Université de Montréal

Approches bio-informatiques protéome-centrées pour l'étude des phénotypes complexes

par

Savandara Ladyson Besse

Département De Biochimie et Médecine Moléculaire
Faculté de Médecine

Thèse présentée en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en Bio-informatique

October 3, 2023

Université de Montréal

Faculté de Médecine

Cette thèse intitulée

Approches bio-informatiques protéome-centrées pour l'étude des phénotypes complexes

présentée par

Savandara Ladyson Besse

a été évaluée par un jury composé des personnes suivantes :

Guillaume Lettre, Professeur titulaire (Dpt. Médecine, UdeM)
(Président-rapporteur)

Adrian Serohijos, Professeur agrégé (Dpt. Biochimie et Médecine Moléculaire, UdeM)
(Directeur de recherche)

Julie Hussin, Professeure ajointe (Dpt. Médecine, UdeM)
(Co-directrice)

Rafaël Najmanovich, Professeur titulaire (Dpt. Pharmacologie et Physiologie UdeM)
(Membre du jury)

Marie Brunet, Professeure adjointe (Dpt. Pédiatrie UdeS)
(Examinatrice externe)

Claude Perreault, Professeur titulaire (Dpt. Médecine, UdeM)
(Représentant du doyen de la FESP)

Résumé

Parmi les différents acteurs impliqués dans le dogme de la biologie moléculaire, les protéines sont des unités biologiques fonctionnelles contribuant à de nombreux processus biologiques. Dans la compréhension de la relation génotype-phénotype, il est important d'étudier l'influence de gènes, ou de variants génétiques, sur des mécanismes moléculaires spécifiques, permettant d'expliquer la variance phénotypique de traits dits complexes. Dans cette thèse nous allons démontrer l'intérêt de proposer différentes stratégies bio-informatiques protéome-centrées pour l'étude de phénotypes complexes. Dans une première étude, nous mettons en avant comment l'utilisation de la génomique comparative, couplée à l'analyse de la propension d'agrégation des protéines, permet d'identifier certains groupes de protéines avec des différences significatives entre espèces dans leurs propriétés intrinsèques contribuant à la protéostase cellulaire. Ce mécanisme est proposé dans cette thèse comme hypothèse de travail pour étudier les différences d'espérance de vie chez les rongeurs: ce travail est réalisé sur deux espèces phylogénétiquement proches, le rat taupe-nu et la souris, mais possédant des différences phénotypiques dans le contexte du vieillissement. Dans une seconde étude, nous proposons une nouvelle méthodologie s'appuyant sur l'étude quantitative des réseaux d'interaction protéine-protéine afin d'identifier les déterminants génétiques qui seraient responsables de la variation de ces interactions, suite à une stimulation médicamenteuse dans une population de levures génétiquement diversifiées. Ces travaux de recherche étudient le protéome et ses interactions et permettent de proposer une abstraction originale des phénotypes complexes.

Mots-clefs : **Protéines, Interactions Protéine-Protéine, Phénotypes complexes, Bio-informatique, Biologie des systèmes, Biologie évolutive, Génétique statistique**

Abstract

Among the different actors involved in the dogma of molecular biology, proteins are functional biological units contributing to many biological processes. In the understanding of the genotype-phenotype relationship, it is important to study the influence of genes, or genetic variants, on specific molecular mechanisms, allowing to explain the phenotypic variance of so-called complex traits. In this thesis we will demonstrate the interest of proposing different proteome-centric bioinformatics strategies for the study of complex phenotypes. In a first study, we highlight how the use of comparative genomics, coupled with the analysis of the aggregation propensity of proteins, allows to identify some groups of proteins with significant differences between species in their intrinsic properties contributing to cellular proteostasis. This mechanism is proposed in this thesis as a working hypothesis to study differences in life expectancy in rodents: this work is performed on two phylogenetically related species, the mole rat and the mouse, but with phenotypic differences in the context of aging. In a second study, we propose a new methodology based on the quantitative study of protein-protein interaction networks in order to identify the genetic determinants that would be responsible for the variation of these interactions, following a drug stimulation in a genetically diversified yeast population. This research studies the proteome and its interactions and proposes an original abstraction of complex phenotypes.

Keywords: **Proteins, Protein-protein interaction, Complex phenotypes, Bioinformatics, Systems Biology, Evolutionary Genomics, Statistical Genomics**

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et des abréviations

- 5-FC : 5-FluoroCytosine
- ACP : Analyse en Composantes Principales
- ADN : Acide Désoxyribo-Nucléique
- ARN : Acide Ribo-Nucléique
  - ARNnc : ARN non codants
  - ARNm : ARN messager
  - ARNr : ARN ribosomaux
  - ARNt : ARN de transfert
  - lncARN : ARN long non-codant
  - miARN : micro-ARN
  - snoARN : small nucleolar ARN
- CDS : de l'anglais *CoDing Sequences*, en français Séquences codantes
- CUT : de l'anglais *Cryptic Unstable Tanscript*, en français transcrits cryptiques instables
- DSSP : de l'anglais *Define Secondary Structure of Proteins*
- ELISA : de l'anglais *Enzyme-linked immunoSorbent Assay*
- ERAD : de l'anglais *Endoplasmic-Reticulum-Associated protein Degradation*
- ERO : Espèces Réactives de l'Oxygène
- FDR : de l'anglais *False Rate Discovery*, en français Taux de Découverte Fausse
- FRET : de l'anglais *Fluorescence Resonance Energy Transfer*
- GATK : de l'anglais *Genome Analysis ToolKit*
- GFP : de l'anglais *Green Fluorescent Protein*
- GWAS : de l'anglais *Genome-Wide Association Studies*, en français études d'Association Pangénomique
- Hsp : de l'anglais *Heat-shock proteins*, en français protéines résistantes au choc thermique
- ICAT : de l'anglais *Isotope-Coded Affinity Tagging*
- Indel : Région courte d'insertion/Délétion
- IIS : de l'anglais *insulin/IGF-1 signaling*

- LC-MS/MS : de l'anglais *Liquid Chromatography-Tandem Mass Spectrometry*, en français Spectométrie de Masse en tandem après Chromatographie Liquide
- LD : de l'anglais *Linkage Desequilibirum*, en français déséquilibre de liaison
- MAF : de l'anglais *Minor Allele Frequency*, en frnaçais Fréquence d'Allèle Mineure
- MTX : Methotrexate
- NGS : de l'anglais *Next-Generation Sequencing*, en français méthodes de séquençage à haut débit
- OR : de l'anglais *Odds Ratio*, en français rapport des cotes
- PCA : de l'anglais *Protein-fragment Complementation Assay*
- PPI : de l'anglais *Protein-Protein Interaction*, en français Interaction Protéine-Protéine
- QTL : de l'anglais *Quantitative Trait Locus*, en français Locus de Caractères Quantitatifs
  - eQTL : expression Quantitative Trait Locus
  - pQTL : protein Quantitative Trait Locus
  - meQTL : methylation Quantitative Trait Locus
  - piQTL : protein interaction Quantitative Trait Locus
- UAS : de l'anglais, *Upstream Activator Sequence*
- UMAP : de l'anglais, *Uniform Manifold Approximation and Projection*
- UPR : de l'anglais, *Unfolded Protein Response*
- RNA-seq : de l'anglais, *RNA-Sequencing*
- SIFT : de l'anglais, *Sort Intolerant From Tolerant*
- SILAC : de l'anglais, *Stable Isotope Labeling with Amino acids in Cell culture*
- SMRT : de l'anglais, *Single Molecule, Real-Time*
- SNP : de l'anglais *Single Nucleotide Polymorphism*, en français Polymorphismes de Nucléotides Uniques
- SUT : de l'anglais *Stable Uncharacterized Transcript*, en français transcrits stables non-caractérisés
- WES : de l'anglais *Whole-Exome Sequencing*, en français Séquençage de l'Exome Entier
- WGS : de l'anglais *Whole-Genome Sequencing*, en français Séquençage du Génome Entier
- Y2H : de l'anglais *Yeast Two-Hybrid*, en français technique du double hybride

Remerciements

Dans cette page, je tiens à remercier toutes les personnes qui m'ont permises de mener à bien (et à terme !) ce doctorat de 5 ans au sein de l'université de Montréal.

Je remercie Professeur Guillaume Lettre, Professeur Rafael Najmanovich, Professeure Marie Brunet et Professeur Claude Perreault de faire partis de mon jury d'évaluation de thèse. Merci aux rapporteurs d'avoir pris le temps pour lire et commenter ma thèse, et de m'accompagner dans ce long processus qu'est la soumission du manuscrit et l'organisation de la soutenance de thèse. Je remercie également chaleureusement toutes les personnes qui ont assisté à ma soutenance, qui a eu lieu Mardi 5 Septembre 2023, à 8H30 à sur le campus de l'université de Montréal. Plus d'une vingtaine de personnes sont venues écouter en présentiel, mais de nombreuses personnes étaient également présentes sur Zoom, à travers divers fuseaux horaires et de multiples endroits, incluant les États-Unis, la France et le Japon. Ce fut pour moi un privilège et une joie de partager ma recherche avec autant de personnes.

L'accomplissement de ce doctorat n'aurait pas été possible sans la présence et le soutien de mes deux co-encadrants et mentors, Professeur Adrian Serohijos et Professeure Julie Hussin. Je ne vous remercierai jamais assez d'avoir accepté de m'avoir "adopté" dans vos laboratoires suite à la démission de mon premier directeur de thèse au milieu de mon doctorat. Merci de m'avoir fait confiance, et de m'avoir accompagné dans mes premiers accomplissements dans la recherche, de la rédaction à la soumission de mes travaux scientifiques ainsi que celle de ma thèse. Je ressors de ces 5 ans de doctorat avec de nombreuses expertises dans différents domaines de la bio-informatique, qui font de moi un véritable couteau-suisse ! Vous avez toujours été à mon écoute et soucieux de me mettre dans les meilleures conditions pour que je puisse m'épanouir dans ma recherche. Merci pour votre bienveillance et vos encouragements. Adrian, je pense avoir hérité de beaucoup de tes réflexes et habitudes de travail. Nous sommes bien les seuls à apprécier l'adrénaline des dates limites pour les applications de fond de recherche. Julie, j'ai beaucoup apprécié nos discussions scientifiques comme non scientifiques. J'espère réaliser un parcours aussi inspirant que le tien dans la

recherche. Merci à vous deux d'avoir été toujours disponible, que ce soit en présentiel ou en ligne, pour parler de science mais aussi de sujets plus personnels, quand cela a été nécessaire.

Travailler en co-direction de deux professeurs m'a donné la chance de pouvoir travailler, discuter et me lier d'amitié avec de (très) nombreuses personnes. Du laboratoire Serohijos, les post-doctorants Tatsuya Sakaguchi et Arnab Barua, les techniciens de laboratoire Zahra Sahaf, Etienne Ossona de Mendes et Cassandre Clermont, les doctorants Hemant Kumar, Melis Gencel et Nazlı Kocatuğ ainsi que les maîtrises Chloé Matta, Xavier Castellanos-Girouard et David Gagne-Leroux. Il m'a été très enrichissant de travailler aux côtés de bio-informaticiens comme des biochimistes. A ceux qui sont toujours dans le laboratoire, prenez bien soin de la machine à café que je vous ai laissé en héritage. Du laboratoire Hussin, la post-doctorante Holly Trochet, les ingénieurs de recherche Jean-Christophe Grenier, Pamela Mehanna et Raphaël Poujol, les doctorants Camille Rochefort-Boulanger, Fatima Mostefai, David Hamelin, Alexis Nolin-Lapalme, Cantin Baron, Matthew Scicluna et Isabel Gamache (toujours en compagnie d'Haru), les maîtrises (diplômés !) Emad Takla, Alex Richard-St-Hilaire, Justin Pelletin et Dominique Fournelle, ainsi que les nouveaux arrivants Simon Paquette et Yonglin Zu. Je n'ai pas pu passer autant de temps avec vous en présentiel mais je me suis toujours sentie incluse chez les gens du Montreal Heart Institute, sans jamais y avoir mis les pieds (chose corrigée pour ma répétition finale de soutenance de thèse, début septembre dernier) !

Ma thèse ne serait pas d'une aussi belle qualité sans les nombreuses relectures de Jean-Christophe Grenier, Raphaël Poujol, Isabel Gamache, Camille Rochefort-Boulanger, Mélanie Lemaire, Xavier Castellanos-Girouard, Flaminia Zane et Lisa Perus. Pour la préparation de ma soutenance de thèse, je tiens à remercier les membres de mes deux laboratoires qui m'ont donné de précieux retours et suggestions pour rendre ma présentation accessible au plus grand nombre. Tout particulièrement, Adrian et Julie, encore merci pour vos précieux conseils. En plus de mes collègues, je souhaite également remercier Professeur Stephen Michnick, Michaël Rera, Flaminia Zane, Lisa Perus, Amélie Laporte et Katharina Kloppenberg qui m'ont aussi beaucoup aidé dans les différentes itérations de la présentation finale.

En ce qui concerne mon implication associative auprès des étudiants en bio-informatique à l'université de Montréal, je ne peux malheureusement pas citer toutes les personnes avec qui j'ai collaboré au risque d'avoir un paragraphe aussi conséquent que celui dédié à mes collègues de laboratoire, mais je dédie cette petite phrase à Caroline Labelle, qui a toujours

été à mes côtés pendant toutes mes années de bénévolat à l'AEBINUM.

Je remercie ma famille, mes parents Jacques (Sène) et Bouasavanh, ainsi que Sarah et Sem pour leur soutien sans faille et leurs encouragements tout au long de mon doctorat. Merci d'avoir crue en moi et d'être très très fière de moi. Pardon d'être partie loin de la maison pendant plus de 5 ans, je ne serai plus trop loin maintenant, c'est promis !

Je tiens à remercier les personnes qui m'ont épaulée et encouragée dans mes moments de stress, de doute et de panique, sans vous je n'aurais pas été capable de terminer ce doctorat. Je remercie :

- les copains du master bio-info de Bordeaux: Alexia Souvane, Sapho Aupetit, Florian Lasalle, Kristina Kastano et Julien Estebeteguy,
- Etienne et Elisabeth (Chouchoune et Maple !) qui sont toujours là pour moi, malgré le décalage horaire avec la Nouvelle-Zélande,
- mes frères de coeur Christian Té, Alexandre de Barros, Quentin Cavaillé, pour leur soutien virtuel et présentiel depuis Bordeaux,
- Kaan Altıntaş et Dollar Vora, mes compagnons de voyage dans l'exploration du Québec, de Montréal et j'espère de nombreuses autres destinations,
- Pedro Bordignon Do Couto, mon premier labmate préféré.

Je remercie aussi, avec beaucoup d'affection, David Hamelin et Roméo Gilleron pour leur écoute quand je suis en panique et en stress. Enfin, je tiens à remercier encore une nouvelle fois Amélie Laporte et Lisa Perus, mes deux meilleures amies, qui ont toujours été là dans les moments plus difficiles de ma thèse, comme par exemple, les dernières heures avant la soumission de thèse et avant la soutenance de thèse. Merci Lisa d'avoir toujours pris le temps de relire tous mes travaux écrits, et d'avoir toujours les bons mots pour me rassurer. Merci Amélie d'être venue jusqu'à Paris presque tous les mois après mon retour en France pour m'épauler et m'aider à gérer mon retour d'expatriation, ainsi que d'être venue à Montréal pour m'encourager et pour célébrer avec moi la fin de cette thèse. Je vous adore tous et toutes, du plus profond de mon coeur.

Pour finir, je remercie les futurs lecteurs et lectrices de cette thèse, qui j'espère vous fera découvrir différents pans de la bio-informatique, en compagnie du rat-taupe nu (ou le rat nu des sables d'après Raphaël), de la souris et de la levure (ainsi que de leurs protéomes et interactomes respectifs).

Je dédie cette thèse à mon ami feu Mathieu Borel, qui a été d'un soutien sans faille pendant ces années de doctorat. Merci d'avoir été à mes côtés pendant les moments difficiles comme les moments heureux. Continue à m'observer depuis le ciel.
J'aurai aimé que tu puisses lire le fruit de ce travail acharné et je réitère à l'écrit, la promesse que je t'avais faite, de continuer à faire de la recherche innovante et pluridisicplinaire où bio-informaticien·ne·s et biochimistes travaillent ensemble, main dans la main.

Chapitre 1

Introduction

1.1. Préambule sur la biologie moléculaire et la bio-informatique

1.1.1. Le dogme central de la biologie moléculaire

Le domaine de la biologie cellulaire permet l'étude du fonctionnement de la cellule, qui est l'entité biologique, structurelle et fonctionnelle de tout être vivant. L'un des premiers événements évolutifs permettant de différencier les procaryotes (bactéries et archées) des eucaryotes (champignons, métazoaires, plantes etc.), est l'apparition du noyau dans la cellule et d'organelles fonctionnelles (appareil de Golgi, réticulum endoplasmique, mitochondrie - provenant d'un événement d'endosymbiose particulier -) chez les eucaryotes. Le noyau contient la majorité du matériel génétique de la cellule. Ce matériel génétique est nécessaire pour permettre de créer tous les éléments structurels et fonctionnels indispensables à l'intégrité de la cellule. L'étude de ce matériel génétique et des produits qui en découlent est possible grâce à des approches expérimentales et informatiques issues de l'intersection de plusieurs disciplines : la génétique, la biochimie métabolique et la bio-informatique (qui seront définies plus en détails dans la sous-section 1.1.2). Le dogme central de la biologie moléculaire (Crick, 1970) (Figure 1.1) décrit les relations entre les différents entités moléculaires permettant le fonctionnement et l'intégrité de la cellule. Dans sa première itération, le modèle décrit comment l'information génétique se transmet par le biais des trois entités moléculaires clefs, que sont l'acide désoxyribonucléique (ADN), l'acide ribonucléique (ARN) et les protéines. Ces entités sont les constituants des éléments fonctionnels respectifs des gènes, transcrits et protéines qui permettent d'établir la correspondance entre le génotype, qui est l'ensemble des caractères génétiques d'un être vivant et le phénotype, qui est l'ensemble des caractères anatomiques, physiologiques et antigéniques permettant de décrire un être vivant.

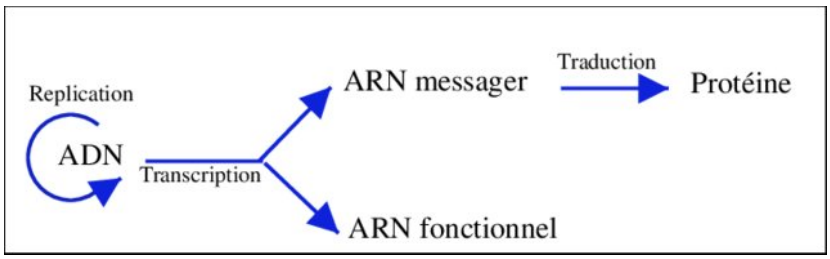

Fig. 1.1. Le dogme central de biologie moléculaire.

(Tirée de Thebault, 2004)
Ce schéma permet de représenter les relations entre ADN, ARN et protéines.

L'**ADN**, est une macromolécule en double hélice, composée de deux brins anti-parallèles, le brin sens étant constitué d'un enchaînement de nucléotides (Adénine (A), Thymine (T), Cytosine (C) ou Guanine (G)), et le brin anti-sens d'une chaîne de nucléotides complémentaires à ceux du brin sens (Figure 1.1). L'ADN est la matrice de départ sur laquelle l'ensemble de l'information génétique est encodée. On appelle génome, l'ensemble du matériel génétique d'un être vivant. Par le mécanisme de transcription, les gènes permettent la synthèse d'ARN messagers (ARNm), où ce sont les régions codantes qui serviront de matrice aux ARNm. Ces ARNm, aussi appelées transcrits, permettent la synthèse de protéines à l'aide du mécanisme de traduction. Les gènes sont répartis dans un ordre défini sur les chromosomes, qui sont une conformation particulière de l'ADN en interaction avec des protéines spécifiques : les histones. La structure chromosomique permet un état de compaction plus ou moins important de l'ADN, c'est à dire qu'il peut être facilement en contact ou non avec d'autres entités moléculaires, comme les ARN ou les protéines permettant la régulation de l'expression des gènes. Le domaine de l'épigénétique étudie les modifications chimiques réversibles de l'ADN ou des histones qui interagissent avec l'ADN, tels que la (dé)méthylation, et la (dés)acéylation. L'étude de ces changements épigénétiques permettent de mieux comprendre comment les facteurs environnementaux peuvent moduler l'expression des gènes. Enfin, les régions codantes du génome restent cependant moindre par rapport aux régions non codantes. Certaines régions sont transcrites en ARN non codants alors que d'autres régions ont des rôles plus structurels (télomères et centromères), ou ont des rôles encore peu élucidés (comme les transposons). Ces régions génomiques font partie du génome sombre (*dark genome*, en anglais), ce sont des régions génomiques dont les fonctions restent encore à déterminer. Les régions non codantes du génome font le sujet de nombreuses études dans le domaine de la génomique (Alexander et al., 2010).

L'**ARN** est une entité moléculaire clef qui permet de faire la transition entre ADN et protéine. Il s'agit d'une séquence simple brin, composée d'un enchaînement de nucléotides (Adénine (A), Uracile (U), Cytosine (C) ou Guanine (G)). Lors de la transcription,

les séquences ADN sont pris comme modèle pour synthétiser les séquences d'ARN à l'aide d'un ensemble de protéines qui comprend l'ARN polymérase II et les facteurs de transcription. Les produits de la transcription des régions codantes sont les ARN messagers (ARNm). Cette catégorie d'ARN va permettre la synthèse de protéines. Comme mentionné précédemment, le génome possède également des régions non codantes, dont certaines peuvent être transcrites en ARN non codants (ARNnc). Parmi les catégories d'ARNnc les plus abondantes, on retrouve les ARN ribosomaux (ARNr), qui sont les principaux constituants des ribosomes et qui sont transcrits principalement par l'ARN polymérase I, et les ARNs de transfert (ARNt), transcrits par l'ARN polymérase III. Ces deux types d'ARN sont impliqués dans le mécanisme de traduction. Il existe également d'autres types d'ARN non codants, de différentes tailles et dont les rôles moléculaires restent encore à déterminer. Chez l'humain, des recherches approfondies se concentrent sur des petits ARN non codants, tels que les micro-ARN (miARN) ou les ARN interférants (piARN et siARN), mais également sur les ARNs non codants plus longs (lncRNA, pour long non coding RNA, en anglais). On appelle transcriptome l'ensemble des ARN exprimés au sein d'un organisme, qu'ils soient codants ou non codants. Chez la levure, organisme modèle d'intérêt que nous allons étudier dans le chapitre 3, on distingue différents types de longs transcrits non codants : les transcrits stables non-caractérisés (Stable Uncharacterized Transcript (SUT)) qui sont dégradées par la voie de dédégradation des ARN non sens puis par l'exosome nucléaire, les transcrits cryptiques instables (Cryptic Unstable Tanscript (CUT)) qui sont dégradées par l'exosome nucléaire par l'activation de la voie Nrd1-Nab3-Sen1 (Marquardt et al., 2011) et les transcrits instables sensibles à Xrn1 (Xrn1-sensitive unstable transcript (XUT)), XRn1 étant une ARN 5'-3' exonuclease .

Les **protéines** sont synthétisées grâce à l'étape de traduction, où les ribosomes vont "lire" la séquence d'ARN par triplet de nucléotides (appelés codons) du codon initiateur, où débute la traduction, et permettant la synthèse d'une méthionine, jusqu'au codon stop qui signe l'arrêt de la traduction. En suivant les règles du code génétique, la traduction des codons permet la synthèse d'acides aminés, qui sont les composés élémentaires des protéines. Le code génétique permet la synthèse de 20 acides aminés "standards" (Alanine, Arginine, Asparagine, Aspartate, Cystéine, Glutamate, Glutamine, Glycine, Histidine, Isoleucine, Leucine, Lysine, Méthionine, Phénylalanine, Proline, Sérine, Thréonine, Tryptophane, Tyrosine, Valine), qui permettent la constitution de toutes les protéines des êtres vivants. On appelle protéome l'ensemble des protéines d'un organisme. Une fois les protéines synthétisées, après l'étape de traduction, certaines d'entre elles vont subir des étapes de repliement et des étapes de modifications post-traductionnelles. Ces modifications chimiques sont réalisées par des protéines spécifiques appelées enzymes, permettent l'ajout et/ou le retrait de composés chimiques. Parmi ces réactions enzymatiques, on peut citer les mécanismes de (dé)méthylation,

de (dé)phosphorylation, ainsi que les additions d'acides gras (phénomène d'acylation) ou de composés glucidiques (phénomène de glycosylation). Ces modifications vont permettre d'influencer la fonction de la protéine, que ce soit au niveau de son action, de sa demi-vie ; c'est le temps que met la protéine pour perdre la moitié de son activité fonctionnelle, ou de sa localisation cellulaire.

1.1.2. La bio-informatique dans le traitement des données moléculaires

Les protéines ont été les premières entités moléculaires dont on a déterminé la séquence. En 1953, Frederick Sanger a déterminé la séquence en acides aminés de l'insuline bovine (Sanger and Thompson, 1953). Une dizaine d'années plus tard, Robert Holley détermine la séquence de nucléotides de l'ARN (Holley et al., 1965), qui nécessite une étape de rétro-transcription en ADN complémentaire (ADNc). Enfin en 1973, Frederick Sanger détermine la séquence complète d'un génome complet de bactériophage phi X174 d'une longueur de $\sim$ 5,000 nucléotides (Sanger et al., 1977). Ces méthodes expérimentales sont les jalons ayant permis la mise en place des méthodes de séquençage à haut débit (NGS, pour *Next-Generation Sequencing* en anglais) dans le début des années 2000. Ce sont les méthodologies moléculaires pour réaliser le séquençage rapide de milliers ou millions de molécules d'ADN ou d'ARN simultanément, en déterminant l'ordre unique et spécifique des acides nucléiques. Quant aux protéines, c'est grâce à la spectrométrie de masse (dont une description plus détaillée se trouve dans la section 1.1.3) qu'on peut déterminer leurs séquences en grand nombre. Ces méthodes de caractérisation à haut débit permettent entre autre de représenter les informations moléculaires contenues dans l'ADN, l'ARN (les séquences de nucléotides) et les protéines (les séquences en acides aminés), en chaînes de caractères qui peuvent être automatiquement traitées et manipulées. C'est ce traitement systématique de données en grande quantité qui a mené au développement rapide du domaine de la bio-informatique ces dernières années. Ce domaine pluridisciplinaire allie la mise en application des connaissances théoriques issues de la biologie moléculaire et l'élaboration de méthodes informatiques pour identifier les fondements moléculaires et fonctionnels du génome (la génomique), du transcriptome (la transcriptomique) et du protéome (la protéomique).

Dans ce chapitre, nous mettrons en évidence l'importance d'élaborer de nouvelles méthodes informatiques protéome-centrées pour étudier la relation génotype-phénotype. Dans un premier temps, nous mettrons l'emphase sur l'importance d'étudier les protéines dans le but de mieux comprendre leur rôle dans l'apparition d'un phénotype, plus particulièrement le vieillissement. Dans un second temps, nous mettrons en avant les méthodes bio-informatiques

permettant l'étude des phénotypes complexes et sur l'importance d'étudier les protéines dans un contexte d'interaction pour proposer une abstraction adéquate du phénotype sous la forme d'un réseau d'interaction protéine-protéine.

1.1.3. Des définitions et principes fondamentaux sur les protéines

Niveaux d'organisation des protéines dans la cellule

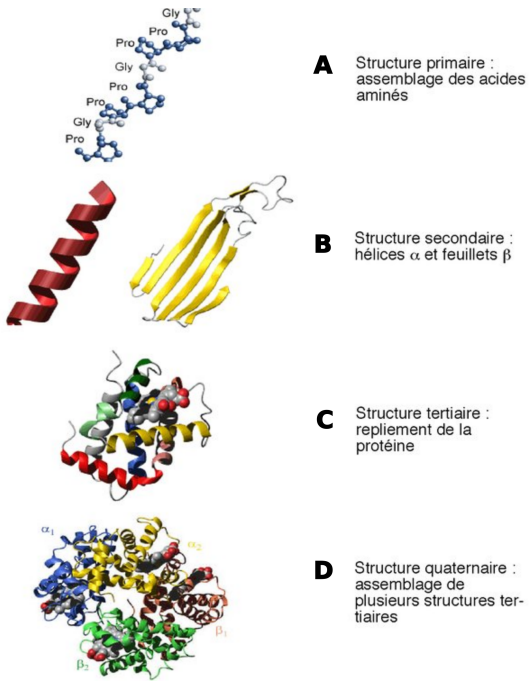

Fig. 1.2. Les différents niveaux de structures des protéines.

(Tirée d'Abraham, 2008)

Une protéine peut être décrite à différents niveaux de structure, telle que :

La **structure primaire** (Figure 1.2A) correspond à la séquence en acide aminés. On compte 20 acides aminés "standards" nécessaires à la composition des protéines de l'ensemble des êtres vivants. Un acide aminé est un composé dont la structure est généralement décrite de la manière suivante : un groupement amine (-NH2) à son extrémité gauche, un groupement carboxyl (-COOH) à son extrémité droite, tous deux accrochés à un carbone α qui est relié à un atome d'hydrogène (H) et une chaîne latérale R qui change pour chaque acide aminé. Les chaînes latérales des acides aminés d'une protéine permettent de déterminer les propriétés physico-chimiques de la séquence primaire d'une protéine. Elles définissent les caractéristiques d'une protéine qui vont permettre de déterminer ses fonctions moléculaires et biologiques, et qui conditionneront également son emplacement au sein de la cellule. Ces caractéristiques sont très importantes pour la stabilité de la protéine et permettent de définir sa taille, sa forme, son pH, mais aussi sa polarité qui permet de déterminer le sens de la protéine et son hydrophobicité qui est sa tendance de répulsion face à l'eau. Elles permettent également de déterminer les types d'interactions que ces protéines pourront avoir avec d'autres entités moléculaires (ADN, ARN ou peptides/protéines). Ces chaînes latérales appartiennent à différentes classes chimiques qui sont décrites dans le

Tableau 1. Pour constituer une chaîne linéraire d'acides aminés, qu'on appellera chaîne polypeptidique, les acides aminés sont reliés par le biais de liaisons peptidiques, qui sont des liaisons covalentes entre le groupement carboxyle porté par le carbone α d'un acide aminé et le groupement amine porté par le carbone α de l'acide aminé suivant dans la chaîne. Ainsi, on lira toujours un peptide à partir du groupement amine libre (ayant une polarité positive) du premier acide aminé (souvent un acide aminé Méthionine), qu'on appellera son extrémité N-terminale, jusqu'au groupement carboxyle libre (ayant une polarité négative) de son dernier acide aminé, qu'on appellera son extrémité C-terminale. Une protéine peut être constituée d'une ou de plusieurs chaînes peptidiques. Chez les eucaryotes, les protéines ont une taille moyenne de ~ 450 acides aminés (Tiessen et al., 2012).

La **structure secondaire** (Figure 1.2B) décrit l'arrangement des résidus d'acides aminés. Ces arrangements sont stabilisés à l'aide de liaisons hydrogènes et les arrangements les plus répandus sont les hélices α et les feuillets β (Reeb and Rost, 2019). Au sein de la structure secondaire, il est aussi possible de trouver des structures dites désordonnées, c'est à dire sans arrangements spatiaux définis. Ces régions sont tout aussi importantes que les régions structurées car elles influencent le repliement final de la protéine (Toto et al., 2020).

La **structure tertiaire** (Figure 1.2C) correspond à la forme finale et fonctionnelle de la protéine. Elle décrit comment les éléments de la structure secondaire interagissent entre eux pour établir la conformation la plus stable de la macromolécule. Ce processus est appellé repliement de la protéine et dépend principalement de protéines spécialisées appelées protéines chaperones (décrites plus en détails dans la section 1.1.4). Grâce au repliement de la protéine, la structure tertiaire aura tendance à enfermer les acides aminés hydrophobes en son coeur, pour éviter tout contact avec le milieu hydrophile de la cellule, et les acides aminés hydrophiles se retrouveront en surface. C'est sur cette structure, appelée également forme native de la protéine, que se produiront les modifications post-traductionnelles, mentionnées précédemment dans la section 1.1.1, qui modifieront les propriétés physico-chimiques des protéines en fonction du contexte cellulaire dans lequel ces dernières se trouvent. Cela leur permettra d'avoir des affinités d'interactions plus ou moins fortes avec d'autres constituants moléculaires de la cellule (ADN, ARN, protéines, métabolites). La structure tertiaire permet de distinguer les différents domaines protéiques fonctionnels,c'est-à-dire les différentes régions de la protéines avec des rôle moléculaires spécifiques (région de contact, activité enzymatique etc.). Ces domaines sont évolutivement très conservés : leurs séquences en acides aminés ne varient que très peu entre les espèces possédant des protéines dîtes "homologues", qui présentent des similarités de séquence et de fonction. Les régions des protéines n'ayant pas de rôle fonctionnel n'ont généralement pas de conformation particulière, et sont dites désordonnées. Ces dernières années, les

études sur le désordre protéique, visant à comprendre son rôle, notamment dans les processus de repliement et de l'agrégation protéique, ont été de plus en plus nombreuses (Y. Liu et al., 2019). Ces deux phénomènes seront décrits plus en détails dans la section 1.1.4.

Enfin, la **structure quaternaire** (Figure 1.2D) concerne surtout les complexes protéiques, constitués de plusieurs sous-unités protéiques, et permet de décrire comment ces différentes sous-unités interagissent physiquement entre elles. L'identification de cette structure quaternaire est notamment primordiale dans la compréhension moléculaire et fonctionnelle des interactions protéine-protéine (PPI, pour *Protein-Protein Interaction* en anglais) (Bertoni et al., 2017).

Il existe différentes approches expérimentales et informatiques pour déterminer ces différents niveaux d'organisation de la protéine, elles seront décrites plus en détails dans la section 1.1.3.

Différentes classes de protéines et leur rôles au sein de la cellule

Les différents niveaux d'organisation des protéines, en particulier les structures tertiaires et quaternaires, mettent en évidence que les protéines ne sont pas des molécules statiques. En fonction de leur localisation cellulaire et de leurs états de stabilité suite à leurs interactions avec d'autres molécules, elles sont susceptibles d'adopter différentes conformations pour réaliser leurs fonctions biologiques. La transition d'une conformation à une autre est appelée changement conformationnel. En fonction des changements conformationnels de leur structure tertiaire et quaternaire et de leur profil physico-chimique global déterminé par leur séquence primaire, on distingue 3 grands groupes de protéines : les protéines fibreuses, membranaires ou globulaires.

Les **protéines fibreuses** sont des protéines en forme de filaments qui sont insolubles dans l'eau. Ces protéines constituent les éléments structurels de tous les êtres vivants. Elles se situent surtout à l'extérieur de la cellule au sein de la matrice extra-cellulaire. Chez l'humain, la protéine fibreuse la plus abondante est le collagène (K. Smith and Rennie, 2007). Les protéines fibreuses sont particulièrement étudiées pour des applications très diverses, entre autre dans les domaines du biomédical ou de la cosmétique (Sionkowska et al., 2017)).

Les **protéines membranaires** sont des protéines situées au sein des membranes des organelles et de la membrane externe de la cellule. Dans les années 2000, on estimait que la proportion du protéome associée à cette classe de protéine était entre 25% à 30% dans les organismes modèles eucaryotes (Stevens and Arkin, 2000). Les propriétés

d'hydrophobicité et de polarité de la séquence primaire des protéines membranaires permettent de déterminer leur conformation au sein des membranes. On différencie les protéines membranaires non-polaires qui sont strictement hydrophobes et les protéines membranaires polaires qui sont plus hydrophiles. L'étude et l'identification des protéines membranaires ainsi que de leur conformation fonctionnelle est particulièrement difficile car ces dernières sont très instables et peu flexibles à cause de leur surface partiellement hydrophobe. Cependant, les innovations technologiques de séquençage à haut-débit en biologie structurale permettent de proposer des nouvelles stratégies pour étudier l'expression, la solubilisation, la purification et la cristallisation de ces protéines. Ces stratégies sont énumérées dans la revue de littérature de Carpenter et collègues (Carpenter et al., 2008).

Enfin, les **protéines globulaires** sont des protéines qui ont des profils plus hydrophiles en comparaison des protéines fibreuses et des protéines membranaires. Ce sont les protéines les plus étudiées et elles assurent des rôles très divers dans la cellule dont les *fonctions enzymatiques* impliquées dans les cascades de signalisation cellulaire ou dans les voies métaboliques permettant la synthèse des différents métabolites nécessaires au fonctionnement de la cellule (nucléotides, acides aminés, acides organiques, antioxydants, vitamines etc.). Les protéines permettant ces réactions sont les enzymes et ont une organisation structurale particulière, leur repliement permet de créer des conformations spécifiques et de constituer des domaines fonctionnels, qu'on appelle site actif. Un site actif est la partie de l'enzyme qui va interagir avec le(s) substrat(s) pour permettre la formation de produits. Les protéines globulaires peuvent aussi avoir des *fonctions de transport* de messages hormonaux, ou d'autres molécules, comme par exemple les gaz nécessaires à la respiration. Les protéines avec ces fonctions interviennent dans les processus de signalisation cellulaire, ce sont les différentes étapes permettant la communication inter- ou intra-cellulaire et le transit des différentes entités moléculaires transportées. Enfin, il existe des *fonctions plus spécifiques*, on peut citer les anticorps qui sont impliqués dans la réponse immunitaire, les histones qui assurent le rôle de compaction des chromosomes, nécessaires aux modulations épigénétiques de l'ADN, et les protéines chaperonnes qui permettent de réguler les repliements des autres protéines.

De par la diversité des rôles structurels et fonctionnels de l'ensemble des protéines au sein d'une cellule, son protéome a besoin d'être finement contrôlé pour maintenir l'intégrité cellulaire, c'est à dire maintenir l'état structurel et fonctionnel de la cellule. Pour permettre la biogenèse ainsi que la maintenance à la dégradation des protéines, différents acteurs et processus biologiques permettent de réguler l'homéostasie des protéines, ou protéostase cellulaire. Nous détaillerrons ces prossessus dans la section 1.1.4.

Étapes préliminaires à l'identification des protéines

Grâce à la protéomique, on peut identifier et quantifier les protéines pour caractériser leur abondance, leur localisation cellulaire, leurs interactions, les modifications post-traductionnelles qui leurs sont associées, mais aussi leur cycle de vie dans un type cellulaire, dans un contexte biologique particulier (méiose ou mitose cellulaire), ou dans le cas d'un phénotype donné (un état physiologique, ou une maladie). C'est ce qui fait de l'analyse protéomique une étude dynamique. En effet, un même génome peut conduire à différents protéomes en fonction des cellules ou tissus étudiés, des conditions physiologiques ou environnementales ou de l'état physiopathologique. Ce sont les protéines globulaires qui sont les plus abondamment étudiées car leurs structures tertiaires et quaternaires sont relativement stables en milieu aqueux. Ici, nous détaillerons exclusivement les méthodes d'identification des protéines globulaires qui sont solubles dans l'eau et qui sont les plus communément étudiées.

Pour étudier la composition d'un protéome, différentes étapes préliminaires à l'analyse des protéines doivent être effectuées. Les **étapes de préparation** inclut des étapes d'extraction et de concentration des protéines. Pour l'étape d'extraction, si les protéines proviennent d'un tissu, il sera nécessaire de séparer les cellules du tissu avant de détruire les parois des cellules pour pouvoir isoler les protéines. Pour obtenir un échantillon pur de protéines, il sera important de contrôler les conditions environnementales des échantillons, notamment :

- leur état de dénaturation : il faut que les protéines restent dans leur état natif pour qu'on puisse identifier les propriétés physico-chimiques de leur séquence primaire. Pour s'assurer que leurs conformations tertiaires ou quaternaires soient intactes, il faudra contrôler les conditions environnementales d'extraction, comme la température, le pH, ainsi que de la force ionique.
- leur état de conservation : à des températures ambiantes ou proches des valeurs physiologiques des espèces étudiées, les protéines peuvent être dégradées par des protéases, il faut donc les maintenir à basse température, ou bien les conserver dans un milieu sans eau, sous forme déshydratée, comme les protéases sont elles-mêmes des protéines qui fonctionnent en milieu aqueux.

À la suite de cette étape de préparation, il peut être nécessaire de réaliser une **étape de séparation** des protéines. C'est le cas par exemple, si on souhaite exclure les protéines trop abondantes ou de non d'intérêt dans un échantillon. En fonction de la problématique de recherche associée au protéome étudié et des méthodes d'analyse choisies en aval de étapes de préparation, les étapes de séparation seront réalisées avec des techniques

expérimentales spécifiques. On peut séparer les protéines en fonction de certaines caractéristiques physiques comme leur taille (par chromatographie par gel de filtration, ou par électrophorèse SDS-PAGE), physico-chimiques comme leur charge (par chromatographie échangeuse d'ions ou par électrophorèse), ou fonctionnelles comme leur spécificité de liaison (par chromatographie d'affinité). Pour cette étape de séparation des protéines, il sera aussi important de contrôler les conditions environnementales dans lesquelles les protéines évoluent et qui contribuent à leur solubilité et leur précipitation. Pour cela, il faudra prendre en compte l'influence de la concentration en sel dans la solution, celle des solvants organiques utilisés pour conserver les protéines, et celle du pH. Vient ensuite l'**étape d'analyse** dont les deux méthodes expérimentales les plus utilisées sont expliquées dans la section 1.1.3.

Méthodes d'identification et de quantification des protéines

Les **méthodes immuno-enzymatiques**, dont la plus connue est la méthode *Enzyme-Linked Immuno Sorbent Assay* (ELISA), permettent d'identifier et de quantifier des protéines, mais aussi des peptides, des anticorps ou des hormones. La technique ELISA repose sur le principe de reconnaissance des protéines (ici considérées comme des antigènes) par des anticorps primaires spécifiques. Ces anticorps peuvent eux-mêmes être liés à des bio-marqueurs (des gènes rapporteurs ou des enzymes rapportrices, ou bien des éléments fluorescents) ou être la cible d'anticorps secondaires liés à ces bio-marqueurs. Les étapes principales de l'ELISA consistent en la succession de 4 étapes : 1) la phase d'enrobage, qui permet d'immobiliser les anticorps ou les antigènes grâce à des interactions spécifiques antigène/anticorps, 2) la phase de blocage, où une protéine exogène à l'échantillon étudié va se lier aux anticorps non impliqués dans une liaison avec les protéines d'intérêt, 3) la phase de détection, qui consiste à ajouter un substrat permettant d'activer les bio-marqueurs liés aux anticorps et 4) la phase de *"screening"*, qui permet de caractériser qualitativement et quantitativement les protéines identifiées par interprétation des mesures de bio-marqueurs. En fonction de la configuration expérimentale des protéines cibles ou des anticorps permettant la reconnaissance des protéines cibles et des entités moléculaires utilisées pour la phase de détection (les anticorps, les antigènes ou le complexe antigène/anticorps), on distingue plusieurs types d'ELISA : direct, indirect, en sandwich et par compétition. Ces différences sont mises en avant dans la revue de littérature d'Aydin (Aydin, 2015). Différents types de bio-marqueurs peuvent être utilisés pour détecter les protéines d'intérêt, comme les gènes rapporteurs et les enzymes rapportrices qui vont permettre de mesurer l'abondance des protéines par mesure colorimétrique, de luminescence ou de fluorescence. Il existe aussi des méthodes de quantification sans ajout de bio-marqueurs, on qualifie ces quantifications de "*label-free*". Pour étudier l'intégralité d'un protéome, l'une des limitations principales de la méthode ELISA est le nombre exponentiel d'anticorps à utiliser pour identifier toutes

les différentes protéines contenues au sein de l'échantillon étudié, le temps nécessaire à la vérification de la spécificité d'interaction entre l'anticorps primaire avec sa protéine cible, ainsi que les potentiels biais de mesure relatifs à l'utilisation des bio-marqueurs. De plus, l'utilisation de la méthode ELISA ne permet pas d'identifier de nouvelles protéines. En effet, la détection des proteines nécessite le design d'anticorps associé à chaque protéine. Cela signifie que celles-ci auront été préalablement identifiées dans des études de biologie moléculaire ou de biochimie antérieures. Ainsi, pour l'identification en grand nombre des protéines au sein d'un échantillon, il sera nécessaire d'investir des coûts conséquents dans la construction de librairies d'un nombre important d'anticorps, chacun spécifique aux protéines d'intérêt. C'est pourquoi des méthodes sans à priori ont été mise en place, dont l'identification des protéines par la spectrométrie de masse.

La **spectrométrie de masse en tandem après chromatographie liquide** est l'autre méthode standard très utilisée pour étudier massivement les protéines. Elle se fait en plusieurs étapes via l'utilisation de deux appareils, une chromatographie en phase liquide, qui permet la séparation des protéines, suivie par une analyse de spectrométrie de masse en tandem (plus couramment appelé l'approche LC-MS/MS). Lors de l'**étape de séparation** les protéines sont dissoutes dans un mélange dans une phase liquide qui va traverser une phase stationnaire et permettre aux protéines d'être séparées en fonction de propriétés physiques (leur taille), physico-chimiques (leur hydrophobicité) ou encore fonctionnelles spécifiques (interactions avec des molécules particulières) en fonction de la variation des conditions expérimentales. Dans l'approche "bottom-up", une **étape de digestion protéolytique** est réalisée, elle permet de découper les protéines en plusieurs peptides composés d'un nombre réduit d'acides aminés à l'aide d'agents enzymatiques ou chimiques. La trypsine est l'agent enzymatique le plus couramment utilisé, elle permet de briser les liaisons peptidiques entre les résidus de lysine ou d'arginine, sauf si ces résidus sont imédiatemment suivis d'une proline. Dans l'approche "top-down", on garde les protéines intactes pour ensuite les analyser avec un spectromètre de masse alors que l'approche "middle-down", est une stratégie alternative aux deux approches précédentes : elle analyse des peptides un peu plus grands que ceux étudiés dans l'approche "bottom-up". La génération de ces peptides se fait par une étape de digestion limitée et des protéases plus spécifiques que la trypsine.

