). "Specifications of standards in systems and synthetic biology: Status and developments in 2019". In: *J. Integr. Bioinform.* 16.2.

Schurch, N. J., P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes, M. Blaxter, and G. J. Barton (Oct. 2016). "Erratum: How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?" In: *RNA* 22.10, p. 1641.

Seyednasrollah, F., A. Laiho, and L. L. Elo (Jan. 2015). "Comparison of software packages for detecting differential expression in RNA-seq studies". In: *Brief. Bioinform.* 16.1, pp. 59–70.

Sheinerman, K. S., J. B. Toledo, V. G. Tsivinsky, D. Irwin, M. Grossman, D. Weintraub, H. I. Hurtig, A. Chen-Plotkin, D. A. Wolk, L. F. McCluskey, L. B. Elman, J. Q. Trojanowski, and S. R. Umansky (Nov. 2017). "Circulating brain-enriched microRNAs as novel biomarkers for detection and differentiation of neurodegenerative diseases". In: *Alzheimer's Research & Therapy* 9.1, p. 89. DOI: 10.1186/s13195-017-0316-0.

Singh, A., C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, and K.-A. Lê Cao (Sept. 2019). "DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays". In: *Bioinformatics* 35.17, pp. 3055–3062.

Sivade Dumousseau, M., D. Alonso-López, M. Ammari, G. Bradley, N. H. Campbell, A. Ceol, G. Cesareni, C. Combe, J. De Las Rivas, N. Del-Toro, J. Heimbach, H. Hermjakob, I. Jurisica, M. Koch, L. Licata, R. C. Lovering, D. J. Lynn, B. H. M.

Meldal, G. Micklem, S. Panni, P. Porras, S. Ricard-Blum, B. Roechert, L. Salwinski, A. Shrivastava, J. Sullivan, N. Thierry-Mieg, Y. Yehudi, K. Van Roey, and S. Orchard (Apr. 2018). "Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions". In: *BMC Bioinformatics* 19.1, p. 134.

Soliman, R., N. O. Mousa, H. R. Rashed, R. R. Moustafa, N. Hamdi, A. Osman, and N. Fahmy (Sept. 2021). "Assessment of diagnostic potential of some circulating microRNAs in Amyotrophic Lateral Sclerosis Patients, an Egyptian study". In: *Clinical Neurology and Neurosurgery* 208, p. 106883. DOI: 10.1016/j.clineuro.2021.106883.

Soneson, C. and M. Delorenzi (Mar. 2013). "A comparison of methods for differential expression analysis of RNA-seq data". In: *BMC Bioinformatics* 14.1, p. 91.

Sreedharan, J., I. P. Blair, V. B. Tripathi, X. Hu, C. Vance, B. Rogelj, S. Ackerley, J. C. Durnall, K. L. Williams, E. Buratti, F. Baralle, J. de Belleroche, J. D. Mitchell, P. N. Leigh, A. Al-Chalabi, C. C. Miller, G. Nicholson, and C. E. Shaw (Mar. 2008). "TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis". In: *Science* 319.5870, pp. 1668–1672.

Stephens, Z. D., S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson (July 2015). "Big Data: Astronomical or genomical?" In: *PLoS Biol.* 13.7, e1002195.

Svensson, V., R. Vento-Tormo, and S. A. Teichmann (Apr. 2018). "Exponential scaling of single-cell RNA-seq in the past decade". In: *Nat. Protoc.* 13.4, pp. 599–604.

Szklarczyk, D., A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, L. J. Jensen, and C. von Mering (Jan. 2021). "The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets". In: *Nucleic Acids Res.* 49.D1, pp. D605–D612.

Takahashi, I., Y. Hama, M. Matsushima, M. Hirotani, T. Kano, H. Hohzen, I. Yabe, J. Utsumi, and H. Sasaki (Oct. 2015). "Identification of plasma microRNAs as a biomarker of sporadic Amyotrophic Lateral Sclerosis". In: *Molecular Brain* 8.1, p. 67. DOI: 10.1186/s13041-015-0161-7.

Tarazona, S., F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa (Dec. 2011). "Differential expression in RNA-seq: a matter of depth". In: *Genome Res.* 21.12, pp. 2213–2223.

Tasca, E., V. Pegoraro, A. Merico, and C. Angelini (Jan. 2016). "Circulating microRNAs as biomarkers of muscle differentiation and atrophy in ALS". In: *Clinical Neuropathology* 35.01, pp. 22–30. DOI: 10.5414/NP300889.

The Gene Ontology Consortium (Jan. 2019). "The Gene Ontology Resource: 20 years and still GOing strong". In: *Nucleic Acids Res.* 47.D1, pp. D330–D338.

Thomas, P. D., D. Ebert, A. Muruganujan, T. Mushayahama, L.-P. Albou, and H. Mi (Jan. 2022). "PANTHER: Making genome-scale phylogenetics accessible to all". In: *Protein Sci.* 31.1, pp. 8–22.