Après l'étape d'hydrolyse, les fragments protéiques sont analysés par un spectromètre de masse en tandem où deux analyses de spectre successives sont réalisées. Il existe différents types d'analyseurs de masse (Temps de Vol, Quadripôle, Piège à Ions) qui permettent de déterminer le ratio m/z selon différentes méthodes et qui ont chacun leurs propres caractéristiques (vitesse de traitement, fenêtre de tolérance, résolution, précision et prix). Les choix

et différentes combinaisons d'analyseurs de masse seront pris en fonction du type d'analyse à réaliser. L'analyse de spectométire de masse en tandem permet de générer des spectres de masse MS/MS. Un spectre de masse contient des pics d'intensité variable à différentes positions (abscisses) qui représentent le ratio entre la masse m et la charge z des molécules (m/z) des molécules étudiées. Contrairement à un spectre MS (généré à la suite de la lecture d'un seul analyseur au sein du spectromètre de masse), où chaque pic significatif correspond au ratio m/z d'un peptide, les pics dans les spectres MS/MS correspondent à la liste des masses des fragments ionisés constituant un peptide d'une protéine analysée. Les intensités des pics représentent leur abondance. Pour obtenir ces spectres MS/MS, les peptides issus de l'étape d'hydrolyse sont **vaporisés et ionisés** par une source d'ionisation (ionisation par électronébuliseur ou désorption-ionisation laser assistée par matrice) et sont ensuite analysés par un premier analyseur de masse. Cette première analyse permet de déterminer le ratio m/z des peptides intacts. Par la suite, les peptides détectés lors de la première analyse sont individuellement sélectionnés et subissent une **étape de fragmentation** réalisée dans une chambre de collision avec un gaz à basse pression. Cette phase de fragmentation se déroule en fonction d'un schéma spécifique où le squelette peptidique sera fragmenté en différents types d'ions dont la nomenclature spécifique a été proposé par Roepstorff et Fohlman (Roepstorff and Fohlman, 1984). La détection de chacun de ces ions va se traduire par un pic dans le spectre de masse MS/MS lu par le seconde analyseur de masses. Après la génération du spectre MS/MS, il sera nécessaire d'effectuer plusieurs étapes de pré-traitement avant de pouvoir les interpréter tels que :

- l'élimination du bruit,
- le calcul des centroïdes qui permet la discrétisation du spectre brut,
- l'ajustement du calibrage, où on réalise le décalage de tous les m/z pour compenser le mauvais calibrage de l'appareil,
- le déisotopage des spectres qui consiste à retirer les pics marquants des isotopes.

Ces étapes sont réalisées par un logiciel propriétaire, fourni par le fabricant de l'appareil qui prendra en compte les spécificités des différents types d'analyseur de masse utilisés dans le spectromètre de masse.

L'identification des protéines. Tout d'abord, il est nécessaire de retrouver et d'associer chaque spectre MS/MS à son peptide correspondant. Pour cela, on dispose de différentes informations comme la masse du peptide obtenu grâce à la lecture du premier analyseur de masse et des informations complémentaires liées à la composition en acides aminés du peptide qu'on pourra retrouver grâce à l'identification des pics associés aux fragments ionisés des peptides. En combinant les peptides identifiés à partir des spectres expérimentaux, il est ensuite possible de retrouver les protéines qui ont été analysées. Il existe deux grandes familles de méthodes permettant d'associer un peptide à un spectre :

- l'interprétation de spectre *de novo*, où on cherche à reconstituer le peptide à partir des informations contenues dans un spectre MS/MS sans s'appuyer sur des informations contenues dans des banques de protéines connues. Cette approche nécessite la génération de spectres MS/MS de bonne qualité et peut prendre un temps considérable en fonction du nombre de protéines à identifier. Différents études (Bringans et al., 2008; Pevtsov et al., 2006) ont mis en avant l'importance de la qualité des spectres MS/MS pour obtenir des résultats satisfaisants et le fait que ces derniers varient en fonction des appareils utilisés. Cette approche semble donc peu adéquate pour l'identification exhaustive d'un protéome contenant de nombreuses protéines.
- l'identification par comparaison de spectres avec des protéines connues, qui est l'approche la plus communément utilisée. Nous détaillerons plus en détails cette stratégie dans les prochaines lignes.

Dans l'approche par comparaison de spectres, on utilise des banques de données de protéines pour lesquelles on génère les spectres théoriques des peptides obtenues par une digestion *in-silico* de ces protéines. La création de ces différents peptides théoriques permettront d'inférer leurs spectres de masse MS/MS. Ces spectres sont des simulations de la fragmentation des peptides théoriques. Chaque fragment du peptide va se traduire par la création de plusieurs pics dans le spectre théorique, chaque pic correspondant à un ion différent, ou à une perte neutre. Ces spectres théoriques décrivent toute l'information disponible, ne contiennent aucun bruit et sont considérés comme "parfaits". Pour limiter le nombre de spectres théorique à comparer avec les spectres expérimentaux produits par le spectromètre de masse, l'utilisation de filtres est nécessaire. Le filtre le plus couramment utilisé repose sur la sélection des spectres théoriques des peptides ayant une masse proche de la masse du peptide précurseur analysé par le spectromètre de masse, réduisant ainsi le nombre de comparaison de spectres à réaliser.

Pour comparer les spectres théoriques générés à partir des banques de protéines et les spectres expérimentaux issus du spectromètre de masse, il sera nécessaire d'évaluer la similarité entre les spectres. Cette similarité va souvent reposer sur le calcul d'un score avec un fort pouvoir discriminant pour pouvoir proposer une liste restreinte de peptides candidats. Parmi les scores de comparaison de spectre, on peut citer :

- l'**identification du nombre de pics en commun** : pour cette méthode on considère que deux spectres sont similaires s'ils possèdent de nombreux pics à la même position. Ce score peut prendre en compte l'absence ou la présence d'un pic à une position donnée, ainsi que leur intensité. Ce type de score est la plus couramment utilisé, on le retrouve par exemple dans le programme Sequest (Eng et al., 1994).

- la **corrélation croisée** : dans cette méthode, les deux spectres sont étudiés comme étant des signaux. La corrélation croisée consiste à mesurer la similitude entre deux signaux. Cette méthode est également utilisée dans Sequest pour recalculer le score des meilleurs peptides candidats identifiés par la méthode d'identification du nombre de pics en commun.

Il existe différentes approches pour identifier la liste des peptides associés aux spectres MS/MS expérimentaux en les comparant à des bases de données de protéines dont les différences sont détaillées dans les travaux de Marquioni et collègues (Marquioni et al., 2021). On peut notamment citer :

- l'identification de peptides par la recherche des ions dans les spectres MS/MS (*ie. ion MS/MS search*, en anglais)
- l'identification de peptides par comparaison à des banques spectrales (*ie. spectral library search*, en anglais)
- l'identification de peptides par l'identification de tag peptidique au sein de la séquence (*ie. peptide sequence tag search*, en anglais)

Ces approches permettent d'établir la liste des peptides candidats les plus probables qui permettront de retrouver les protéines qui leur sont associées. Les étapes de comparaison de spectre, d'identification des peptides et d'identification des protéines sont associées à des méthodes de correction statistiques pour s'assurer de réduire le nombre de faux positifs inférés à chacune des étapes.

Autres approches pour l'identification à large échelle des protéines. Parmi les technologies les plus récentes, on peut citer :

L'identification des protéines à l'aide des études d'extension de proximité: Olink Proteomics a développé une technologie reposant sur le principe d'essai d'extension de proximité (Lundberg et al., 2011) qui combine la reconnaissance des protéines par des anticorps hautement spécifiques liés à des séquences d'ADN utilisées comme tag qui seront analysées par séquençage à haut débit. Elle permet l'identification d'un grand nombre de protéines dans des échantillons de plasma ou de sang. Les différentes étapes de cette approche sont les suivantes :

- L'étape de conjugaison anticorps-ADN : Chaque anticorps spécifique à la reconnaissance d'une protéine donnée est couplé à une construction ADN où on retrouve une séquence d'ADN contenant un barcode qui permettra l'identification de l'anticorps ainsi qu'une séquence d'ADN servant comme site de liaison pour des amorces spécifiques nécessaires à la PCR.

- L'étape de liaison à la protéine : les combinaisons anticorps-ADN sont ajoutées au sein de l'échantillon et les anticorps se lient de manière spécifique aux protéines pour lesquelles ils possèdent une grande affinité et une grande spécificité. Cela crée des complexes protéine-anticorps.
- L'étape d'extension de proximité : En présence des complexes protéine-anticorps formés, des réactions d'extension et de ligation d'ADN sont déclenchées pour permettre la formation d'amplicons d'ADN du barcode associé au complexe protéine-anticorps.
- L'étape d'amplification PCR : Les amplicons ADN sont amplifiés par PCR grâce à la liaison d'amorces spécifiques au site de liaison des amorces contenues dans la séquence ADN associée à chaque complexe protéine-anticorps.
- L'étape de séquençage à haut débit des amplicons d'ADN et leur analyse : pour chaque complexe protéine-anticorps, les amplicons ADN sont séquencés et permettent d'obtenir des informations quantitatives sur l'abondance de la protéine à laquelle les amplicons ADN sont associés. Chaque barcode séquençé pourra permettre l'identification de la protéine qui lui est associée en retrouvant le complexe anticorps-ADN auquel il est associé.

Des études récentes reposant sur la technologie Olink ont permis l'identification simultanée de 1500 protéines dans une centaine d'échantillons (Wik et al., 2021).

L'identification des protéines à l'aide des aptamères peut être réalisée à l'aide de la technologie SOMAscan (Gold et al., 2010) développé par l'entreprise SOMAlogics. C'est une technologie de protéomique mutltiplexée reposant sur l'utilisation d'aptamères appelés SOMAmers (pour SLow Off-rate Modified Aptamers). Les aptamères sont de courts segments d'ADN ou d'ARN qui peuvent se lier de manière spécifique à une cible moléculaire avec une grande affinité et un grande spécicificité. La sélection des aptamères se fait à l'aide de la méthode SELEX (pour *Systematic Evolution of Ligands by EXponential enrichment*, en anglais) dont les premières utilisations remontent aux années 1990 (Tuerk and Gold, 1990). La méthode se déroule en plusieurs étapes :

- la sélection des aptamères : un grand nombre de séquences nucléotidiques aléatoires sont générées puis mises en présence de la molécule d'intérêt afin d'identifier les séquences se liant de manière spécifique à celle-ci. Les séquences nucléotidiques ne se liant pas à la molécule d'intérêt sont éliminées par des phases de lavage.
- l'étape d'amélioration d'affinité de liaison : les aptamères sélectionnés sont séparés de la protéine d'intérêt et sont amplifiés par des réactions de polymérase en chaîne (PCR)

pour les séquences d'ADN, ou par des réaction de polymérase en chaîne après transcription inverse (RT-PCR) pour les séquences d'ARN. Cette étape permet d'augmenter le nombre de séquences nucléotidiques ayant une grande affinité avec la molécule d'intérêt.

- des phases de sélection répétées : une fois amplifiées, les séquences nucléotidiques subissent des étapes supplémentaires de sélection et d'amplification où les conditions de lavage sont modulées (variations en température ou en concentration de sels) pour améliorer les affinités de liaison et leur spécificité.
- le séquençage et la caractérisation des aptamères ayant les plus grandes affinités et spécificités de liaison à la molécule d'intérêt.

L'approche SomaScan permet d'identifier un grand nombre de protéines, qu'elles soient présentes en faible ou forte abondance au sein de l'échantillon. Elle permet l'identification simultanée de plusieurs milliers de protéines. Par exemple, dans sa version la plus récente, la technologie 7k SomaScan v4.1 peut identifier jusqu'à 7300 protéines chez l'humain de manière simultanée. Des études récentes mettent à profit cette technologie pour réaliser des études de protéomique à large échelle sur des miliers d'échantillons (Candia et al., 2022).

Quantification des protéines. Il existe différentes méthodes de quantification des protéines décrites dans diverses revues littéraires (Anand et al., 2017; Y. Zhang et al., 2013).

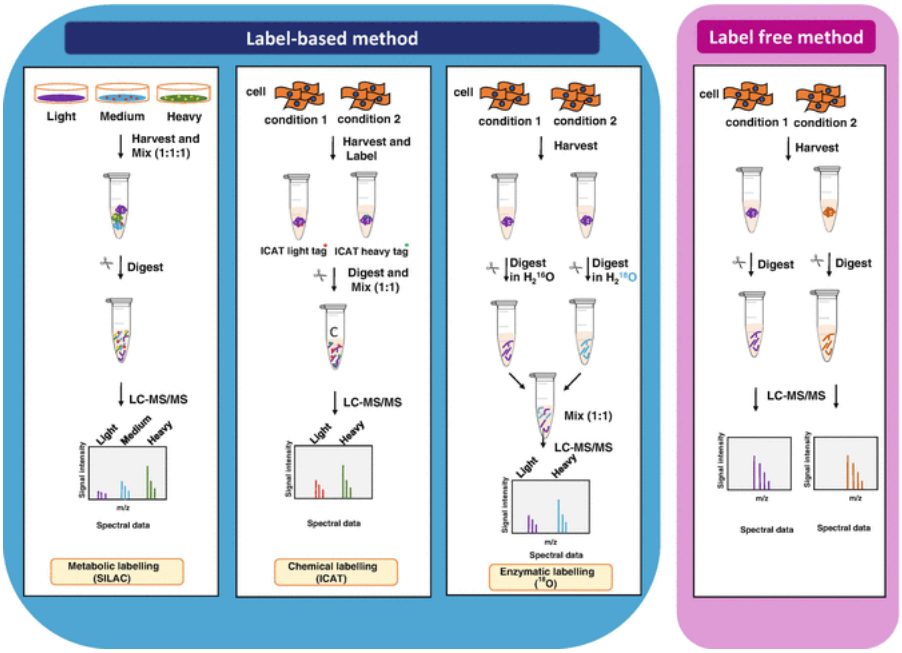

Fig. 1.3. Les différentes méthodes de quantification des protéines.

(Tirée de Anand et al., 2017)

On distingue deux grandes approches de quantification des protéines :

Les méthodes avec étiquettes. On incorpore des isotopes stables (des éléments chimiques dont la radioactivité n'est pas décelable) au sein des peptides avant l'étape d'ionisation dans le spectromètre de masse. Cela permet d'induire des différences de masse au sein des spectres de masse qui permettront de quantifier par la suite les protéines. Ces étiquettes peuvent être lues soit pendant la lecture des spectre de masse MS, ou bien lors de la lecture des spectre de masse MS/MS. Nous nous attarderons surtout sur les quantifications issues de la lecture des spectres de masse MS.

Pour ces méthodes, différents types d'isotopes stables peuvent être utilisés. L'utilisation de ces isotopes permet principalement de faire des mesures de quantification relative, où on établit la différence entre les aires sous la courbe des spectres de masse MS pour un échantillon donné (celui qui sera isotopiquement marqué) avec celles des quantifications obtenues d'un échantillon de référence (un échantillon ayant été généré sous les mêmes conditions expérimentales mais n'ayant pas été marqué). Différents types d'étiquettes peuvent être employés :

- les étiquettes métaboliques, qui sont utilisées dans un contexte expérimental où il est possible d'intervenir lors de l'étape de synthèse des protéines, par exemple dans le cas de culture cellulaire, ce qui permet de minimiser la variabilité expérimentale. La méthode la plus répandue utilisant ces étiquettes métaboliques est la méthode *Stable Isotope Labeling with Amino acids in Cell culture* (SILAC) (Ong et al., 2002), qui consiste à incorporer dans des milieux de culture des acides aminés marqués pour qu'ils soient incorporés dans la séquence primaire en acides aminés. Ce sont les éléments d'hydrogène, d'azote et de carbone (H, 14N et 12C) qui sont substitués par leurs isotopes stables respectifs (Y2H, 15N, and 13C). L'ajout de ces isotopes stables dans les séquences des protéines va modifier les poids moléculaires associés aux différents spectres de masse des fragments de peptides issus de ces protéines.
- les étiquettes chimiques, dont les plus utilisées sont celles issues de la méthode *Isotope-Coded Affinity Tagging* (ICAT) (Gygi et al., 1999). Cette méthode permet d'ajouter des étiquettes isotopiques sur les résidus cystéine via une réaction chimique faite avec un réactif chimique d'iodoacétyle lié, par un "linker" isotopiqement stable, à un groupe biotine. Le groupe biotine est une vitamine qui sera utilisée pour isoler les peptides marqués après purification par affinité via son interaction avec l'avidine. Les peptides ainsi isolés seront par la suite analysés par le spectromètre de masse pour générer leur spectre de masse.
- les étiquettes enzymatiques, qui sont ajoutées pendant l'étape de digestion enzymatique des protéines en peptides. Une enzyme de trypsine modifiée permet de digérer les protéines et de remplacer les atomes d'oxygène 16O des peptides résultants par l'isotopes stable 18O, ce qui, comme la méthode SILAC, modifiera les spectres de

masse des peptides marqués.

Il existe des méthodes d'ajouts d'étiquettes pour une lecture au niveau des spectres de masse MS/MS telles que les méthodes *Isobaric tags for relative and absolute quantification* (Ross and Poirier, 2004), et *Tandem Mass Tags* (Thompson et al., 2003). Ces méthodes permettent également de faire des mesures de quantification relative. En ce qui concerne les stratégies de quantification absolue, la principale différence est qu'on utilise comme étiquettes des peptides synthétiques marqués isotopiquement, qui peuvent être synthétisés chimiquement (méthode *Absolute QUAntification*) (Kirkpatrick et al., 2005) ou biologiquement (méthode *Quantification conCATamer* (Beynon et al., 2005). On peut aussi utiliser des protéines étalons (méthode SILAC absolu (Hanke et al., 2008) ; méthode *Protein Standard for Absolute Quantification* (Kaiser et al., 2011). Pour ces approches, on connaît préalablement le poids moléculaire de ces étiquettes, qui pourront être soustraits aux spectres de masse des peptides sur lesquels ces dernières se seront fixés.

Les méthodes sans étiquette. On estime la quantité relative des protéines par différentes stratégies :

- le comptage spectral (Washburn et al., 2001) est la méthode la plus simple. Elle repose sur l'observation suivante : plus la protéine est abondante, plus les peptides qui lui sont associées sont nombreux. Cependant, il faut aussi prendre en compte que plus une protéine est grande, et plus elle sera susceptible de créer un grand nombre de peptides. C'est pourquoi il existe des méthodes alternatives au comptage spectral qui tentent d'équilibrer les potentiels biais dûs à cette contrainte conceptuelle, incluant les scores suivants :
  - le *Protein Abundance Index* (Rappsilber et al., 2002), où on compte le nombre de peptides identifiés pour une protéine normalisé par le nombre de peptides tryptiques (issues de l'étape de digestion protéique par la trypsine) théoriquement observable, et sa version modifiée *exponentially modified Protein Abundance Index*(Ishihama et al., 2005), qui prend en compte les pourcentages de fraction molaire et de poids moléculaire.
  - l'*Absolute Protein EXpression* (Lu et al., 2007)
  - le *Normalized Spectral Index* (Griffin et al., 2010)
- les méthodes de mesure de pics associés aux spectres de masse, car il a été observé qu'il existe une relation linéaire entre l'abondance protéique et leurs régions présentant des pics dans les spectres de masse associés (Al Shweiki et al., 2017)

Les approches de quantification avec ou sans étiquettes ont chacune des avantages et des limites. Le choix des méthodes de quantification dépend surtout de la question biologique,

des budgets alloués et du type de technologie disponible pour réaliser les mesures de spectrométrie de masse.

Pour faire le traitement, l'analyse et les mesures de quantification relative ou absolue à partir des spectres de masse issus de la méthode LC-MS/MS, dans le but d'identifier et de quantifier les protéomes, différents pipelines bio-informatiques ont été développés par différentes compagnies (ProteoIQ, ProteomeDiscoverer, Scaffold *etc.*) Ces pipelines sont pour la plupart des suites de programme propriétaire, dont les algorithmes sont soumis à brevet et dont le code n'est pas disponible pour modification. Mascot (Brosch et al., 2009) et Maxquant (Cox and Mann, 2008) sont les suites logiciels les plus utilisés et sont disponibles gratuitement.

1.1.4. La protéostase cellulaire et son importance dans l'intégrité cellulaire

Acteurs de la protéostase cellulaire

Le concept d'homéostasie des protéines a été introduit pour la première fois en 2008 par Balch et ses collègues (Balch et al., 2008). On définit l'homéostasie des protéines comme l'ensemble des processus biologiques permettant de maintenir l'équilibre entre biogenèse et dégradation de l'ensemble du protéome au sein de la cellule. Cela se traduit, entre autre, par le contrôle de l'abondance protéique, des états de conformation des protéines, de leurs interactions, et de leurs localisations. L'étude de la protéostase est également nécessaire à la compréhension des maladies résultant de l'accumulation d'agrégats protéiques anormaux, comme les maladies neurodénégératives, telles que les maladies d'Alzheimer et de Parkinson (Irvine et al., 2008). Le maintien de la protéostase cellulaire est notamment nécessaire pour la résistance aux stress environnementaux, comme par exemple le stress oxydatif dû à la respiration cellulaire (D'Amico et al., 2017). Ce maintien est aussi primordial dans la réponse immunitaire face aux pathogènes pour éliminer toute protéine exogène à l'organisme (Zügel and Kaufmann, 1999)

Les principaux mécanismes permettant le maintien de la protéostase sont décrits ci-bas.

Les **mécanismes de régulation de la traduction** sont principalement pris en charge par le complexe ribosomal et les facteurs d'initiation de la traduction (R. J. Jackson et al., 2010). Les étapes clefs de la traduction sont son démarrage au niveau du codon initiateur, et sa terminaison au niveau du codon stop. Un ar-rêt prématuré de la traduction entraîne la formation de protéines tronquées et non fonctionnelles, qui peuvent entraîner des pertes de fonction à l'origine de l'apparition de phénotypes pathologiques. Un arrêt

prématuré de la traduction pour des protéines impliquées dans des processus essentiels au bon fonctionnement de la cellule peut se révéler particulièrement néfaste. À la fin de la traduction, les protéines vont subir un étape de repliement pour devenir fonctionnelles .

Les **mécanismes de repliement** sont médiés par les protéines chaperonnes. Cette famille de protéines, historiquement appelée famille des protéines résistantes au choc thermique (Hsp pour *Heat-shock proteins* en anglais) permet la maturation des protéines et leur repliement (Hsp60, Hsp70, Hsp90 etc.). Le repliement permet aux protéines de trouver une conformation thermodynamiquement stable, où les régions hydrophobes seront enfermées dans le coeur de la protéine, à l'abri de tout contact avec le milieu aqueux de la cellule, alors que les régions hydrophiles seront présentées en surface. Certaines protéines chaperonnes contribuent au repliement des protéines en se fixant sur les régions hydrophobes, exposées quand la protéine n'est pas encore repliée, afin de les masquer. D'autres protéines chaperonnes participent à la modification de la structure tertiaire de la protéine, pour que les régions hydrophiles entourent les régions hydrophobes. En cas de défaut de repliement, les régions hydrophobes des protéines sont exposées à la surface et la structure tertiaire de la protéine est donc compromise, elle sera dite mal-repliée et sera considérée comme thermodynamiquement instable. Les protéines chaperonnes seront capables de remédier à ces défauts de repliement et permettent de retrouver des conformations stables. Cette famille de protéine est d'ailleurs très conservée à travers les espèces (Draceni and Pechmann, 2019; El-Samad et al., 2005) et contribue de manière importante à la robustesse cellulaire, ainsi qu'à la capacité de la cellule à rester structurellement et fonctionnellement intègre, malgré la présence d'entités moléculaires pouvant l'altérer, ici les agrégats protéiques. Sans prise en charge des protéines chaperonnes, ces agrégats protéiques vont essayer de dissimuler ces régions hydrophobes en se greffant à d'autres protéines présentant des défauts de repliements, créant ainsi des agrégats multi-protéiques insolubles dans le milieu intracellulaire, qui peuvent être à l'origine d'altérations structurelles au niveau des différents composants de la cellule (Squier, 2001). En cas de surcharge des protéines chaperonnes et d'une accumulation croissante d'agrégats protéiques, d'autres systèmes de contrôle qualité seront sollicités, il s'agit des voies de dégradation des protéines.

Les **machineries de dégradation des protéines** (dont les mécanismes sont détaillés dans la revue de littérature de Zhao et collègues (Zhao et al., 2022)), sont sollicitées quand il est nécessaire de réduire le niveau global de protéines au sein de la cellule. Parmi les différents substrats ciblés par ces machineries, on distingue les fragments non fonctionnels des protéines issus d'un arrêt prématuré de la traduction, des protéines n'étant pas repliées ou ayant des défauts de repliement qui n'ont pu être pris en charge par les protéines chaperonnes, et les agrégats protéiques et les protéines dont la fonction cellulaire n'est

plus nécessaire. La dégradation des protéines dont l'activité n'est plus nécessaire dans la cellule, peut se faire à la suite d'un signal indiquant qu'une protéine est dépliée (UPR, pour *Unfolded Protein Response* en anglais) ou par le système de dégradation associé au réticulum endoplasmique (ERAD, pour *Endoplasmic-Reticulum-Associated protein Degradation* en anglais). Quant aux protéines mal repliées ou aux agrégats protéiques, ils peuvent être pris en charge par le protéasome. Le protéasome est un complexe multi-protéique dont la fonction principale est de dégrader de manière ciblée les protéines mal repliées, dénaturées ou obsolètes au sein de la cellule. Le protéoasome favorise la résistance au stress oxydant et limite le vieillissement cellulaire prématuré (Pickering and Davies, 2012). Quand les protéines mal repliées sont marquées de molécules d'ubiquitine, un peptide de 70 acides aminés qui permet d'identifier les protéines à éliminer, elles sont prises en charge par le système ubiquitine-protéasome pour être éliminées. Le système ubiquitine-protéasome a d'autres rôles spécifiques détaillés dans la revue de littérature de Qu et collègues (Qu et al., 2021). Enfin l'autophagie, processus d'auto-digestion qui consiste en une dégradation de composants intracellulaires par le lysosome, et la phagocytose, processus d'endocytose par des cellules spécialisées appelées phagocytes, sont d'autres mécanismes aboutissant à la dégradation de ces reliquats protéiques (M. P. Jackson and Hewitt, 2016; J. Li et al., 2018).

Il est important de souligner que des dérèglements à n'importe quelle étape de la protéostase cellulaire peuvent entraîner la formation d'agrégats protéiques. De nombreuses études ont montré une corrélation entre l'accumulation d'agrégats protéiques et l'apparition de phénotypes pathologiques associées à la dégénérescence cellulaire, suite à une perte d'homéostasie des protéines. Par exemple, les amyloïdoses, sont des maladies causées par l'accumulation d'agrégats protéiques présents dans les tissus nerveux. On peut notamment citer la sclérose latérale amyotrophique, les maladies d'Alzheimer et de Parkinson, ainsi que les maladies à prions (Aguzzi and O'Connor, 2010; Ross and Poirier, 2004). Le déclin de la protéostase cellulaire fait également partie des caractéristiques de l'apparition du vieillissement (López-Otín et al., 2013). Nous allons détailler, dans la prochaine section, les approches expérimentales et computationnelles permettant d'étudier l'origine de la formation de ces agrégats, à savoir la propension d'agrégation des protéines.

Méthodes pour l'étude de la propension d'agrégation des protéines

La formation d'agrégats protéiques peut apparaître en fonction de différentes conditions environnementales (Fink, 1998). D'un point de vue expérimental, l'identification des agrégats protéiques peut être réalisée grâce à des approches utilisant de la fluorescence, comme la méthode de transfert d'énergie par résonance de fluorescence (FRET, pour *Fluorescence Resonance Energy Transfer* en anglais) (De and Klenerman, 2019). On peut également identifier ces agrégats avec des approches utilisant des microscopes (à force

atomique ou électronique) (Ruggeri et al., 2019). Cependant, la caractérisation de ces agrégats protéiques, notamment leurs séquences primaires en acides aminés, reste difficile à effectuer car ils sont très instables. Il est par exemple très difficile de déterminer leur structure 3D avec les méthodes de cristallographie à rayons X, méthode couramment utilisée pour déterminer la structure tertiaire, voire quaternaire, des protéines globulaires.

Il existe des méthodes bio-informatiques pour identifier les agrégats protéiques, qui s'intéressent précisément aux structures primaires et secondaires des protéines pour évaluer la propension d'agrégation. La propension d'agrégation des protéines dépend principalement de leur composition en acides aminés. Les protéines avec de nombreux résidus hydrophobes auront une plus grande probabilité de s'agréger que les protéines présentant une proportion moindre de résidus hydrophobes. L'ordre et les propriétés des acides aminés vont aussi permettre de déterminer quel type de structures secondaires sera formé. Les arrangements spatiaux des protéines peuvent également influencer la propension d'agrégation des protéines. La prédiction de ces structures secondaires peut être fait à l'aide de l'algorithme DSSP (pour Define Secondary Structure of Proteins en anglais) (Kabsch and Sander, 1983). Il a été notamment démontré que la formation de feuillets β favorise la formation de fibres amyloïdes, qui sont des agrégats de protéines formés par de nombreuses copies d'une même protéine qui possède des défauts de repliement, entraînant leur agglutination sous la forme de fibrilles insolubles dans l'eau (Luheshi et al., 2007). Ici, nous citerons ci-bas trois méthodes permettant d'étudier la propension d'agrégation, qui utilisent toutes les séquences primaires des protéines comme données d'entrée. Ces algorithmes vont être capable d'estimer des scores de propension d'agrégation par résidu d'acide aminé.

Tango. (Fernandez-Escamilla et al., 2004) Ce logiciel permet d'estimer le taux de propension d'agrégation par acide aminé. On fournit en entrée une séquence primaire de la protéine d'intérêt. Pour chaque acide aminé, on calcule un score de propension d'agrégation : il s'agit d'un pourcentage de chance pour lequel ce dernier peut appartenir à une structure secondaire donnée (conformations natives stables comme l'hélice-α, feuillet-β ou conformations plus aléatoires comme le tour-β). Ce score prend en compte les propriétés physico-chimiques de chaque résidu et de ses voisins adjacents. Il tient également compte du contexte environnemental de la protéine (température, pH, et la force ionique). L'algorithme de Tango permet donc de prédire la β-agrégation dans les peptides et les protéines par la détermination de la structure secondaire la plus probable. Pour prédire les segments β-agrégants d'une protéine, TANGO calcule la fonction de partition qui englobe les propriétés statistiques associés aux résidus de la protéine étudiée se trouvant dans un contexte environnemental donné. En fonction du type de structure secondaire auquel l'acide aminé sera rattaché, son taux de propension d'agrégation sera plus ou moins important. Les acides aminés impliqués dans un

feuillet-β auront des taux de propension d'agrégation plus importants que ceux impliqués dans les autres structures secondaires.
Aggrescan. (Conchillo-Solé et al., 2007) Ce logiciel permet d'estimer des valeurs de taux de propension d'agrégation par acide aminé en se basant sur des mesures de différence d'hydrophobicité entre des séquences sauvages et mutantes de peptides ayant des propriétés d'agrégation établies (comme le peptide β-amyloïde humain). Cet algorithme va permettre de distinguer des régions au sein de la protéine qui auront des fortes propensions d'agrégation, la taille de ces régions étant estimée par comparaison avec des tailles de régions de référence au sein d'une base de données de 57 protéines amyloïdogéniques où les régions *"hot spot"* sont expérimentalement validées.
PASTA. (Walsh et al., 2014) Ce logiciel repose sur l'hypothèse que le mécanisme de formation de feuillets-β par les protéines est à l'origine de la formation d'agrégats. PASTA prédit les régions potentiellement impliquées dans la formation de feuillets-β en examinant si les acides aminés qui les composent et se faisant face ont des propriétés physico-chimiques adéquates à la formation d'un feuillet-β. Les valeurs prédites pour ces acides aminés sont comparés à des valeurs de référence pour des acides aminés au sein de protéines globulaires dont les structures natives sont déjà connues et qui ont été caractérisées dans des contextes pathologiques.

Pour estimer les taux de propension d'agrégation d'une protéine, il suffit de faire la moyenne des scores de propension d'agrégation par résidu, normalisé par la taille de la protéine. Les logiciels cités précédemment sont habituellement utilisés pour identifier les régions des protéines pouvant potentiellement s'agréger, et donc identifier des candidats susceptibles d'être impliqués dans la formation d'agrégats protéiques, ou encore permettent de prédire les conséquences des mutations d'une protéine au niveau de sa propension d'agrégation. Certains chercheurs, comme Tsolis et collègues (Tsolis et al., 2013), ont réalisé des tests de performance entre les différents logiciels cités plus hauts et d'autres outils similaires, et ont montré que pour un même groupe de protéines à analyser, la moyenne entre sensitivité et spécificité pour ces logiciels est comprise entre 53 et 62%, mettant en évidence que ces logiciels ont des performances égales pour l'étude des propriétés de propension d'agrégation. C'est pourquoi, ces auteurs proposent d'établir un logiciel consensus, appelé **AmylPred**, qui utilise plusieurs de ces logiciels afin de calculer des scores de propension d'agrégation plus robustes que ceux d'un unique logiciel.
Parmi les limitations de ces logiciels, la plus importante est que ces derniers sont exclusifs au calcul de propension d'agrégation pour des protéines globulaires, car les processus d'agrégation pour les protéines membranaires ou fibreuses, qui sont insolubles dans l'eau, sont plus difficiles à caractériser. Certains logiciels comme Aggrescan ont des limites concernant la taille maximale des protéines pour lesquelles le logiciel peut estimer le taux de propension

d'agrégation. Enfin, la plupart de ces logiciels sont actuellement disponibles uniquement sous la forme de serveur web, rendant leur utilisation dans un contexte à large-échelle (c'est à dire pour l'analyse de protéome entier) plus difficile à mettre en place, soit nécessitant des ressources informatiques conséquentes ou un besoin d'accès au serveur web avec des données d'entrée à fournir en en grand nombre.

Étude de la tolérance de mutations des protéines et son lien avec la propension d'agrégation

Face à différents facteurs environnementaux (changements de température, changement de pH, ou stress oxidatif), des altérations moléculaires peuvent se produire au niveau de l'ADN. Il s'agit de mutations génétiques. Lorsque ces altérations ont lieu dans les cellules germinales d'un organisme, susceptibles de former les gamètes (spermatozoïdes et ovocytes chez les animaux), et ne sont pas corrigées par les systèmes de réparation de l'ADN, elles peuvent se transmettre à la descendance. Si ces dommages ont lieu au niveau d'un gène et de sa séquence codante, cela peut engendrer différents types de mutations, plus particulièrement des substitutions de nucléotides dans les séquences d'ADN, qui peuvent affecter, ou non, la séquence en acides aminés de la protéine du gène altéré. Comme le code génétique est redondant, certaines substitutions peuvent être synonymes, c'est à dire que le changement du nucléotide dans un codon résultera en la synthèse du même acide aminé que dans la séquence dite de référence. Dans le cas où les substitutions sont non-synonymes, le codon modifié va entraîner la synthèse d'un autre acide aminé que celui de référence, ce qui entraînera une modification de la composition primaire de la protéine. Les changements physico-chimiques de la séquence en acides aminés de cette dernière peuvent entraîner des changements d'interactions entre les chaînes latérales, ce qui peut avoir des conséquences sur sa conformation tertiaire, et donc sur le repliement de la protéine. Cela peut entraîner l'exposition des régions hydrophobes de la protéine, qui seront alors à l'origine de la formation d'agrégats protéiques. En plus des mutations affectant les protéines elles-mêmes, la formation d'agrégats protéiques peut aussi être causée indirectement lors de différentes étapes de la protéostase cellulaire. Par exemple, elle peut se produire lors de leur prise en charge par les protéines chaperonnes si les protéines chaperonnes sont inactives ou surchargées. Leur apparition peut aussi faire suite si des protéines non-repliées ou mal repliées ne sont pas prises en charge par le système ubquitine-protéasome pour être dégradées.

Il est donc primordial de comprendre les principes généraux permettant d'expliquer la tolérance de mutations par les protéines. Pour quantifier la tolérance des protéines face à des changements aléatoires, il est nécessaire d'estimer la probabilité qu'un remplacement aléatoire d'acide aminés mène à l'inactivation complète de la fonction de la protéine.

L'inactivation fonctionnelle de la protéine se traduit en général par un changement conformationnel qui rend la molécule instable, et qui peut *in fine* être à l'origine de la formation d'agrégats protéiques. L'étude de la tolérance des mutations peut donc renseigner sur les probabilités de propension d'agrégation des protéines. Guo et collègues ont notamment démontré l'intérêt d'étudier expérimentalement la tolérance aux mutations aléatoires dans les protéines par la mise en place de protocoles expérimentaux permettant de réaliser de la mutagenèse à large échelle chez la bactérie *E. coli* (Guo et al., 2004). Leur méthode a permis de mettre en évidence l'importance des effets de mutations sur des résidus spécifiques pour la structure et la fonction d'enzyme, en fonction du nombre et des types de substitutions tolérées. Il est intéressant de noter que les indices de substituabilité des résidus individuels, soit leur tolérance de mutations en rapport avec les substitutions, peuvent être obtenus indépendamment des informations sur la conservation ou sur la structure et sont généralement cohérents avec les deux. D'autres études ont permis la mise en place de bases de données qui listent les conséquences des différentes mutations possibles, obtenues par des méthodes expérimentales, dans l'ensemble du génome de différents organismes modèles, comme par exemple MaveDB, mis en place par Esposito et collègues (Esposito et al., 2019) qui est un répertoire public qui référence des mesures à large-échelle pour l'étude de l'impact des variants au niveau des gènes. Plus récemment, Høie et collègues (Høie et al., 2022) ont étudié plus de 150,000 effets de variants qui ont été déterminés à partir de méthodes expérimentales sur un ensemble de 29 protéines. Pour ces protéines, ils ont aussi analysé leur stabilité et leur conservation au niveau de leur séquence primaire. Plus de la moitié des variants causant des pertes de fonction sont associés à des pertes de stabilité.

Ces différentes études nécessitent la construction de librairies expérimentales, demandant des ressources matérielles et financières conséquentes. Il serait donc pertinent de proposer des stratégies informatiques à large échelle qui permettraient d'étudier toutes les substitutions possibles au sein du génome codant ou du protéome, pour estimer la tolérance de mutation globale d'une espèce. Des outils comme SIFT (pour *Sort Intolerant From Tolerant* en anglais)) (Ng and Henikoff, 2001) se base sur l'homologie entre séquences pour déterminer si la substitution des acides aminés au sein d'une protéine aura un effet neutre ou délétère sur sa fonction. Cet algorithme part du principe que dans l'alignement multi-séquences de la protéine étudiée et de ses homologues, les résidus impliquées dans la fonction de la protéine seront très conservés et donc que leurs mutations seront peu tolérées, alors que les autres résidus seront plus divergeants, donc les mutations pour ces résidus seront plus tolérées. SIFT va par conséquent identifier les régions tolérantes et non tolérantes aux substitutions au sein des protéines étudiées. L'implémentation R de SIFT, SIFTr proposé par Omar Wagih, permet par exemple de mettre en place une stratégie computationnelle à large échelle pour réaliser une mutagenèse *in silico* afin d'étudier des

protéomes entiers. La prédiction des effets de mutation selon cette méthode est cependant uniquement basée sur la conservation des acides aminés au sein de la séquence. Cette méthode ne permet pas de renseigner sur l'impact des mutations sur le repliement des protéines ou sur leur propension à l'agrégation.

Plus récemment, Schwersensky et collègues ont proposé une méthode pour réaliser une mutagenèse *in silico* à large échelle (Schwersensky et al., 2020). Ils ont étudié les changements d'énergie libre impliqués dans le repliement, pour l'ensemble des substitutions possibles pour chaque acide aminé au sein de 20,000 structures secondaires de protéines. À l'échelle des acides aminés, ils ont observé que les résidus impliqués au niveau de la surface des protéines sont souvent plus robustes aux mutations aléatoires que des résidus situés au coeur de la protéine. Cette différence est particulièrement marquée dans les protéines de petite taille. Les mutations déstabilisantes et neutres sont plus nombreuses au niveau du coeur et de la surface des protéines, en comparaison des mutations dites stabilisantes qui forment moins de 4% des mutations observées pour les deux régions (Schwersensky et al., 2020). L'ensemble de ces résultats montre la non-universalité de la tolérance des mutations et ses liens avec les caractéristiques des protéines, le code génétique et l'utilisation des codons.

1.1.5. L'intérêt d'une approche protéome-centrée pour l'étude du vieillissement

Du vieillissement de la cellule sommatique au vieillissement de l'organisme

Dans les organismes, on distingue les cellules somatiques, qui sont les éléments unitaires constituants les différents types de tissus cellulaires, et les cellules de la lignée germinale, qui sont les cellules qui sont transmises à la descendance. Le vieillissement de l'organisme, dans ce contexte, est lié a l'arrêt du renouvellement des cellules somatiques, ou du soma tandis que la lignée germinale ne vieillit pas. Ici, on s'intéressa donc plus particulièrement au vieillissement du soma, l'ensemble des cellules somatiques de l'organisme qui est causé par des dysfonctionnements cellulaires. Le vieillissement est un processus physiologique qui touche tout organisme et ses fonctions dans leur ensemble. Au niveau cellulaire, le vieillissement cellulaire (ou sénescence) se traduit par l'arrêt du cycle cellulaire et de leur prolifération, même dans des conditions de croissance optimale. Par ce phénomène, les informations génétiques sont dupliquées, ce qui permet de maintenir des copies intègres du génome. La division cellulaire contribue à la prolifération cellulaire qui permet la création de tissus, ensemble de cellules ayant des fonctions spécialisées. Quand une cellule entrant en phase de sénescence, subira une mort programmée, qu'on appelle apoptose, les fonctions cellulaires sont ralenties puis mises à l'arrêt avant que le cellule ne soit détruit après une succession de différents processus biologiques (Elmore, 2007). L'apoptose est une étape

nécessaire au renouvellement cellulaire et contribue au maintien de l'homéostasie cellulaire.

A l'échelle de l'organisme, le vieillissement peut se décrire comme le déclin progressif de la santé et des différentes fonctionnalités biologiques, conduisant *in fine* à la mort. C'est un phénomène qui affecte la plupart des organismes vivants. Ce phénomène est caractérisé par le dysfonctionnement de différents processus biologiques, appelées caractéristiques du vieillissement (*hallmarks of aging*, en anglais). Encore aujourd'hui, il reste difficile d'avoir une image complète du vieillissement, où l'on propose des mécanismes moléculaires communs permettant l'origine des dysfonctionnements des différents processus biologiques associés au vieillissement. Dans cette thèse, nous proposerons d'adopter une stratégie protéome-centrée pour étudier le phénomène du vieillissement. Dans les paragraphes suivants, nous mettrons en avant les différents travaux nous permettant de justifier cette stratégie dont les résultats seront présentés dans le chapitre 2 de cette thèse.

Le vieillissement, un déclin progressif de différents processus biologiques

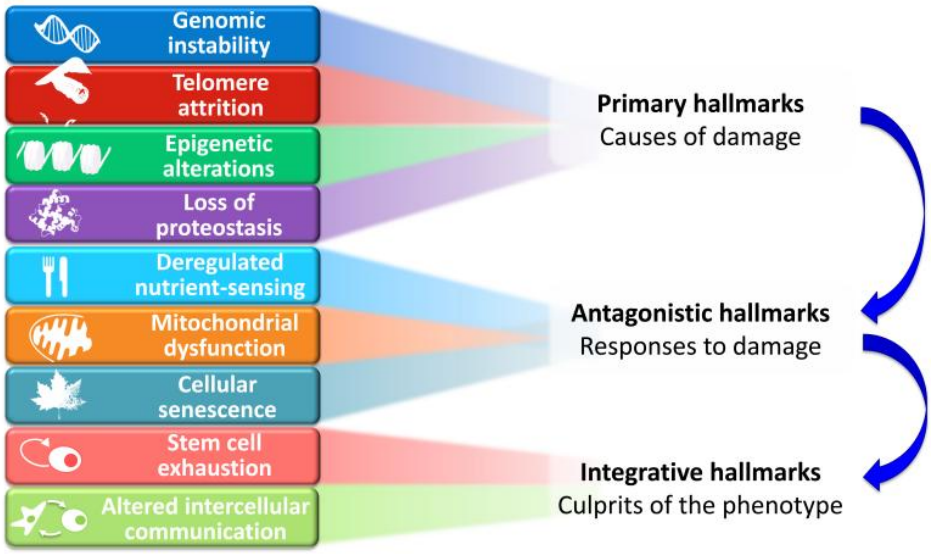

Fig. 1.4. Les différentes caractéristiques du vieillissement.

(Tirée de López-Otín et al., 2013)

Le dysfonctionnement de plusieurs processus biologiques a permis de définir différentes caractéristiques qui permettent de décrire le vieillissement (López-Otín et al., 2013). Ces dernières ont été établies à partir de différents critères, soit leur présence lors du vieillissement normal et leur capacité, après manipulation des gènes associés à ces processus, à retarder ou à accélérer le vieillissement. Elles sont classées en différentes catégories (dommages moleculaires, perturbations des fonctions cellulaires, perturbations extra-cellualires) (Figure 1.4), chacun contribuant à l'apparition du vieillissement. Ces mécanismes ont

été identifiés par différentes méthodes expérimentales dans différents organismes modèles, incluant, la levure, le nématode, la drosophile, la souris et l'humain.

Les caractéristiques primaires. Ce sont les processus biologiques qui peuvent être à l'origine des dysfonctions macro-moléculaires qui s'accumulent progressivement au cours du temps et qui peuvent provoquer différents types de dommages au sein de la cellule.

Le phénomène d'instabilité génomique et le raccourcissement des télomères. Il est à l'origine de l'accumulation de dommages génétiques, plus particulièrement les mutations somatiques, au cours de la vie. Leurs origines peuvent être dues soit à des erreurs de réplication de l'ADN, des réactions hydrolytiques spontanées, où à la suite d'interaction avec des espèces réactives de l'oxygène (ERO) (Hoeijmakers, 2009). Ces mutations sont normalement prises en charge par les systèmes de réparation de l'ADN (dont les mécanismes sont détaillées dans la revue de littérature de Chatterjee et collègues (Chatterjee and Walker, 2017). Leur non-prise en charge peut entraîner l'apparition de phénotype de vieillissement prématuré, ceci étant démontré dans différents organismes modèles, dont l'humain (Maynard et al., 2015; Moskalev et al., 2013). L'accumulation des mutations somatiques peut avoir lieu au niveau du génome nucléaire, comme du génome mitochondrial (Park and Larsson, 2011).
Bien que l'accumulation des dommages de l'ADN apparaisse de manière aléatoire sur l'ensemble du génome lors du vieillissement, il a été constaté que certaines régions génomiques sont particulièrement sensibles aux détériorations liées à l'âge. Il s'agit des télomères, des séquences de nucléotides hautement répétitives qui se situent aux extrémités des chromosomes (E. H. Blackburn et al., 2006). Le raccourcissement progressif survient à la suite d'un problème de réplication au niveau de ces régions. En temps normal, ce phénomène est contré par la télomérase qui rallonge préférentiellement les télomères les plus courts. En son absence, ou en cas de dysfonctionnement de cette enzyme, les télomères courts sont à l'origine de la sénescence réplicative. Ce phénomène a été mis en évidence, notamment chez l'humain et la souris (Blasco, 2007). Ces observations sont à l'origine de la théorie du vieillissement dû aux télomères (Aubert and Lansdorp, 2008).

Les altérations épigénétiques. Elles affectent toutes les cellules au cours de la vie (Talens et al., 2012), elles sont causées par l'action d'enzymes permettant d'ajouter ou de retirer des groupements chimiques au niveau de l'ADN ou des histones. Certains changements associés à la régulation du génome et sa stabilisation à travers des mécanismes épigénétiques ont été précédemment démontrés comme associés au vieillissement. Plus particulièrement, les modifications des histones (par des phénomènes de méthylation ou d'acétylation) ont été observés. Par exemple, il a été observé que l'inhibition de l'histone déméthylase

(pour H3K27) permet l'expansion de la longévité chez *C.elegans*, en ciblant des acteurs moléculaires impliqués dans la voie de signalisation *insulin/IGF-1 signaling* (IIS) (Jin et al., 2011), une voie permettant de réguler la signalisation cellulaire via l'intermédiaire d'hormones incluant l'insuline et des peptides "insulin-like".
Bien que le rôle de la méthylation de l'ADN n'a pas encore été clairement établi, des études ont mis en avant que les événements de méthylation peuvent être utilisés comme des bio-marqueurs de l'âge. Dans la recherche du vieillissement, l'observation de signatures spécifiques de méthylation, a permise la création d'une horloge biologique associée au taux de méthylation global chez différents organismes comme la baleine, le chien ou la souris. Cette horloge épigénétique associée à l'âge permettrait de prédire l'âge biologique, l'âge réel du corps en fonction de l'état fonctionnel des différents systèmes qui le compose (cardio-respiratoire, pulmonaire, osseux, immunitaire etc.) qui le rapproche ou l'éloigne de la mort. Des études ont montré que cette horloge moléculaire serait fortement corrélé à l'âge chronologique ($r \geq 0.8$) (Bell et al., 2019; Horvath and Raj, 2018), qui décrit simplement le temps écoulé depuis de la naissance de l'organisme. Cette horloge épigénétique associée à l'âge aiderait à prédire le risque d'apparition de maladies associées à l'âge. Par exemple, cette horloge épigéntique serait bonne pour prédire les risques de crise cardiaque (Soriano-Tárraga et al., 2021; Soriano-Tárraga et al., 2016).

Le déclin de la protéostase, précédemment défini dans la section 1.1.4, est un marqueur important du vieillissement d'un organisme. On a notamment démontré le rôle important des protéines Hsp dans la longévité. Par exemple, l'inactivation de la protéine régulatrice aux protéines Hsp (Hsf1 chez l'humain), entraîne un raccourcissement de longévité chez *C. elegans* (Walker and Lithgow, 2003). Il a été aussi montré que les protéines Hsp sont sur-exprimées chez les nématodes (Frenk and Houseley, 2018) et chez les drosophiles (Tower, 2011) lors du vieillisement. Ces études mettent en évidence que l'augmentation du stress des cellules avec l'âge est lié á la formation de protéines non-repliées et la formation d'agrégats protéiques.

Les caractéristiques antagonistes. Ici sont regroupées les processus biologiques suivants : 1) la voie de détection des nutriments, 2) la sénescence cellulaire, et 3) les fonctions associées à la mitochondrie. Ces processus biologiques sont nécessaires au bon fonctionnement de la cellule. Cependant, si ces processus sont trop sollicités, en partie par les dommages créés via les caractéristiques primaires, ces processus biologiques deviendront progressivement néfastes. Par exemple, il a été montré que la dérégulation des voies de détection des nutriments impacte la longévité chez différents organismes modèles. Par exemple, la voie mTOR intervient à différents niveaux de régulation pour les processus de prolifération

cellulaire, d'autophagie et d'apoptose, est un régulateur central de la longévité et du vieillissement (Papadopoli et al., 2019). De manière spécifique, l'inhibition de la rapamycine par mTORC1 (une des sous-unités de la protéine mTOR) permet l'allongement de la longévité chez la souris (Harrison et al., 2009). De plus, la modulation des gènes impliqués dans la voie IIS, partie intégrante de la voie de détection des nutriments, permet de compenser ce dérèglement contribuant à l'extension de la durée de vie chez le nématode, la drosophile et la souris (Fontana et al., 2010). La perte de l'intégrité et de l'activité mitochondriale sont aussi associés au vieillissement sain, mais sont également à l'origine de différentes maladies associées à l'âge (Sun et al., 2016). Les liens entre sénescence cellulaire, vieillissement et cancers ont été mis en avant dans le revue de littérature de Campisi et collègues (Campisi, 2013).

Les caractéristiques intégratives. Cette dernière catégorie regroupe les processus biologiques qui apparaissent suite à l'accumulation des dommages créés par les caractéristiques primaires, et pour lesquelles les caractéristiques antagonistes n'ont pas pu compenser par le biais des mécanismes d'homéostasie tissulaire. Parmi ces caractéristiques, on peut citer l'épuisement des cellules souches (Goodell and Rando, 2015). Les cellules souches sont des cellules qui ont la capacité de se différencier en n'importe quel type de cellules spécialisées, et contribuent au renouvellement tissulaire tout au long de la vie. Avec l'âge, il a été mis en évidence que les cellules souches âgées possèdent un taux élevé de dommages ADN, ainsi qu'une déficience de la régulation métabolique, avec une balance altérée entre glycolyse et respiration mitochondriale (Ermolaeva et al., 2018). On peut également citer l'altération de la communication intercellulaire, menant à l'apparition des réponses inflammatoires (B. K. Kennedy et al., 2014; López-Otín et al., 2013). Les réponses inflammatoires sont une réponse biologique après stimulation du système immunitaire à la suite de dommages cellulaires, mais également suite à l'exposition à des composés toxiques ou des pathogènes. Comme discuté dans la revue de littérature de Calder et collègues, différentes études chez l'humain ont mis en évidence la présence de bio-marqueurs de l'inflammation dans le sang de personnes âgées (Calder et al., 2017). Il a été également démontré qu'on retrouve une production élevée de cytokines pro-inflammatoires, des protéines impliquées dans la communication intercellulaire, est retrouvée dans des populations d'individus âgées et sains. (Fagiolo et al., 1993; Franceschi et al., 1995). Enfin, des signatures d'expression génique et épigénétique spécifiques au vieillissement ont été identifiées dans des jeunes souris après induction de réponses inflammatoires (Benayoun et al., 2019).

Différentes théories sur le vieillissement

Il existe différentes théories du vieillissement, on différencie 1) les théories qui décrivent le vieillissement comme un phénomène stochastique ou non-programmé, et 2) les théories

qui le décrivent comme un phénomène programmé.

Les théories non-programmées du vieillissement.
La théorie des liens croisés. Proposé par Bjorksten en 1942, il décrit que le déclin observé lors du vieillissement serait dû à l'interaction accidentelle entre différentes protéines, qui aboutiraient à l'apparition éventuelle de dommages cellulaires et tissulaires.

La théorie de mutation-accumulation. Introduit par Medawar en 1952 (Medawar, 1952), cette théorie affirme que le vieillissement est le résultat d'une accumulation aléatoire de mutations dont les effets n'apparaîtraient que tard au cours de la vie, après la période de reproduction de l'organisme.

Les théories de compromis. En 1957, Williams (Williams, 2001) propose que le vieillissement soit défini par la *théorie de la pléiotropie antagoniste*. Dans cette thoérie, il propose que les gènes influençant différents traits, impliqués dans le phénomène de pléiotropie, peuvent être à l'origine d'avantages évolutifs chez l'organisme au début de sa vie, mais peuvent devenir ensuite néfastes plus tard dans la vie et donc avoir un rôle antagoniste. Ici, on sous-entend que ces gènes auraient été sélectionnés activement au cours de l'évolution, mais que leur implication au sein du vieillissement ne serait qu'un effet collatéral. C'est un prolongement de la théorie de Medawar.

Plus tard en 1977, Kirkwood propose la *théorie du soma jetable* (Kirkwood, 1977). Il souligne que la longévité de l'organisme (des cellules somatiques) n'est qu'un changement de véhicule du point de vue de la lignée germinale. Il pointe la corrélation négative entre activité métabolique et longévité. En conséquence, l'investissement énergétique dans la reproduction ne nécéssite pas une grande maintenance fonctionnelle, ce qui conduit à un vieillissement hâtif.

Ici dans ces différentes théories, ce sont des évènements aléatoires qui contribueraient à l'accumulation de dommages au fil du temps, ce qui à terme, conduit à l'apparition du vieillissement. Cependant, le vieillissement aurait toutefois un intérêt évolutif puisqu'il existe une sélection des processus biologiques contribuant à la protection contre ces événements.

Les théories programmées du vieillissement.
La théorie du coût de la vie. Proposée en 1908 (Ferrucci et al., 2012), cette théorie de Rubner affirme que la longévité d'un organisme est inversement proportionnelle à son taux métabolique de base, quantité d'énergie que brûle l'organisme au repos de manière

quotidienne. Cette hypothèse repose sur l'observation que les animaux ayant un taux métabolique important, vivent plus longtemps que les animaux ayant un taux métabolique plus lent. Toutefois, certains organismes ne suivent pas cette tendance décrite, comme par exemple le rat-taupe nu (Ruby et al., 2018).

La théorie des radicaux libres et des EROs. Cette théorie d'abord proposée Harman dans les années 1950 (Harraan, 1955), permet de décrire comment les organismes, et les cellules qui les composent, vieillissent suite à l'accumulation de dommages oxidatifs au sein de leurs composants, causés par un grand nombre de radicaux libres (comme le dioxygène). Malgré les mécanismes permettant la prise en charge de ses radicaux libres dans la cellule, ces derniers finissent surchargés, résultant à l'accumulation de dommages sur les macromolécules de la cellule (ADN, ARN, protéine).

La longévité programmée. Cette théorie affirme que le génome est intrinsèquement programmé à devenir obsolète au cours du temps. Dans ce contexte, les individus vivant longtemps seraient privilégiés, possédant un génome intrinsèquement plus stable, ce qui retarderait leur vieillissement. Quant aux individus normaux, ils posséderaient un génome intrinsèquement "normal", et présenteraient des signes du vieillissement considérés comme normal (Davidovic et al., 2010).

Dans ces théories, le vieillissement serait dépendant de la régulation de l'expression des gènes dans les processus biologiques associés à la maintenance et la réparation de l'instabilité génomique, ainsi que les mécanismes de réponse au stress et à la prise en charge des dommages oxidatifs. La régulation de ces processus biologiques deviendrait moins performante au cours du temps, ce qui expliquerait l'apparition du vieillissement, potentiellement causé par l'accumulation de dommages, au niveau moléculaire et structural de la cellule.

Bien que toutes ces théories du vieillissement mettent en avant différents mécanismes pour expliquer son apparition, toutes s'accordent sur le fait que l'accumulation des dommages contribue à l'apparition de ce phénotype. L'étude des différents processus biologiques associés aux caractéristiques du vieillissement montre qu'il s'agit bien d'un phénotype complexe multi-génique. Il reste encore beaucoup d'interrogations sur les mécanismes moléculaires entraînant l'apparition du vieillissement. A l'heure actuelle, il n'existe pas de modèle consensus permettant d'interconnecter l'ensemble des caractéristiques du vieillissement. Selon les organismes modèles étudiés, les mécanismes de régulation des gènes associés au vieillissement et leurs conséquences au niveau des phénotypes varient grandement d'un organisme modèle à un autre. Il est donc pertinent de trouver une définition plus systémique du vieillissement, qui ne se concentre pas sur des processus biologiques spécifiques, mais sur des principes

généraux responsables de la perte d'intégrité cellulaire et qui mettent en avant les mécanismes permettant de ralentir / retarder l'accumulation des dommages résultant au déclin fonctionnel contribuant à l'apparition du vieillissement.

1.1.6. Les études de génomique comparée pour étudier le vieillissement

Les études de génomique comparée sont des études qui s'appuient sur la conservation des séquences (ADN, ARN et protéines) à travers l'ensemble des espèces grâce à des mécanismes spécifiques de l'évolution. La génomique comparée permet d'identifier les similitudes et différences entre espèces, afin de comprendre comment les comportements et la biologie des êtres vivants changent au cours du temps. Les études de génomique comparée repose principalement sur l'identification de séquences orthologues, ce sont des séquences ayant une similarité de séquence supérieure à 50% et dont les fonctions moléculaires sont similaires (Remm et al., 2001). Ces séquences proviendraient de deux espèces ayant subi un événement de spéciation, un phénomène évolutif au cours duquel la descendance d'un ancêtre commun permettra la création de deux espèces génétiquement différentes. Les méthodes informatiques permettant l'identification des séquences orthologues reposent sur des algorithmes d'alignement comme Blast (Altschul et al., 1990). De nombreuses bases de données existent pour établir des listes de protéines orthologues pour différentes espèces comme Inparanoid (Remm et al., 2001), ou OrthoDB (Kriventseva et al., 2019).

Le vieillissement est un phénomène régulé par des processus conservés à travers les espèces. L'inhibition de l'expression de certains gènes des voies cellulaires impliquées dans l'apparition du vieillissement, comme la voie IIS et mTOR, entraînent une augmentation de durée de vie chez la levure, le nématode, la drosophile et la souris (Partridge and Gems, 2007; Piper et al., 2005; E. D. Smith et al., 2007). Plusieurs études ont démontré la conservation de certaines caractéristiques du vieillissement à travers les espèces eucaryotes (López-Otín et al., 2013; Singh et al., 2019). Par exemple, différents gènes associés aux caractéristiques du vieillissement ont une expression génique conservée, notamment ceux liés au métabolisme mitochondrial, à la réparation de l'ADN et à la dégradation des protéines (McCarroll et al., 2004). On retrouve également une conservation d'expression génique dans les processus d'inflammation, de réponse immunitaire et de sénescence cellulaire chez la souris, le rat et l'humain (de Magalhães et al., 2009). Il a également été démontré que certaines caractéristiques du vieillissement sont spécifiques aux vertébrés, comme la sénescence, l'épuisement des cellules souches, et les réponses inflammatoires (Singh et al., 2019).

Les données de longévité des espèces (Tacutu et al., 2018) couplées aux données issues des technologies à haut débit sont des informations importantes pour élaborer une stratégie et comprendre la diversité des espérances de vie. L'espérance de vie maximale mesure la durée maximale pendant laquelle un organisme vit. Des études ont démontré qu'il existait une corrélation positive entre l'espérance de vie des espèces et leur ratio taille / poids (Magalhães et al., 2007). Alors que les caractéristiques associées au vieillissement ont tout d'abord été identifiées dans des organismes modèles conventionnels tels que la levure, le nématode, la drosophile et la souris (voir section 1.1.5), il serait toutefois intéressant d'étudier des espèces qui ont des longévités dites "exceptionnelles". En effet, des signatures moléculaires spécifiques contribuant à l'apparition tardive de leur vieillissement ont été identifiées dans des espèces avec de espérance de vie plus élevées, tels le rat-taupe nu (plus de 35 ans), l'humain (plus de 120 ans), la baleine (plus de 200 ans). Ces résultats sont détaillés dans la revue de littérature de Ma et Gladyshev (Ma and Gladyshev, 2017).

1.1.7. L'étude des propriétés intrinsèques du protéome pour comprendre les différences d'espérance de vie

Comme mentionné précédemment, les acteurs de la protéostase (voir section 1.1.4) sont importants dans l'apparition du vieillissement (voir section 1.1.5). En effet, la surcharge ou le dysfonctionnement de ces systèmes de contrôle qualité de la protéostase contribuent à l'apparition du vieillissement (Morimoto and Cuervo, 2009), de certaines maladies associées à l'agrégation des protéines, comme la maladie d'Huntington et la sclérose latérale amyotrophique (Ross and Poirier, 2004) et d'autres maladies associées à l'âge dont le diabète de type II ainsi que les maladies d'Alzheimer et de Parkinson (Chiti and Dobson, 2017; Irvine et al., 2008).

L'identification des principes fondamentaux qui caractérisent les systèmes de contrôle qualité de la protéostase est nécessaire pour comprendre ce qui explique le déclin de ces systèmes dans le contexte du vieillissement (Morimoto and Cuervo, 2014). Le maintien de l'homéostasie cellulaire peut s'étudier au niveau moléculaire en identifiant les propriétés associées à la robustesse intrinsèque des protéines. La tolérance aux mutations d'une protéine peut se quantifier par sa capacité à maintenir son état fonctionnel en dépit de la présence de mutations dans sa séquence. La sensibilité à ces dommages peut entraîner à terme la formation d'agrégats dont il faut comprendre les mécanismes d'apparition. En effet, le changement des propriétés physico-chimiques par la substitution d'acides aminés dans la séquence d'une protéine peut changer sa conformation structurelle, pouvant entraîner la formation d'agrégats (David, 2012). La tendance de formation des agrégats, ainsi que la tolérance des mutations, sont deux propriétés intrinsèques des protéines, qui semblent

être des principes fondamentaux clefs pour le maintien de la protéostase. Leurs rôles mécanistiques restent encore à approfondir dans le contexte du vieillissement.

Dans le chapitre 2 de cette thèse, nous mettrons en avant l'importance d'étudier ces caractéristiques intrinsèques au niveau de l'ensemble du protéome, à l'aide de différentes approches informatiques (voir la section 1.1.4) pour comprendre les différences d'espérance de vie entre deux organismes phylogénétiquement proches en utilisant des méthodes de génomique comparée.

Dans la suite de ce chapitre, nous allons présenter les approches modernes pour étudier les phénotypes à partir des données moléculaires.

1.2. Concepts et approches bio-informatiques pour l'étude des phénotypes

1.2.1. La relation entre génotype et phénotype

Par opposition au génotype, qui décrit les informations génétiques d'un individu qu'il hérite de la génération parentale, on définit le phénotype comme l'ensemble des traits observables d'un organisme. Certains phénotypes sont héritables ou sont acquis au cours de la vie, c'est à dire qu'ils sont transmis d'une génération parentale à leur descendance directe. Il a été démontré que l'ADN est une source de l'héritabilité, ce qui signifie que les phénotypes héritables sont encodés par les gènes.

En génétique des populations, on étudie la distribution et les changements de la fréquence des versions de l'ADN (qu'on appelle allèles) dans les populations d'êtres vivants, sous l'influence des mécanismes de "pressions évolutives" comme :

- les mutations, qui sont les variations sur laquelle agissent des facteurs stochastiques et déterministes,
- la recombinaison méiotique, qui permet l'échange de matériel génétique entre deux individus parentaux et qui permettra la génération d'une descendance avec des combinaisons d'allèles différentes de celles observées chez les individus parentaux,
- la dérive génétique, processus par lequel les fréquences alléliques changent dans les populations à cause de biais aléatoires d'échantillonnage dans la transmission des allèles d'une génération à l'autre,
- la migration, déplacement géographique d'une population d'un point A à un point B, favorisant le métissage des populations, et donc une plus grande diversité génétique,
- la sélection naturelle, où à chaque génération, les génotypes qui favorisent la reproduction, c'est à dire la survie jusqu'à la procréation ainsi que le succès reproducteur, contribuent de manière disproportionnée à la prochaine génération.
- la selection purificatrice qui va faire disparaître les variations qui empêchent la survie/reproduction

.

La valeur adaptative (ou *fitness* en anglais) est une quantité qui décrit la capacité d'un individu à se reproduire. A l'échelle d'une population, les variations phénotypiques sont le résultats d'une relation complexe entre les variations génétiques et l'environnement. Les variations génotypiques les plus favorisés, ou en d'autres mots qui ont la plus grande valeur adaptative, sont celles qui permettent la reproduction et la survie des espèces, c'est ce qu'on appelle l'adaptation évolutive qui se traduit soit 1) par le maintien d'une mutation bénéfique (sélection positive) ou 2) l'élimination d'une mutation délétère (sélection négative) au cours des prochaines générations (Gray, 1860). Ces phénotypes héritables regroupent

un ensemble de traits d'histoire de vie propre à chaque espèce (ex : espérance de vie, taux métabolique basale, fin de la maturité sexuelle etc.), influencés par la contribution de milliers de caractères dits phénotypiques.

Vischer et collègues proposent de définir la relation entre le phénotype P et le génotype G d'un individu par l'équation suivante (Visscher et al., 2008) :

$$P = G + E \tag{1.1}$$

où E représente l'effet de l'environnement sur le phénotype.

Au niveau de la population, une relation similaire peut être définie pour la variance phénotypique Var(P), qui représente la variance phénotypique dans la population :

$$\sigma^2(P) = \sigma^2(G) + \sigma^2(E) \tag{1.2}$$

où $\sigma^2(G)$ représente la variance génétique dans la population
et $\sigma^2(E)$ la variance attribuée à l'environnement.

Cette relation peut être complexifiée en tenant compte, par exemple, des interactions entre le génotype et l'environnement, sous la forme d'un terme de covariance entre G et E ($\sigma^2(GxE)$) (Moore et al., 2019). Dans cette thèse, nous nous concentrons principalement sur les méthodes permettant d'identifier la contribution de la variance génétique dans la variabilité phénotypique.

Phénotypes mendéliens

Gregor Mendel, connu pour avoir établi les premières théories sur la génétique en utilisant les pois (Mendel, 1865), a établi 3 principes fondamentaux pour expliquer l'héritabilité d'un trait qui sont :

- la loi de dominance (ou loi d'uniformité des hybrides de première génération),
- la loi de ségrégation (ou loi de disjonction des allèles),
- la loi de l'assortiment indépendant.

Les traits qui sont hérités en respectant ces 3 lois de Mendel, sont appelés traits mendéliens. Les traits héritables mendéliens sont forcément associés à des allèles qui ont une relation dominance-récessivité établie. Quand un allèle est dominant, cela signifie que le trait ou phénotype héritable dépend intégralement de la version d'un gène, sans contribution de la seconde version. La première version sera appelée allèle dominant, alors que la seconde version sera appelée allèle récessif. Cela signifie qu'un individu hétérozygote (présentant

une copie dominante et une copie récessive du gène) exprimera le trait dominant puisqu'il en possède un allèle dominant permettant l'expression du trait. L'expression d'un trait héritable ne dépendant que de l'allèle dominant est appelé dominance complète, et dans ce cas de figure, l'allèle dominant masque l'expression de l'allèle récessif chez les organismes hétérozygotes.

Pour exprimer le phénotype récessif (phénotype moins répandu dans la population), l'individu devra alors présenter deux allèles récessifs provenant chacun de ses parents. Les traits purement mendéliens représentent une infime minorité de tous les traits, puisque la plupart des traits phénotypiques présentent une dominance, une codominance et des contributions incomplètes de nombreux gènes. Comme exemple de traits mendéliens, on peut citer l'albinisme, le daltonisme, la distrophie musculaire de Duchenne ou la maladie d'Huntington (Klug and Ward, 2013). Les traits hérités selon la transmission mendélienne peuvent être représentés à l'aide du modèle de Wright-Fisher (Ewens, 2004), où on observe que la distribution du trait au sein de la population se fait en fonction d'une distribution normale, où le phénotype dominant sera présent dans la majorité de la population alors que le phénotype récessif sera présent dans un nombre plus restreint d'individus. Ce modèle est seulement applicable dans le cadre de l'étude d'une population idéale qui se définit en fonction de différentes propriétés :

- les individus sont diploïdes,
- la reproduction est sexuelle et les fréquences sont égales entre humains et femmes,
- les générations sont discrètes, et ne se chevauchent pas.

De par ces propriétés, on considère que les fréquences génétiques sont ici intrinsèquement stables.

Phénotypes non-mendéliens

Par définition, ces traits ne suivent pas au moins l'une des trois lois de Mendel, c'est pourquoi on ne peut pas déterminer les traits non-mendéliens par des allèles dominants ou récessifs. De plus, ils impliquent souvent plus d'un gène. Parmi les différents types d'héritabilité non-mendélienne, on distingue :

La dominance incomplète (Frizzell, 2013). Cette héritabilité résulte d'un croisement où la contribution de chaque parent est génétiquement unique, ce qui fait que la progéniture présentera un phénotype intermédiaire mélangeant des caractéristiques des phénotypes parentaux. On appelle cette héritabilité, la demi-dominance ou la dominance partielle. Contrairement à la codominance, aucun des deux allèles n'a l'ascendant sur le phénotype résultant au niveau de la descendance. On retrouve des exemples de dominance incomplète dans la nature, notamment dans la variation phénotypique des couleurs de fleurs, ou pour

la couleur de certaines robes de chats, où les couleurs des pétales et pelages des phénotypes des descendants directs sont des mélanges de couleurs plus ou moins diffus des couleurs initialement présentes chez les phénotypes parentaux.

La codominance (Xia, 2013). Il s'agit du phénomène génétique dans lequel les produits géniques (ARN) des deux allèles d'un individu hétérozygote sont produits en quantités à peu près égales. À la suite de l'expression des deux copies co-dominantes du gène donné, on pourra avoir deux transcrits différents, ou la production de protéines différentes suite à la transcription de ces différents transcrits, ou encore la production de différents métabolites associés à l'activité enzymatique des transcrits ou des protéines allèle-spécifiques. L'exemple de codominance le plus connu chez l'humain est le groupe sanguin AB, où le locus permettant de produire les antigènes fixées aux globules rouges (locus ABO) exprime à la fois des antigènes A et des antigènes B.

Les allèles multiples. Dans l'héritabilité mendélienne, on considère qu'il ne peut y avoir plus de deux versions d'allèle possibles pour un trait spécifique. Mentionné plus haut, le locus ABO possède trois allèle différents, qui permet d'encoder les groupes sanguins distincts A, B et O.

L'héritabilité liée au sexe. Chez l'humain, certains traits dépendent de locus situés au niveau des chromosomes sexuels (les chromosomes X et Y). Souvent les traits influencés par ce mode d'héritabilité sont associés à des gènes situés sur le chromosome X (comme le chromosome Y est beaucoup plus petit en taille). Ce sont plus souvent les hommes qui seront souvent sujets à l'expression des phénotypes liés au sexe car ces derniers n'ont qu'une seule copie du chromosome X, alors que les femmes présentent deux copies de ce chromosome, mais dont une copie est inactive.

Les traits polygéniques. Aussi appelés traits complexes, ces traits s'expriment comme le résultat de l'interaction d'expression entre différents gènes et l'effet additif de plusieurs gènes. Ici le modèle de Wright-Fisher ne peut pas permettre d'expliquer l'héritabilité des traits complexes, car il a été constaté que la contribution de chaque gène devient proportionnellement plus petite, ce qui conduit à une limite au "modèle infinitésimal" de Wright-Fisher (Barton et al., 2017). Parmi les exemples de traits complexes, on peut citer, des phénotypes associés à la morphologie, comme la taille, le poids et la couleur de peau. Des phénotypes associés à des états physiologiques, comme le vieillissement, ou des états pathologiques, comme le diabète de type II, ou la schizophrénie sont considérés comme des traits non-mendéliens. Cette thèse s'intéressera plus spécifiquement à ce type de phénotypes.

Pour étudier les phénotypes non-mendéliens, les approches ciblées sur quelques gènes, dont la résolution est de seulement quelques polymorphismes de nucléotides uniques (SNP, pour *Single Nucleotide Polymorphisms* en anglais), ne suffisent pas. C'est pourquoi, il est nécessaire d'utiliser des approches non biaisées plus systémiques qui permettent d'identifier un nombre importants de SNPs associés à ces traits, au sein d'une grande population. La section suivante va donc s'intéresser plus particulièrement aux études d'association pangénomiques (GWAS, pour *Genome-Wide Association Studies* en anglais).

1.2.2. Les GWAS pour l'identification des gènes associés aux phénotypes complexes

Avec les avancées technologiques et la réduction du coût du séquençage et génotypage, il est maintenant possible d'analyser la variation génomique de nombreux individus au sein d'une population. Les GWAS (Uffelmann et al., 2021) permettent l'analyse de nombreux variants génétiques chez de nombreux individus, afin d'étudier leurs associations avec des traits phénotypiques. La stratégie GWAS peut se diviser en différentes étapes (Figure 1.5), que nous allons ici détailler dans les différents paragraphes suivants.

Collecte des données

Pour étudier les gènes associés à un phénotype donné, que ce soit un trait ou une maladie mendélienne comme non-mendélienne, il sera nécessaire d'obtenir différents types d'information permettant de caractériser des individus issus de deux populations distinctes, 1) la première présentant le phénotype d'intérêt (population cible), 2) la seconde ne présentant pas le phénotype d'intérêt (population de contrôle) :

Les données génotypiques de chaque individu dans les deux populations. Il s'agit de l'ensemble des informations sur les variants génétiques, qu'on appelera génotype. Le génotype permet d'établir quels allèles sont présents pour chaque variant génétique localisé dans une région génomique spécifique (i.e. un locus) sur l'ensemble du génome (Adams, 2022). Cette notation des variants peut se faire de différentes façons : 1) avec les nucléotides de la séquence ADN sur lequel se situe le variant (i.e. CC, CT, TT) 2) avec des symboles ayant une syntaxe définie (i.e. BB, Bb, bb), 3) sous la forme de chiffres (i.e. 1/-1 ou 1/0 etc.). En fonction de l'espèce étudiée, on peut avoir 1) des individus haploïdes comme chez la levure au stade de spores, ce sont des individus qui n'ont qu'une seule copie de gène provenant d'un des deux parents, ou on peut avoir 2) des individus diploïdes comme chez l'humain, où les individus possèdent, eux, les deux copies de gènes parentales. Dans le premier cas, on aura deux possibilités de notation, l'une représentant le génotype de référence, et le génotype alternatif. Dans le second cas, on aura trois possibilités de notation,

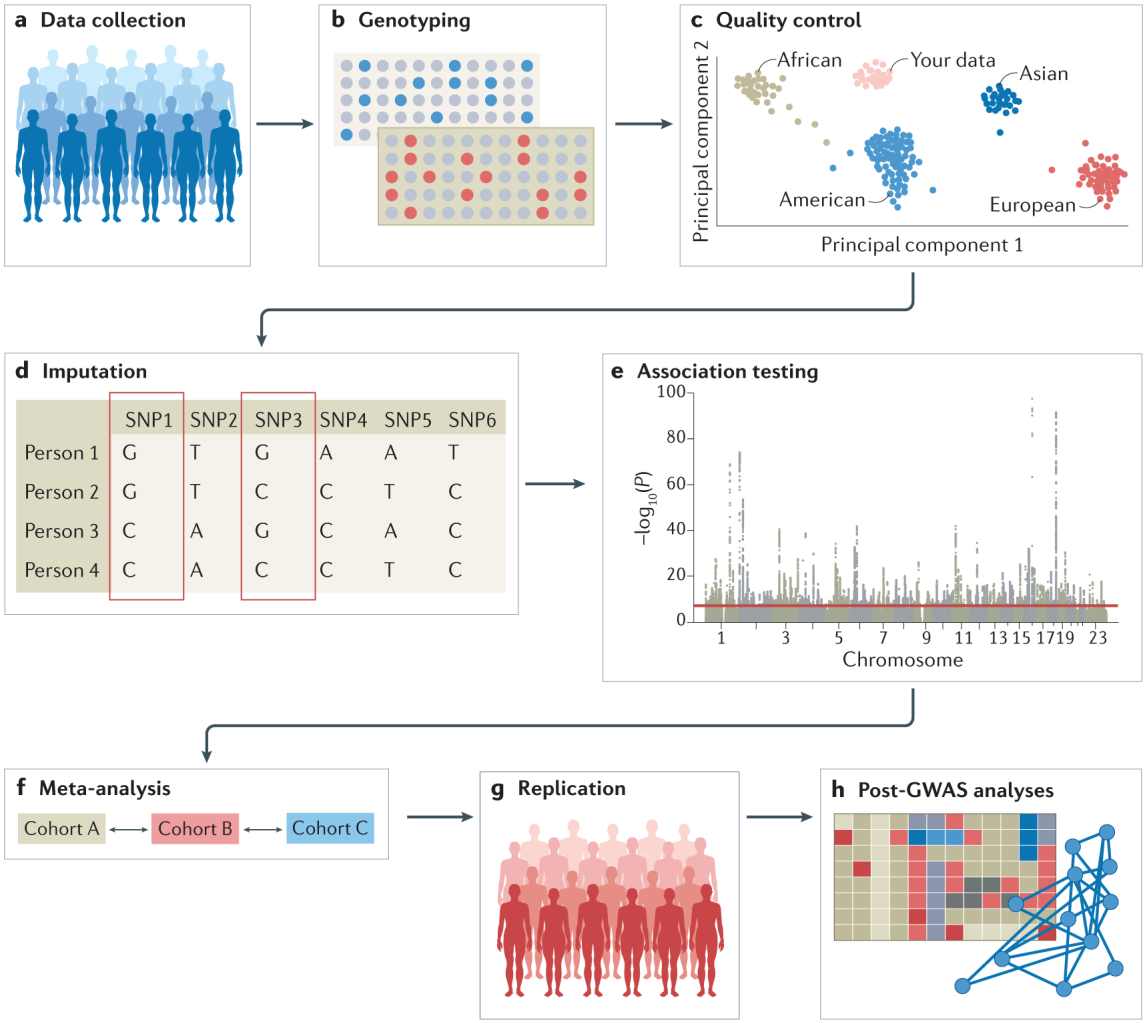

Fig. 1.5. Les différentes étapes d'un GWAS.

(Tirée de Uffelmann et al., 2021)

qui correspondent respectivement au génotype homozygote dominant (BB), au génotype hétérozygote (Bb) et au génotype homozygote récessif (bb). Par exemple, le projet 1000 Génomes (1000 Genomes Project Consortium et al., 2015) est un catalogue des variations génétiques les plus communes, provenant des génomes d'individus se déclarant en bonne santé et ayant consenti à mettre à disposition leurs données génotypiques pour la science. Elle contient ∼2,500 génomes non-apparentés provenant de 26 populations différentes qui ont été séquencés à l'aide de différentes techniques de séquençage. Ce catalogue contient plus de 88 millions de variants génétiques, dont 84.7 millions de SNPs, 3.6 millions de régions courtes d'insertion/délétions (indels) et 60,000 variants structuraux. Cette base de données peut être utilisé pour établir l'ethnicité génétique des populations cible et contrôle étudiées dans les GWAS.

Les données phénotypiques de chaque individu dans les deux populations. Ce sont les données qui permettent de décrire le phénotype étudié. Dans le cas d'un GWAS de type population cible comparé à la population de contrôle où on veut identifier les variants génétiques potentiellement lié à un phénotype ou une maladie étudiée, la population cible présentera le trait d'intérêt alors que la population de contrôle ne le possédera pas. Dans ce cas-ci, les données phénotypiques seront représentées par une variable qualitative binaire (ex : présence *vs.* absence du phénotype, traitement *vs.* placebo etc.) Le phénotype peut aussi être décrit à l'aide de variables quantitatives (ex : des traits moléculaires mesurables comme l'expression de gènes spécifiques, ou leur abondance protéique). Ce sont ces variables qui seront comparées à la variance associées aux données génotypiques avec lesquelles on souhaite connaître leur association. Par exemple, la base de données GTEx (The GTEx Consortium, 2017; The GTEx Consortium et al., 2020) peut être utilisée pour étudier les niveaux d'expression des gènes préalablement établis comme étant associés à des maladies à travers les tissus (ex : le gène *CETP* pour les maladies coronariennes d'origine artérielle (Gamache et al., 2021)). Un total de 54 tissus ont été prélevés sur presque 1,000 individus décédés. L'expression de ces gènes provient principalement d'études moléculaires, permettant la quantification de l'expression de gènes par les techniques de micro-puces et ou de séquençage d'ARN (abrégé en *RNA-seq*, en anglais).

Les données pouvant être utilisées comme des co-variables. Ce sont des valeurs de variables quantitatives (ex : âge, poids, taille etc.) ou qualitatives (ex : sexe, consommation de tabac etc.). Ces variables n'ont pas de lien préalablement établi avec le phénotype d'étude, mais peuvent avoir un effet sur ce dernier. Elle pourront être inclus dans les modèles stastistiques permettant de décrire la relation génotype-phénotype pour corriger leur effet. On peut aussi les utiliser pour séparer la population en sous-population afin de faire des analyses par stratification. Par exemple, la base de données UK Biobank (Sudlow et al., 2015), en plus de fournir les données génotypiques et des données phénotypiques associées à certaines maladies (ex : présence de bio-marqueurs pour les maladies cardiaques), pour des individus provenant de la population du Royaume-Uni, rapporte aussi différentes informations pouvant êtres utilisées comme co-variables (ex : sexe, âge, consommation d'alcool, ethnicité etc.).

Lors de cette étape de collecte des données, il est important de minimiser l'introduction de biais (des facteurs externes ou internes qui modifient une association entre les variants génétiques et le phénotype étudié). Il faudra par exemple être vigilant à contrôler la structure populationnelle des cohortes étudiées, qui doivent être idéalement des échantillons d'individus provenant de population dite idéale. Cela signifie que pour ces deux cohortes, il est nécessaire

de vérifier leur structure populationnelle, ainsi que différentes conditions pré-établies telles que :

- les individus doivent provenir d'une très grande population,
- les événements d'accouplement doivent être aléatoires,
- il ne doit pas avoir d'événement de migration ou de métissage au sein de la population,
- la population est stable en ce qui concerne la fréquence des allèles et des génotypes, qui ne changera pas à travers les générations. On appelle ce principe l'équilibre de Hardy-Weinberg. (Abramovs et al., 2020).

En pratique, ces conditions sont très difficiles à respecter, et les études doivent mentionner toute évidence de violation de ces hypothèses de base, qui pourraient influencer les résultats.

Génotypage

Les données génotypiques, mentionnées lors de l'étape de collecte des données, peuvent être obtenues par le biais de différentes stratégies expérimentales.

Les puces de génotypage. Les puces de génotypage, aussi appelées puces à SNPs, sont des plaques à ADN qui permettent de fournir des informations sur plusieurs centaines à milliers de variants génétiques. Pour identifier les SNPs d'une population cible, des amorces d'oligonucléotides synthétiques sont créées en complémentarité des allèles cibles étudiés. Les séquences modèles de ces amorces proviennent du séquençage d'un panel d'individus de référence, où les positions génétiques variant à une fréquence spécifique sont définies comme des SNPs. Ces amorces sont fixées à une plaque, sur laquelle on effectuera une étape d'hybridation avec des fragments de séquences d'ADN provenant de la population cible, qui auront été préalablement étiquetés avec des éléments fluorescents, afin de permettre la quantification des séquences hybridées. Ainsi, pour chaque SNP étudié sur cette plaque à génotypage, on pourra identifier pour les différents individus de la population cible leurs différents allèles, qu'ils soient similaires à ceux de la population de contrôle, où qu'ils soient différents, et seront donc considérés comme des allèles alternatifs.
Les puces de génotypage sont caractérisées par leur nombre de variants génétiques cibles qui peut atteindre entre 1.9 et 2.2 millions de SNPs (Ceballos et al., 2018), ne constituant que 2% des SNPs identifiés dans le génome humain (LaFramboise, 2009). Les puces de génotypage sont idéales pour faire des GWAS sur des phénotypes ou maladies où les SNPs potentiellement associés à leurs traits héritables sont communs, et donc déjà identifiés au sein d'une population de référence.

Une population de référence est une population d'individus qui doit représenter de manière impartiale l'ensemble de la population humaine. Elle permet d'identifier les variants

génétiques et leurs fréquences qui seront utilisés comme statistiques de référence lors de comparaison avec des populations d'individus plus spécifiques. Comme exemple, on peut citer les groupes d'individus présentant des phénotypes particuliers comme les "centenaires" qui sont des groupes d'individus dont l'espérance de vie excède la centaine d'années (Perls, 2007), ou des groupes ethniques présentant des maladies qui leur sont uniques, par exemple comme l'ataxie récessive spastique de Charlevoix-Saguenay, une maladie unique à la population canadienne française des régions Charlevoix et Saguenay, au Québec (Thiffault et al., 2013). C'est donc à partir de la population de référence, qu'on va identifier les régions génomiques où se trouvent les SNPs à étudier et qui permettront le design des amorces utilisées dans les puces de génotypage. En revanche, ce type de technologie ne permet pas d'identifier de nouveaux SNPs (non présents dans la population de référence) qui pourraient être associées à ces traits.

La technique de WGS et WES. La technique WGS permet de séquencer le génome complet d'un individu, en déterminant l'entièreté de la séquence d'ADN en une fois. Dans le cadre de la médecine personnalisée, cette technologie permet d'identifier l'ensemble des variants génétiques spécifiques à un individu, dans des régions codantes comme non codantes du génome. On peut choisir de ne s'intéresser qu'aux variants génétiques présents dans les régions codantes du génome, dans ce cas-ci, il s'agit de la technique WES, qui étudie l'exome entier d'un individu, soit l'ensemble des exons, incluant les CDS à l'origine des protéines, les entités moléculaires structurelles et fonctionnelles permettant de décrire un phénotype, et leurs régions non-transcrites 5' et 3' . L'utilisation de cette technique, revient à étudier seulement 1.5% du génome humain, et à l'identification d'environ 3 millions de SNPs (Green, 2022). Pour ces deux techniques, les technologies standards de séquençage à haut débit sont couramment utilisées, telles que les appareils de la plateforme Illumina, la technique de pyséquençage et le séquençage *Single Molecule, Real-Time* (SMRT). Toutes ces méthodes reposent sur la stratégie *Shotgun* (Green, 2022), qui consiste à fragmenter de manière aléatoire le génome en plusieurs segments (appelés lectures), qui seront séquencés de manière individuelle. Par la suite, des algorithmes d'alignement vont permettre de reconstituer le génome, en chevauchant les différentes séquences d'ADN dans le bon ordre, en comparant ces lectures à un génome de référence. Plus récemment, de nouvelles technologies, comme le nanopore, permettent également de faire du séquençage entier d'exome ou de génome. Elles permettent de générer des lectures plus longues, en comparaison aux méthodes mentionnées précédemment.

Pour les trois méthodes de génotypage mentionnées, il est possible d'utiliser des stratégies informatiques standardisées pour identifier les différents variants génétiques, dont les SNPs et les indels. Celle proposée par la Broad Institute est connue sous l'appelation anglaise

Genome Analysis ToolKit (GATK) qui consiste en différentes étapes pour chaque individu au sein de la population étudiée (Figure 1.6):

(1) l'alignement de lectures généré par la stratégie *Shotgun* à un génome de référence;

(2) la détermination des positions qui sont des SNPs, c'est l'étape d'appel de SNPs;

(3) la cartographie de ces SNPs à des régions génomiques annotées (ex : gènes, exons, régions non transcrites (ex: promoteurs, introns régions 5' et 3') etc.).

Si la dite région n'est pas annotée, le SNP sera caractérisé par les éléments génomiques les plus proches (ex : le(s) gène(s) le(s) plus proche(s) pour les SNPs dans les régions intergéniques).