Thomas, P. D., A. Kejariwal, M. J. Campbell, H. Mi, K. Diemer, N. Guo, I. Ladunga, B. Ulitsky-Lazareva, A. Muruganujan, S. Rabkin, J. A. Vandergriff, and O. Doremieux (Jan. 2003). "PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification". In: *Nucleic Acids Res.* 31.1, pp. 334–341.

Toro, N. del, M. Dumousseau, S. Orchard, R. C. Jimenez, E. Galeota, G. Launay, J. Goll, K. Breuer, K. Ono, L. Salwinski, and H. Hermjakob (2013). "A new reference implementation of the PSICQUIC web service". In: *Nucleic Acids Res* 41.Web Server issue, W601–606.

Trapnell, C., D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter (Jan. 2013). "Differential analysis of gene regulation at transcript resolution with RNA-seq". In: *Nat. Biotechnol.* 31.1, pp. 46–53.

Uetz, P, L Giot, G Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D Lockshon, V Narayan, M Srinivasan, P Pochart, A Qureshi-Emili, Y Li, B Godwin, D Conover, T Kalbfleisch, G Vijayadamodar, M Yang, M Johnston, S Fields, and J. M. Rothberg (Feb. 2000). "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae". In: *Nature* 403.6770, pp. 623–627.

Ullmann, T., S. Peschel, P. Finger, C. L. Müller, and A.-L. Boulesteix (2022). "Over-optimism in unsupervised microbiome analysis: Insights from network learning and clustering". In: *bioRxiv*. DOI: 10.1101/2022.06.24.497500. eprint: https://www.biorxiv.org/content/early/2022/06/28/2022.06.24.497500.full.pdf.

Valdeolivas, A., L. Tichit, C. Navarro, S. Perrin, G. Odelin, N. Levy, P. Cau, E. Remy, and A. Baudot (Feb. 2019). "Random walk with restart on multiplex and heterogeneous biological networks". In: *Bioinformatics* 35.3, pp. 497–505.

Venkatraghavan, V., E. E. Bron, W. J. Niessen, and S. Klein (Feb. 2019). "Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling". In: *NeuroImage* 186, pp. 518–532. DOI: 10.1016/j.neuroimage.2018.11.024.

Villaveces, J. M., R. C. Jiménez, P. Porras, N. del Toro, M. Duesbury, M. Dumousseau, S. Orchard, H. Choi, P. Ping, N. C. Zong, M. Askenazi, B. H. Habermann, and H. Hermjakob (Feb. 2015). "Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study". In: *Database* 2015. DOI: 10.1093/database/bau131. eprint: https://academic.oup.com/database/article-pdf/doi/10.1093/database/bau131/7298497/bau131.pdf.

Wagner, A., A. Regev, and N. Yosef (Nov. 2016). "Revealing the vectors of cellular identity with single-cell genomics". In: *Nat. Biotechnol.* 34.11, pp. 1145–1160.

Waller, R., E. F. Goodall, M. Milo, J. Cooper-Knock, M. Da Costa, E. Hobson, M. Kazoka, H. Wollff, P. R. Heath, P. J. Shaw, and J. Kirby (July 2017). "Serum miRNAs miR-206, 143-3p and 374b-5p as potential biomarkers for amyotrophic lateral sclerosis (ALS)". In: *Neurobiology of Aging* 55, pp. 123–131. DOI: 10.1016/j.neurobiolaging.2017.03.027.

Wang, Z., M. Gerstein, and M. Snyder (Jan. 2009a). "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10.1, pp. 57–63. DOI: 10.1038/nrg2484.

— (Jan. 2009b). "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nat. Rev. Genet.* 10.1, pp. 57–63.

Wijeratne, P. A., E. B. Johnson, S. Gregory, N. Georgiou-Karistianis, J. S. Paulsen, R. I. Scahill, S. J. Tabrizi, and D. C. Alexander (Aug. 2021). "A Multi-Study Model-Based Evaluation of the Sequence of Imaging and Clinical Biomarker Changes in Huntington's Disease". In: *Frontiers in Big Data* 4, p. 662200. DOI: 10.3389/fdata.2021.662200.

Wilkinson, M. D., M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons (2016). "The FAIR Guiding

Principles for scientific data management and stewardship". In: *Scientific data* 3, p. 160018.

Xu, Q., Y. Zhao, X. Zhou, J. Luan, Y. Cui, and J. Han (Feb. 2018). "Comparison of the extraction and determination of serum exosome and miRNA in serum and the detection of miR-27a-3p in serum exosome of ALS patients". In: *Intractable & Rare Diseases Research* 7.1, pp. 13–18. DOI: 10.5582/irdr.2017.01091.

Young, A. L., N. P. Oxtoby, P. Daga, D. M. Cash, N. C. Fox, S. Ourselin, J. M. Schott, D. C. Alexander, and on behalf of the Alzheimer's Disease Neuroimaging Initiative (Sept. 2014). "A data-driven model of biomarker changes in sporadic Alzheimer's disease". In: *Brain* 137.9, pp. 2564–2577. DOI: 10.1093/brain/awu176.