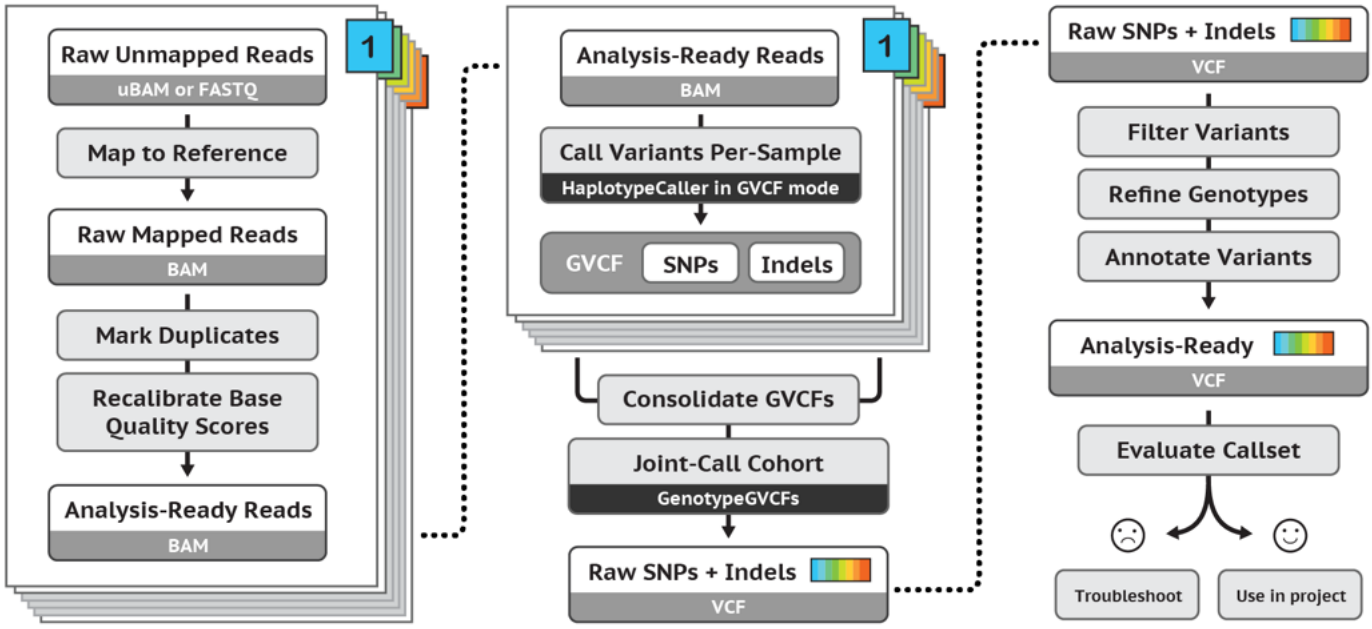

Fig. 1.6. Les étapes d'identification de SNPs et indels selon les pratiques GATK.
(source : GATK - Broad Institute)

Conditions pour un génotype de bonne qualité

Phasage et d'imputation. Après l'étape de génotypage, il est important de contrôler la qualité des génotypes de tous les individus. En effet, une mauvaise qualité pourrait générer des biais qui peuvent influencer la structure populationnelle (voir section 1.2.2). Dans de nombreux cas il convient de phaser les génotypes, c'est à dire d'établir quels génotypes se trouvent sur le même chromosome parental. Ainsi, l'étape de phasage permet d'inférer des haplotypes à partir de données de génotypes (A. N. Blackburn et al., 2020). Cela consiste à identifier de quel parent provient chaque allèle génotypé au sein de l'individu. Un haplotype représente les combinaisons des allèles pour différents variants génétiques qui ont été hérités ensemble depuis le génome d'un des deux parents(Biesecker, 2022). Browning et Browning proposent une revue de littérature des différentes méthodes informatiques pour l'obtention

d'haplotypes (Browning and Browning, 2011).

La génération des haplotypes permet d'identifier les régions génomiques identiques par descendance et la détection des erreurs de séquençage et leur correction. Ils sont également très utilisés pour faire l'imputation de génotypes. L'imputation a pour but de déterminer des génotypes ambiguës ou inconnus grâce à des prédictions qui utilisent l'agrégation de données génotypiques provenant de populations déjà génotypées. Souvent, on tentera de réaliser cette imputation à partir d'une population de référence. Il existe différentes stratégies informatiques pour imputer les génotypes d'une population. Parmi les plus populaires, on retrouve les algorithmes reposant sur l'utilisation de modèles cachés de Markov (Fernández et al., 2013; Marchini and Howie, 2010), qui sont des modèles statistiques permettant de prédire les génotypes en utilisant les variations de fréquence au sein des haplotypes et qui prennent en compte le concept de déséquilibre de liaison (Clouard et al., 2022). Les variants génétiques de faible qualité après les étapes de phasage et d'imputation sont en général exclus de l'analyse GWAS.

Élagage des SNPs à l'aide du concept de déséquilibre de liaison. En génétique des populations, on définit le déséquilibre de liaison (LD, pour *Linkage Desequilibirum*, en anglais) dans une population comme l'association non aléatoire d'allèles à des locus différents (Ytournel, 2008). On parle de LD si la fréquence des allèles de deux loci différents A et B est différente de ce que donnerait une association aléatoire de ces allèles. Pour illustrer cette définition, Pritchard et Przeworski (J. K. Pritchard and Przeworski, 2001) proposent l'exemple suivant : si on suppose qu'un allèle A à un locus X, et un allèle B à un locus Y ont des fréquences définies, respectivement , f_A et f_B au sein d'une population. Si ces deux loci sont indépendants, cela signifie qu'on aura l'apparition de l'haplotype AB au sein de la population, avec une fréquence théorique $f_{AB} = f_A * f_B$. Si on observe que la fréquence de l'haplotype AB au sein de la population est soit plus haute ou plus basse que cette fréquence théorique, alors les deux allèles ne sont pas considérés comme indépendants, et sont dits en LD.

Le déséquilibre de liaison est un signe qu'il y a association préférentielle entre deux allèles. Ce phénomène est créé par différentes forces évolutives dont, la sélection, les mutations, le mélange des populations (ou métissage), la dérive génétique et les goulots d'étranglement (ou réduction de la taille de population). L'étude du déséquilibre de liaison permet de renseigner sur l'histoire de la population, et donc de son évolution. Il existe de nombreuses statistiques qui permettent de mesurer cette valeur de LD, chacune possédant ses avantages, en fonction du contexte d'étude. Ces stratégies sont plus largement détaillées dans différentes revues de littérature Devlin and Risch, 1995; Hudson, 2004; Jordá and Puig,

2020). La plupart des mesures de LD reposent sur la quantification du degré d'association entre une paire d'allèles. Les deux méthodes les plus couramment utilisées sont la statistique D' Lewontin, 1964 et le coefficient de corrélation r^2.

La statistique r^2. Il s'agit du coefficient de corrélation entre les paires de loci (r^2), introduit par Hill & Robertson. (Hill and Robertson, 1968). Ce coefficient est compris entre 0 et 1, 0 qui équivaut à un équilibre de liaison (les allèles sont indépendants) et 1 qui équivaut à un déséquilibre de liaison (les allèles sont dépendants).

$$r^2 = \left(\frac{D}{\sqrt{f_A f_B f_a f_b}} \right)^2 \tag{1.3}$$

où D est le déséquilibre de liaison défini par l'équation suivante : $D = f_{AB} - f_A * f_B$
avec f_{AB} qui est la fréquence de l'haplotype porteur des allèles A et B
f_A/f_B est la fréquence des allèles de référence
f_a/f_b est la fréquence des allèles alternatifs
et r représente le taux populationnel p

Cette métrique est couramment utilisée dans les étapes préliminaires aux tests d'association des GWAS afin de ne prendre en compte que les variants génétiques qui ne sont pas impliqués dans un déséquilibre de liaison, et qui sont donc indépendants les uns des autres.

Étapes de contrôle de qualité des SNPs

Il est important de s'assurer de la qualité des données génotypiques utilisées dans les GWAS. Grâce à l'obtention du génotype d'une population donnée, il sera possible de calculer les fréquences de variants spécifiques à cette population, et de calculer différentes métriques pour caractériser la structure de la population étudiée. Certaines de ces métriques sont décrites ci-bas.

L'hétérozygocité. Dans le cas d'une population avec des individus diploïdes, cela permet d'estimer la proportion des individus hétérozygotes au sein de la population (Butler, 2015). Cela permet de décrire la variance au sein de la population comment celle-ci se distribue au niveau des allèles dans les loci contenant des variants génétiques.

L'indice de fixation F_{ST}. La statistique F de Wright (Wright, 1951) permet de mesurer l'écart de proportion d'hétérozygote Aa (H) par rapport à la fréquence de l'allèle dominant (p) et celle de l'allèle récessif (q), correspondant à l'équation (1.4). Grâce à F et p, on pourra

donc spécifier la structure génétique d'une population, aussi appelée la structure de Wright.

$$\begin{aligned} F &= 1 - H/2pq \\ \iff H &= (1-F)2pq \end{aligned} \tag{1.4}$$

Avec la statistique F de Wright, on peut calculer l'indice de fixation (F_{ST}) (Hudson et al., 1992) . F_{ST} (équation (1.5)) compare les niveaux les moins inclusifs (H_S pour des sous-populations différentes) aux plus inclusifs (H_T pour la même sous-population), en mesurant tous les effets de la sous-structure de la population combinée :

$$F_{ST} = 1 - \frac{H_S}{H_T} = \frac{H_T - H_S}{H_T} \tag{1.5}$$

En génétique des populations moderne, les données de génotypage pouvant provenir de différentes technologies (comme les puces de génotypage ou via les technologies de séquençage à haut débit), il est important de considérer les différents problèmes à résoudre pour caractériser la structure populationnelle comme :

- l'estimation de F_{ST} pour un SNP unique
- la combinaison de F_{ST} de plusieurs SNPs
- la choix des SNPs utilisées pour l'estimation du F_{ST}

La Fréquence d'Allèle Mineure (MAF, pour *Minor Allele Frequency* en anglais). Pour chaque locus, il correspond à la fréquence d'allèle la plus basse (Chanock and Ostrander, 2014). Cette valeur peut varier pour chaque variant génétique en fonction des populations. Cette mesure est souvent utilisée pour minimiser les erreurs dues au séquençage de l'ADN pour le génotypage d'une population de plusieurs centaines voire de milliers d'individus. Il est important de choisir consciencieusement les valeurs seuil de FAM pour éliminer les variants génétiques provenant d'erreurs de séquençage. En effet, certaines valeurs basses de FAM peuvent être simplement associées à des variants génétiques présents en très faible fréquence dans la population étudiée. Par exemple, pour la détermination des variants génétiques caractérisant la maladie d'Alzheimer, Chouraki et Seshadri (Chouraki and Seshadri, 2014) proposent différents seuils de MAF pour classer les variants en 3 catégories :

- les variants fréquents (MAF $> 5\%$)
- les variants peu fréquents ($1\% <=$ MAF $<= 5\%$)
- les variants rares (MAF $< 1\%$)

En fonction de la problématique de recherche et du pouvoir de découverte de l'étude, les variants peu fréquents et variants rares peuvent être inclus ou non dans l'analyse. Ces seuils sont régulièrement utilisé pour trier les SNPs dans les études GWAS (Manolio et al., 2009).

Méthodes de vérification de la structure d'une population

Il est possible d'étudier la structure d'une population graphiquement en utilisant différentes méthodes de réduction de données du génome complet des individus telles que l'**analyse en composantes principales** (ACP). L'ACP est une technique permettant de faire une analyse multivariée (Patterson et al., 2006) qui permet de décomposer la variation génétique totale en axes K appelés composantes principales (CP). Ces CPs permettent de mettre en avant les processus évolutifs comme la divergence génétique permettant de différencier les populations (McVean, 2009). Par exemple, l'ACP réalisée sur le projet 1000 Génomes par Gaspar & Breen permet de différencier 20 groupes populationnels en utilisant les 2 premières CPs (Figure 1.7). Cette méthode de visualisation permet de préserver la structure de la population à large échelle.

La structure de la population des 1000 Génomes a été récemment visualisée à l'aide de la méthode ***Uniform Manifold Approximation and Projection*** (UMAP) qui permet d'observer des différences plus locales au sein de la population et distinguer des divergences génétiques entre plusieurs sous-populations (Diaz-Papkovich et al., 2021). En général, cette méthode est utilisée de manière complémentaire à la méthode de l'ACP, on pourra par exemple lui fournir comme données d'entrée les CPs avec les variances expliquées les plus importantes, ce qui permettra de faire une projection 2D de plusieurs CPs d'importance à l'aide de cette réduction de données supplémentaire (Figure 1.8).

Dans certains cas, ces visualisations de structure de données génétiques peuvent être faussées, par exemple si les données génotypiques présentent un grand nombre de données manquantes. Ce cas de figure peut arriver si les étapes d'imputation et d'élagage des SNPs, ne sont pas correctement réalisées. Ces étapes permettent d'éviter d'avoir des biais dans la structure populationnelle qui serait due à des problèmes techniques liés au génotypage (erreurs de séquençage, couverture de séquençage pauvre).

Statistiques, visualisation des résultats et méthodes de validation

Les tests d'association et la nécessité d'appliquer des corrections multi-tests. Pour chaque variant génétique identifié lors de l'étape de génotypage, les tests d'association statistiques permettent d'établir un lien entre le trait phénotypique et les génotypes existants au sein des deux populations pour identifier les variants et leurs régions génomiques (ou loci) qui pourraient expliquer ces différences phénotypiques. La relation entre la variance génotypique et la variance phénotypique peut être modélisée en fonction de plusieurs modèles (ex :modèles additifs et non additifs). Dans ces modèles, il sera important de corriger les variables confondantes (comme les variables influençant la structure de la population), qui sont des variables aléatoires pouvant influencer à la fois la variable dépendante et les variables explicatives. De plus, un grand nombre de tests statistiques sera réalisé, comme

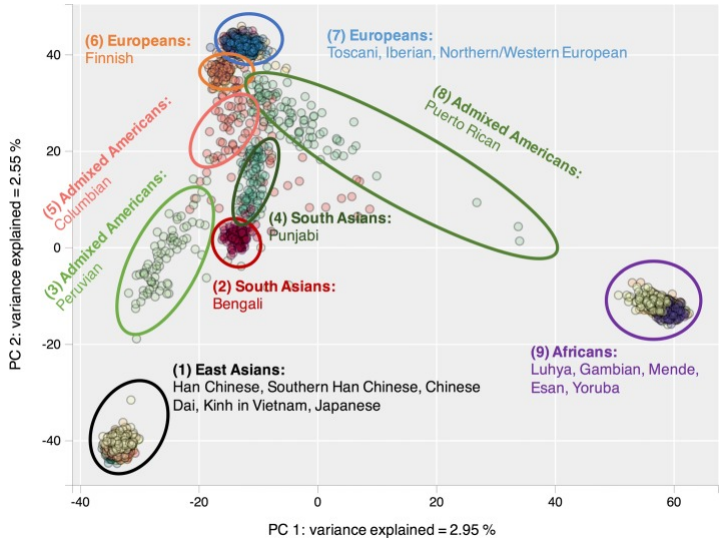

Fig. 1.7. ACP sur les données génotypiques du projet 1000 Génomes.

(Tirée de Gaspar and Breen, 2019)

Fig. 1.8. UMAP sur les données génotypiques du projet 1000 Génomes.

(Tirée de Diaz-Papkovich et al., 2021)

il faut comparer les variances génotypiques de tous les SNPs (d'une dizaine de milliers chez la levure, à plusieurs millions chez l'humain) avec les variances phénotypiques au sein des deux populations étudiées, ce qui engendra un problème de comparaisons multiples. Cette problématique arrive quand un grand nombre de tests stastistiques est réalisé, pouvant créer des inférences statistiques (p-valeurs) faussées. Pour corriger ces p-valeurs, il existe différentes méthodes statistiques pour la correction de tests multiples (ex : la méthode de FDR, la correction de Benjamini–Hochberg, etc.).

Parmi les tests d'association possibles, le plus simple consiste à calculer un rapport de cote (OR, pour *odds ratio* en anglais), qui permet d'établir quelles sont les chances d'apparition de maladie pour les individus ayant un allèle spécifique et les chances d'apparition de maladie pour les individus qui n'ont pas ce même allèle. Lorsque la fréquence de l'allèle dans la population cible est beaucoup plus élevée que dans la population témoin, le rapport de cotes est supérieur à 1, et vice versa. De plus, une valeur de P pour la significativité du OR est typiquement calculée en utilisant un Test du χ^2. Les OR significativement différents de 1 reportent une association pangénomique entre un SNP et une maladie. Il est important de rappeler que les GWAS ne donnent pas une relation de causalité du SNP à son caractère mais renseignent plutôt sur les variants corrélés avec ce trait.

Outre les p-valeurs associées à chaque test de corrélation entre les variances génotypiques et les variances phénotypiques, on pourra aussi s'intéresser à la métrique de taille d'effet (*effect size*, en anglais). Cette mesure permet de quantifier les différences entre la moyenne des deux groupes étudiées, et d'établir une relation entre les variables. Dans un modèle linéaire simple (où ici ne seront pas pris en compte les co-variables) représentant la relation génotype et phénotype, l'effet de taille associé représente la pente de la courbe (formule modifiée basée sur l'équation 1) de Uffelmann et al., 2021), qu'on appelera β :

$$\boldsymbol{P} \sim \boldsymbol{X}_s \boldsymbol{\beta}_s \tag{1.6}$$

où P représente un vecteur contenant les données phénotypiques,
X_s est le vecteur contenant les données génotypiques pour tous les SNPs individuels
β_s est la taille d'effet fixe associé à chaque SNP s
Les valeurs d'effets de taille peuvent être négatives ou positives. La taille d'effet décrit l'ampleur de l'effet du variant génétique sur la variabilité du phénotype.

Les méthodes de visualisation des résultats de GWAS. Généralement les résultats des GWAS sont visualisés à l'aide de graphiques appelés *Manhattan* (*Manhattan plot*),

permettant de visualiser l'ensemble des variants génétiques et leurs positions chromosomiques sur l'axe des x, et le logarithme négatif de la valeur P d'association sur l'axe des y. Chez l'humain, le seuil de significativé des variants génétiques associés à un phénotype complexe, équivaut à la p-valeur 0.05 divisée par le nombre total de variations génétiques indépendantes, soit environ 1,000,000 de variants, ce qui donne une valeur de $5 * 10^{-8}$ (Dudbridge and Gusnanto, 2008). Bien que ce seuil de significativité soit consensus au sein des experts en génétique, il reste toutefois des discussions autour de la sévérité de ce seuil.

Ces Manhattan plots sont également accompagnés de graphiques quantile-quantile (QQ plots), des graphiques montrant la distribution réelle des valeurs P en fonction de leur distribution théorique. Ces graphiques permettent, à l'aide du calcul de leur pente (appelé coefficient d'inflation génétique) de déterminer la qualité des résultats de GWAS. Des profils anormaux de QQ plots révèlent des défauts de puissance statistique dans l'analyse (dû à un faible nombre d'individus dans la cohorte), ou mettent en évidence des événements évolutifs particuliers qui affectent la structure populationnelle (Voorman et al., 2011).

Les stratégies de validation des GWAS. Pour que les GWAS aient une puissance statistique suffisante, il est important que les cohortes soient de tailles suffisamment importantes, idéalement quelques milliers d'individus, afin d'avoir un échantillon populationnel représentatif. Il est aussi important d'identifier un grand nombre de variants génétiques. On peut également regrouper les résultats de plusieurs analyses GWAS portant sur le même phénotype étudié, afin d'augmenter *a posteriori* la puissance statistique de ces études, ce sont des approches de méta-analyses (qui seront définies plus en détails ici, McCarthy et al., 2008). Pour réaliser ces méta-analyses, il est nécessaire de vérifier que les approches statistiques utilisées dans chaque étude puissent être normalisées de la même manière afin d'être intégrées ensemble.
Pour valider un GWAS, on peut également reproduire cette étude en réalisant un sous-échantillonnage des populations initialement étudiées et vérifier qu'on retrouve les mêmes résultats. On peut aussi faire la réplication de l'étude en reproduisant le GWAS sur une autre cohorte indépendante. Kraft et collègues proposent une revue de littérature qui récapitule les étapes clefs pour réaliser des stratégies robustes de validation de GWAS (Kraft et al., 2009; McCarthy et al., 2008).

Analyses post-GWAS

Une fois les SNPs candidats identifiés à l'aide des GWAS, des stratégies informatiques peuvent être implémentées pour obtenir plus d'informations sur les variants génétiques identifiés et leurs influences directes ou indirectes sur le phénotype héritable étudié. Comme plusieurs centaines de SNPs candidats sont identifiés, il sera important de proposer

des stratégies de priorisation de SNPs, afin de mettre en évidence un ensemble de loci indépendants permettant d'expliquer l'héritabilité de ce phénotype. L'identification des rôles biologiques de ces loci et leur implication dans des contextes plus systémiques, comme leur implication dans des voies métaboliques ou des processus biologiques, est essentielle pour mieux comprendre leur contribution à la variance phénotypique. Pour cela, l'étude intégrative des données -omiques (ex : transcriptomique, protéomique, métabolomique, interactomique etc.) avec ces SNPs et leurs loci associés, permet de proposer différentes approches de priorisation (Paik et al., 2012).

Les études de cartographie fine des loci associés à des traits quantitatifs. (ou cartographie QTL, pour *Quantitative Trait Locus Mapping* en anglais). Ces méthodes ont pour but d'identifier les loci influençant des traits quantitatifs, elles permettent d'inférer la relation entre les variances génétiques et les changements associés à des caractéristiques phénotypiques en s'appuyant sur des méthodes statistiques. Plus particulièrement, on s'intéressera à des métriques quantitatives directement liées à des entités moléculaires permettant de proposer des abstractions plus ou moins complexes du phénotype, tels que l'expression des gènes (eQTL, pour *expression Quantitative Trait Locus*, en anglais), l'abondance protéique (pQTL, pour *protein Quantitative Trait Locus* en anglais), et bien d'autres, dont les avantages et les limites sont mis en avant dans différentes revues de littérature (Cookson et al., 2009; Gauthier et al., 2020; Molendijk and Parker, 2021).

Parmi les autres approches permettant de caractériser fonctionnellement les SNPs identifiés dans les études GWAS, on peut citer l'annotation des SNPs avec :

- l'identification des gènes sur lesquels se trouvent ces SNPs ou ceux qui se trouvent à proximité des SNPs,
- l'identification de la fonction de ces gènes et des conséquences fonctionnelles des mutations associées à leurs variations génétiques,
- l'identification de voies métaboliques ou processus biologiques dans lesquels ces gènes sont impliqués et leur importance fonctionnelle, par des analyses d'enrichissement de gènes.

Si ces SNPs sont situés sur des gènes ou à proximité de gènes, on pourra tenter de proposer différentes hypothèses fonctionnelles à l'aide de différentes approches :

- les analyses de corrélation génique, qui permettent d'identifier les gènes potentiellement co-exprimés et l'inférence de réseaux de co-expression de gènes,
- l'identification de processus d'épistasie entre plusieurs gènes, qui caractérise l'interaction existant entre deux ou plusieurs gènes, l'un d'entre eux (ou plusieurs) masquant ou empêchant l'expression des autres,

- la détermination du risque d'apparition des phénotypes chez les individus en considérant leurs variants génétique

Pour valider les hypothèses fonctionnelles proposées par ces différentes approches, des validations expérimentales seront nécessaires. Ces validations expérimentales peuvent être réalisées sur des lignées cellulaires humaines pour lesquelles des modèles de phénotypes / maladies ont été préalablement établis. On peut également valider ces hypothèses en les étudiant dans des contextes évolutifs moins complexes que l'humain, en recherchant les gènes homologues des gènes précédemment identifiées chez différents organismes modèles pour lesquels on a identifié des contextes phénotypiques proches du phénotype ou de la maladie étudiée. Parmi les organismes modèles les plus populaires, on peut citer la levure, le nématode, la drosophile et la souris.

1.2.3. Le problème de l'héritabilité manquante dans l'étude des phénotypes complexes avec les GWAS

Depuis plus d'une dizaine d'années, les GWAS ont permis d'identifier des centaines de variants génétiques contribuant à l'apparition des phénotypes complexes chez l'humain et ont permis une meilleure compréhension de leur influence sur l'architecture génétique (Manolio et al., 2009). Cependant, il a été mis en évidence que la plupart des variants génétiques identifiés, considérés de manière individuelle ou de manière combinée, ne contribuent que très peu aux facteurs de risque d'apparition de ces phénotypes, et expliquent seulement une petite proportion de l'héritabilité de ces phénotypes complexes (Hindorff et al., 2009), c'est la problématique de l'héritabilité manquante.

D'abord décrite par Manolio et collègues en 2009, elle met en évidence qu'il existe un écart entre l'estimation de l'héritabilité faite à partir des données issues de génotypage et celle estimée dans des études réalisées sur des cohortes de jumeaux. Les premières mesures d'héritabilité ont été réalisées à partir d'études s'intéressant à la transmission d'un trait au sein de familles, et plus particulièrement dans des cohortes de jumeaux (Boomsma et al., 2002). Celles-ci ont permis de mettre en évidence les différences d'héritabilité en comparant la similarité phénotypique entre des jumeaux monozygotiques (dont les génotypes sont strictement identiques) et des jumeaux dizygotiques (dont les génotypes sont identiques à seulement 50%). Avec ces deux types cohortes, il a été démontré que les estimations d'héritabilité entre des jumeaux monozygotiques ont une plus grande similarité, que chez des jumeaux dizyogotiques, ce qui met en évidence la contribution des variations génétiques sur la variabilité du phénotype. Grâce à ces études, on a pu définir l'héritabilité "au sens étroit" (*narrow-sense heritability*, en anglais) (Falconer, 1996) dont on suppose qu'elle peut

varier en fonction des environnements. Chez l'humain, cette mesure d'héritabilité peut être influencée à la suite des effets génétiques non additifs (tels que la dominance, l'épistasie ou les interactions gène-gène), en fonction des contextes familiaux (fratrie, consanguinité), et par les interactions ou corrélations entre génotypes et environnements. Au niveau de l'échelle populationnelle, cela met en évidence qu'il reste une part conséquente de l'héritabilité qui reste encore inexpliquée.

Par exemple, l'étude de Yang et collègues démontre en 2010 que la taille, un trait complexe héritable, dont à l'époque, on identifie une quarantaine de variants génétiques fréquents (MAF $> 5\%$) qui influencent le trait (Yang et al., 2010). Cette proportion de variants estimait l'héritabilité à seulement 5% pour expliquer la variance phénotypique à partir de données génotypiques provenant d'une population de plusieurs dizaines d'individus, alors qu'on l'estime à plus de 80% dans les études de cohorte de jumeaux. Une dizaine d'année plus tard, différentes études de consortium (Marouli et al., 2017; Wainschtein et al., 2022) ont démontré que l'héritabilité manquante attribuée à ce trait pouvait être expliquée par des variants génétiques peu fréquents, voir rares dont les effets de taille sont plus importants, en comparaison avec ceux estimés pour les variants fréquents (Figure 1.9). Ces variants génétiques n'avaient auparavant pas été identifiés, car les p-valeurs auxquelles ils étaient associés n'excédaient pas les seuils de significativité. Il faut donc mettre en avant des stratégies pour pouvoir étudier ces variants génétiques peu fréquents et rares, ainsi que les caractériser fonctionnellement. Une étude récente par Yengo et collègues met en avant que l'héritabilité manquante pour la taille chez l'humain peut être résolue en réalisant des GWAS sur des cohortes populationnelles de grande taille (Yengo et al., 2022). Ils identifient plus de 10,000 SNPs communs permettant d'expliquer la variance phénotypique (soit $\sim$ 40%) dans une population avec une ancestralité européenne. Toutefois, cette étude ne quantifie l'apport de l'héritabilité des variants plus rares ni la contribution des variants génétiques situées dans des régions non codantes du génome.

1.2.4. Les défis vers l'identification et la caractérisation fonctionnelle des SNPs influençant les phénotypes complexes

Dans la population humaine qui possède plusieurs millions de SNPs à travers son génome et le déséquilibre de liaison entre ces SNPs, il est plus que probable qu'une fraction non négligeable des résultats des études GWAS soient en LD avec des variants génétiques responsables des traits, qu'on appelle les variants génétiques causaux (Tam et al., 2019). Par conséquent, les loci identifiés ne permettent pas forcément d'informer sur les mécanismes fondamentaux pour mieux comprendre la biologie derrière ces traits.

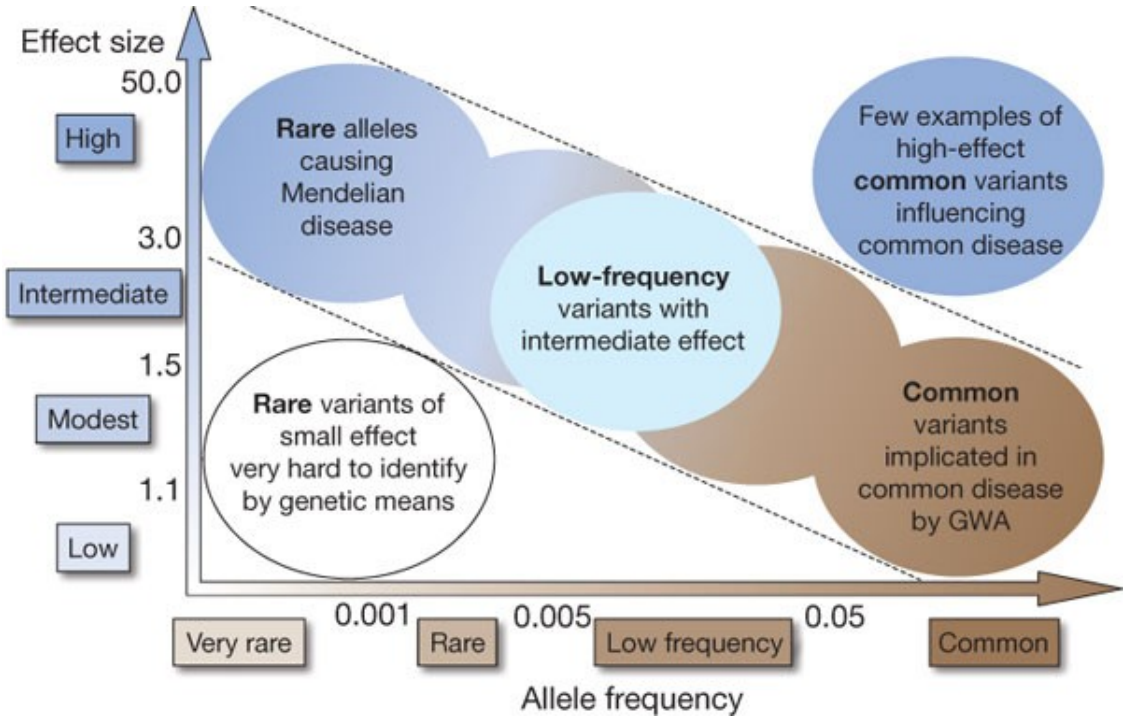

Fig. 1.9. Relation entre les effets de tailles et les catégories de variants.
(Tirée de Manolio et al., 2009)

Des étapes de validation empiriques seront donc nécessaires pour réduire le nombre de SNPs candidats, et de proposer ceux qui sont véritablement impliqués fonctionnellement au niveau du phénotype (Gauthier et al., 2020).

Élimination du LD dans une population d'individus co-sanguins

Différentes études ont démontré qu'il était possible de se débarasser du LD dans des populations dites "co-sanguines" (*inbreds*, en anglais). Par exemple, l'étude de She & Jarosz propose une stratégie de cartographie fine des variants génétiques responsable des changements phénotypiques suite à à une stimulation médicamenteuse (She and Jarosz, 2018), chez *S. cerevisiae*. Pour s'abstraire du LD, les auteurs ont réalisé un croisement successif sur 6 générations, entre une souche sauvage (individus de laboratoire issus d'un prélèvement dans la nature) et une souche clinique (individus existants uniquement en laboratoire) de *S. cerevisiae*, afin de favoriser les événements de recombinaison méiotique, permettant d'augmenter la diversité génétique au sein de la population. Le croisement de ces souches a permis de générer une descendance d'environ 1000 souches haploïdes. Cette population possède un degré de polymorphisme d'environ 1 polymorphisme toutes les 1,000 paires de base. Cette distance entre les variants génétiques assure que chacun des variants puisse être considéré comme des variants génétiques indépendants de tout LD. Des études similaires ont été conduites chez

d'autres organismes modèles, comme la souris (Collaborative Cross Consortium, 2012) ou la drosophile (Mackay et al., 2012).

Cartographie QTL, concepts théoriques et enjeux statistiques

Précédemment défini dans la section 1.2.2, l'approche de cartographie QTL est indispensable pour comprendre comment les variants génétiques influencent les traits quantitatifs (Members of the Complex Trait Consortium, 2003). L'apparition de ces traits quantitatifs est dû principalement grâce à différentes entités moléculaires (expression de gènes, abondance protéique etc.) qui peuvent varier en fonction des variations génétiques, que nous allons mettre en avant dans les prochains paragraphes.

Les approches étudiant les eQTLs. Grâce aux GWAS, nous avons pu détecter de nombreux variants génétiques localisés à proximité ou sur des gènes (Võsa et al., 2021). Tout naturellement, on a pu donc se demander si ces variants génétiques ne pourraient pas influencer l'expression de ces gènes. Un eQTL est donc une région génomique influençant l'expression des gènes. Pour identifier ces loci, il faut réaliser des tests d'association statistiques entre des données génotypiques obtenus par génotypage (voir section 1.2.2), et des données d'expression génique, provenant de la quantification du transcriptome via puces d'expression ou RNA-seq, préalablement normalisées. L'étape de normalisation permet de réaliser une série de calculs qui vont permettre de réduire la variabilité des données dont l'origine ne serait pas biologique. Cette variabilité technique proviendrait des étapes en amont des mesures de quantification, que ce soit au niveau de l'étape de préparation des échantillons ou bien de l'étape de séquençage. L'expression d'un gène est une mesure quantifiable du nombre de molécules d'ARN sous des conditions spécifiques. Ici, un eQTL est locus polymorphique où le changement d'un allèle pour un autre produit en moyenne un changement dans l'expression du gène influencé par ce locus.

Chez l'humain, on pourra par exemple conduire des analyses eQTL sur l'expression génique de 10 à 30,000 gènes pour plusieurs milliers à millions de variants génétiques, mesurées sur plusieurs centaines à milliers d'individus issus d'une population, nécessitant plusieurs dizaines de millions de tests statistiques. En conséquence, comme pour les études GWAS, les p-valeurs de ces tests statistiques devront être corrigées par le biais des approches de correction multi-tests (voir section 1.2.2), pour s'assurer de leur indépendance statistique.

On peut utiliser différents types de modèles pour tenter de définir la relation entre la variation génétique et les changements d'expression génique :

- le modèle linéaire simple

- le modèle linéaire généralisé

L'étude des eQTL permet entre autre de caractériser des éléments permettant de réguler l'expression génique et permet d'inférer des hypothèses sur les relations de régulation entre les régions contenant ces variants génétiques et l'expression du gène influencé (Gilad et al., 2008; Nica and Dermitzakis, 2013). Ils permettent entre autre d'identifier deux types d'eQTLs :

- les cis-eQTLs (ou eQTLs locaux) : ce sont des eQTLs situés proches du gène dont l'expression est modulée par la variation génétique, il peut se trouver au niveau de sa séquence codante, ou bien de ses régions régulatrices d'environ 10kb en amont ou en aval de sa région codante (les régions 5' et 3'),
- les trans-eQTLs (ou eQTLs distaux) : ces sont des eQTLs situées loin du gène dont l'expression est modulée, ils se trouvent souvent sur des chromosomes différents.

Les approches analogues aux eQTLs. Il existe d'autres approches QTL permettant d'étudier d'autres traits moléculaires quantifiables, tels que :

- l'abondance protéique (pQTL, pour *protein QTL*) dont les mesures de quantification proviennent des données de spectrométrie de masse (voir section 1.1.3)
- le taux de méthylation de l'ADN (meQTL, pour *methylation QTL*)

Ces approches ont des considérations statistiques similaires à celles mentionnées pour l'approche eQTL. Leurs différences résident principalement sur le type de trait moléculaire étudié et sur les méthodes de normalisation qui sont spécifiques à leur mesure de quantification. Chacune de ces méthodes permet d'inférer des hypothèses fonctionnelles sur les variants génétiques identifiés préalablement par les études GWAS, mais celles-ci ne sont pas suffisantes pour caractériser l'ensemble des candidats inférés influençant des phénotypes complexes (Yao et al., 2020). En effet, la plus grande majorité de ces SNPs ne se situent pas nécessairement sur des régions génomiques associées à l'expression et la régulation des gènes, mais peuvent se trouver dans des régions non codantes (Visscher et al., 2017).

L'identification de nouveaux traits quantificatifs moléculaires est donc essentielle pour expliquer l'héritabilité manquante de ces variants génétiques présents dans des régions génomiques dont le rôle n'a pas été clairement établi. Dans cette thèse, nous allons nous intéresser plus spécifiquement aux interactions protéine-protéine, et mettre en avant comment l'étude de protéines dans ce contexte fonctionnel particulier (ici, une interaction physique ou fonctionnelle entre deux protéines), peut contribuer à l'établissement d'un modèle permettant une abstraction plus précise des phénotypes complexes.

1.2.5. L'importance des réseaux d'interactions protéine-protéine dans la résolution du problème d'héritabilité manquante pour les phénotypes complexes

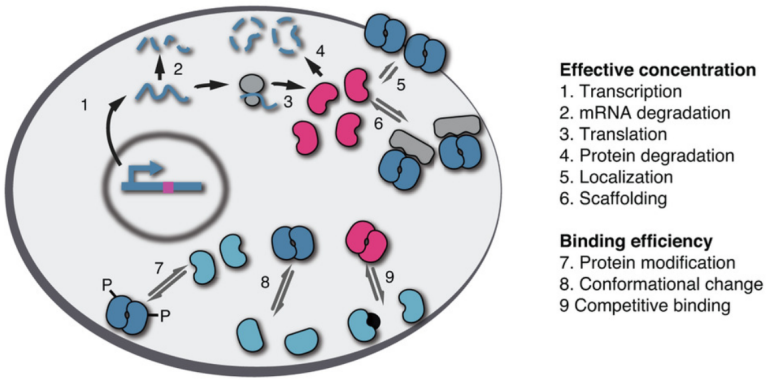

Fig. 1.10. Les informations rapportées par le protéotype.
(Tirée de Gauthier et al., 2020)

L'étude des différents états des protéines, de leur localisation et de leurs interactions permet de mieux appréhender le concept de protéotype (Gauthier et al., 2020), qui permet de décrire un état cellulaire donné avec une emphase particulière sur les protéines. Comme le souligne le schéma de Gauthier et collègues (Figure 1.10), l'étude des protéines, que ce soit leur concentration ou leur efficacité de liaison, permet d'obtenir des informations sur différents processus biologiques de la cellule, directement impliqués dans la protéostase. Elle permet également d'identifier les conditions menant à l'interaction des protéines, et dans quel contexte ces dernières sont identifiées. Le protéotype permet également d'étudier indirectement le renouvellement des ARNm (Gauthier et al., 2020).

Dans un contexte d'interaction, l'étude des protéines permet de comprendre les mécanismes dans lesquels elles sont impliquées au sein des processus biologiques. On définit une interaction entre protéines comme un phénomène au cours duquel un contact physique, régi par des forces intermoléculaires, est établi entre deux protéines. L'interaction entre protéines peut être considérée comme une forme d'épistasie fonctionnelle, où les protéines, qui sont les produits de deux gènes distincts, ont besoin d'interagir physiquement ensemble pour réaliser leur fonction biologique, et donc participer à la caractérisation d'un phénotype. Dans la prochaine section, nous allons présenter différentes méthodes d'identification et de quantification de ces PPIs dans le but de caractériser des phénotypes.

Identification et la quantification des PPIs

L'identification des protéines sans contexte d'interactions a déjà été détaillée dans la section 1.1.3. Sous certaines conditions particulières, la spectrométrie de masse permet d'identifier des PPIs (Richards et al., 2021). Différentes méthodes *in vitro* ont permis l'identification et la quantification des PPIs, comme la purification par affinité suivie d'une spectrométrie de masse (C.-M. Lee et al., 2017) ou la spectrométrie de masse après étiquettage de proximité (Bosch et al., 2021). Ces méthodes proposent différentes stratégies pour isoler les complexes protéiques avant de les analyser par spectrométrie de masse. Ces méthodes ont comme avantage de pouvoir identifier un grand nombre de complexes d'interaction protéine-protéine de manière non biaisée. En ce qui concerne la quantification des PPIs, où plus précisément la quantification du nombre de complexes protéiques formées, il est possible d'utiliser les différentes méthodes de quantification des protéines isolées, précédemment décrites dans la section 1.1.3. Cependant pour obtenir des quantifications exactes des PPIs avec un spectromètre de masse, il faudra veiller à l'intégrité des complexes protéiques avant d'effectuer les mesures de quantification après dissociation des complexes. Ces méthodes d'identification et de quantification permettent d'étudier les PPIs dans des contextes *in vitro* qui ne reflètent pas forcément la réalité biologique complète. C'est pourquoi des méthodes d'identification des PPI et de quantification dans des contextes *in vivo* ont été développées.

La technique du double hybride. (Y2H, pour *Yeast Two-Hybrid* en anglais) (Fields and Song, 1989; Fromont-Racine et al., 1997). Historiquement implémentée chez la levure, cette méthode permet de détecter une interaction physique entre deux protéines, rapportée à l'aide d'un gène rapporteur (comme LacZ) où une séquence d'activation (UAS, pour Upstream Activator Sequence en anglais) a été greffée en amont pour moduler son expression. La protéine issue de l'expression du facteur de transcription Gal4 (impliqué dans la production de la galactosidase) peut se fixer spécifiquement à la région UAS, ce qui permettra l'expression du gène LacZ. Parallèment, on va créer deux protéines hybrides :

- une protéine appât (*bait*, en anglais) qui sera fusionnée au domaine protéique BD, issu de l'expression du facteur de transcription Gal4
- une protéine proie (*prey*, en anglais) qui sera fusionnée au domaine protéique AD, issu également de l'expression du facteur de transcription Gal4

Pour que le facteur de transcription Gal4 soit fonctionnel, et qu'il active l'expression du gène LacZ, les deux domaines BD et AD doivent se retrouver tous deux à proximité de la séquence UAS. Ainsi, si les deux protéines interagissent physiquement, ces deux protéines hybrides vont se rapprocher et permettre la transcription du gène rapporteur. Dans cette approche, les protéines appât et proie sont à priori connues. Cette méthode a permis

l'identification de nombreuses interactions protéine-protéine, notamment chez la levure, la drosophile et l'humain (Brückner et al., 2009).

Ses avantages et limites sont discutés dans la revue de littérature de Brükner et collègues (Brückner et al., 2009). Parmi ses limites, on peut mentionner que cette technique ne permet pas de proposer une quantification directe des PPIs. En effet, le niveau d'expression du gène rapporteur permet de mesurer seulement qualitativement la formation du complexe d'interaction protéique étudiée (par la lecture de l'intensité des couleurs émises par le système rapporteur). Pour obtenir une mesure quantitative, il sera nécessaire de coupler cette stratégie à une étape de cytométrie de flux, qui permettra d'obtenir une mesure de quantification de la formation des PPIs (J. Chen et al., 2012). Cependant, tous les laboratoires de recherche n'ont pas forcément à disposition cette technologie.
De plus, il sera nécessaire de vérifier que les mesures d'expression du gène rapporteur dépendent uniquement de la formation spécifique du PPI étudiée, et non de l'activation du gène par d'autres mécanismes externes de régulation, afin de minimiser le nombre de faux positifs. Par ailleurs, cette technique implique, par nécessité, que l'expression des protéines d'intérêt se produise dans le noyau de la cellule. Or, le noyau n'étant pas le milieu natif pour certaines interactions, celles-ci ne seront pas détectées. Enfin, les modifications nécessaires aux constructions expérimentales pour mesurer la formation du complexe protéique peuvent éventuellement empêcher la formation de l'interaction (en cas de problème de conformation suite à l'ajout des domaines protéiques aux protéines appât et à la proie).

La technique par complémentation de fragments protéiques. (PCA, pour *Protein-fragment Complementation Assay* en anglais) proposé par Michnick (S. W. Michnick, 2003) est une autre technique de biologie moléculaire qui permet d'identifier une interaction entre deux protéines et de quantifier l'abondance relative du complexe formé par celles-ci.. Pour cela, les protéines sont chacune étiquetées avec des fragments complémentaires d'une protéine rapporteur. Lorsque les protéines d'intérêt interagissent, les fragments complémentaires de la protéine rapporteur sont rapprochés, permettant ainsi la reconstitution de sa forme complète fonctionnelle. Inversement, s'il n'y pas d'interaction entre les protéines d'intérêt, la forme complète n'est pas reconstituée. La protéine rapporteur fonctionnelle génère un signal mesurable, par lequel il est possible de quantifier une interaction, c'est-à-dire que le signal augmente en fonction de la concentration intracellulaire du complexe formée par l'interaction. Parmi les enzymes rapporteurs utilisées dans cette stratégie, on peut citer différentes enzymes comme la dihydrofolate reductase, différents types de luciférase, mais aussi la protéine fluorescente GFP (pour *Green Fluorescent Protein* en anglais). Les interactions mesurées se produisent dans leur compartiment cellulaire natif. Elle permet de mesurer des interactions faibles et transitoires (avantage important sur les méthodes basées

sur la spectrométrie de masse). Les mesures sont de nature quantitative, i.e. des quantités relatives du complexe formé par l'interaction binaire peuvent être inférées. Elle a notamment permis l'inférence d'un interactome regroupant plus de 2,500 PPIs de haute qualité chez la levure (Tarassov et al., 2008). Elle a permis également d'identifier une fraction importante de PPIs chez la drosophile et le nématode (B. Chen et al., 2007; Hudry et al., 2011).

Enfin, il existe d'autres méthodes permettant l'identification de PPIs, elles sont listées dans la revue de littérature de Buntru et collègues (Buntru et al., 2016).

Stratégies de phénotypage d'une population par la quantification de PPIs

Les premières stratégies de phénotypage qui s'appuient sur des quantifications de PPIs ont été réalisées dans des populations de levures (Yachie et al., 2016, Celaj et al., 2017, Schlecht et al., 2017). Elles se basent sur une mesure de *fitness* des individus qui ont été mis dans des conditions de croissance compétitive, pour des contextes environnementaux particuliers (ex : exposition à un médicament) qui caractérisent différents phénotypes en réponse à ces environnements. La mesure d'un *fitness* permet de quantifier l'habilité d'un individu d'un certain génotype à survivre et se reproduire, elle permet donc d'estimer le taux de succès reproductif de ce génotype.

Pour mesurer le *fitness* de chaque individu de la population, des amplicons / codes-barres uniques, des courtes séquences d'ADN, sont ajoutées au sein de région de réplication de chaque individu. Ainsi la quantification de ces codes-barres permettra d'établir un proxy du nombre d'événements de division cellulaire, reflétant le *fitness* de chaque individu. Ensuite, pour identifier et rattacher les différentes fréquences de codes-barres à chaque individu de la population étudiée, il sera nécessaire de réaliser une étape d'extraction d'ADN des cellules résultantes de l'étape de croissance compétitive, qui sera par la suite séquencée. Les algorithmes qui réalisent ces étapes reposent principalement sur la recherche de motifs qui caractérisent les codes-barres pour la partie extraction, et l'implémentation de différents types d'algorithmes de regroupement (ex: calcul de similarité de distance, k-means, etc.) pour proposer des groupes de codes-barres par individu. Cette étape de regroupement est nécessaire puisque lors de l'étape d'intégration du code-barre au sein de l'individu, les étapes d'amplification ont pu potentiellement introduire des erreurs de réplication altérant le motif initial du code-barre.

Après avoir attribué chaque groupe de codes-barres à chaque individu, on pourra donc estimer leur *fitness*. Deux méthodes sont possibles pour estimer une mesure relative du *fitness* des individus.

La méthode calculant des scores d'enrichissement. (*fold enrichment*, en anglais) permet de calculer le *fitness* relatif en mesurant la fréquence de codes-barres par individu lors de deux points de temps précis :

- un à temps zéro (T0), c'est le temps après insertion des codes-barres au sein des individus et avant introduction des conditions environnementales favorisant la croissance compétitive,
- un à temps final (TF). En fonction des organismes étudiés, ce temps représentera la durée minimale nécessaire à la production de plusieurs générations d'individus.

Ainsi, la définition du *fitness* relatif par individu se définira par la formule suivante :

$$\log \frac{Barcode\ Frequency\ _{TF}}{Barcode\ Frequency\ _{T0}} \tag{1.7}$$

Une implémentation de cette méthode est proposée par l'algorithme de Faure et collègues (Faure et al., 2020). Cependant, bien que précise, cette mesure de *fitness* dépend du temps de croissance propre à chaque individu, de leurs génotypes et de la distribution du *fitness* pour ces génotypes au sein de la population en fonction des différents environnements.

Pour s'abstraire de ces limites, Li et collègues (F. Li et al., 2018) proposent une méthode d'estimation reposant sur une **approche de plausibilité maximum** (*Maximum Likelihood*, en anglais) qui permet de calculer le *fitness* relatif par l'estimation des paramètres pour modéliser la distribution optimale de *fitness* en fonction de l'abondance des génotypes au sein de la population. Pour calculer ces paramètres, les fréquences de codes-barres à plusieurs points de temps devront être estimées afin de permettre une construction robuste du modèle de distribution. Différents modèles de *fitness* sont proposés :

- le *fitness* malthusien mesure le taux de croissance (x_m) pour lequel on assume que les cellules ont un *fitness* leur permettant de se multiplier de manière exponentielle, il se définit par l'équation suivante :

  $$n_t = n_0 * ex_m^t \tag{1.8}$$

  où n_t représente la taille de la population à un temps t
- le *fitness* wrightien permet de décrire le nombre moyen de descendants d'une cellule par génération (x_w^t), sous l'hypothèse que les cellules se soient multipliées de manière différentielle et se définit par l'équation suivante :

  $$n_t = n_0 * x_w^t \tag{1.9}$$

- ces deux mesures de *fitness* sont différentes, en effet leur relation est la suivante :

  $$x_m = \ln(x_w) \tag{1.10}$$

Grâce à ces différentes méthodes d'estimation de *fitness*, il sera possible de proposer une quantification relative des PPIs qui pourra être utilisée dans des stratégies d'association entre variance génotypique et phénotypique, ici approximées par les mesures de formation de complexe d'interaction protéique, qui dépendent de la restauration de l'activité de l'enzyme rapporteur utilisée dans la stratégie PCA. Le choix de calcul pour l'estimation du *fitness* dépendant principalement des stratégies expérimentales choisies (qui sont évidemment dépendantes des coûts et des ressources technologiques disponibles).

Dans le chapitre 3, nous décrirons la mise en place et l'implémentation d'une stratégie informatique de ces quantifications, reposant sur des mesures de *fitness* relative, dans des individus génétiquement diversifiés, préalablement étiquetés, qui poussent sous sélection dans différentes conditions environnementales.

1.2.6. La compréhension des phénotypes complexes à l'aide des réseaux de PPIs

Interprétation des phénotypes complexes à l'aide du modèle omnigénique

Dans la revue de littérature de Boyle et al. (Boyle et al., 2017), les auteurs résument comment les GWAS ont mis en avant les rôles des différents types de variants génétiques dans l'héritabilité des phénotypes complexes. Les premiers GWAS ont permis de mettre en évidence que la contribution des SNPs communs à un trait complexe représente une fraction modeste de la variance génétique (Manolio et al., 2009), leur taille d'effet étant relativement faible. Depuis, des études plus récentes ont mis en évidence que certains SNPs peu fréquents ou rares, ne dépassant pas le seuil de significativité pangénomique (p-valeur $> 5 * 10^{-8}$) ont pourtant des tailles d'effet importants qui doivent être pris en compte dans l'héritabilité de différents traits complexes (Yang et al., 2010). Par exemple, l'étude de la schizophénie a démontré que des SNPs rares contribuent également à la variation génétique liée au phénotype, avec des tailles d'effet importants (Manolio et al., 2009).

Grâce à ses observations, Boyle et al. propose ainsi un nouveau modèle pour comprendre les phénotypes complexes, qu'ils appellent modèle omnigénique (Figure 1.11). Ce modèle propose comme postulat qu'un trait complexe, phénotype ou maladie, est le reflet des réseaux biologiques permettant de réguler la cellule. Au sein de ce réseau hautement connecté, on distingue deux sous-ensembles de gènes :

- les gènes centraux et leurs régulateurs directs (*core genes*, en anglais) sont les gènes à l'origine des mécanismes moléculaires qui sont caractéristiques du phénotype. L'identification de ces gènes apporte une compréhension mécanistique et des processus biologiques à l'origine du phénotype. La prédiction de perte de fonction ou d'autres

mutations délétères dans ces gènes pourraient permettre d'anticiper l'apparition du dit phénotype dans ses formes les plus sévères. Bien qu'important fonctionnellement, ces gènes contribuent finalement peu à l'héritabilité du phénotype complexe.

- les gènes périphériques (*peripherical genes*, en anglais), sont des gènes indirectement liés aux gènes centraux, qui ne jouent pas de rôle direct dans l'apparition du phénotype. Du fait, que le réseau biologique soit hautement connecté, ces gènes peuvent toutefois influencer de manière indirecte la régulation des gènes centraux auxquels ces derniers seraient connectés. Bien plus nombreux que les gènes centraux, ce sont ces gènes qui contribueraient à la majeure partie de l'héritabilité du phénotype.

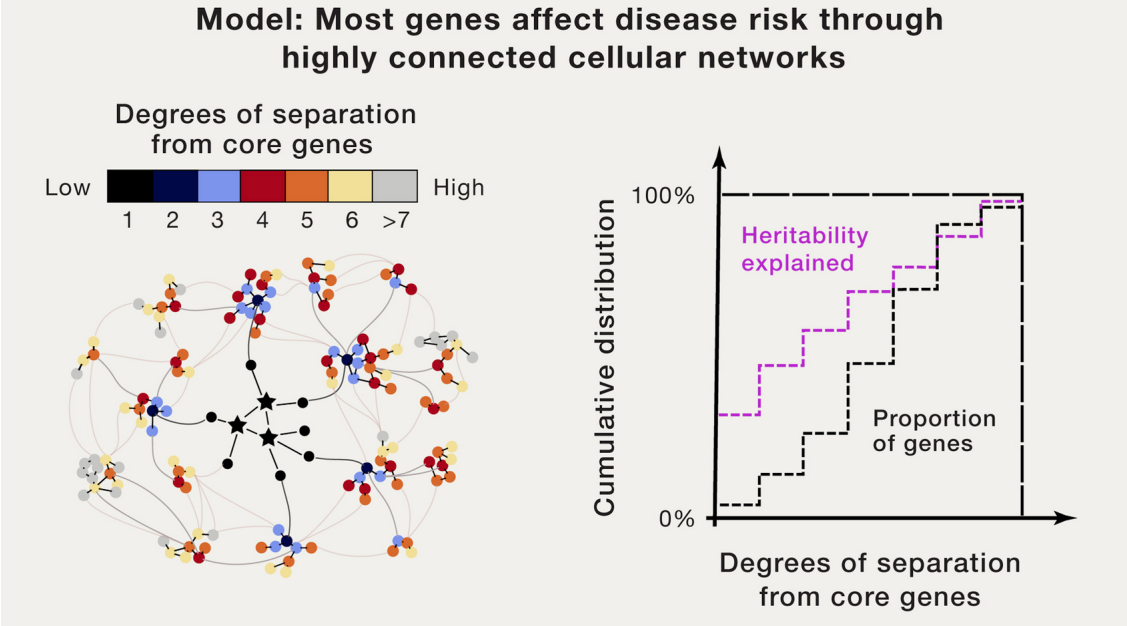

Fig. 1.11. Conceptualisation du modèle omnigénique.

(Tirée de Boyle et al., 2017)

Ce modèle met l'emphase sur l'importance des variants génétiques qui sont présents au sein des séquences codantes et cis-régulatrices des gènes. Ce modèle repose principalement sur la complexité des réseaux biologiques et de leur architecture, constituant une abstraction adéquate pour décrire les relations entre génotype et phénotype. De nombreux types de réseaux biologiques existent, ces derniers mettent en avant différents niveaux d'interactions moléculaires, incluant les réseaux transcriptionnels, les modifications post-traductionnelles; comme notamment les réseaux de phosphorylation et déphosphorylation; les réseaux d'interactions de PPIs, ou encore les réseaux de signalisation intercellulaire. La reconstitution complète de ces réseaux biologiques sont donc des étapes clefs pour l'exploitation du modèle omnigénique.

Cependant le modèle omnigénique présente également des limites, en effet il considère ici que tous les gènes pourraient directement (avec les gènes centraux) ou indirectement (avec les gènes périphériques) contribuer à la variance phénotypique. Il peut être plus probable que seulement certains gènes enrichis dans des voies de signalisation ou dans des processus biologiques spécifiques soient impliqués. Pour prendre en compte de cette limite, il sera donc important de s'intéresser à l'identification des sous-graphes dans les réseaux biologiques qui sont enrichis avec des gènes qui sont impliqués dans la variance phénotypique, on parlera de "hub", groupe de gènes hautement connectés au sein du dit réseau (Chakravarti and Turner, 2016; Furlong, 2013). De plus, ce modèle ne permet pas l'incorporation des variants génétiques situées sur des régions non codantes du génome, ces derniers pouvant être toute aussi importants dans l'héritabilité du phénotype (French and Edwards, 2020).

Dans cette thèse, nous émettons l'hypothèse que l'étude des interactions protéine-protéine permettrait d'établir une abstraction plus directe et précise des phénotypes, en comparaison à l'expression des gènes et l'étude isolée des protéines (Figure 1.12). Plus spécifiquement, on s'intéressera à l'étude des réseaux de PPIs à l'aide d'approches bio-informatiques pour comprendre comment ces types de réseaux biologiques peuvent caractériser des phénotypes complexes.

Dans la prochaine section, nous présenterons les bases de données régulièrement mises à jour permettant l'extraction des réseaux d'interactions protéine-protéine.

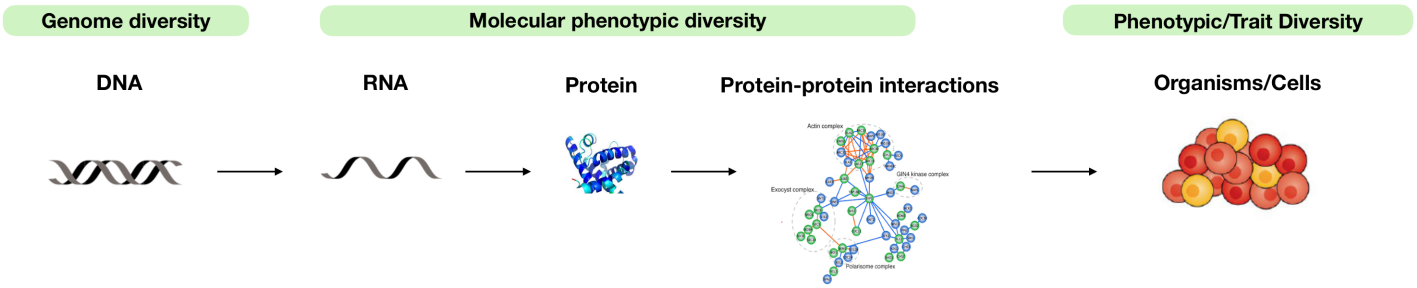

Fig. 1.12. Liens entre la relation génotype-phénotype et les acteurs du dogme central de la biologie moléculaire.

(Créé par Adrian Serohijos)

Bases de données pour l'inférence de réseaux de PPIs

Les réseaux d'interaction protéine-protéine (ou réseau de PPIs) sont des réseaux biologiques permettant de décrire les interactions physiques et fonctionnelles entre les protéines. Deux types d'interactions sont distingués :

- les interactions avec des évidences directes, pour lesquelles les protéines interagissent physiquement ensemble. La formation d'un complexe protéique est un exemple d'interaction directe entre différentes protéines.
- les interactions avec des évidences indirectes, pour lesquelles les protéines sont fonctionnellement associées ensemble. Les interactions avec des évidences indirectes sont inférées par des prédictions informatiques, commme par exemple le transfert de connaissance entre espèces, où l'on assume que certaines interactions protéine-protéine sont conservées à travers l'évolution.

STRING (https://string-db.org/). STRING est une base de données regroupant exclusivement des informations sur des données d'IPP. Elle permet d'obtenir des informations pour environ 14,000 organismes sur près de 67.6 millions de protéines et caractérise plus de 20 dizaines de millions d'interactions protéine-protéine (pour la version 5.11, consultée en Novembre 2022). Cette base de données est dite agglomérante car elle répertorie les PPIs identifiées ou prédites provenant de différentes sources d'informations, entre autres 1) des prédictions issues du contexte génomique ou inférence de PPI par orthologie, 2) des données de séquençage à haut débit, 3) de données de co-expression, 4) de la recherche automatique dans la littérature scientifique (ou *text-mining*, en anglais) et 5) des connaissances précédemment établies de bases de données plus générales manuellement validées telles que BIOGRID (Oughtred et al., 2019), KEGG (Kanehisa et al., 2007) ou Reactome (Jassal et al., 2019).

A partir de l'agglomération de ces informations, STRING propose des scores de confiance combinée pour qualifier la véracité ou l'existence des PPI. (von Mering et al. 2005) Le calcul de ce score, pour une interaction protéine-protéine donnée, consiste à additionner la probabilité d'existence de ce PPI associée à chaque source d'information disponible pour celle-ci. Chaque probabilité issue d'une source d'information est d'abord corrigée par un score de probabilité *a priori* que cette interaction puisse être observée par un phénomène aléatoire avant d'être ajoutée à la somme des scores combinés. Ces scores pouvant aller de 0 à 1. La base de données propose plusieurs seuils de confiance :

- Confiance basse ($STRING_{score} <= 0.2$)
- Confiance moyenne ($STRING_{score} >= 0.4$)
- Confiance haute ($STRING_{score} >= 0.7$)
- Confiance très haute ($STRING_{score} >= 0.9$)

Il est bien important de souligner que l'utilisation de ce réseau pour interpréter le contexte biologique, pour une interaction protéine-protéine donnée, reste avant tout indicatif. En effet, comme les sources d'origine de ces interactions sont multiples et sont

plus ou moins manuellement validées, ces informations n'ont pas forcément le même poids face à la réalité physique ou fonctionnelle de l'interaction (von Mering, 2004).

BIOGRID (https://thebiogrid.org/). BioGrid est un répertoire exhaustif d'interactions moléculaires, plus particulièrement les interactions génétiques et protéine-protéine, constituant le plus grand nombre d'interactions répertoriées (2.5 millions pour la version 2.4.213, consultée en Novembre 2022). Il répertorie également d'autres types d'interactions telles que les interactions chimiques et les modifications post-traductionnelles pour différents organismes modèles (pour plus de détails, consulter la page des Stastistiques de la base de données BioGrid). La validation de l'existence d'une interaction par BIOGRID se fait d'abord par une vérification du contenu de la publication dans laquelle cette dernière aura été identifiée. Les interactions identifiées directement par des expériences de séquençage à haut débit sont directement considérées comme véridiques. Dans certains cas, quand les méthodes d'identification des interactions ne sont pas communiquées, l'interaction passe une étape de validation interne par une équipe d'experts propre à la plateforme.

Ainsi, contrairement à STRING, BIOGRID répertorie exclusivement des interactions ayant été validées expérimentalement ou par le biais d'experts du domaine. Plus particulièrement pour les interactions protéine-protéine, la base de données répertorie les interactions validées par une dizaine de méthodes expérimentales différentes (Liste exhaustive des expériences biochimiques pour l'identification des interactions protéine-protéine). Les protocoles et particularités de chaque expérience apportent des informations pertinentes dans la compréhension du contexte de capture des interactions identifiées.

Stratégies informatiques pour étudier les réseaux de PPIs

L'étude des réseaux biologiques de leurs propriétés topologiques et structurelles, se fait à l'aide des principes fondés sur la théorie des graphes. Dans le cas des réseaux de PPIs, ces réseaux sont représentés par des modèles abstraits composés de noeuds (les protéines) et d'arêtes (la présence d'une interaction entre protéines). Ce sont des graphes non-orientés, c'est-à-dire que les arêtes entre les noeuds n'ont pas de sens de direction, car les méthodes expérimentales ayant permis l'identification de ces PPIs n'apportent pas forcément d'information sur les sens d'interactions entre les deux protéines interagissant ensemble.

L'étude de la topologie des réseaux biologiques permet de caractériser leurs propriétés, 1) par l'identification de leur nombre de noeuds et d'arêtes, 2) mais aussi par l'étude de leur répartition spatiale, ce qui permet d'étudier la hiérarchisation ou l'organisation du réseau. Des études ont mis en avant le lien entre la topologie des réseaux biologiques et une meilleure compréhension des fonctions biologiques associées à un groupe de molécules

interagissant ensemble, mais mettent aussi en avant leur évolution (Barabási and Oltvai, 2004). Par exemple, dans les réseaux d'interaction de PPIs, les protéines d'une même famille peuvent se retrouver regroupées dans un sous-ensemble de noeuds ayant beaucoup d'arêtes en commun. Le développement du domaine de l'étude de la biologie à l'aide d'approches réseaux (*network biology*, en anglais) s'est développé pour proposer différentes approches qui ont deux angles différents majeurs.

L'**identification et la description au sein des réseaux biologiques de sous-sytèmes fonctionnels**, où on veut identifier des modules, un sous-réseau de noeuds qui seraient plus proches les uns des autres, par rapport au reste du réseau. On peut également identifier des chemins entre molécules qui représenteraient la succession d'interactions nécessaire pour caractériser un processus biologique. Dans cette approche, on s'intéressera plutôt à la topologie locale du réseaux plutôt qu'à sa topologie globale. Elle permet entre autre de tester des hypothèses de robustesse topologique. La robustesse topologique est une question permettant de définir le sous-ensemble minimal de noeuds essentiel à la topologie du réseau. Si ces derniers sont retirés du réseau, on suppose que le réseau se décomposerait en plusieurs composants (un composant est un ensemble de noeuds et d'arêtes) déconnectés des uns et des autres (Qi et al., 2020).

L'**utilisation de la topologie des réseaux pour faire la prédiction de fonctions spécifiques ou bien de phénotypes**. Pour cet angle d'approche, il est intéressant d'utiliser différentes types de données (ex : informations sur les termes ontologiques, données issues de la transcriptomique, de la protéomique, de la métabolomique etc.) pour faire de la projection de données afin mieux les contextualiser. Chez la levure, Costanzo et collègues ont reconstitué le réseau d'interaction génique de la levure et ont pu identifié des regroupements de gènes avec des fonctions biologiques similaires (Costanzo et al., 2016). Par exemple pour les réseaux de PPIs, on pourrait proposer des prédictions de fonction ou d'évolution commune pour des groupes de protéines dont les liens fonctionnels n'ont pas encore été déterminé de manière expérimentale.

Grâce à ces approches d'analyse de réseaux, on peut tester si le modèle omnigènique avec des réseaux de PPIs est une bonne abstraction pour la représentation d'un phénotype complexe. Nous montrerons à travers le chapitre 3, comment on peut mettre à profit les réseaux de PPIs pour étudier des phénotypes complexes par la projection de données issues de cartographie QTL reposant sur la quantification de ces PPIs.

1.3. Préambule des travaux présentés dans cette thèse

Dans cette thèse, notre objectif est d'utiliser des méthodes bio-informatiques centrées sur l'étude du protéome pour étudier différents phénotypes complexes comme le vieillisement et la réponse aux médicaments dans différents organismes modèles.

Le chapitre 2 propose une approche de génomique comparée pour étudier les propriétés intrinsèques des protéines, la propension d'agrégation et la tolérance des mutations, pour étudier les différences d'espérance de vie entre le rat-taupe nu et la souris. Ce travail a été publié dans la revue *Genome Biology and Evolution* en avril 2022.

Le chapite 3 met en avant la conception et l'implémentation d'une cartographie fine des variants génétiques qui influencent la formation des PPIs, dans une population de levure génétique diversifiée, qui est exposée à différents médicaments. Ce travail est sous la forme d'un manuscrit prêt à être soumis.

Le chapitre 4 offre une discussion portant sur l'originalité des méthodes proposées par les deux précédents chapitres, leurs limites et perspectives. Enfin il détaillera également quels sont les avancements que cette thèse propose dans le domaine de la bio-informatique.

Matériel Supplémentaire pour l'introduction

Noms complets	Code à 3 Lettres	Polarité	Hydrophobicité	Propriétés spécifiques
Glycine	Gly (G)	Polaires	Hydrophile	Plus petit acide aminé
Sérine, Thréonine	Ser (S), Thr (T)			Non chargées avec groupement hydroxyle
Asparagine, Glutamine	Asn (N), Gln (Q)			Non chargées avec groupement amine
Acide aspartique, Acide glutamique	Asp (D), Glu (E)			Chargées +
Histidine, Lysine, Arginine	His (H), Lys (K), Arg (R)			Chargées -
Cystéine, Méthionine	Cys (C), Met (M)	Non-Polaires	Hydrophobe	Non chargées contenant un élément soufre
Phénylalanine, Tryptophane, Tyrosine	Phe (F), Trp (W), Tyr (Y)			Non chargées contenant un cycle aromatique
Alanine, Isoleucine, Leucine, Proline, Valine	Ala (A), Ile (I), Leu (L), Pro (P), Val (V)			Aliphatiques

Tableau 1. Les différentes propriétés physico-chimiques des acides aminés

Premier article.

Etude comparative de la tendance d'agrégation et de la tolérance aux mutations des protéines entre le rat-taupe nu et la souris

par

Savandara Besse[1,2,*], Raphaël Poujol[3] et Julie Hussin[3,4]

([1]) Centre Robert-Cedergren en Bioinformatique et Génomique, Université de Montréal - Montréal, Québec, Canada

([2]) Département de Biochimie et Médecine Moléculaire, Faculté de Médecine, Université de Montréal - Montréal, Québec, Canada

([3]) Institut de Cardiologie de Montréal - Montréal, Québec, Canada

([4]) Département de Médecine, Faculté de Médecine, Université de Montréal - Montréal, Québec, Canada

État du manuscrit : Publié dans **Genome Biology and Evolution**.

Correspondance : savandara.besse@umontreal.ca

Disponible dans Genome Biology and Evolution (May 2022)

Résumé. Les mécanismes moléculaires du vieillissement et de la longévité ont été étudié dans des organismes modèles avec une courte espérance de vie. L'étude des espèces vivant plus longtemps contribuerait à l'élaboration de stratégies avantageuses pour vieillir en bonne santé comme des traitements thérapeutiques pour des maladies associées à l'âge. Le rat-taupe nu, connu pour son extrême longévité, possède des signes phénotypiques atténués du vieillissement en comparaison à la souris. Leur résistance au stress oxydatif est une caractéristique d'un vieillissement sain, suggérant un meilleur maintien de leur homéostasie cellulaire, au niveau des protéines. Pour identifier les principes généraux permettant la robustesse de la protéostase, nous avons comparé la tendance d'agrégation et la tolérance de mutations entre les protéines orthologues du rat-taupe nu et de la souris. Notre analyse a montré qu'il n'existe pas de différence globale du protéome pour ces deux propriétés, mais nous avons identifié des groupes de protéines qui ont des différences significatives de tendance d'agrégation. Nous avons trouvé un enrichissement de protéines avec une tendance d'agrégation plus élevée chez le rat-taupe nu, dont certaines sont impliquées dans les complexes d'inflammasome et dans l'interaction avec les acides nucléiques. Nous avons aussi observé que les protéines avec une basse tendance d'agrégation chez le rat-taupe nu ont une meilleure tolérance de mutation que le reste des protéines ; certaines étant relatives à des maladies associées à l'âge. Ces découvertes corroborent l'hypothèse que le protéome du rat-taupe nu aurait la capacité de ralentir son vieillissement grâce aux propriétés de son architecture.
Mots clés : Rat-taupe nu, Longévité, Vieillissement, Homéostasie des Protéines, Tendance d'aggrégation protéique, Tolérance aux mutations

ABSTRACT. The molecular mechanisms of aging and life expectancy have been studied in model organisms with short lifespans. However, long-lived species may provide insights into successful strategies for healthy aging, potentially opening the door for novel therapeutic interventions in age-related diseases. Notably, naked mole-rats, the longest-lived rodent, present attenuated aging phenotypes compared to mice. Their resistance toward oxidative stress has been proposed as one hallmark of their healthy aging, suggesting their ability to maintain cell homeostasis, specifically their protein homeostasis. To identify the general principles behind their protein homeostasis robustness, we compared the aggregation propensity and mutation tolerance of naked mole-rat and mouse orthologous proteins. Our analysis showed no proteome-wide differential effects in aggregation propensity and mutation tolerance between these species, but several subsets of proteins with a significant difference in aggregation propensity. We found an enrichment of proteins with higher aggregation propensity in naked mole-rat, and these are functionally involved in the inflammasome complex and nucleic acid binding. On the other hand, proteins with lower aggregation propensity in naked mole-rat have a significantly higher mutation tolerance compared to the rest of the proteins. Among them, we identified proteins known to be associated with neurodegenerative and age-related diseases. These findings highlight the intriguing hypothesis about the capacity of the naked mole-rat proteome to delay aging through its proteomic intrinsic architecture.

Significant statement: The molecular mechanisms behind naked mole-rat longevity are still poorly understood. Here, we address how the proteome architecture can help delay the onset of aging in naked mole-rat by studying properties that modulate protein aggregation. We identify 1,000 proteins with significant differences in aggregation propensity and mutation tolerance involved in processes known to be dysfunctional during aging. These findings highlight how evolutionary adaptations in protein aggregation in distinct biological processes could explain naked mole-rat longevity. **Keywords:** naked mole-rat, longevity, aging, protein homeostasis, protein aggregation propensity, mutation tolerance

Contributions personnelles à ce chapitre

Mes contributions à l'article inclus dans ce chapitre sont les suivantes en tant que première auteure :

- Choix de la thématique d'étude sous un angle de génomique comparée
- Formulation des hypothèses de recherche
- Conception et implémentation des méthodologies informatiques présentées
  - Recherche automatique des protéines orthologues à analyser
  - Établissement des critères d'exclusion des protéines de basse qualité et des protéines membranaires
  - Développement d'un pipeline permettant le calcul de scores d'agrégation, à l'échelle de la séquence entière et des domaines protéiques, pour l'ensemble du protéome partagé entre les organismes étudiés ($>$ 20,000 protéines)
  - Implémentation de l'algorithme permettant de réaliser la mutagenèse computationnelle à large échelle (plusieurs millions de séquences générées)
  - Définition et implémentation des différentes équations relatives à la mesure de tolérance de mutation se basant sur des scores de propension d'agrégations
- Analyses bio-informatiques et interprétation des résultats
- Rédaction de l'intégralité de l'article (introduction, méthodes, résultats, discussion) en collaboration avec Professeure Julie Hussin (co-auteure),
- Conception de toutes les figures en collaboration avec Professeure Julie Hussin (co-auteure) et Professeur Adrian Serohijos
- Réalisation de toutes les figures
- Révisions, réponses aux arbitres et analyses supplémentaires pour parution finale de l'article.

L'ensemble des contributions informatiques de ce projet sont disponibles sur le répertoire Github suivant : https://github.com/ladyson1806/NKR_lifespan

Ce travail n'aurait pu être réalisé sans l'aide de :

- Professeur Sebastian Pechmann qui a supervisé les analyses préliminaires associées à ce chapitre,
- Professeure Julie Hussin (co-auteure) et Professeur Adrian Serohijos qui ont tout deux activement participé :
  - à la supervision des analyses finales présentes dans la publication de cet article
  - à la phase d'analyse et d'interprétation des résultats
  - à la relecture et aux corrections des différentes versions de l'article jusqu'à sa publication finale

- M. Raphaël Poujol (co-auteur) dont j'ai sollicité l'expertise technique à plusieurs reprises pour consolider mes analyses bio-informatiques.

Chapitre 2

Comparative study of protein aggregation propensity and mutation tolerance between naked mole-rat and mouse

2.1. Introduction

Understanding the mechanism of aging and life longevity is a major biological problem. The hallmarks of aging describe the dysfunction of several biological processes such as genomic instability, telomere attrition, loss of protein homeostasis (proteostasis), epigenetic alterations, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, deregulated nutrient-sensing pathways, and altered intercellular communication (López-Otín et al., 2013). The aggravation of these hallmarks usually leads to an early manifestation of aging while their amelioration contributes to its delay and an increase of healthy lifespan. However, all the hallmarks are not yet fully supported by experimental interventions that succeeded in improving aging and extending lifespan. The genetics behind the hallmarks of aging have been identified through genetic perturbation studies in multiple model organisms such as yeast, nematodes, flies, and mice (reviewed in Singh et al., 2019; Taormina et al., 2019). These model organisms have been critical in our understanding of aging thanks to their short lifespan that aids tractable experimentation, relatively cheap maintenance, and possibilities for genetic manipulation. However, there is a need to study organisms with longer lifespans to understand the mechanisms behind their longevity better. Recent whole-genome sequencing efforts allowed the study of organisms with a longer lifespan. Cross-species "omics" studies of these long-lived species, such as transcriptomic, metabolic, and lipidomic profiles associated with long-lived species, highlighted molecular signatures that could be important to aging (reviewed in Ma and Gladyshev, 2017; Tian et al., 2017). One notable example is the naked mole-rat, the longest-lived rodent among those with a known maximum lifespan and a model organism for studies on healthy aging

and longevity (Buffenstein and Ruby, 2021). Indeed, this organism presents attenuated age-related changes, suggesting the presence of anti-aging mechanisms contributing to its longevity (Buffenstein, 2005). Several comparative studies between naked mole-rat and mice reported significant differences in their maintenance of protein homeostasis. Naked mole-rats show high oxidative damage levels from young ages (Andziak et al., 2006). Still, their ubiquitinylated proteins are maintained at lower levels at both young and old ages, suggesting less accumulation of damaged and misfolded proteins during aging (Pérez et al., 2009). The low levels of damaged and misfolded proteins could also be explained by their high proteasome activity (Rodriguez et al., 2012). Taken together, these observations emphasize the importance to study the general principles contributing to the robustness of protein homeostasis in the naked mole-rat. Nevertheless, these general principles have not been established at the proteome level. Thus, in this paper, we propose identifying the proteomic features that contribute to protein homeostasis maintenance.

In naked mole-rats, several works have studied the molecular key players of protein homeostasis and their potential role to rodent longevity. For example, proteostasis-centered theories of aging propose that aging results from the decline of quality-control systems involved in protein synthesis, degradation, and chaperoning that normally contribute to protein turnover (Balch et al., 2008; Powers et al., 2009; Proctor and Lorimer, 2011; Taylor and Dillin, 2011). Proteostasis is essential for protein stability through the protection of their structures and functions against environmental perturbations. Impaired proteostasis leads to the appearance of phenotypic aging markers and age-related diseases such as Alzheimer's and Parkinson's diseases, known to be characterized by the accumulation of protein aggregates of specific proteins (Hipp et al., 2019; Irvine et al., 2008; Powers et al., 2009). Indeed, there is an increase in the expression of chaperones with higher proteasome and autophagy activities in naked mole-rat (Tian et al., 2017). From a system biology perspective, the maintenance of proteostasis is essential for delaying the onset or slowing down the process of aging (Koga et al., 2011). In addition, the protein aggregates are processed by quality control systems such as chaperones and protein degradation pathways (proteasome and autophagy) (Morimoto and Cuervo, 2009). These mechanisms are robust in young individuals but tend to decline with age, leading to an increase of protein aggregates within the cell, thus participating in the dysfunction of multiple biological processes (Labbadia and Morimoto, 2015). A recent study in C. elegans describes the proteostasis decline with age and observed an exponential increase of protein aggregates in old cells (Santra et al., 2019).

Our study focuses on intrinsic protein properties that could contribute to proteostasis maintenance by reducing the formation of protein aggregates. Causes of protein aggregation can arise from protein features and cell features. Protein aggregation propensity is a

protein sequence feature that characterizes the ability of the protein to aggregate and is estimated based on the physicochemical properties of the amino acid sequence. However, this property of intrinsic aggregation propensity alone does not fully determine whether a protein will aggregate in vivo, which is determined by confounding cellular factors (e.g., cellular concentration, recruitment by chaperones). Whether a sequence that has high aggregation propensity will in fact aggregate will need to account for cellular features. In the cell, cumulative damage through non-enzymatic post-translational modifications from reactions with metabolites or reactive oxygen species (Golubev et al., 2017), leads to protein instability, and subsequently to the formation of protein aggregates. Alternatively, the formation of protein aggregates could result from destabilizing mutations. The accumulation of somatic mutation burden has been proposed as a driver of aging (Vijg, 2014). Several studies previously demonstrated the importance of mutation accumulation in the onset of aging and the reduction of lifespan (M. B. Lee et al., 2019; Lodato et al., 2018). However, it is still unclear whether the accumulation of mutation would contribute to the formation of protein aggregates. To tackle this question, we also propose to study "mutation tolerance" or the ability of proteins to tolerate the potential effects of mutations on their aggregation propensity.

Here, we performed a comparative analysis on protein aggregation propensity and the mutation tolerance between the naked mole-rat and the mouse. From the study of these two properties, we aim to understand how they might contribute to explaining the difference in lifespan between these two species. First, we estimated their aggregation propensity between the two species at the level of whole-protein sequences (entire ORFs) and at the level of individual folding domains. We performed a random and exhaustive computational mutagenesis to estimate the mutation tolerance of these proteins. We found that although there is no global difference of aggregation propensity in the proteome shared between naked mole-rat and the mouse, we identified groups of proteins that significantly differ in their aggregation propensity. This observation holds both at the level of individual domains and the level of entire protein sequences. By performing gene set enrichment analyses, we retrieve several biological processes, some of them were already reported to be potentially involved in the naked mole-rat longevity, notably processes associated with the immune system. We also highlight their inflammation's versability, as we found proteins with high and low aggregation propensities from this process. We also report proteins, previously reported as involved in neurodegenerative diseases in human, that has not yet been considered as aging gene markers. Furthermore, these subsets of proteins have different distributions of mutation tolerance in the naked mole-rat, but not in the mouse, suggesting specific adaptations of these properties in the longest-lived rodent.

2.2. Methods

2.2.1. Definition of the orthologous dataset and subsets

Orthologous sequences are homologous sequences that share similarities from a speciation event. The orthologous amino acid (AA) sequences shared between naked mole-rat and mice were retrieved using the Inparanoid algorithm (version 4.1) (Remm et al., 2001) with default parameters. As initial inputs, we use the naked mole-rat and mouse latest proteome assemblies, downloaded from Uniprot (https://www.uniprot.org/, accessed April 2019). The Inparanoid algorithm performs a reciprocal best-hit search to cluster the orthologous and in-paralog proteins, to identify the orthologous groups between the two species. For our analysis, each orthologous group was represented by a pair of proteins with the highest mutual best hit score, yielding 13,806 orthologous pairs. Mouse and naked mole-rat Uniprot protein identifiers are available in Supplementary Table S2 (Besse_et_al_SM.xlsx). To assess the quality of these orthologous pairs, we computed their local alignments with Matcher (Waterman and Eggert 1987; Huang and Miller 1991) and collected the percentage of similarity and the percentage of gaps within the pairwise alignments. Orthologous pairs with a percentage of similarity below 60% or a percentage of gaps above 20% were removed, altogether keeping a total of 13,513 pairs.

For the estimation of aggregation propensity from Tango software (see below), we excluded transmembrane proteins. To identify the proteins with transmembrane regions to exclude, we first parsed mouse gene annotations available in the proteome FASTA file and defined the ones containing the keyword "transmembrane" as transmembrane proteins and excluded them. Additionally, we also predicted transmembrane regions in the remaining sequences with TMHMM (Krogh et al., 2001). All mouse and naked mole-rat proteins with at least one transmembrane region predicted were removed, restricting our analyses to 9,522 protein pairs. We also collected their associated protein-coding nucleotide sequences for our computational large-scale mutagenesis analysis (see below). Moreover, we identified a specific subset, containing all the proteins known to interact with chaperone proteins. For this specific dataset, we used the human chaperone client proteins (annotated with their ENSEMBL identifiers) from a recent study (Victor et al., 2020) to infer the mouse chaperone clients. The human ENSEMBL identifiers were converted to their corresponding Uniprot identifiers for mapping them towards the mouse Uniprot ortholog identifiers. Similarly, we then mapped the mouse Uniprot identifiers to the naked mole-rat ortholog identifiers. This specific subset of orthologs is composed of 1,298 protein pairs.

2.2.2. Identification of protein domains in naked mole-rat and mouse

To obtain mouse and naked mole-rat domain definitions, we first collected mouse domain information from the Pfam database (http://pfam.xfam.org/, version 33.1). Within a given protein, we considered any peptide as a functional domain when their entire sequence matched domain annotations, corresponding to the start and end positions in PFAM protein alignments. For the naked mole-rat, the domain definitions were inferred using the reciprocal best hit method where the mouse annotated domains are used as reference. We collected a total number of 19,413 annotated domains available for 8,475 protein pairs, representing 89% of our initial dataset.

2.2.3. Phylogenetic tree and data related to longevity

The evolutionary distances between rodent species were determined using TimeTree (Kumar et al., 2017), through the available webserver. This method retrieved all existing phylogenetic trees for the given species and provided the concatenation of these trees to determine the median time when species diverged. These phylogenetic trees were built based on gene alignments. The available information on maximum lifespan, adult weight, female maturity, and metabolic rate for rodent species was retrieved from the AnAge database (Tacutu et al., 2018, build 14) and are given in Table S1 (Besse_et_al_SM.xlsx). We reported more recent maximum lifespans for naked mole-rat (Buffenstein and Ruby, 2021) and damaraland naked mole-rat (Rodriguez et al., 2016).

2.2.4. Computation of aggregation propensity scores

To predict the propensity of proteins to aggregate, we used the Tango software (Fernandez-Escamilla et al., 2004). Tango assigns per-residue aggregation propensity scores based on the amino acid physicochemical properties. For each orthologous protein pair, we computed the per-residue aggregation score with Tango for each sequence independently and then calculated their whole-protein sequence aggregation and domain aggregation. Per-domain aggregation score is defined as the sum of the per-residue aggregation propensity score for a defined functional domain divided by the domain length (Agg_D, Equation (2.1)). The whole-protein sequence aggregation propensity score is defined as the sum of per-residue aggregation propensity scores for the entire sequence divided by the protein length (Agg_P, Equation (2.2)).

$$Agg_D = \frac{\sum Per-residue\ Aggregation\ propensity\ score\ (for\ domain\ sequence)}{Domain\ Length} \quad (2.1)$$

$$Agg_P = \frac{\sum Per-residue\ Aggregation\ propensity\ score\ (for\ whole-protein\ sequence)}{Protein\ Length} \tag{2.2}$$

2.2.5. Identification of proteins with significant difference of aggregation propensity

To compare mouse and naked mole-rat protein aggregation propensity scores, we computed their difference at the domain (ΔAgg_D, Equation (2.3)) and the whole-protein sequence (ΔAgg_P, Equation (2.4)) levels with the following formulas:

$$\Delta Agg_D = Agg_{D;Naked-Mole\ Rat} - Agg_{D;Mouse} \tag{2.3}$$

$$\Delta Agg_P = Agg_{P;Naked-Mole\ Rat} - Agg_{P;Mouse} \tag{2.4}$$

The difference of aggregation propensity scores was normalized to obtain z-scores. Proteins with z-scores exceeding 2 times the standard deviation are considered significantly different from each other. Both for whole-sequence and domain aggregation propensity analyses, two groups were defined as: (i) proteins with ΔAgg z-scores > 2 being considered to have a higher aggregation in naked mole-rat compared to mouse; (ii) proteins with ΔAgg z-scores < -2 being considered to have a lower aggregation in naked mole-rat compared to mouse.

2.2.6. Functional enrichment analyses

With the previously identified subsets of proteins, we investigated in which cellular components, molecular functions, and biological processes from GO annotations, these proteins are over or under-represented. To do so, we used hypergeometric tests implemented on the Panther database (Mi et al., 2019). As the protein annotations for naked mole-rat were not proposed in the database, we used the annotations from the mouse, assuming the naked mole-rat proteins have similar annotations to their mouse orthologs. The subsets from the domain analysis were compared to the set of proteins with annotated domains within the shared proteome (n=8,475). The subsets from the whole-protein sequence analyses were compared to all the proteins of the shared proteome (n=9,522). Raw p-values of Fisher's exact tests were computed to identify the gene ontologies significantly over- or under-represented for each subset, corrected by a False Discovery Rate (FDR). Only GO terms associated with at least 5 proteins are shown in Figure 2.3. The entire list of GO terms with FDR < 0.05 and, for the domain and the whole-protein sequence analyses, are available in Table S3 and S4, respectively (Besse_et_al_SM.xlsx). The list of proteins within the groups and their

annotations are available in Table S5 (Besse_et_al_SM.xlsx). To identify which GO terms where the chaperone client proteins are differently distributed compared to the rest of the proteins, we computed chi-square tests, corrected by a Benjamini/Hochberg FDR.

2.2.7. Quantification of protein mutation tolerance

This quantification of mutation tolerance was initially performed on 9,522 protein pairs. However, 176 proteins (mostly proteins with more than 10,000 amino acids) were removed as the calculation of their mutation tolerance score was too computationally expensive, thus, reducing the dataset to 9,346 protein pairs. We also removed protein pairs where naked mole-rat coding sequences were truncated, obtaining a final dataset of 7,939 proteins.

We designed a large-scale in silico mutagenesis experiment to estimate the mutation tolerance of the proteins shared between naked mole-rat and mouse. Specifically, the mutation tolerance score is a ratio from 0 to 1 that quantifies the ability of a protein to tolerate mutations. We mutated one nucleotide at a time within the DNA sequence to all 3 other possible nucleotide mutations (self-substitution is excluded). For example, for a coding sequence of X nucleotides, we would generate X3 possible substitutions that would engender X3 mutated sequences. All these DNA sequences are then translated into amino acid sequences. We kept only non-redundant protein sequences (resulting from non-synonymous changes), different from the wild-type sequence (WT), for predicting their protein aggregation propensity using Tango, as described in the section Computation of aggregation scores. Whole-protein sequence aggregation scores for mutated (MT) sequence were then computed and are used to calculate the difference of aggregation propensity (mutational aggregation propensity score, Mutational Agg_P - Equation (2.5)) between MT and WT sequences:

$$Mutational\ Agg_P = Aggregation_{P;MT} - Aggregation_{P;WT} \tag{2.5}$$

We defined 3 categories of proteins, according to their change in aggregation propensity:

- Mutational $Agg_P = 0$:
  No change in aggregation propensity of the mutated sequence
- Mutational $Agg_P > 1$:
  High increase in aggregation propensity of the mutated sequence
- Mutational $Agg_P < -1$:
  High decrease in aggregation propensity of the mutated sequence

For a given protein, these scores were used to define their mutation tolerance. It calculated the ratio of the number of mutations with no impact on protein aggregation normalized by the number of all possible mutations (Strict Mutation tolerance, Equation (2.6)). The

total number of mutations corresponds to the number of protein sequences which result from a non-synonymous substitution that does alter the length of the protein. Therefore, we exclude the truncated sequences resulting from the change of the first methionine of the amino acid sequence and the ones that contain premature codon stop by checking that the lengths of the wild-type protein sequence and the mutated sequence are equal.

$$Strict\ Mutation\ Tolerance = \frac{Number\ of\ (Mutational\ Agg_P = 0)}{Total\ Number\ of\ Mutations} \tag{2.6}$$

For identifying proteins with a significant difference in their strict mutation tolerance, we calculated the difference of strict mutation tolerance between naked mole-rat and mouse (ΔMutTol, Equation (2.7)).

$$\Delta MutTol = Strict\ Mutation\ tolerance_{Nakedmole-rat} - Strict\ Mutation\ tolerance_{Mouse} \tag{2.7}$$

All the ΔMutTol scores were normalized to obtain ΔMutTol z-scores. Proteins with ΔMutTol z-scores exceeding 2 times the standard deviation are considered significantly different from each other: (1) proteins with a ΔMutTol z-score $>$ 2 are considered to have higher strict mutation tolerance in naked mole-rat compared to mouse, (2) proteins with a ΔMutTol z-score $<$ -2 are considered to have a lower mutation tolerance in naked mole-rat compared to mouse. We tested a second definition of mutation tolerance that includes not only neutral mutations (no impact on aggregation propensity), but also mutations that would decrease aggregation propensity (Lenient Mutation tolerance, Equation (2.8)).

$$Lenient\ Mutation\ Tolerance = \frac{Number\ of\ (Mutational\ Agg_P <= 0)}{Total\ Number\ of\ Mutations} \tag{2.8}$$

Additionally, we computed a metric that estimates the proportion of mutations (Proportion of beneficial mutation, Equation (2.9)) resulting in a decrease of aggregation propensity, which we call "beneficial mutations". For $Mutational\ Agg_P$ scores below -1, we consider a mutation as beneficial.

$$Proportion\ of\ beneficial\ mutations = \frac{Number\ of\ (Mutational\ Agg_P < -1)}{Total\ Number\ of\ Mutations} \tag{2.9}$$

2.2.8. Pairwise comparison of aggregation propensity and mutational aggregation propensity for ATX proteins

We first generated the multiple sequence alignment (MSA) of ATX-10 and ATX-3 using Muscle with the default parameters for protein alignment (Edgar, 2004). We detected the presence of gaps in the mouse sequence compared to the naked mole-rat sequence. To determine if these gaps correspond to insertion or deletion events, we blasted the naked mole-rat amino acid sequence across Uniprot database (Blast), which were not found with a match in any other rodent species than naked mole-rat (e-value < 0.001). We mapped the per-residue aggregation propensity scores, generated with Tango, to the MSA positions. Similarly, the mutational aggregation propensity scores (Mutational Agg_P - Equation (2.5)) were also mapped to the MSA positions. Concretely, for a specific amino acid, we associated the different mutational aggregation propensity score corresponding to all mutated sequence that include a substitution event at this position. If the Mutational Agg_P score is above 1 we consider the introduced random mutations to be detrimental, as it increased aggregation propensity. If Mutational Agg_P score is below -1, we consider the introduced random mutations as beneficial, as it decreased aggregation propensity.

2.2.9. Figure generation and statistical analysis

The different plots were generated with Python graphic libraries, Matplotlib (version 3.2.1), Seaborn (version 0.10.0), and Plotnine (version 0.8.0). All statistical analyses were performed using the Scipy stats module (version 1.6.2), unless specified otherwise. The FDR correction were computed with the statsmodels module (0.12.2), unless specified otherwise. Significance thresholds for p-values and FDR were set at 0.05. Statistical tests and p-values are reported in the figure legends can be found as outputs of the Python3 scripts that generate the figures.

2.3. Results

2.3.1. Analysis of the orthologous proteome shared between naked mole-rat and mouse

To check the lifespan variability across rodents (Figure 2.1A), we collected maximum lifespan data available in the AnAge database (Tacutu et al., 2018) and retrieved information for 18 species. Furthermore, we extracted several metrics describing life-history traits such as body mass, basal metabolic rate, and female maturity available in AnAge, previously shown to be correlated with maximum lifespan in mammals (Fushan et al., 2015). On the reconstructed rodent phylogenetic tree, we observed that indeed the naked mole-rat is the longest-lived rodent (Ruby et al., 2018) and shares a common ancestor with other rodents living more than 12 years (Figure 2.1A, blue). This group is separated from a larger monophyletic group, which include a large cluster (Figure 2.1A, red) with rodents with a shorter maximum lifespan, less than 10 years, including mouse. The remaining two groups (Figure 2.1A, in green and orange) contain a low number of species with no clear tendency in their maximum lifespan. We plotted life-history traits metrics against the maximum lifespan (Figure 2.1B-D), confirming that naked mole-rat is an outlier from the rest of the rodents. These observations support the fact that the naked mole-rat is an appropriate organism to study aging because of its unexpectedly long lifespan among rodents, in contrast to mouse that is a good representative for short-lived species.

To identify the general principles behind the naked mole-rat longevity, we compared the orthologous proteome shared between naked mole-rat and mouse. The mouse has a well-curated and annotated genome and has also been extensively studied in the field of aging (Mitchell et al., 2015). Our comparative analysis between naked mole-rat and mouse focuses on 13,806 ortholog pairs collected from the orthologous mapping database Inparanoid (see Methods). We considered two properties among these orthologous proteins, specifically: 1) their aggregation propensity and 2) their mutation tolerance, to determine if they could partly explain the higher maintenance of protein homeostasis in naked mole-rat compared to the mouse. To study these properties within the two species, we estimated the aggregation propensity of the ortholog pairs using the software Tango (Fernandez-Escamilla et al., 2004) (see Methods), which scores the per-residue aggregation propensity of protein sequences. With this tool, the property of protein aggregation propensity is accurately predicted on proteins with no transmembrane regions; therefore, we excluded the transmembrane proteins (see Methods), leaving a total of 9,522 ortholog protein pairs.

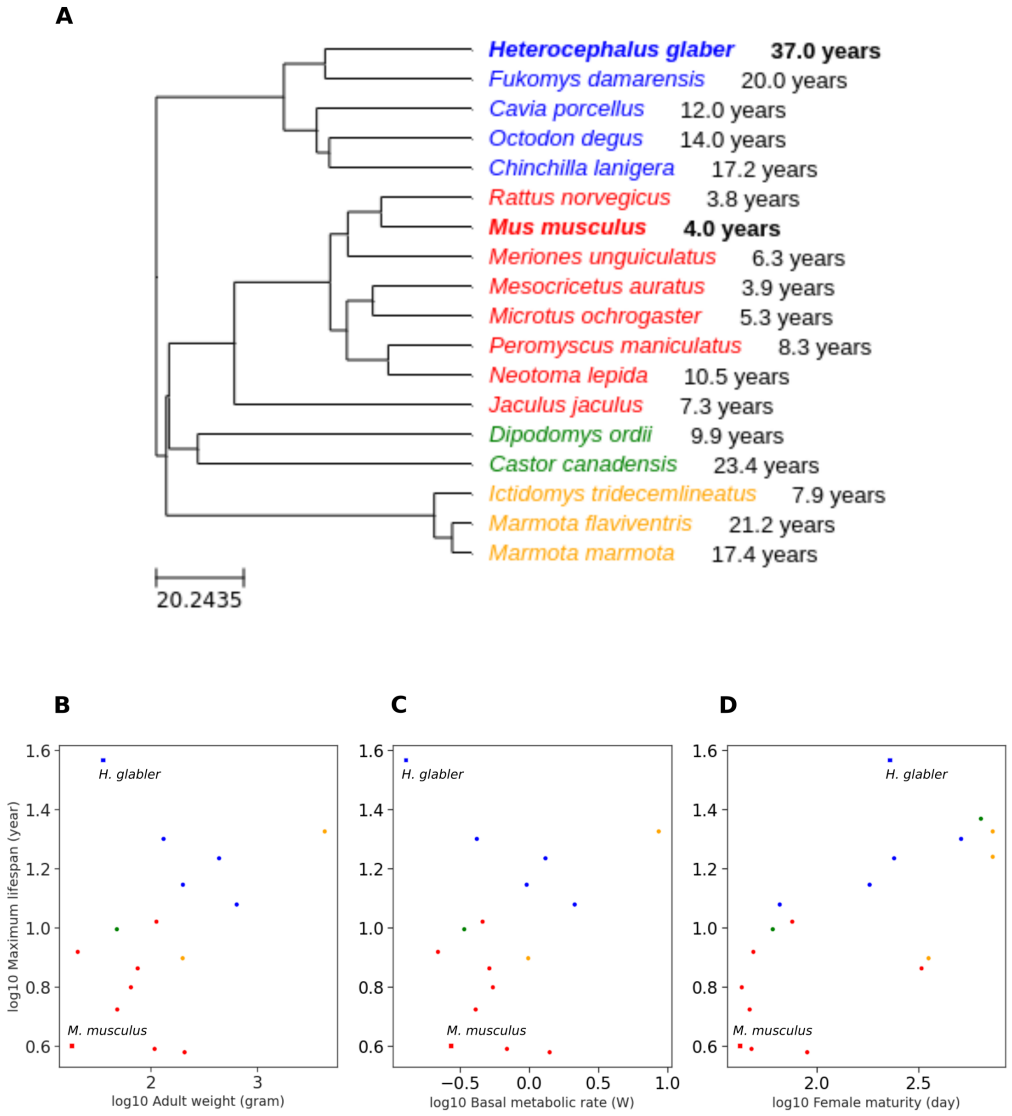

Fig. 2.1. Maximum lifespan variation across rodents

A), Phylogenetic distribution of rodent species with known maximum lifespan. The tree was generated with TimeTree using rodent species with known maximum lifespan. Four groups were colored according to their closest common ancestor. Mouse (Mus musculus) and naked mole-rat (Heterocephalus glaber), highlighted in bold, are the selected organisms for our comparative study, as they have a drastic difference of maximum lifespan. Mouse can live up to 4 years while naked mole-rat can live up to 37 years. Rodent maximum lifespan compared to **B)**, adult weight, **C)**, basal metabolic rate, and **D)**, female maturity, for the rodents mentioned in A. Maximum lifespan, adult weight, female maturity, metabolic rate data are extracted from the AnAge database. All values were log10-transformed. Mouse and naked mole-rat are represented with a square shape.

Since different regions of an ORF could have different folding properties, the aggregation propensity scores were also computed at the domain level. To do so, we retrieved 19,413 annotated domains available for 8,475 proteins (see Methods). Moreover, we looked more closely at a specific subset of proteins, the chaperone client proteins, which are the proteins interacting with chaperones in known protein-protein interaction networks. This subset is composed of 1,298 protein pairs (see Methods).

2.3.2. Specific subsets of proteins display significant differences in aggregation propensity

The accumulation of protein aggregates is potentially toxic to cells (Stefani and Dobson, 2003) and results from the decline of protein homeostasis. Protein aggregation tends to increase with age and initiate amyloid-beta aggregation in nematodes and mice (Groh et al., 2017). Since such protein aggregates are found in specific tissues and cause age-related diseases such as Alzheimer's and Parkinson's diseases, we asked whether the systematic presence of proteins with higher chance to aggregate within the cells could be correlated to the onset of aging. In the naked mole-rat, despite high levels of oxidation, they maintain low rates of ubiquitylated proteins (Pérez et al., 2009), suggesting a reduced formation of protein aggregates. To identify if there is a proteome-wide difference in protein aggregation propensity between naked mole-rat and mouse, we first estimated the protein aggregation propensity on the ortholog proteins using Tango (see Methods). For a given protein sequence, this approach estimates the per-residue aggregation propensity scores based on their physicochemical properties with specific environmental parameters. With these scores, we computed two metrics, (1) an aggregation score for the whole-protein sequence and (2) an aggregation score for each annotated domain of the proteins (see Methods). We compared the aggregation scores between naked mole-rat and mouse, in the whole-protein sequence, and their domains (Figure 2.2).

Overall, whole-protein sequence propensity scores are low (Naked mole-rat Agg_P=3.48 +/- 2.77, Mouse Agg_P=3.37 +/- 2.73) and per-domain aggregation propensity scores have higher variance than whole-protein sequence (Naked mole-rat Agg_D=3.79 +/- 4.60, Mouse Agg_D=3.76 +/- 4.57). We observed a high correlation in aggregation propensity between naked mole-rat and mouse at the whole-protein sequence (r=0.89, p-value=$2 * 10^{-16}$) and the domain (r=0.91, p-value=$2 * 10^{-16}$), indicating no proteome-wide global differences in aggregation propensity between these two species (Figure 2.2A,B). In parallel, we focused on the chaperone client proteins to see if they have specific aggregation propensity and mutation tolerance compared to the rest of the proteins since they interact with the chaperones. Their whole-protein sequence and per-domain aggregation propensity scores

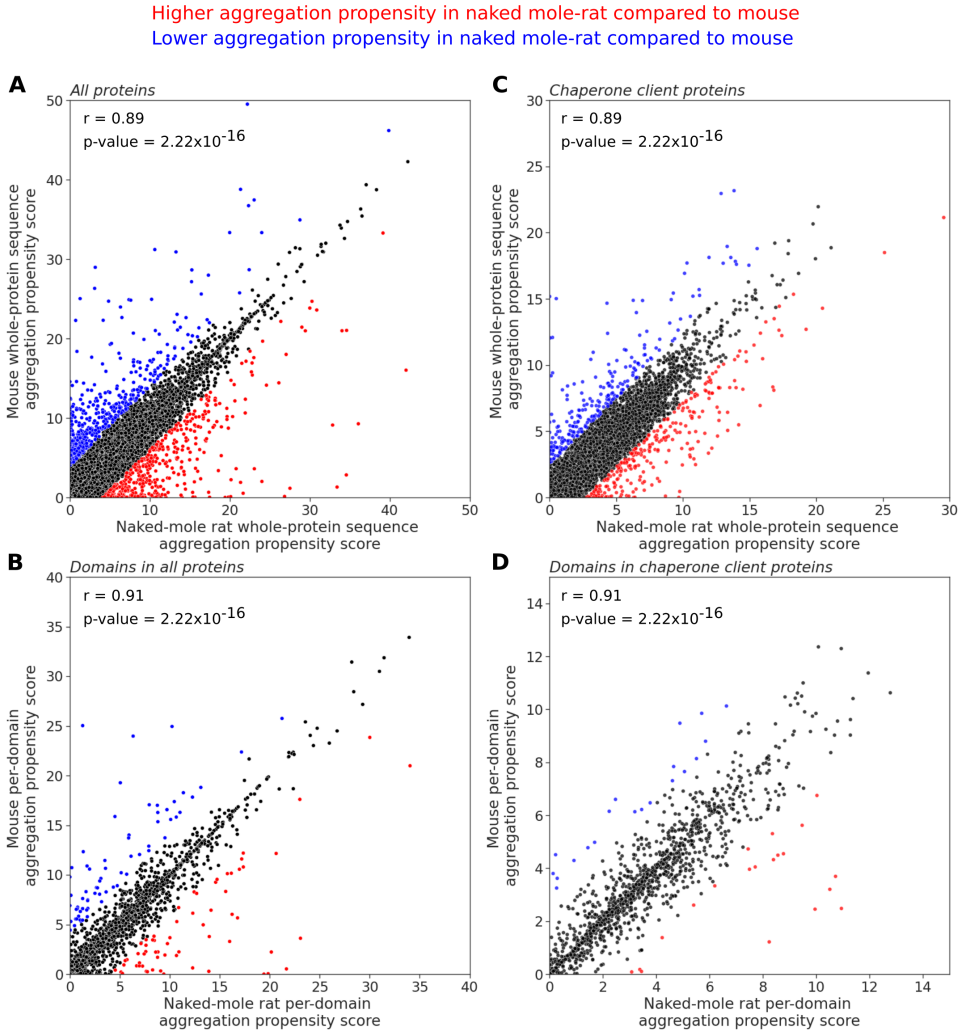

Fig. 2.2. Study of aggregation propensity in naked mole-rat and mouse

Comparison of aggregation propensity scores in orthologous proteins from naked mole-rat and mouse. Each point represents an ortholog pair. Whole-protein sequence aggregation propensity scores (Agg_P) **A)** for the whole dataset (n=9,522), **B)**, for the subset of chaperone client proteins (n=1,298). Per-domain aggregation propensity scores (Agg_D) **C)**, for the whole dataset (n=19,413 domains), **D)**, for the subset of chaperone client proteins (n=3,126 domains). see Methods for details on calculations Agg_D and Agg_P. Pearson correlations coefficients (r) between the naked mole-rat and mouse aggregation propensity scores are reported. Domains and proteins with a higher aggregation propensity in naked mole-rat compared to mouse are in red; and proteins with lower aggregation propensity are in blue.

are also low (Naked mole-rat Agg_P=3.48 +/- 2.37, Mouse Agg_P=3.37 +/- 2.31). We observed a high correlation in aggregation propensity, as in the all-proteins dataset, at the whole-protein sequence level (r=0.89, p-value=$2 * 10^{-16}$) and the domain level (r=0.91, p-value=$2 * 10^{-16}$) (Figure 2.2C,D), suggesting that chaperone client proteins do not differ

in terms of aggregation propensity between these two species.

We computed differences of aggregation propensity (ΔAgg) to identify proteins differing significantly between the species. Altogether, we found 269 proteins (including 20 chaperone clients) with higher whole-protein sequence aggregation propensity (z-scores > 2, see Methods) in naked mole-rat compared to mouse, and 247 proteins (including 21 chaperone clients) with lower aggregation propensity (z-scores < -2). In proteins with annotated domains (n=8,475), we found 904 protein domains with significantly different aggregation propensity scores within 754 different proteins. Specifically, 452 protein domains (including 63 domains from chaperone clients) have higher aggregation propensity (z-scores > 2) in naked mole-rat compared to mouse, and 452 protein domains (including 70 domains from chaperone clients) have lower aggregation propensity (z-scores < -2). In total, in combining the whole-protein sequence and per-domain analyses, we identified 1,155 distinct proteins with differences in their aggregation propensity. Additionally, we see no significant difference when comparing the distribution of ΔAgg z-scores from chaperone client proteins to proteome-wide values for the whole-protein sequence (p-value=0.72, Student t-test) and per-domain analyses (p-value=0.90, Student t-test). The proportion of proteins with a significant difference of aggregation propensity is similar in chaperone client proteins and the other proteins, indicating the chaperone client subset is not enriched in proteins with a significant difference of aggregation propensity between the naked mole-rat and mouse.

2.3.3. Function of proteins with a significant difference of aggregation propensity

We investigated the over- and under-representation of specific Gene Ontology (GO) annotation terms associated with protein subsets with either significantly high or low aggregation propensity in naked mole-rat (see Methods). We computed and sorted enrichment scores associated with each GO term (Figure 2.3).

We found enriched or depleted groups having proteins with low aggregation propensity in naked mole-rat (in blue). These groups are associated with GO terms within Biological Process (Figure 2.3A) and Cellular Component (Figure 2.3B) categories. Depleted groups in Biological Process category are cell organization ($5 * 10^{-7} <$ p-value $< 7 * 10^{-5}$), regulation of different macromolecule biosynthesis ($2 * 10^{-8} <$ p-value $< 8 * 10^{-5}$) and regulation of gene expression (p-value=$3 * 10^{-5}$). Proteins with significantly low aggregation propensity are under-represented in these processes. In contrast, enriched groups are related to immune response (p-value=8*10-6) and lipid metabolism (p-value=$4 * 10^{-5}$). The amid ceramidase (ASAH1), an enzyme involved in lipid metabolism, is associated with

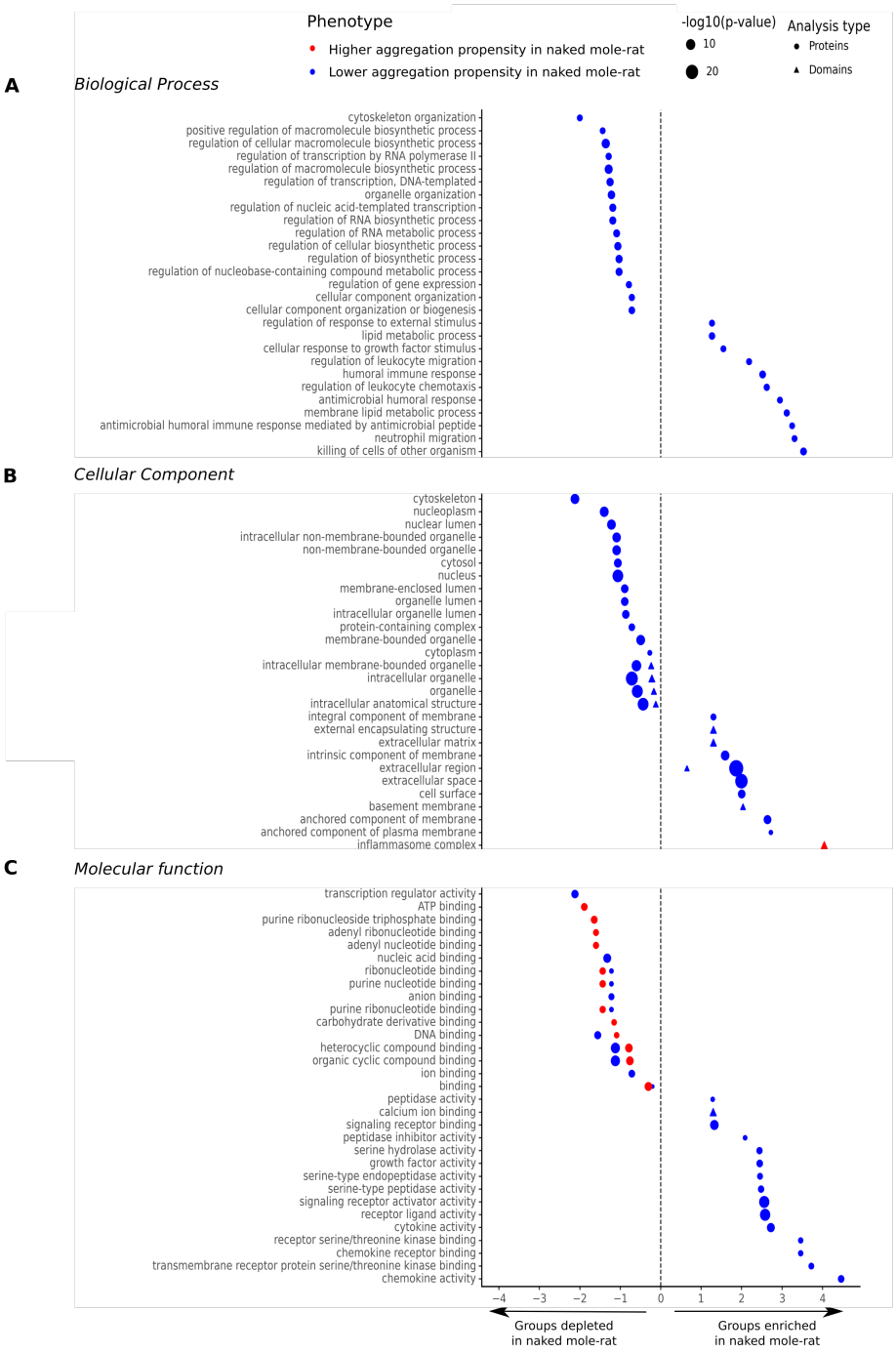

Fig. 2.3. Significant Gene Ontology (GO) terms associated with domains and proteins with higher and lower aggregation propensity in naked mole-rat

Log2 fold enrichment (FE) values indicate which GO terms are depleted (log2 FE < 0) or enriched (log2 FE > 0) in proteins ($\bullet$ shape) and domains ($\triangle$ shape) with a higher (in red) or a lower (in blue) aggregation propensity in naked mole-rat. The GO terms are grouped by categories: **A)**, Molecular Function, **B)**, Cellular Component, and **C)**, Biological Process. The size of the dots is proportional to their -log10 p-values. Only GO terms with at least 5 proteins and FDR < 0.05 are shown.

age-related diseases (Parveen et al., 2019). Depleted groups in Cellular Component category, are intracellular compartments, while enriched groups are membrane (p-value=$5 * 10^{-9}$ & p-value=$7 * 10^{-5}$) and extracellular components (Figure 2.3B), such as the extracellular matrix (p-value=$1 * 10^{-28}$) and the cell surface (p-value=$4 * 10^{-7}$). Notably, in these compartments, we found numerous metalloproteases from the matrixin family such as MMP3, MMP10, MMP13, MMP19 containing several hemopexin repeats; MMP7 and MMP25 with a peptidase M10 domain. These metalloproteases can degrade proteins from the extracellular matrix.

Additionally, we noticed that proteins in the inflammasome complex (p-value=$3 * 10^{-6}$) contain domains with significantly high aggregation propensity in the naked mole-rat. Particularly, we identified the peptidase C14 domain of CASP-1 (Caspase 1) and CASP-12 (Caspase 12) from the caspase family, the NOD2-WH domain of NLRP-1A (Nod-Like Receptor Pyrin domain-containing 1A), NLRP-3 (Nod-Like Receptor Pyrin domain-containing 3), and NLRP-6 (Nod-Like Receptor Pyrin domain-containing 6), the functional domain of GSDMDC1 (Gasdermin Domain-Containing protein 1), and the CARD domain of NLRC4 (Nod-Like Receptor CARD domain-containing protein 4). All these proteins are involved in inflammation. Surprisingly, only 2 of the 7 identified inflammasome complex proteins have higher aggregation propensity in naked mole-rat compared to mouse (z-scores > 2) at the whole-sequence level, despite all of them having domains with higher aggregation propensity in naked mole-rat compared to mouse (Figure 2.4). When investigating further the domains with higher aggregation propensity of the inflammasome proteins, we observed that similar protein domain families are shared across the ORFs. The domain peptidase C14 is restricted to the caspase family (Figure 2.4C), and usually has higher aggregation propensity scores in naked mole-rat than mouse. However, we observed that only peptidase C14-containing proteins with a significant difference of aggregation propensity between species are involved in the formation of inflammasome complexes, such as CASP-1, CASP-12 and CASP-4. The CARD domain is also shared among proteins but the difference of domain aggregation propensity is less consistent across proteins with CARD domain (Figure 2.4D).

The lack of enriched GO terms for the subset of proteins with high aggregation propensity in naked mole-rat than in mouse across all GO categories suggest this may be a random group of proteins. However, in the Molecular Function category (Figure 2.3C), we identified depleted groups in proteins with significantly high aggregation propensity (in red) related to ATP binding and its sub-categories (p-value=$1 * 10^{-5}$). The associated proteins with these functions have a more conserved aggregation propensity than expected by chance. Interestingly, we found depleted groups containing both proteins with higher and lower aggregation propensity in naked mole-rat, related to various binding functions. This

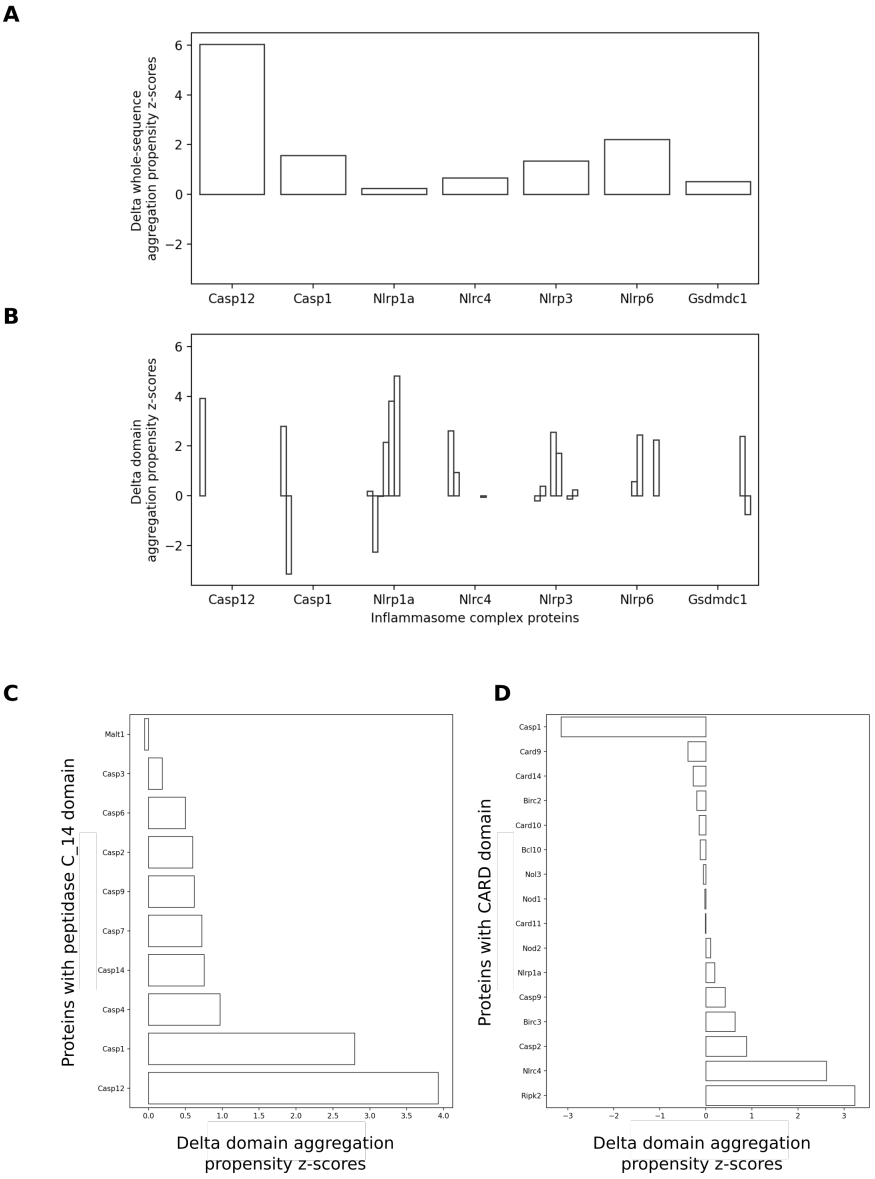

Fig. 2.4. Aggregation propensity in inflammasome proteins

A), Distribution of the difference of aggregation propensity z-scores between naked mole-rat and mouse in inflammasome proteins at the whole-sequence level. **B)**, Distribution of the difference of aggregation propensity z-scores between naked mole-rat and mouse in inflammasome proteins at the domain level. The number of bars per protein represents the number of shared domains between naked mole-rat and mouse. **C)**, Distribution of the difference of aggregation propensity z-scores between naked mole-rat and mouse in proteins with peptidase C14 domain. **D)**, Distribution of the difference of aggregation propensity z-scores between naked mole-rat and mouse in proteins with CARD domain.

observation supports the fact that proteins with specific and well-defined molecular functions are generally more structurally conserved across species and are less likely to have significant differences of aggregation propensity between species. Nevertheless, only one group (calcium ion binding, p-value=$3 * 10^{-6}$) contains proteins with differences of aggregation propensity from domains. The other enriched groups from the Molecular Function category have proteins with different enzymatic activities (serine-type peptidase, p-value=$3 * 10^{-5}$; serine hydrolase, p-value=$3 * 10^{-5}$). Among them, we identified Chymotrypsin-C, which contributes to proteolysis, the breakdown of proteins as polypeptides. Finally, we found enriched groups of proteins associated with chemokine and cytokine activity (p-value=1*105; p-value=$1 * 10^{-7}$, respectively). We identified several members of the chemokine family, the immunoglobulin receptor IL-40, the interferon-alpha IFNA13, the Cerberus and Wnt-2b proteins from the Wnt pathway. All annotations of the proteins associated with specific GO terms are shown in Table S5 (Besse_et_al_SM.xlsx).

Furthermore, the distribution of the number of proteins per GO terms within each category (Figure 2.5) is similar for chaperone client proteins and the other proteins, except for the ones associated with immune response and extracellular components (marked with an asterisk, corrected p-values < 0.05, chi-square test, Figure 2.5), indicating there are few or no proteins with lower aggregation in naked mole-rat that need chaperones to fold in these groups.

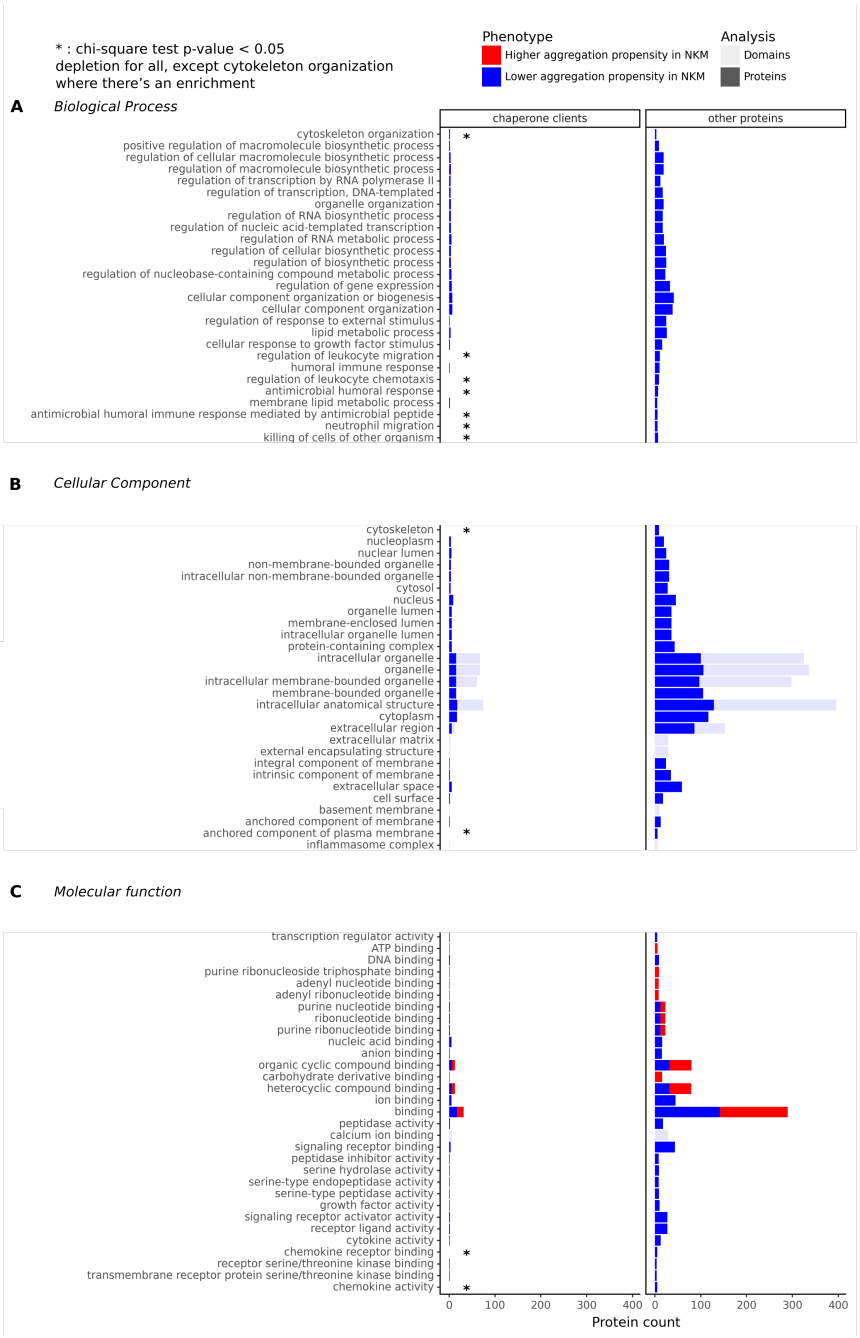

Fig. 2.5. Protein count per GO terms associated to proteins and domains with significant difference of aggregation propensity

This graph shows the number of proteins within each GO term. Redundant protein identifiers were not removed. **A)**, shows only chaperone client proteins and **B)**, shows the rest of the proteins. Counts of proteins from the domain analysis have lighter coloration than the counts of proteins from the whole-protein sequence analysis. Only GO terms with at least 5 proteins are shown. The groups of chaperone client proteins with significant differences of distribution compared to the rest of the proteins are marked (*).

2.3.4. Proteins with lower aggregation propensity in naked mole-rat better tolerate mutations

Finally, we explored the somatic mutation theory of aging by studying mutation tolerance in naked mole-rat and mouse orthologous proteins. This theory hypothesizes that mutation accumulation is an essential player in the onset of aging (S. R. Kennedy et al., 2012) and influences longevity. We designed a large-scale in silico mutagenesis experiment by generating all possible 1-nucleotide mutations on gene sequences for 9,346 protein pairs (of length below 10,000 amino acids) and then estimated the aggregation propensity of these mutants (see Methods). The difference between the aggregation propensity from mutated sequences and the aggregation propensity from the original sequences allows us to predict if a substitution would increase, maintain, or decrease this property. Assuming that proteins would preferably tolerate substitutions that do not significantly change their aggregation propensity, we derive a mutation tolerance score defined as a ratio of the number of substitutions with no change on the aggregation propensity (Mutational $Agg_P = 0$) divided by the total number of generated substitutions (strict mutation tolerance, see Methods – Equation (2.6)). These values range from 0 to 1, representing weak to strong tolerance to substitutions. In this definition, being tolerant correspond to the ability of the protein to strictly maintain the property of aggregation propensity. This score allows us to study the relationship between whole-protein sequence aggregation propensity and strict mutation tolerance of orthologous proteins in the two rodents (Figure 2.6).

There is a high correlation between strict mutation tolerance scores between naked mole-rat and the mouse (r=0.89, p-value$2 * 10^{-16}$), suggesting no global difference in strict mutation tolerance between their proteomes (Figure 2.6A). In both species, we observed a negative correlation between the sequence aggregation propensity and the strict mutation tolerance (r=-0.58, p-value $= 2 * 10^{-16}$), suggesting that proteins with a low aggregation propensity tend to be more resistant to substitutions. Moreover, it also suggests that proteins with high aggregation propensity scores contain more residues with non-zero aggregation propensity scores, suggesting that these residues might be more affected by random mutations (Figure 2.6B, C). Importantly, we identified subsets of proteins with significant differences in strict mutation tolerance between the species (Figure 2.6A). We tested whether proteins with significant differences in strict mutation tolerance between the species have a similar aggregation propensity to the rest of the dataset. The distribution of aggregation propensity for proteins with higher strict mutation tolerance compared to the distribution of other proteins is significantly different in both species (Figure 2.6B, Mouse p-value=$1*10^{-19}$; Figure 2.6C Naked mole-rat p-value=$2*10^{-15}$, Kolmogorov–Smirnov test), with proteins with higher strict mutation tolerance in a species having lower aggregation

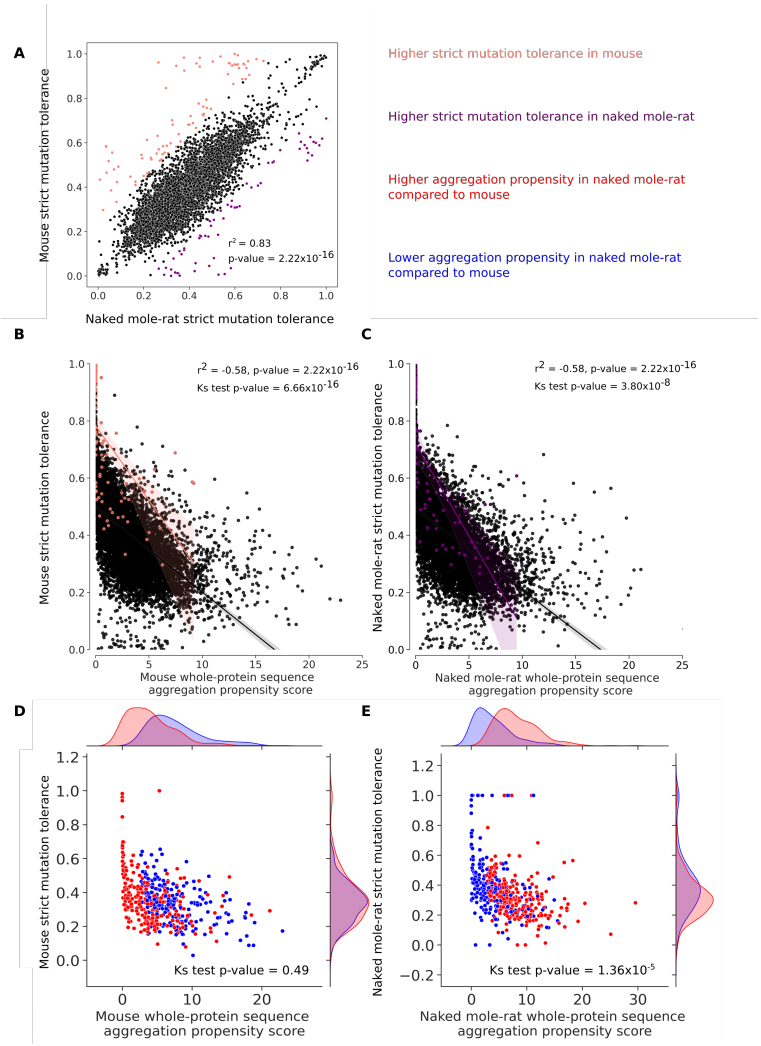

Fig. 2.6. Study of mutation tolerance in naked mole-rat and mouse

A), Comparison of mutation tolerance scores in orthologous proteins between naked mole-rat and mouse (n=9,346 proteins). Proteins in the naked mole-rat with higher mutation tolerance are in purple, the ones with lower mutation tolerance are in pink. Correlation between mutation tolerance and whole-protein sequence aggregation propensity scores in **B)**, mouse and **C)**, naked mole-rat. Protein pairs with significant differences in mutation tolerance are colored, using the color code from panel **A)**. Pearson correlation (r) between mutation tolerance and aggregation propensity are reported in both organisms. Kolmogorov–Smirnov test is used to assess the difference of distribution between proteins with mutation tolerance scores similar in mouse and naked mole-rat, and the ones which are different. Scatterplots of mutation tolerance against whole-protein sequence aggregation propensity scores in **D)**, mouse and in **E)**, naked mole-rat, restricted to the subsets of proteins identified with significant difference of aggregation propensity (n=510 proteins). Proteins with higher aggregation in naked mole-rat compared to mouse are in red, proteins with lower aggregation are in blue. Kolmogorov–Smirnov (KS) test is used to assess differences in mutation tolerance distributions between the two subsets in each organism.

propensity compared to the rest of the proteins. This result implies that the proteins with low aggregation propensities better tolerate mutations which is not surprising, given that our strict mutation tolerance score itself is based on the whole-protein sequence aggregation propensity. We investigated the function of these proteins by performing an enrichment analysis as previously described, but no specific GO term was under- or over-represented in these subsets.

Moreover, we investigated the strict mutation tolerance scores of the proteins with higher and lower aggregation propensity in naked mole-rat compared to a mouse. In the mouse (Figure 2.6D), the distributions of strict mutation tolerance scores between higher and lower aggregation propensity proteins are not significantly different (p-value=0.49, Kolmogorov–Smirnov test), indicating that the distributions of the strict mutation tolerance of the two subsets are similar. However, in naked mole-rat (Figure 2.6E), we find a significant difference in the strict mutation tolerance scores between higher and lower aggregation subsets (p-value=$2 * 10^{-8}$, Kolmogorov–Smirnov test). In naked mole-rat, proteins with lower aggregation propensity better tolerate substitutions than proteins with higher aggregation propensity. These proteins are found in biological processes or pathways shown in Figure 2.3, which we will discuss as potential players towards naked mole-rat longevity. We next tested if our results hold up when mutations that decrease aggregation propensity are included in the computation of another mutation tolerance score (referred as lenient mutation tolerance, see Methods, Equation (2.8)). In this definition, we include not only neutral mutations (no change in aggregation propensity), but also mutations that would decrease aggregation propensity (Mutational $Agg_P <= 0$). Indeed, it could be assumed that these mutations would be beneficial and should therefore be considered as tolerated. With this alternative definition, although the lenient mutation tolerance of naked mole-rat and mouse are less correlated (r= 0.77 vs. r=0.83), the main results described above hold (Figure 2.7).

In mouse, the distributions of "lenient" mutation tolerance scores between higher and lower aggregation propensity proteins are still not significantly different (p-value=0.56 Kolmogorov–Smirnov test) whereas, in naked mole-rat (Figure 2.7C), a similar difference in the lenient mutation tolerance scores between higher and lower aggregation subsets is seen (suggestive p-value=0.06, Kolmogorov–Smirnov test). Since including beneficial mutations weakens the difference seen in naked mole-rat, we further explored the specific signature associated with beneficial mutations (Figure 2.7D). We computed the proportion of beneficial mutations in proteins in both species (see Methods) and observed a low correlation between proportions from naked mole-rat and mouse (r: 0.43, p-value=$2.22 * 10^{-16}$), suggesting that the proteins that can possibly improve their property of aggregation propensity

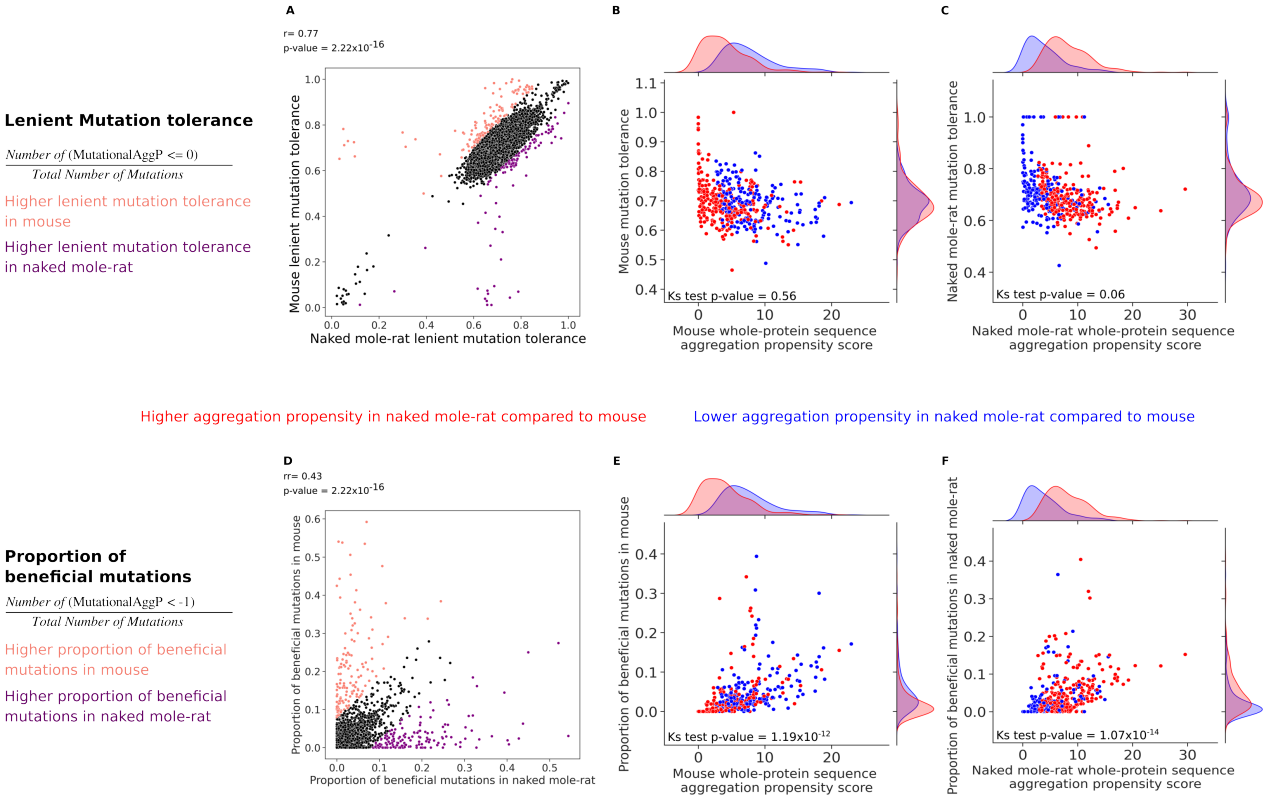

Fig. 2.7. Signatures found from the lenient mutation tolerance and the proportion of beneficial mutations

A), Comparison of lenient mutation tolerance scores in orthologous proteins between naked mole-rat and mouse (n=7,939 proteins). This definition takes account of mutations with no impact in aggregation propensity and the ones that reduce the aggregation propensity. Pearson's correlation (r) was used to compare the lenient mutation tolerance scores between the two species. Scatterplots of mutation tolerance against whole-protein sequence aggregation propensity scores in **B)**, mouse and in **C)**, naked mole-rat, restricted to the subsets of proteins identified with significant difference of aggregation propensity (n=510 proteins). Kolmogorov–Smirnov (KS) test is used to assess differences in mutation tolerance distributions between the two subsets in each organism. **D)**, Comparison of proportion of beneficial mutations in orthologous proteins between naked mole-rat and mouse (n=7,939 proteins). Pearson's correlation (r) was used to compare the proportion of beneficial mutations between the two species. Scatterplots of proportion of beneficial mutations against whole-protein sequence aggregation propensity scores in **E)**, mouse and in **F)**, naked mole-rat, restricted to the subsets of proteins identified with significant difference of aggregation propensity (n=510 proteins). Kolmogorov–Smirnov (KS) test is used to assess differences in distributions of proportion of beneficial mutation between the two subsets in each organism.

with beneficial mutations are not the same in naked mole-rat and in mouse. Furthermore, in both species (Figure 2.7E,F), we find a significant difference in the distribution of these proportions between higher and lower aggregation subsets (p-value=$1.19 * 10^{-12}$ for mouse, p-value=$1.07 * 10^{-14}$, Kolmogorov–Smirnov test), with proteins with high aggregation propensity in a given species having a higher potential for accumulating beneficial mutations compared to proteins with low aggregation propensity in the respective species. Similar trends have been previously observed for protein stability (Serohijos et al., 2012).

2.3.5. Evolutionary changes specific to naked mole-rat influence local differences in aggregation propensity in ATX proteins

Among the proteins we identified with significant difference of aggregation propensity between the species, several of them have been reported to be associated with human age-related diseases. Specifically, Ataxin-10 (ATX10) and Ataxin-3 (ATX3) are both responsible for different forms of spinocerebellar ataxia in humans, a type of neurodegenerative disease. Both proteins have a lower whole-protein sequence aggregation propensity in naked mole-rat, compared to mouse (ATX-10: 6.63 (naked mole-rat) $<$ 8.19 (mouse) ; ATX-3: 2.40 (naked mole-rat) $<$ 6.16 (mouse)). We investigated the origin of these differences of aggregation propensity in these ATX proteins by comparing the distribution of their per-residue aggregation propensity scores and mutational aggregation propensity score between the two species along the aligned sequences (Figure 2.8).

In ATX-10 (Figure 2.8A,B,C) and in ATX-3 (Figure 2.8D,E,F), regions with high peaks of aggregation propensity (Agg $>$ 50), are co-localized with negative mutational aggregation hotspots in both species whereas residues with positive mutational aggregation regions are found in low aggregation propensity regions (Agg 0), with fewer random mutations resulting in increased aggregation (detrimental mutations) than resulting in decreased aggregation (beneficial mutations). This is in line with the idea that regions of high aggregation propensity would benefit more from mutations than low aggregation propensity regions would be affected by mutations, suggesting that the low aggregation propensity regions are possibly in an optimal state favoring protein robustness. In both proteins, we observed insertions in naked mole-rat (dashed line, Figure 2.8A,D), which are not found in any other rodent species (blast e-value $<$0.001, see Methods). Interestingly, in both cases, these insertions are located within a mouse aggregation propensity peak (positions 416-474 of the ATX10 alignment and positions 106-140 of the ATX3 alignment). Overall, naked mole-rat and mouse ATX-10 (Figure 2.8A, in orange and grey respectively) have similar aggregation propensity profiles, meaning that the residues with high and low aggregation propensity are located in the same regions, but with mouse having higher values between position 200 and 300. In this

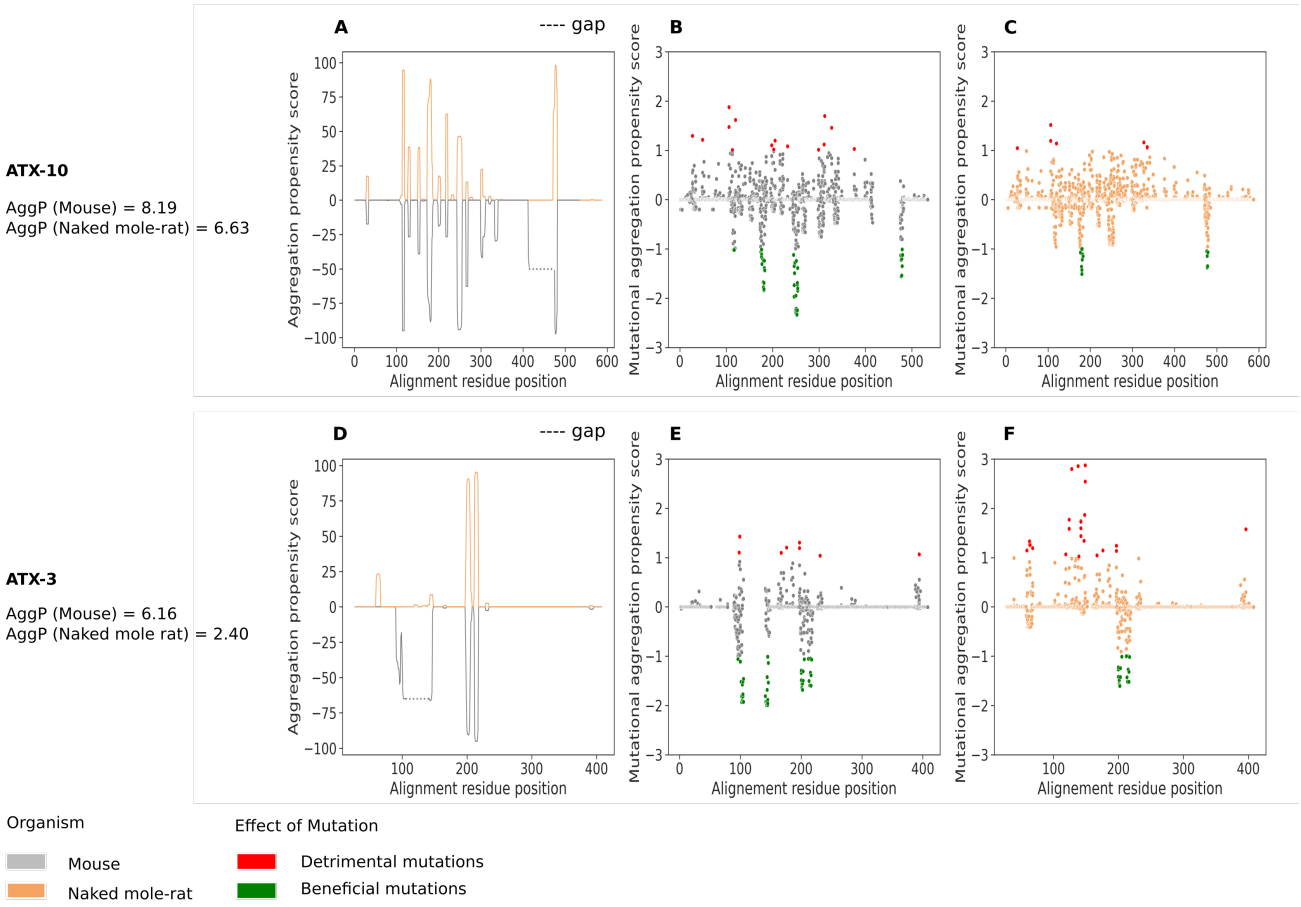

Fig. 2.8. Aggregation propensity and mutational aggregation propensity profiles of ATX proteins

Aggregation propensity profile along the positions from the sequence alignment of the two rodent protein sequences for **A)**, ATX-10 and **D)**, ATX-3 proteins. Mouse profile is in in grey and naked mole-rat profile is in orange. We plot the opposite of the mouse aggregation value to facilitate comparison. Regions with gap are represented with a dashed line. Mutational aggregation profile of mouse and naked mole-rat for (B,C) ATX-10 and (E,F) ATX-3 proteins. Beneficial and detrimental mutations are annotated in red (detrimental, Mutational $Agg_P > 1$) and green (beneficial, Mutational $Agg_P <$ -1), respectively.

region, we indeed observe a hotspot of random mutations beneficial in mouse only (Figure 2.8B, green coloring), indicating that this region has the potential to acquire lower aggregation propensity in mouse, whereas in naked mole-rat (Figure 2.8C), the region's aggregation propensity is not significantly improved by mutations. In the case of ATX-3 (Figure 2.8D), naked mole-rat and mouse have distinct aggregation propensity profiles before amino acid 150, with naked mole-rat having lower aggregation propensity compared to mouse in the region of the naked mole-rat specific insertion. In mouse only, we observed hotspots of beneficial mutations on both side of the insertion (Figure 2.8E, green coloring), and in naked

mole-rat, random mutations within the insertion tend to significantly increase aggregation propensity (Figure 2.8F, red coloring). This result indicates that the insertion is likely the determining factor of the decrease in aggregation propensity in ATX3 in naked mole-rat compared to mouse. This region is already stabilized for mutations. Interestingly, this insertion is located in the functional domain Josephin, that is known to contribute to ubiquitin chain binding and cleavage for ATX-3 (Nicastro et al., 2009). In the case of the ATX10 insertion, which split a PF09759 domain seen in mouse sequence into two sub-domains in the naked mole-rat sequence, we did not observe detrimental mutations in the insertion in naked mole-rat, and both species show an increase in aggregation propensity at the insertion breakpoint.

2.4. Discussion

Aggregation propensity and mutation tolerance are two intrinsic properties of proteins that could contribute to the better maintenance of protein homeostasis. In this study, we designed a computational strategy to estimate these properties at the scale of the whole-proteome in naked mole-rat and mouse using a comparative genomic framework. Among their orthologous proteome (n=9,522 proteins), we did not identify global differences in aggregation propensity, but about 1,000 proteins showed significant differences from their domains or their whole-protein sequences. Our analyses specifically study chaperone client proteins to determine whether this subset has differing intrinsic properties but did not find significant differences. Previous studies have shown that chaperone client proteins evolve slower and have a lower aggregation propensity compared to non-client proteins (Victor et al., 2020). Still, our study shows that these properties remain similar between naked mole-rat and mouse. As for caveats, we inferred the naked mole-rat and mouse chaperone clients from human orthologs based on information reported in the BioGRID database. The data from this database does not necessarily indicate actual chaperone dependence. Therefore, it is possible that the subset of proteins we defined as chaperone client proteins is highly incomplete or does not interact with chaperones in naked mole-rat and/or in mice. Moreover, we do not specify which specific chaperones were interacting with those chaperone client proteins, which may also bias the results.

From the gene-enrichment analysis, we observed that the proteins of naked mole-rat with less aggregation propensity are over-represented mainly in the extracellular compartments, within several specific biological processes related to immune response and lipid metabolism, and have functions associated with binding and protein degradation. The proteins with more aggregation propensity are not enriched in a particular biological process, except in the inflammasome complex, known to contain aggresomal complexes. Among the proteins

we identified with significant differences in aggregation propensity, we identified several previously known proteins in neurodegenerative and age-related diseases. For instance, ATX3 is a poly-glutamine tract-containing protein that contributes to cytoskeleton organization and is involved in protein inclusion bodies (Burnett and Pittman, 2005). The accumulation of ATX3 in brain cells causes a proteostasis impairment that leads to the Machado-Joseph disease, or spinocerebellar ataxia-3 (Dantuma et al., 2020). Particularly, ATX3 is associated with double-stranded DNA binding. Previously, the study of ATX3-mutant in mouse brain cells showed an impairment of DNA repair efficiency, leading to the accumulation of DNA damage (Gao et al., 2015). ATAX10 was also identified here, which is associated with pentanucleotide disorder SCA10 (Bampi et al., 2017). Identifying lower aggregation propensity in these poly-glutamine proteins in naked mole-rat could contribute to resistance towards certain types of neurodegenerative diseases, leading to premature death (Dantuma et al., 2020). Moreover, we also identified proteins related to lipid metabolism with lower aggregation propensity in naked mole-rat, such as the acid ceramidase ASAH1. This protein is involved in the intra-lysosomal ceramide homeostasis and is known to be associated with Alzheimer's disease, cancer, and diabetes (Parveen et al., 2019). Furthermore, a recent study highlighted specific lipidic signatures in naked mole-rat that confer neuroprotective mechanisms against oxidative damage (Frankel et al., 2020). The lower aggregation propensity of the lipid metabolism proteins may contribute to protein stability and discharge of quality control systems of proteostasis.

Our study also highlighted the versatility of the aggregation propensity within inflammation pathways in naked mole-rats. Indeed, these rodents have a unique immune system able to better resist bacterial infection. They have a unique myeloid cell subset that highly expressed genes for the antimicrobial response (Hilton et al., 2019). Genes involved in the NOD-like receptor signaling pathway can activate pyroptosis, which is cell death after exposure to a bacterial infection. Interestingly, the NLRP-3 inflammasome pathway, which we found to have a higher aggregation propensity at the level of protein domains in our study, is known to be regulated by the ubiquitin system. However, the exact molecular mechanisms of its non-canonical activation remain unclear (Lopez-Castejon 2020). The increase of domain aggregation propensity within proteins associated with the inflammasome complex might explain their affinity with the ubiquitin system, however, we note that this result could be explained by specific domains (eg. peptidase C14) overrepresented in proteins involved in the formation of inflammasome complexes. Moreover, naked mole-rat's immune system is more frequently solicited during bacterial infection than in the mouse (Cheng et al., 2017). Our study observed that proteins with chemokine and cytokine activity have significantly lower aggregation propensity. This suggests that the intrinsic

properties of these naked mole-rat proteins adapt to be less prone to aggregate.

We also identified several metalloproteases having domains with lower aggregation propensity in naked mole-rat. Metalloproteases are known to degrade extracellular matrix proteins. Interestingly, the naked mole-rats highly produce the high-molecular-mass hyaluronan (Tian et al., 2013), a component of the extracellular matrix, known to have anti-inflammatory properties (Takasugi et al., 2020). These proteins might facilitate the hyaluronan turnover and balance the pro-inflammatory responses from the high activity of the inflammasome. Recently, two studies highlighted the importance of MMP13 as a therapeutic target for Alzheimer's and Parkinson's disease (Sánchez and Maguire-Zeiss, 2020; Zhu et al., 2019). Tight regulation of inflammatory responses in naked mole-rat seems essential to maintain protein homeostasis, particularly during bacterial infection. Naked mole-rats are known to maintain proteasomal proteolytic activities in their late stages of life (Pérez et al., 2009). These adaptations could indirectly promote healthy aging in naked mole-rat, increasing its maximum lifespan.

Mutation tolerance is another intrinsic property of proteins that could contribute to maintaining protein homeostasis. It indicates the ability of the protein to maintain its aggregation propensity despite mutations. We used the difference of aggregation propensity between mutated and wild-type sequences to estimate whether a substitution event in the coding sequence would later drastically change or not the aggregation propensity of a protein. In the definition of our mutation tolerance score, synonymous substitutions favor protein stability and avoid the formation of protein aggregates. Despite no global differences in mutation tolerance between the two species' proteomes, proteins with lower aggregation propensity in naked mole-rat better tolerate mutation than proteins with higher aggregation propensity. Such a difference is not seen in the mouse, which suggests these proteins in naked mole-rat have intrinsic properties that slow down the overload of the quality control systems of proteostasis, thus might contribute to its longevity.

We further studied ATX-10 and ATX-3, from ATX-family, known to be associated with the neurodegenerative disease, spinocerebellar ataxia in humans. We were able to highlight evolutionary events that occurred only naked mole-rat sequences and that helped to improve the aggregation propensity in that species. We observed peaks of aggregation propensity closed to insertion breakpoints in the two ATX proteins. Notably, for ATX-3, we observed that an insertion in a functional domain decreases the aggregation propensity profile of the naked mole-rat compared the mouse sequence, suggesting that this event contributes to the stability of the protein. The combined observation of the aggregation propensity and mutational aggregation profiles of these specific regions thus informs on

the consequences of evolutionary events such as insertions on aggregation propensity and protein stability. These regions are candidates that could be further studied for optimized protein design toward stability. It would be also interesting to see if other proteins with significant differences in aggregation propensity might also contain insertion events specific to naked mole-rat. Another future study could try to identify if the specific insertions in naked mole-rat are systematically co-localized in with aggregation propensity, as well as in other species. These analyses are of course not exhaustive but are good examples of the potential use of our different metrics to perform comparative analyses of proteins that could explain the difference in protein stability between the two species, with possible implications for differences in lifespan.

Studying the diversity of lifespan within eukaryotes with comparative genomic approaches requires well-curated genome assemblies and reliable maximum lifespan measurements. In this study, we restricted our analysis to two species from the same taxonomic order, with a drastic difference of maximum lifespans, to identify the proteomic features explaining their lifespan difference. Working with closed-related species helps to identify subsets of proteins associated explicitly with biological processes related to longevity in the two species, without taking account of the complications arising from comparing from evolutionary-distant species. Although these results could be specific to rodents, the pathways and genes identified in this study are known to be shared across eukaryotes. Therefore, our study is a step towards a more extensive investigation of these properties across species. In addition to restricting the comparative analysis to only two species, our study has several limitations. First, we only focused on orthologous proteins shared between naked mole-rat and mouse, ignoring proteins unique to naked mole-rat, which could also contribute to its extended longevity. Second, to predict the aggregation propensity of the proteins shared between naked mole-rat and mouse, we used the Tango software, which is a predictive approach that heavily relies on the physicochemical properties of the amino acid sequences and their likelihood to be involved in the formation of beta-sheets structures participating in functional folding. This approach performs well to predict the aggregation propensity of globular proteins (Linding et al., 2004), which resulted in the exclusion of transmembrane and membrane proteins from our analyses. Moreover, the aggregation propensity scores are predicted for a given set of environmental parameters. They may not represent the dynamic range of aggregation propensity scores that the proteins could adopt in different tissues.

Alternative bioinformatics methods to estimate aggregation propensity based on amino-acid sequences are implemented as web server tools (Santos et al., 2020), incompatible with our high-throughput computational strategy for estimating mutation tolerance by generating

billions of sequences that could only be processed promptly using a command-line software. Therefore, Tango allowed us to build a systematic and highly efficient pipeline to estimate the aggregation propensity of 10,000 proteins in two different organisms. This large-scale experiment is unfeasible to achieve in vitro. However, further molecular investigations will be necessary to validate the role of the identified less aggregation-prone proteins in naked mole-rat in the context of aging. Finally, to validate whether the patterns we identified regarding aggregation propensity and mutation tolerance are not only specific to the comparison of naked mole-rat to mouse, these patterns will need to be more systematically confirmed by comparing long-lived versus short-lived rodents. The challenge of this strategy will be to properly define the long-lived and short-lived groups and verify if the phylogenetic relationships in each group are equally distributed. The use of longevity quotient (Austad and Fischer 1991), which indicates whether a species has an average lifespan or is unusually long- or short-lived relative to its body size, could be used to distinct the groups with extreme longevity.

In conclusion, we investigated the peculiarity of naked mole-rat longevity by studying specific intrinsic properties of the proteome that influence the maintenance of proteostasis. Our study highlighted a trade-off in the regulation of inflammation responses in naked mole-rat, directly encoded in the amino acid composition of the proteins as it relates to its propensity to aggregation. We also identified several proteins with lower aggregation propensity compared to the mouse that has been found to characterize neurodegenerative or age-related diseases in humans. Our findings propose the existence of a successful strategy encoded in the naked mole-rat proteome architecture to delay aging through better maintenance of protein homeostasis in the longest-lived rodent.

Availability of Data and Materials

The processed data and code used to generate the figures are available in the following Github repository: **ladyson1806:NKR_lifespan**. We also provide the different Python3 scripts and notebooks used to collect and pre-process the initial dataset, as well as the code that generates the different scores.

Acknowledgments

This work was supported by funds from the Department of Biochemistry and Molecular Medicine of Université de Montréal and through the access to computational resources provided by Calcul Québec to JGH. We thank Sebastian Pechmann for his supervision on the preliminary analyses. We are grateful to Adrian Serohijos for his mentoring support and the fruitful discussions throughout the project. Finally, we thank all the members of the Hussin

lab for their constructive comments and feedback on the figures for the manuscript. JGH is a Fonds de la Recherche du Québec en Santé (FRQS) Junior 1 Scholar, funded by the Institute for Data Valorization (IVADO).

Deuxième article.

piQTL : Cartographie de QTL par les interactions protéine-protéine pour identifier des déterminants fonctionnels de phénotypes complexes chez la levure

par

Savandara Besse[1,2,3,#], Tatsuya Sakaguchi[1,2,#], Louis Gauthier[1,2], Zahra Safar[1,2], Lidice Gonzales[1,2], Xavier Castellanos-Girouard[1,2], Chloé Matta[1,2], Julie Hussin[3,4], Stephen Michnick[1,2,*] et Adrian Serohijos[1,2,*]

([1]) Département de Biochimie et Médecine Moléculaire, Université de Montréal - Montréal, Québec, Canada

([2]) Centre Robert-Cedergen en Bio-Informatique et Génomique, Université de Montréal - Montréal, Québec, Canada

([3]) Institut de Cardiologie de Montréal - Montréal, Québec, Canada

([4]) Département de Médecine, Faculté de Médecine, Université de Montréal - Montréal, Québec, Canada

État du manuscrit : **En préparation**.

Correspondance:
adrian.serohijos@umontreal.ca
stephen.michnick@umontreal.ca

Résumé. Un objectif central de la génétique est d'expliquer comment les traits phénotypiques sont déterminés par la variation génétique et ont des répercussions sur presque tous les aspects de la biologie et de la médecine. Les études d'association à l'échelle du génome (GWAS) et la cartographie des loci de traits quantitatifs d'expression (eQTL) suggérent que les signaux d'association ont tendance à se propager à travers le génome et incluent de nombreux génes sans lien évident avec le phénotype ou la maladie. Trouver les déterminants génétiques responsables de traits complexes reste un grand défi. Ici, en utilisant 354 souches de levures consanguines contenant environ 12,000 SNPs et 61 paires d'interactions protéine-protéine, nous avons développé une cartographie des loci de traits quantitatifs d'interaction protéique (ou piQTL *in vivo*) qui disséque les facteurs causaux et fonctionnels des phénotypes complexes, ici la croissance de la levure dans différents conditions environnementales : sous anti-fongiques (5-Fluorocystosine et fluconazole), sous metformine, un médicament anti-diabétique, et trifluopérazine, un composé anti-psychotique. Nous avons constaté que les piQTL sont répartis dans tout le génome, y compris les régions codantes pour les protéines, leurs promoteurs et dans les 3'UTR. Étonnamment, les piQTLs comprennent également une fraction importante de régions intergéniques et d'ARN non codants, en particulier ceux classés comme SUT (transcrits stables non définis). Dans l'ensemble, cette étude démontre que les conséquences cellulaires ultimes de la relation génotype-phénotype se reflétent dans l'abondance des protéines à l'état d'équilibre, les modifications post-traductionnelles et la localisation sub-cellulaire des protéines et des interactions protéine-protéine (PPI). Cette étude fournit également une carte de route pour un nouveau type de cartographie QTL basée sur des réseaux PPI qui révélent les déterminants fonctionnels et biochimiques des phénotypes complexes.

Mots clés : Levure, Génétique des populations, Cartographie QTL, Phénotypes complexes, Interactions protéine-protéine, Réseaux biologiques

Abstract. Protein-protein interactions (PPI) accurately map environmental perturbations to molecular consequences in the cell, but how PPIs are modulated by genomic variation is unknown. If genome variation also causes PPI perturbations then, in principle, differences in the PPI network integrate both genomic and environmental effects (G×E) and probing PPIs could aid in identifying biochemical mechanisms that result in complex traits. Here, using 354 inbred strains of the yeast *Saccharomyces cerevisiae* having ∼ 12,000 sequence variations in their genomes, we report a "protein-interaction quantitative trait loci" (piQTL) mapping analysis, linking genetic variation in the strains to a network of 61 in vivo PPIs centered around the ergosterol synthesis pathway. We determined piQTLs in the yeast strains treated by fluconazole, which inhibit ergosterol synthesis, and other drugs whose known mode of actions are outside the 61 probed PPIs, the antifungal 5-fluorocytosine, the anti-diabetic metformin and the antipsychotic trifluoperazine. piQTLs were found throughout diverse genomic elements, including protein coding and transcriptional regulator regions (promoters and 3' UTRs). A significant proportion of piQTLs (∼ 5%) include long non-coding RNAs, especially those classified as stable undefined transcripts (SUTs). There are many piQTLs close to the loci of probed PPI partner proteins (cis-piQTL) and trans-piQTLs. Although, trans-piQTLs are broadly distributed in the genome, they are enriched in genes encoding protein components of biochemical pathways probed by the PPIs. In contrast, piQTLs are depleted in functional modules defined from genetic interactions due to pairwise knockouts, highlighting the different effects of gene deletion versus genetic variation on PPI. piQTL identifies known mechanisms of antifungal resistance and unforeseen effects of metformin on heavy metal homeostasis that may explain its pleotropic effects on cellular metabolism. Altogether, our study shows how piQTL can be applied to identify the biochemical drivers of complex phenotypes.
Keywords: Yeast, Population Genetics, piQTL mapping, Complex phenotypes, Protein-protein Interaction, Biological networks

Contributions personnelles à ce chapitre

Mes contributions pour cet article en préparation sont les suivantes, en tant que première co-auteure :

- Discussion et validation des hypothèses établies par Professeur Adrian Serohijos.
- Prise en charge de la conception de la stratégie bio-informatique en collaboration avec Professeur Adrian Serohijos.
- Implémentation de toutes les méthologies bio-informatiques présentées dans ce chapitre.
- Traitement des données de séquençage issues de la stratégie expérimentale incluant :
  - Prise en charge et nettoyage des données génotypiques provenant de l'étude GWAS par She and Jarosz (She and Jarosz, 2018)
  - Transposition de la cartographie architecturale des SNPs sur l'ensemble du génome de la levure via l'intégration des annotations génomiques disponibles sur la base de données SGD (pour *Saccharomyces Genome Database*) et des

annotations génomiques provenant de l'étude de Rossi et al. (Rossi et al., 2021)
    - Pré-traitement automatique et reproductible des fichiers FASTQ après contrôle de leur qualité (retraits des amorces et fusion des lectures avant l'étape d'extraction des codes-barres)
    - Adaptation et optimisation de la stratégie bio-informatique pour l'extraction des codes-barres au sein de la population de levures incluant
        * la cartographie des codes-barres extraits à leurs souches correspondantes
        * le calcul de *fitness* relatif de chacune de ces souches pour toutes les conditions expérimentales (361 souches * 62 PPIs * 5 médicaments * 2 environnement de sélection)
- Implémentation du calcul permettant la quantification des PPIs.
- Adaptation et optimisation des algorithmes utilisés pour réaliser l'étape cartographie fine entre PPIs et variants génétiques.
- Définition des critères d'exclusion des variants génétiques non-spécifiques à la réponse des différents médicaments.
- Analyses bio-informatiques post-piQTL
    - Études de co-localisation entre piQTL et GWAS chez la levure
    - Croisement des résultats de piQTL et GWAS chez l'humain pour des maladies complexes (diabète de type II)
    - Étude de la relation entre les mesures statistiques associées aux piQTLs et la topologie de différents réseaux biologiques étudiées dans le chapitre.
- Création d'une interface utilisateur pour la visualisation interactive des graphiques reliés aux analyses des piQTLs (Manhatthan et QQ plots + navigateur de génome) en R/Shiny
- Rédaction de toutes les sections de l'article sur la méthodologie bio-informatique en collaboration avec Professeur Adrian Serohijos (co-auteur).
- Conception et réalisation des figures principales de l'article ainsi que de toutes les figures supplémentaires à l'exception des figures 3.6 et 3.7 . Cela a été fait en collaboration avec Professeur Adrian Serohijos (co-auteur).
- Relectures et révisions du reste de l'article (Introduction, Méthodes expérimentales, Résultats, Discussion)

L'ensemble de mes contributions bio-informatiques de ce projet sont disponibles sur le répertoire Github suivant : https://github.com/ladyson1806/piQTL_mapping

Ce travail n'aurait pu être réalisé sans l'aide de :

- Professeur Adrian Serohijos (co-auteur), qui a supervisé l'ensemble du projet, de la conception des stratégies expérimentales et bio-informatiques, à l'analyse des résultats préliminaires et finaux. Professeur Adrian Serohijos a participé à l'écriture de l'introduction, des résultats, et de la discussion de l'article. Il a conçu l'ensemble des figures présentes dans l'article et réalisé la Figure 3.1.
- Professeur Stephen Michnick (co-auteur), qui a également supervisé l'ensemble du projet. Professeur Stephen Michnick a participé activement à l'analyse et l'interprétation des résultats, ainsi qu'à la rédaction des résultats de l'article. Il nous a également fourni de précieux conseils et suggestions au niveau de la conception des stratégies expérimentales.
- Professeure Julie Hussin (co-auteure), dont l'expertise en génétique des populations a été grandement sollicitée. En particulier, Professeure Julie Hussin nous a fourni de précieux conseils et ainsi que des suggestions dans les étapes de vérification des données génotypiques, dans l'implémentation de différents modèles pour les tests d'association entre la variance des génotypes et la variance de la quantification des PPIs, ainsi que dans l'interprétation des analyses post-piQTL.
- M. Tatsuya Sakaguchi (premier co-auteur) qui a pris en charge l'ensemble de la stratégie expérimentale relative à ce chapitre, incluant le choix des 62 PPIs à étudier dans l'itération finale de l'ensemble de la stratégie piQTL pour la démonstration de concept. M. Tatsuya Sakaguchi a également participé activement à l'analyse des résultats, à la conception et à la réalisation des figures suivantes : Figure 3.3, Figure Supplémentaire 3.7, Figure Supplémentaire 3.8. Il m'a également apporté son aide dans certaines analyses bio-informatiques, incluant les analyses d'enrichissement des protéines impliqués dans les 62 PPIs, le choix final du calcul de *fitness* relatif et la création de visualisation du réseaux d'interactions protéine-protéine avec les 62 PPIs.
- M. Louis Gauthier (co-auteur) et de Mme. Zahra Safar (co-auteure), qui ont tous deux initié les analyses préliminaires expérimentales relatives à ce chapitre. M. Louis Gauthier est également à l'origine de l'initiation du projet en collaboration avec Professeur Adrian Serohijos. Il a contribué à l'élaboration des hypothèses de départ, à la construction de la librairie des barres-codes associées à chaque souche de la population, ainsi que le design expérimental des constructions génétiques permettant l'étiquetage des protéines impliquées dans chaque PPI par le système rapporteur. Il m'a également suggéré des approches bio-informatiques à tester lors de ma phase de familiarisation au projet. Mme. Zahra Safar a aidé M. Louis Gauthier dans la mise en place et l'exécution des protocoles expérimentaux nécessaires à la quantification des PPIs lors des étapes préliminaires du projet.
- Mme Lidice Gonzales (co-auteure), pour son aide sur la réalisation des expériences présentées dans la Figure 3.4.

- M. Xavier Castellanos-Girouard (co-auteur), dont j'ai sollicité à plusieurs reprises l'expertise afin de mieux comprendre les stratégies expérimentales du projet. Il a également activement participé à l'interprétation des résultats et proposé des hypothèses fonctionnelles pour certains piQTLs.
- Mme Chloé Matta (co-auteure), pour son aide sur les analyses réseaux présentées dans le manuscrit.

Chapitre 3

---

piQTL: Protein-interaction QTL to dissect the functional drivers of complex phenotypes in yeast

3.1. Introduction

A central question in genetics is how genomic sequences influence complex traits and phenotypes. The overall goal is to trace the functional and mechanistic consequences of genomic variations to the various levels of cellular processes. Genome-wide association studies (GWAS) suggest that genomic loci with significant statistical associations to the variation of complex or quantitative phenotypes (quantitative trait loci or QTLs) are spread across the genome. These include many genes without an obvious connection to the phenotype (Eichler et al., 2010; Manolio et al., 2009). There are now over 400,000 unique associations spanning over 5,000 traits in humans (**watanabe_global_2019**; Sollis et al., 2023). However, despite advances, assigning function and causality to the resulting QTLs remains a challenge. Advances in transcriptomics have led to using gene expression as a proxy molecular trait and a functional intermediate between genomic sequence and phenotype, thus identifying genomic loci that regulate the mRNA levels or "expression QTLs" (eQTLs). Several types of association mapping strategies based on other genomic molecular readouts - such as chromatin accessibility (caQTL), DNA methylation (meQTL), transcription factor binding (bQTL) by ChIP-Seq, and metabolomics (mQTL) - all aim to bridge the functional consequences of genetic variation to cellular mechanisms and eventually phenotype.

Recent studies suggest that the ultimate cellular consequences of the genotype-phenotype-environment relationship (G×E) are reflected in the steady-state levels, post-translational modifications, and subcellular localization of proteins and the other

proteins they interact with (Hein et al., 2015; Messner et al., 2023; Skinnider et al., 2021). This ability of protein abundance, and potentially PPI, to be accurate reporters of G×E may not be reflected by eQTLs, since gene expression is only weakly correlated with protein abundance (Greenbaum et al., 2003; Wainberg, 2019). Indeed, only $\sim 40\%$ of the variation in protein abundance can be explained by mRNA levels (Buccitelli and Selbach, 2020), highlighting a significant limitation in bridging the G×E relationship simply by looking at the transcriptomic state of the cell. Recent efforts to evaluate the genotype-phenotype relationship at the level of protein abundances (or pQTL) showed that direct measurements of these link genetic variation to molecular mechanisms underlying phenotypes than either eQTL based on whole or translated mRNA by ribosome profiling (Romanov et al., 2019).

Moreover, it has been demonstrated that beyond protein abundances, probing interactions of proteins prove to be accurate reporters of the overall state of the cell across multiple environments (Hein et al., 2015; Skinnider et al., 2021; Stynen et al., 2018). PPIs likewise integrate diverse cellular processes that regulate or control the fate of genes and their protein products (Hein et al., 2015; Skinnider et al., 2021; Stynen et al., 2018). Indeed, the dynamic association, binding, and co-localization of molecules, including proteins, in the cell is the standard operational definition of "function". Here, we report the development of an experimental and computational strategy, which we call piQTL mapping, to correlate genetic variations to in vivo PPI using the budding yeast *S. cerevisiae* as a model system. We use piQTL mapping to probe G×E relationships, first with drugs that target known pathways in yeast, the two antifungals, fluconazole and 5-fluorocytosine and second, based on the conservation of some biochemical pathways between yeast and humans despite $\sim$ 1 billion years of divergence (Kachroo et al., 2015), we also determined the piQTLs when yeast is treated with drugs for human diseases, the type II diabetes drug metformin and the anti-psychotic trifluoperazine.

3.2. Results

3.2.1. Pooled quantification of in vivo PPI in an inbred yeast cohort

The statistical power of GWAS and any QTL mapping method depends on the amount of genetic diversity in the sample population and the recombination rate of a species to break linkage disequilibrium (LD). We wanted to use a model that minimizes the co-segregation of multiple high-frequency alleles due to LD, thus we used a collection of 353 haploid yeast strains used in a previous yeast GWAS study that resulted from the inbreeding of two divergent strains for 6 generations (Figure 3.1a) (She and Jarosz, 2018). These strains have

been fully sequenced (genotype map in Extended Data Figure 3.5a). The 6 generations of inbreeding resulted in the reduction of LD among these strains (pairwise LD score between SNPs in Figure 3.1d). The population contains $\sim$ 12K high-frequency SNPs spread across the 16 chromosomes and mitochondrial genome (Extended Data Figure 3.5a). There is at least 1 SNP every $\sim$ 1 kilobase; thus, genetic diversity covers coding DNA sequences (CDS) and their promoter and regulatory structure (Extended Data Figure 3.5b). In addition, these SNPs also cover almost all the essential features of the yeast genomic architecture, including transcribed non-coding RNA transcripts, which are a significant subject area in eQTL mapping (Goede et al., 2021), and non-transcribed yet functional regions of the genome (Extended Data Figure 3.5b). Altogether, because of the density and distribution of these SNPs, we will be able to perform association mapping and determine the consequences of diverse architectural features of the genome on the dynamics of PPI networks, including genomic loci which are not coding for proteins or even transcribed, but potentially contribute to the phenotype. Based on power calculations (data not shown), our sample size of 354 inbred strain strains is well-above the statistical requirement to perform QTL studies. In comparison, the early development of eQTL was also performed in cross-bred strains (e.g., genome-wide eQTL of $\sim$ 100 strains of the Hybrid Mouse Diversity Panel (Bennett et al., 2010) and 205 inbred lines of Drosophila (Cannavò et al., 2017; W. Huang et al., 2014)).

To measure the strength of PPI in vivo, we used the protein-fragment complementation assay (PCA) based on the reporter enzyme methotrexate (MTX)-resistant murine dihydrofolate reductase (mrDHFR) (Figure 3.1c, ii) (Stynen et al., 2018; Tarassov et al., 2008). In each of the 354 inbred strains, we integrated complementary N- and C-terminal fragments encoding for mrDHFR into the 3' ends of the ORFs coding for the interacting bait-prey proteins. When the bait and prey interact, they bring the complementary mrDHFR fragments into proximity ($\sim$ 10 nm) to fold into an active enzyme, which confers growth in cells in which the endogenous DHFR is inhibited by MTX13. Several features of mrDHFR PCA make it an ideal in vivo reporter of PPI for piQTL mapping. 1, The assay provides a direct readout of the number of protein complexes formed between bait and prey proteins; 2, It is sensitive enough to detect as few as 25 complexes per cell. Thus, genes can be expressed under their native promoters, assuring natural expression levels at appropriate phases of the cell cycle and under conditions that genes are typically expressed; 3, proteins are expressed in appropriate cellular compartments with correct post-translational modifications; 4, importantly, mrDHFR PCA is fully reversible and does not alter normal equilibrium or kinetics of association of proteins, as demonstrated in vivo and in vitro23; 5, mrDHFR PCA is scalable, important for association mapping where the same PPI needs to be measured in many strains with different genomic backgrounds. mrDHFR PCA was applied to determine the first and only in vivo protein interactome of *S. cerevisiae* (Tarassov et al., 2008) and

to map environmental perturbation by drugs to known and novel biochemical pathways (Messier et al., 2013; S. Michnick et al., 2007; Stynen et al., 2018; Tarassov et al., 2008). It has not, however, been used to measure the effects of gene variation on cellular processes in strains with different genomic backgrounds.

We measured the PPI of 61 bait-prey pairs that cover 44 unique genes chosen based on their diverse coverage of molecular function, biological processes, and cellular localization and which have been validated in a previous proteome-wide mrDHFR PCA screen in yeast (Tarassov et al., 2008) (Table S1, GO enrichment, available on request). The phenotypes that we investigated include yeast's responses to two antifungals, fluconazole and 5-fluorocytosine (5-FC). PPIs of fluconazole's target, the lanosterol 14-alpha-demethylase (Erg11), a key enzyme in the ergosterol synthetic pathway, were specifically included in our set to provide direct reporters of the effect of the drug and, simultaneously, of genome variation on the ergosterol synthetic pathway (Figure 3.1b). This allows for a direct measure of the contributions of G×E variation to a known pathway. 5-FC is a prodrug that is activated when metabolized via the pyrimidine salvage pathway, where it acts as a subversive substrate with the subsequent production of toxic nucleotides and disruption of DNA and protein synthesis. Despite well-known mechanisms, the mode of action and the mechanism for the emergence of antifungal resistance are pleiotropic (Jakobson and Jarosz, 2019; She and Jarosz, 2018), and its consequences on our 61 PPI proves are not fully known. We also tested two human drugs:, metformin, which is a first-line treatment for type 2 diabetes, and trifluoperazine, which is an antipsychotic used to treat schizophrenia. In the case of metformin, no specific target is known in humans or yeast so no specific PPIs can be probed, however in a previous mrDHFR PCA screen in yeast, we showed that both known and novel processes that metformin act on could be identified by perturbations of specific PPIs by metformin (Stynen et al., 2018). Finally, trifluoperazine is known to block dopaminergic receptors in humans and calmodulins in yeast. In the cases of the later three drugs, we can think of the 61 PPIs as "antennas" that, due the small-world properties of PPI networks (Goldberg and Roth, 2003), probe changes in the biochemical network of the cell. In the case of genomic variation, we hypothesize that genes whose variation result in changes to the PPI strengths in our sub-network will be mechanistically related to the actions of the drugs.

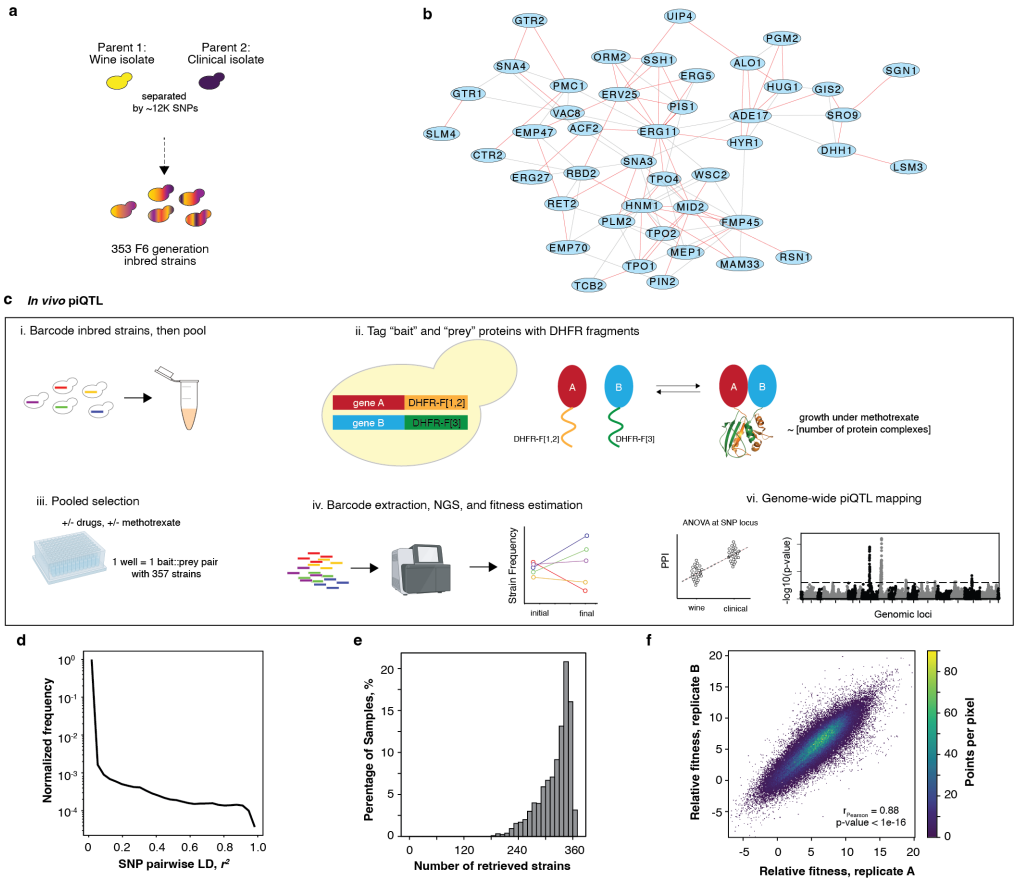

Fig. 3.1. piQTL mapping in an inbred yeast strain cohort.

a, A collection of 353 yeast strains from F6 generation inbreeding of wine and clinical isolates separated by ∼ 12K SNPs. **b**, We measured the protein interaction of 61 bait-prey pairs that span 44 unique genes. These were chosen based on their coverage of diverse functional categories biochemical and metabolic pathways 3.14. All edges are experimentally validated PPI from a previous genome wide PCA screen (Tarassov et al., 2008), in red are edges in this piQTL study. **c**, Pooled screening strategy for *in vivo* piQTL mapping. **i**, Strains are individually barcoded in a neutral chromosomal locus by CRISPR/Cas9 and homologous recombination and then pooled (Extended Data Figure 3.6). **ii**, Step-wise editing on the pooled strains introduce the complementary mouse DHFR N-terminal fragment (or DHFR[1,2]) and C-terminal-fragment (or DHFR[3]) into the C-terminal region of the ORF corresponding to the two bait::prey proteins (Extended Data Figure 3.7). Reconstitution of DHFR confers resistance against methotrexate (MTX). Strength of protein-protein interaction is proportional to growth rate under MTX (Stynen et al., 2018; Tarassov et al., 2008). **iii**, We propagated the pooled 357 strains (353 inbred strains labeled with DHFR fragments and another 4 strains unlabeled without PPI to serve as reference). Selection in +/- MTX as well as under drug perturbations, two antifungals fluconazole and 5-fluorocytosine (5-FC) and two human drugs, anti-diabetes metformin and anti-psychotic trifluoperazine. **iv**, Genomic extraction and PCR amplification of the barcode followed by NGS. Fitness of each strain and, consequently, the magnitude of PPIs, are estimated from changes in barcode frequencies. **vi**, Genome-wide piQTL mapping by performing linear regression (ANOVA) between PPI and genetic composition at each SNP locus. **d**, Distribution pairwise linkage disequilibrium (LD) score among the SNPs. **e**, Number of retrieved strains from barcode mapping within all the samples. **f**, Comparison of relative fitness between the biological replicates.

We had to measure ∼ 350,000 PPIs (61 PPI x 354 strains x 4 drugs x 2 timepoints x 2 replicates) and thus the scalability of mrDHFR PCA was important to our devising a

"pooled" screening strategy (Figure 3.1c). First, we introduced a chromosomal barcode to uniquely identify each of the 354 strains (Table S2, Strains available on request). Second, we inserted the complementary DHFR N-fragment (or DHFR[1,2]) and C-terminal-fragment (or DHFR[3])-coding oligos, including coding sequences for a 10 amino acid flexible linker peptide (Gly4Ser)2 fused 5' to each fragment. These were integrated 3' to the interacting proteins coding sequence so that products would be the endogenous bait or prey proteins with one or the other of the linker-mrDHFR fragments fused to their C-termini' coding sequence. Third, we propagated the strains under MTX and in the presence of drugs (the "environmental" perturbation). Fourth, we performed genomic extraction and PCR amplification of the barcode followed by NGS to estimate the fitness of each strain and, consequently, the magnitude of PPIs. Lastly, we performed piQTL mapping by performing linear regression between PPI strength and genetic identity at each SNP locus.

Our yeast editing approach for barcoding and DHFR tagging used a combination of homologous recombination for accuracy and CRISPR/Cas9 for efficiency (Extended Data Figure 3.6a) (Horwitz, 2015; Jakociunas et al., 2016; Ryan, 2014). To perform the chromosomal barcoding, the donor DNA (Extended Data Figure 3.6b) contained two NNNNN sites that uniquely label each of the 354 yeast strains (Table S2, Strains, available on request) in tandem with a Ura3 gene selection marker that allows selection in minimal media without uracil. The barcode is inserted into the neutral locus YBR209W (Extended Data Figure 3.6b,c) (Kao and Sherlock, 2008; Levy et al., 2015; Venkataram, 2016)). To determine the intrinsic reproducibility of fitness estimates within each pool, 3 strains (17, 40, and 180) were labeled with two unique barcodes. Once barcoded, the 354 strains (with 357 unique barcodes) were pooled before tagging with the DHFR fragments (Figure 3.1c (i)). Similarly, to integrate DHFR[1,2] fragment to the bait protein (Fig. 3.1c (ii)), the donor DNA includes the DHFR[1,2] fragment, an sgRNA targeting the bait ORF, and a selection marker Nat1 to guarantee successful integration (Extended Data Figure 3.7a). To integrate the DHFR[3] fragment to the prey ORF, the donor cassette contained the selection marker hph. This pooling strategy led to a 96-well plate, where 62 wells contained unique PPI bait/prey pairs under 354 genomic backgrounds tagged with DHFR. To serve as a baseline fitness corresponding to zero PPI in each pool, 2 strains (43 and 599), were also barcoded but did not receive the DHFR fragments (Fig. 3.1c (iii)). Throughout the barcoding of the strains and the tagging by DHFR fragments, all strains were maintained to be haploid.

Comparison of the DHFR-labeled strain's growth in the presence of MTX (10 mg/ml) relative to the untagged reference strain would determine the PPI strength of the bait/prey pair under "basal control" conditions (minimal media+supplements). While the comparison of the growth between MTX and MTX+drugs would determine the change in PPI due to

the drug. The concentration of the drugs was determined in a prior GWAS study (She and Jarosz, 2018) and PCA screen (Tarassov et al., 2008): 5-FC (0.2 mM), fluconazole (100 mM), metformin (50 mgM), and trifluoperazine (17.5 mM). To estimate the growth rate of each strain, we extracted the genomic DNA of the pooled samples before and after selection, PCR amplified the locus containing the strain barcode, and then performed NGS analysis (Fig. 3.1d (iv)). The PPI strength of a bait/prey pair π for an inbred strain i is the log-fold-change in barcode count before and after selection, normalized by the response of the reference (untagged) strains:

$$PPI_{\pi_i} = \log_2 \frac{n_{\pi_i,final}}{n_{\pi_i,initial}} - \log_2 \frac{n_{\pi_{reference},final}}{n_{\pi_{reference},initial}} \tag{3.1}$$

We found that that the PPI strength estimate is strongly reproducible between the two biological replicates (Pearson r2=0.88, P-value<1e-16, Figure 3.1f) in the presence of MTX. Furthermore, the within-well internal controls also showed a high correlation between the internal duplicates (Pearson r=0.89, P-value<1e-16; Extended Data Figure 3.8), thus, we had sufficient accuracy for the PPI strength from the pooled estimations. Additionally, across all PPIs and drug conditions, we recovered on average $\sim$ 330 strains out of the 354 ($\sim$ 93%) (Figure 3.1e). Thus, we estimated the protein interactions with sufficient accuracy over enough strains to perform association between PPI and genetic variation.

Principal component analysis across all the conditions determined the largest source of average PPI variation (Extended Data Figure 3.9). Each point in the figure is the estimated change in PPI strength, averaged over all strains (Equation 1, Methods). Expectedly, $\sim$ 65% of the variation (PC1 and PC2) is due to methotrexate, which is our selective drug for determining the effective number of protein complexes formed based on mrDHFR complementation (Extended Data Figure 3.9a). In addition, we found that the drug conditions are largely overlapping, except for fluconazole (Extended Data Fig. 3.9b), which is expected, since the main functional target of the drug, Erg11 as well as other genes of the ergosterol pathway, are included in our set of 61 PPI reporters (Figure 3.1b). In contrast, the known mode of action of the three other drugs, are not in our PPI reporters.

3.2.2. piQTLs are spread across the genome including regulatory regions and non-coding RNAs

To perform QTL analysis, we performed linear regression each SNP locus using the following model (Shabalin, 2012):

$$\vec{y}_{obs} = \hat{\alpha} + \hat{\beta}\vec{s} + \varepsilon \tag{3.2}$$

where $\vec{y}_{obs}$ is the vector of observed PPI values across all strains, $\vec{s}$ is the genotype, $\hat{\alpha}$ is the intercept, and $\hat{\beta}$ is the slope coefficient (the QTL "effect size") and ϵ is a random variable from an independent, normal distribution. As cross-validation for the effect of residual linkage in the cohort, we also performed QTL mapping approach (Yin et al., 2021) that account for kinship between the genotypes ((Methods), but since the cohort is almost under linkage equilibrium (Figure 3.1d), the resulting estimates of effect sizes with and without kinship are strongly correlated (Extended Data Figure 3.10). All association mapping results, and the annotated genomic locations are consolidated in a piQTL webserver (Extended Data Figure 3.11).

Overall, we observed a total of $\sim$ 2084 piQTLs (false-discovery rate (FDR) < 0.05 in both association mapping that corrects for residual kinship) for all 61 PPIs under the 5 conditions ("no drug" and 4 drugs). These piQTLs are in 471 unique SNP loci. The largest number of piQTLs occur for fluconazole, which has the direct target Erg11 in the 62 PPI as well as the other genes in the ergosterol pathway (Figure 3.2a; see also the Extended Data Figure 3.12, where only the most significant SNP within an LD block r2 > 0.70). The piQTLs for all conditions are broadly distributed across the various architectural features of the yeast genome (Figure 3.2a). Since piQTL fine-mapping occurs at the level of the proteome, we expectedly found that a significant number of piQTLs are in protein-coding regions ((Figure 3.2a; Extended Data Figure 3.12). The yeast genome is more compact than other model organisms and humans, and indeed, $\sim$ 8000 out of the 12K SNPs are found to be in the protein coding regions (Extended Data Figure 3.12b). Thus, normalizing for the relative abundance of each genomic feature, we found that $\sim$ 8% of these protein-coding SNPs are piQTLs under fluconazole, and $\sim$ 3% for the other drugs whose known mode of action is outside the 61 PPI probes. We found that a comparable proportion of piQTLs occur in regulatory regions, such as gene promoters and 3'UTRs (Figure. 3.2b). These results suggest that in vivo PPI is sufficiently sensitive to report changes in cellular abundances of the bait and prey due to 5'- and 3'- transcriptional regulatory machineries. Strikingly, the effect sizes of piQTLs in 3'-UTR are comparable, if not greater, than either the promoter or protein-coding regions (Figure 3.2b). This suggest that protein-protein interaction in vivo reflects the post-transcriptional regulation encoded in the 3'-UTR, such as mRNA localization, stability, and rate of translation.

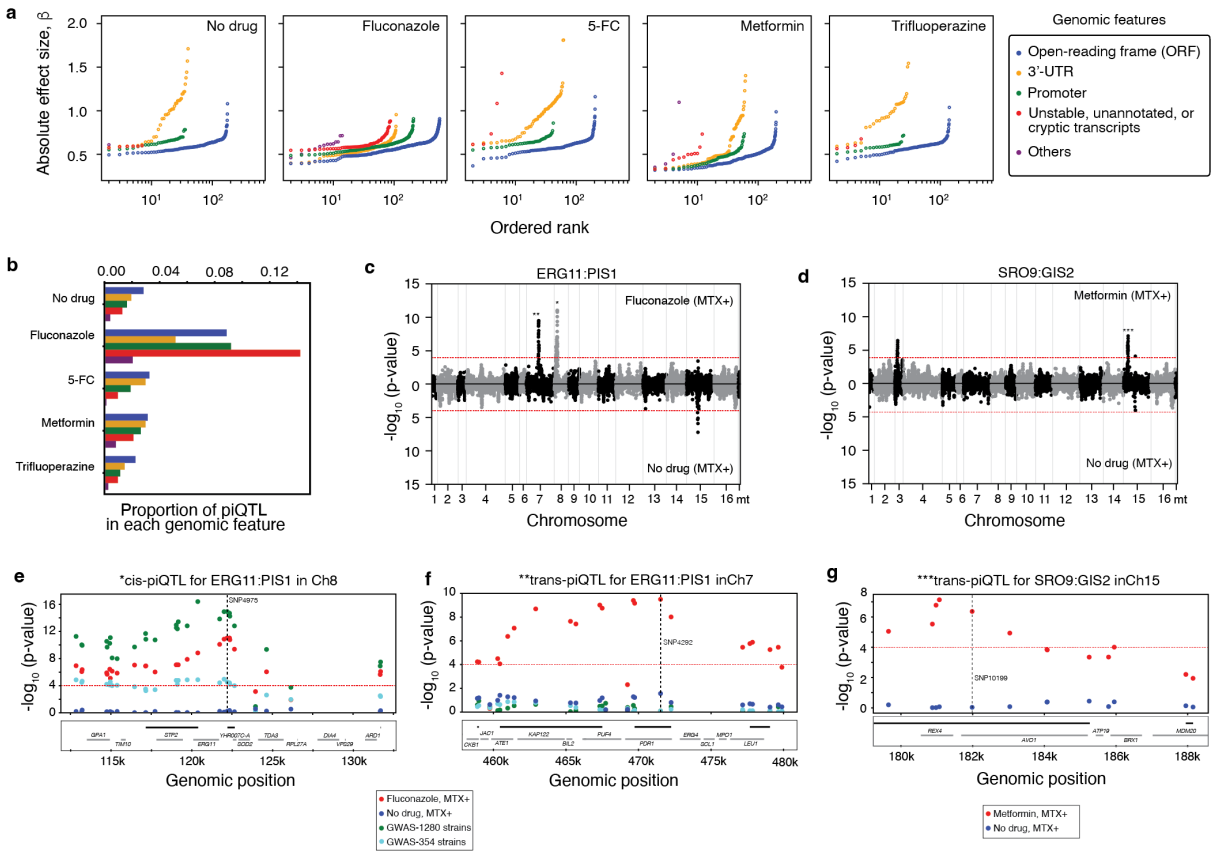

Fig. 3.2. Genomic landscape of piQTLs.

a, Absolute effect sizes of significant piQTLs (FDR < 0.05) and their genomic locations under no drug and under the 4 drugs tested. The total number of piQTLs for each condition are 263, 974, 314, 332, and 201, respectively. **b**, Proportion of SNPs for each genomic feature (color according panel a) that are piQTLs. **c**, Miami plot showing the association for the ERG11:PIS1 protein interaction across all $\sim$ 12K SNPs. Highlighted are two regions showing a "cis-piQTL" in Chr8, where ERG11, is located and a "trans-piQTL" in Ch7, where the lead SNPs do not directly involve either the bait or prey. ERG11 itself is the target of Fluconazole. **d**, Miami plot for SRO9:GIS2 showing a "cis-piQTL" under metformin. **e-g**, Locus plots showing the piQTL strength in the presence and absence of the drug, together with the GWAS study.

Most interestingly, piQTL mapping also captured effects arising from genomic regions containing non-coding RNAs, whose functional consequences are not fully known or even annotated, including SUT (stable unannotated transcripts), CUT (cryptic unstable transcripts), and XUT (XRN1-sensitive unannotated transcripts) (Dijk et al., 2011; Xu et al., 2009) (Figure 3.2a). Although fewer piQTLs are found in non-coding RNAs than in 3' UTRs or protein CDS, their effect sizes are comparable under some drug conditions (Figure 3.2b). Altogether, these demonstrate that piQTL mapping is a sensitive reporter of the cellular consequences of genomic variations, including in those regions that do not encode for proteins and those for which functional effects, such as non-coding RNAs, are not yet known.

3.2.3. Abundance of cis- and trans-piQTLs

Analogous to cis- and trans-eQTLs, we can also classify the piQTLs based on their proximity to the genomic loci of the mrDHFR PCA bait/prey pairs. cis-piQTLs are those that are close to either the bait or prey ORFs, such that the consequences of the piQTL could be to directly affect the abundances of either bait or prey proteins. Additionally, cis-piQTLs could be amino acid substitutions that change the binding affinity of either bait or prey to intermediary proteins. Conversely, trans-piQTLs are those that are distant (operationally defined as 1K bp away) from the bait/prey genomic loci. Figure 3.2c shows the Miami plot comparing the piQTLs of the protein pair ERG11 (lanosterol 14-alpha-demethylase, chromosome 8) and PIS1 (phosphatidylinositol synthase, chromosome 16), with and without fluconazole. Only in the presence of fluconazole, ERG11:PIS1 shows a cis-piQTL in the intergenic region upstream of ERG11 (Figure 3.2c,e). This peak was manifested in a previous GWAS study (She and Jarosz, 2018) using all 1,125 inbred strains, resulting in high association power. When we performed GWAS on the smaller 354 strains used for this piQTL study, the association for this locus was close to the FDR cut-off. Taken together, this means that piQTL can recover GWAS hits with a smaller number of strains, suggesting that piQTL could be more powerful in detecting the potential functional relevance for the locus of interest.

The second peak in chromosome 7 centered around the gene Pdr1 (Pleiotropic Drug Resistance) (Figure 3.2c,f) is an example of a trans-piQTL. This signal was not detected in the GWAS (She and Jarosz, 2018) using fluconazole (Figure 3.2f) but was a hit in the GWAS under a related antifungal, ketoconazole. Pdr1 encodes a transcription factor that regulates the overproduction of transporters (Balzi et al., 1987), conferring pleiotropic drug resistance. The pathogenic yeast Candida glabrata frequently acquires resistance to azole drugs including fluconazole via mutations in Pdr1 (Tsai et al., 2006). Interestingly, this locus is also adjacent to ERG4 (sterol C-24 reductase), the penultimate enzyme in the ergosterol pathway targeted by fluconazole (Fig. 3.2f). However, the absence of SNPs in ERG4 did not allow for piQTL determination in this gene. Another example of a trans-piQTL is the hit at chromosome 15 for the bait/prey pair SRO9 (RNA binding protein suppressor of rho3, chromosome 3) and the zinc finger protein GIS2 (chromosome 14). These piQTLs are located near the gene Avo1 (adheres voraciously to Tor2 protein 1), which regulates cell cycle-dependent polarization of the actin cytoskeleton and cell wall integrity. Tor2, a gene that encodes a serine/threonine kinase, is a direct target of rapamycin, an anti-cancer and immunosuppressant drug, whose functional consequences in the cell are known to overlap with metformin in model organisms (Anisimov et al., 2011; J. Huang et al., 2004; Stynen et al., 2018), including yeast and humans.

3.2.4. pi-QTL reflects the connectivity of biochemical networks

The 61 PPIs we probed form a connected sub-graph of the PPI network (Fig. 3.1b) either because they share a common node, or they are part of multi-protein complexes. We asked if closely related PPIs share similar piQTL association profile. Figure 3.3a is a heatmap showing the significant piQTLs across the PPI probes under fluconazole. The rows of PPIs are ordered based on their proximity on the PPI network (Figure 3.1b, also Extended Data Figure 3.12). The results show that similar piQTLs are likely to occur in closely connected PPIs. To quantify this trend, the number of overlapping piQTL peaks between Manhattan plots of a PPI pair decreases as their distance on the PPI network increases (Figure 3.3b for fluconazole). Furthermore, despite the broad genomic distribution of piQTLs for a given interacting pair, the piQTLs and the reporter bait/prey pair are closely connected on the protein interaction network. Figure 3.3c-d shows the piQTL and bait/prey connectivity on the PPI network defined by Tarassov and colleagues (Tarassov et al., 2008). This observation is robust to other curations of reference PPI networks in yeast. This trend is also robust across all the no drug and 4 drug conditions. The trend is strongest in fluconazole since is the condition that has the largest number of piQTLs.

Next, we determined whether the similarity in association is also reflected by the underlying genetic interactions between the proteins. To this end, we mapped the piQTLs onto the genetic interactions network (GIN) defined by comprehensive double-gene knockouts (Costanzo et al., 2016). Although, 61 PPIs are enriched in some functional modules in the GIN (Extended Data Figure 3.13a; e.g., Module 1 on cell polarity and morphogenesis), the piQTLs themselves are not enriched in such modules. In contrast, most of the piQTLs are statistically enriched in the intervening nodes between the modules. We hypothesize from these results that the gene knockouts reflect drastic effects on protein-protein interactions, in contrast to effects of polymorphisms that are part of the standing genetic variation, such as the $\sim$ 12K SNPs between the two parental strains (Fig. 3.1a). Thus, our piQTL mapping provides a unique measurement of the genetic sensitivity of the PPI network to standing genetic variation, which is complementary to knockouts and/or random mutagenesis by CRISPR/Cas9. We articulate in the Discussion section the potential implication of the distinction between gene-knockouts and standing genetic variation on the topology of protein-interaction networks.

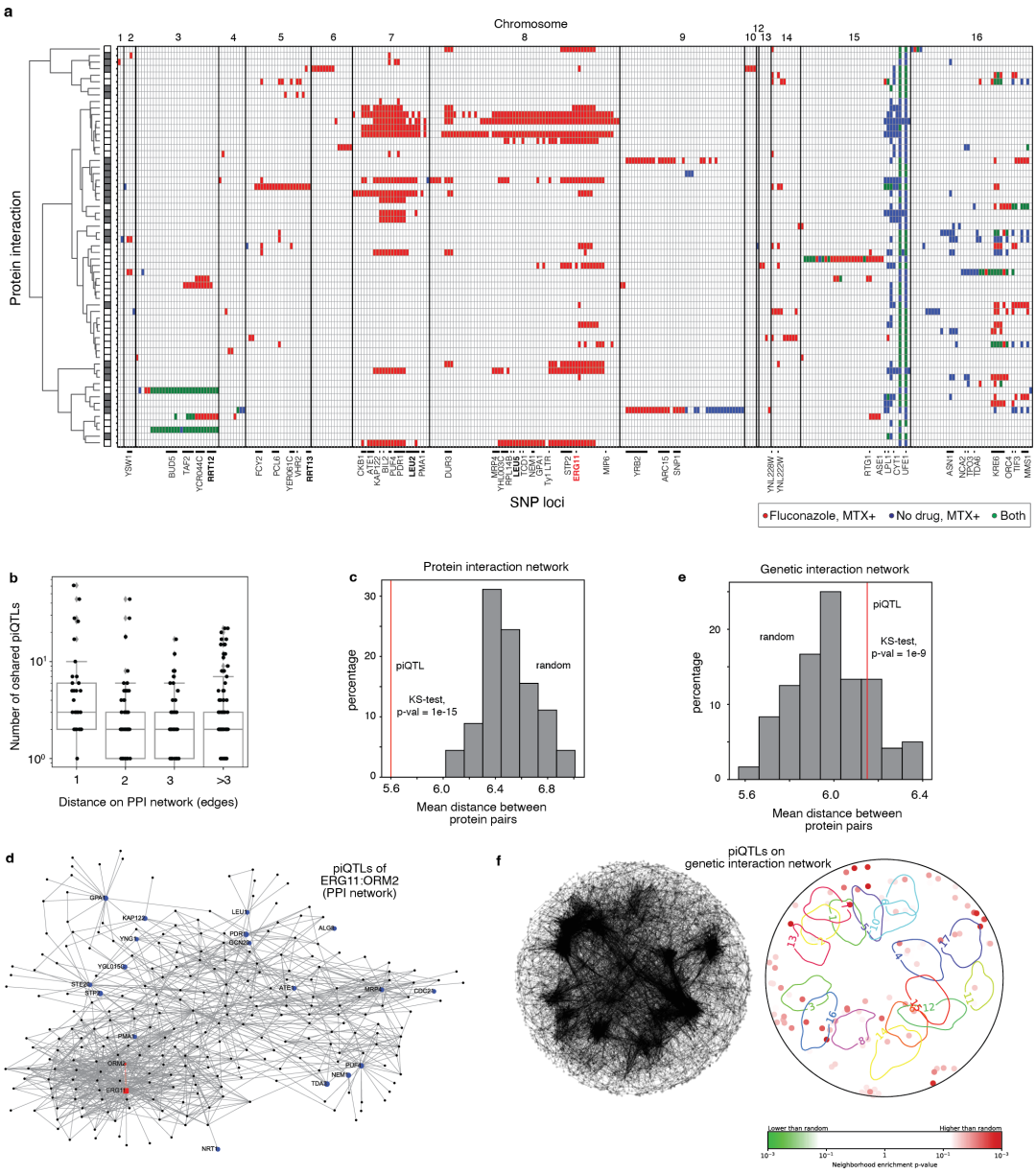

Fig. 3.3. piQTLs reflect the topology and strength of biological networks.

a), SNPs that were piQTLs in at one PPI (red is PPI in fluconazole, blue in no drug, and green in both). The rows are ordered according to the hierarchical clustering of the PPI based on their distance on the PPI network. Identities of the rows are shown in Figure 3.12. **b**, Similarity in genome-wide association between two PPI based on the number of overlapping piQTLs versus their distance on the PPI network. **c-d**, piQTLs tend to cluster in PPI networks compared to random as quantified by their pair-wise distances (panel c) and demonstrated by an example (panel d). **e-f**, piQTLs under fluconazole were projected onto the genetic interaction network defined from double-knockouts (Costanzo et al., 2016). piQTLs tend to be between distant pairs compared to random and enriched to connections between functional modules (panel f). See also Figure 3.13 for the definition of the modules and piQTL enrichment in other conditions.

3.2.5. Abundance of piQTLs in non-coding RNAs

The surprising abundance of piQTLs that we found in non-coding RNAs (ncRNAs) (Figure 3.2b), prompted us to determine how the piQTLs partition into the different classes of non-coding RNAs (Figure 3.4a). In *S. cerevisiae*, there are $\sim$ 2,000 ncRNAs in yeast that make up $\sim$ 20% of its genome (Parker et al., 2018). Two classes of ncRNAs were initially identified according to their half-life in the cell, the stable unannotated transcripts (SUTs) have a relatively long half-life, whereas the cryptic unstable transcripts (CUTs) RNAs have a short half-life and were revealed only after deletion of the exosome complex exoribonuclease Rrp629 (Joo et al., 2017; Parker et al., 2018; Wery et al., 2016; Xu et al., 2009). Deletion of the cytoplasmic exonuclease Xrn1, followed by RNA sequencing, revealed another class of ncRNAs termed Xrn1-sensitive unstable transcripts (XUTs) (Dijk et al., 2011), some of which overlap with either a SUT or CUT. Since, Xrn-1 occurs only in the cytoplasm, SUTs or CUTs that are also XUTs are exported to the cytoplasm and processed like regular mRNAs (Dijk et al., 2011). The full mechanism and role of these transcripts are generally unknown and a subject of debate (Quinn and Chang, 2016).

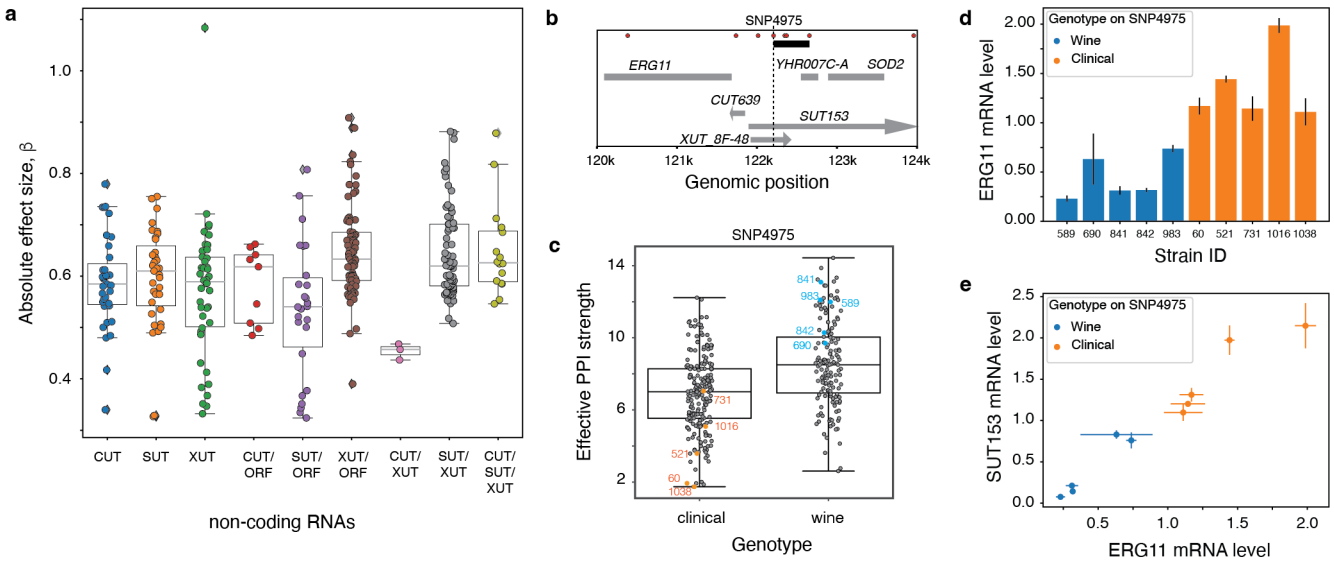

Fig. 3.4. Non-coding RNAs significantly affect protein-protein interactions.

a), Effect sizes of the piQTLs in non-coding RNAs CUT, SUT, and XUT. Grouped separately are SNPs that have multiple annotations either as overlapping with ORFs or other non-coding RNAs. **b**, Zoom in of the genomic locus near SNP4975 showing the most significant piQTL under fluconazole near ERG11. This region contains multiple overlapping non-coding RNAs. Red dots indicate the SNPs (see also Fig. 3.2e). Black line indicates LD > 0.75. **c**, Protein interaction of strains grouped according to their genotype at the SNP4975 locus. **d-e**, We quantified the transcripts for ERG11 and SUT153 by qPCR using five strains from each genotypic background (colored dots in panel c). mRNA level is expressed relative to the Abp1 gene (actin).

We found that the effect sizes are distributed across all types of non-coding RNAs (Figure 3.4a). Interestingly, we also observed that SUTs and XUTs found in the coding regions have larger effect sizes. This could indicate that these SNPs either have dual effects, through the gene's protein product or through the non-coding RNA itself. Additionally, the effect sizes are also higher when a SUT is also an annotated XUT (t-test P-value= 0.008) and when SNPs are triply annotated as SUT/XUT/CUT. This indeed supports the hypothesis that non-coding RNAs (such as SUTs initially in the nucleus) are processed as mRNAs and exported to the cytoplasm (XUTs) and are thus, more likely functional. For instance, one of the strongest piQTLs for fluconazole is SNP4975 in the $\sim$ 2 kb SUT153, which is also annotated to overlap with the shorter $\sim$ 0.5 kb XUT_8F-48 (shown also for ERG11:PIS1 PPI in Fig. 3.2c,e). This locus is also a piQTL in 16 other probe PPIs (Fig. 3.3a) and is the only significant locus around ERG11 for 4 of these PPIs, suggesting that this is the most likely lead causal SNP in the region. Thus, we hypothesized that SUT153 regulates the transcription of ERG11 and consequently its intracellular abundance and interaction with other proteins. This sensitivity was shown for SUT153, where a smaller fragment of it is found in the cytoplasm (it is also annotated as an XUT) (Figure 3.4b) which does not overlap with the adjacent gene (YHR007C-A, a putative protein of unknown function). SUT153 is a potential "cis-piQTL", affecting a gene involved in the bait:prey pair. To test our hypothesis, we quantified by qPCR the mRNA level of ERG11 and SUT153 on 10 strains, half contains the SNP from the clinical parent and the other from the wine parent (Figure 3.4c). The two groups of strains differ in their ERG11 expression (Figure 3.4d), reflecting the strong sensitivity of this locus to fluconazole (Figure 3.4c). Moreover, the ERG11 and SUT153 mRNA levels are strongly correlated (Figure 3.4e; mRNA level is quantified relative to the gene ABP1 (actin)), suggesting a possible active regulation by the loci identified as piQTLs. Overall, this result suggests that piQTL provides complementary information to eQTL by providing the downstream and protein-level functional significance of non-coding RNAs.

3.3. Discussion

Here, we developed a new fine mapping approach based on PPI as molecular phenotypes and accurate reporters of the functional consequences of genomic variation. piQTL has advantages over eQTL and pQTL. First, broad applicability of pQTL faces practical limitations in that proteomics, typically done by mass spectrometry, is technically more challenging and costlier than transcriptomics by RNA-seq, which takes advantage of next-generation sequencing (NGS). The piQTL method is an association at the proteome level, thus potentially a better reporter of GxE than gene transcripts. As demonstrated by our results, piQTL could accurately report on multiple cellular processes. Additionally, our implementation of pooled PPI screen shows that piQTL has the practicality and scalability of RNA-seq and eQTL because the PPI readout is through NGS.

The preponderance of GWAS QTLs and eQTLs across the genome and over many genes without an interpretable connection to the phenotype or disease of interest (Boyle et al., 2017; Y. Li, 2016; Pickrell, 2014; J. Pritchard, 2001; The International Schizophrenia, 2009) led Pritchard and colleagues to propose the "omnigenic" model of complex traits (Boyle et al., 2017; X. Liu et al., 2019). This model postulates that gene regulatory networks are sufficiently interconnected that all genes expressed in disease-relevant cells are liable to affect the functions of core disease-related genes. PPIs are a direct measure of physical interaction between protein gene products compared to gene co-expression. Thus, piQTL will also be a direct reflect of the pleiotropy and the hypothesized "omnigenicity" of complex traits, as shown by the clustering of piQTLs in biological networks.

Additionally, there is the vexing and long-standing problem of "missing heritability" in complex trait genetics, which is the inability of known QTLs to explain a significant proportion of heritable phenotypic variation (Manolio et al., 2009). This missing heritability is partly explained by statistical models based on non-additive interactions between mutations in different genomic loci (epistasis) or on the ability of a single locus to affect multiple phenotypes (pleiotropy). Indeed, the abundance of QTLs led Pritchard and colleagues to propose the "omnigenic" model of complex traits based on the hypothesis that these genetic and phenotypic interactions occur within networks of molecular phenotypes, such as transcriptional regulatory networks (Boyle et al., 2017; X. Liu et al., 2019). These, however, do not capture the effects of genetic variation at the highest level, protein biochemical networks directly connected to phenotypes.

Probing PPI is a direct measure of the physical interaction between gene products, while "interaction" between genes at the gene expression level or protein abundance level is through

correlation. Hence, developing a molecular phenotypic readout based on measuring variations in PPI is potentially a more direct way to quantify epistasis (interactions between mutations in different genomic loci) and pleiotropy (the ability of a single locus to affect multiple phenotypes). This study provides a roadmap for a new type of QTL mapping based on PPI networks that reveals the functional and biochemical drivers of complex phenotypes. The systematic development of piQTL in yeast mirrors the development over the past decade of eQTL, which was also established in more tractable model organisms prior to its broad application in humans (Brem et al., 2002; Stynen et al., 2018) In yeast, we were able to determine new mechanisms for the action of two antifungal compounds 5-FC and fluconazole. Additionally, even in yeast, we could already discern biochemical mechanisms that apply to human phenotypes, such as the action of the type II diabetes drug metformin and an anti-psychotic compound trifluoperazine (data not shown).

3.4. Material & methods

3.4.1. Strain description

The population cohort in this study consisted of 354 haploid inbred strains of *S. cerevisiae* previously constructed and genotyped by She and Jarosz for a GWAS study (She and Jarosz, 2018). The strains were kept haploid during genomic editing to add the barcode and DHFR fragment tags using a mating inhibitory peptide. The strains were F6 progenies from inbreeding of two parental strains RM11 and YJM765. There were 12,054 high-frequency SNPs in the *S. cerevisiae* parental and inbred strains.

3.4.2. Annotation of the SNPs genomic features

To determine the genomic elements on which the 12,054 SNPs are located, we obtained the annotated reference genome of *S. cerevisiae* from the Saccharomyces Genome Database (http://sgd-archive.yeastgenome.org/sequence/S288C_reference/) latest version R 64-3-1 (released on 2021-04-27). The features annotated include all open-reading frames (ORFs), the 5' and 3' UTR of these ORFs, non-coding RNAs, intergenic regions, and other genomic features (long tandem repeats (LTRs), transposons, and telomeric regions). We also mapped the annotations of promoter regions and unstable transcripts performed by Rossi and colleagues (Rossi et al., 2021). The SNPs in the CDS regions were further classified into either synonymous or non-synonymous. Reference datasets and results of the annotation mapping pipeline are available on GitHub (https://github.com/ladyson1806/piQTL_mapping)

3.4.3. Subnetwork of 61 PPIs

The 61 PPIs (Supplementary Table 1, available on request) targeted in this study were selected from a previous proteome-wide *in vivo* interactome by DHFR-PCA (Tarassov et al., 2008). Therefore, the PPIs used in this study are guaranteed to have their PPIs detectable using DHFR-PCA. In addition, the PPIs targeted in this study were manually curated to cover a wide range of molecular functions and biological processes (Table 3.14)

3.4.4. Genome editing for chromosomal barcoding and DHFR fragment tagging

Overview

Our yeast editing approach for barcoding and DHFR tagging used a combination of homologous recombination for accuracy and CRISPR/Cas9 for efficiency (Extended Data Figure 3.6a) (Horwitz, 2015; Jakociunas et al., 2016). Briefly, a "donor DNA" with homology arms to the target genomic locus, an sgRNA targeting the locus of integration, and

a linearized Cas9 expression vector were co-transformed into yeast. Gap repair led to the reconstitution of the Cas9 expression plasmid, followed by the cutting of the genomic target locus by Cas9/sgRNA, and its subsequent repair by homologous recombination.

Competent cell preparation and transformation

Yeast-competent cells were prepared by following the lithium acetate protocol (Gietz and Schiestl, 2007). The frozen stock of the yeast cells was incubated overnight at 30°C in YPD medium supplemented with 50 nM mating inhibitory peptides. The following day, the pre-cultured yeast cells were diluted in pre-warmed 2x YPD media to a cell concentration of 2x106 cells/mL. Yeast cells were incubated until an OD600 range of 0.8-1.0 ($\sim$ 8x106 cells/mL), at which point the cells were harvested by centrifugation. The harvested cells were washed once with sterile water and resuspended in 0.01 volume of frozen competent cell (FCC) solution (5% v/v glycerol, 10% v/v DMSO). The cell suspension was dispensed into microcentrifuge tubes in 50 L aliquots and gradually frozen at -80°C. For transformation, the frozen competent cells were thawed in a 37°C water bath for 15-30 seconds and briefly centrifuged to remove the supernatant. The cells were resuspended in 360 L of transformation mix (10 L salmon sperm DNA, 260 L 50% w/v PEG 3350, 36 L 1.0M LiOAc, transforming DNA, and fill-up with sterile water). The cell suspension was then heat-shocked at 42°C for 30 min and briefly centrifuged to remove the supernatant. The cells were resuspended in 1 mL YPD medium, incubated at 30°C for 4 h, and plated on appropriate selective YPD plates.

Chromosomal barcoding of inbred strains

To perform the chromosomal barcoding, the donor DNA (Extended Data Figure 3.6b) contained two NNNNN sites that uniquely label each of the 354 yeast strains (Table S2, Strains, available on request) in tandem with a Ura3 gene selection marker that allows selection in minimal media without uracil. To determine the intrinsic reproducibility of fitness estimates within each pool, 3 strains (17, 40, and 180) were labeled with two unique barcodes. Thus, the total number of barcodes is 357 corresponding to 354 unique inbred strains.

i) Barcode design. To identify the 357 library strains using next-generation sequencing (NGS), they were labeled with a DNA barcode consisting of two 5-bp short barcodes assembled to form a complete barcode. The complete barcode sequence consisted of two short barcode sequences placed on either side of a 20 bp consensus sequence. There were 24 unique sequences in each of the left and right short barcodes, resulting in 576 possible unique barcode combinations, of which we used 357 (Table S2, Strains, available on request). These barcodes were also designed to have comparable GC content with two or three G/C nucleotides and at least three Hamming distances from other barcodes in the same location.

The strain barcodes were inserted into the YBR209W dubious open reading frame of the library strains along with the URA3 gene as selection marker, which contains its own promoter. The disruption of YBR209 has previously been demonstrated to have no effect on fitness (Kao and Sherlock, 2008; Levy et al., 2015).

ii) Barcode cassette preparation. The URA3-barcode DNA cassettes were prepared by assembling the YBR209W left homology arm (175 bp), the URA3 gene, left barcode oligo , right barcode oligo, and the YBR209W right homology arm (175 bp). The YBR209W homology arms and URA3 gene cassettes were amplified using PCR. The YBR209W homology arms were amplified directly from the library strain genome, and the URA3 cassette was amplified from plasmid pWS158. The amplified DNA fragments and barcode oligos had 20 bp complementary DNA sequences to adjacent DNA fragments or oligos. The URA3-barcode cassettes were constructed by assembling the DNA fragments/oligos from the YBR209W left homology arm to the YBR209W right homology arm in steps (fusion PCR).

iii) Barcode cassette insertion into yeast genome. DNA barcode labeling of the yeast library strains was performed in accordance with the CRISPR yeast genome editing protocol by Shaw and co-workers45 (see also Extended Data Figure 3.6a). To prepare the linearized sgRNA expression cassette for targeting the YBR209W locus, oligos containing the CRISPR sequence of YBR209W were inserted into pWS082 (Addgene), the template plasmid of sgRNA cassette, using BsmbI Golden Gate Assembly . The linearized sgRNA expression cassette targeting YBR209W (Addgene) was then amplified from pWS082-YBR209W using PCR. The linearized Cas9 cassette was amplified from pWS176 (Addgene) by PCR, and the donor DNA was the URA3-barcode cassette. The three cassettes were gel purified by extracting fragments of the following sizes: sgRNA cassette, 1045 bp; Cas9 cassette, 9993 bp; URA3-barcode cassette, 1488 bp. The purified DNA cassettes were mixed in the following ratio and transformed into competent cells: sgRNA cassette, 200 ng; Cas9 cassette, 100 ng; and URA3-barcode cassette, 5 g. The transformed cells were stored at -80°C for downstream processing.

iv) Validation and Barcoding efficiency. After transformation, 25 colonies were picked at random, and all 25 colonies were confirmed to have been transformed with the target DNA cassette by colony PCR (Extended Data Figure 3.6c) and Sanger sequencing.

DHFR-fragment genomic tagging in yeast barcoded library strains

We pooled the barcoded yeast strains at equal densities and then performed DHFR fragments tagging via CRISPR yeast genome editing protocol45. To prepare the linearized sgRNA expression cassettes for targeting loci, oligos containing the CRISPR sequence (Supplementary Table 1, available on request) were inserted into pWS082 (Addgene), the template plasmid of sgRNA cassette, using BsmbI Golden Gate Assembly, as described in the reference protocol. sgRNAs were designed to digest the 3'-end region (within 16 bp of the stop codon) of the coding region of the target genes. The linearized sgRNA expression cassettes were then amplified by PCR from each gene's modified pWS082. Donor DNA cassettes were amplified from the plasmid pAG25 (Addgene) (for DHFR-F[1,2] fragment) or pAG32 (Addgene) (for DHFR-F[3] fragment) using target gene-specific primers with 40 bp homology sequences as overhangs. The linearized Cas9 cassette was amplified from pWS176 by PCR. The three cassettes were gel purified by extracting fragments of the following sizes: sgRNA cassette, 1045 bp; Cas9 cassette, 9993 bp; donor DNA cassette for DHFR-F[1,2], 2009 bp, donor DNA cassette for DHFR-F[3], 1740 bp. The purified DNA cassettes were mixed in the following ratio and transformed: sgRNA cassette, 200 ng; Cas9 cassette, 100 ng; and donor DNA cassette, 5 g. The transformation efficiency depended on the target gene, and DHFR fragment labeling was repeated until a colony count of at least ten times the yeast library size (3,570 colonies) was obtained. The average colony number was 21,879 ($\sim$ 61-fold of the library size). The transformed cells were stored at -80°C for downstream processing.

3.4.5. Adding a spike-in strains as reference for "no PPI" due to DHFR-PCA

As a reference for the growth rate for strains that do will not measurable PPI due to DHFR-PCA complementation, two strains (43 and 599, Supplementary Table 1) were barcoded, but these were not tagged by the DHFR fragments. These two strains were labeled by two unique barcodes to have two estimates of their fitness in each well. To perform the spike-in, the prepared library of tagged DHFR strains and the two control strains were each incubated overnight at 30°C in YPD medium, adjusted to an OD600 value of 2, and mixed with 950 L of the library pool and 12.5 L of each control strain. As a result, 5% control spike-in samples were generated for all target PPI libraries and stored at -80°C for further processing. Altogether, the total number of unique barcodes in the pool is 361 (351 singly-barcoded strains tagged with DHFR + (3x2) from 3 strains doubly-barcoded tagged with DHFR + (2x2) from 2 strains doubly-barcoded but untagged with DHFR) (Supplementary Table 2, Strains, available on request).

3.4.6. Selection and growth conditions for measuring PPI interactions

The constructed DHFR-tagged inbred strain libraries (61 PPIs + no-PPI reference) were grown in serial batch culture under SD media based on five different supplemental drug conditions with 0 or 10 g/mL methotrexate concentrations (minimal SD bases, 2% glucose, 120 mg/L Leucine + supplemental drugs + methotrexate). The concentration of the drugs was determined in a prior GWAS study (She and Jarosz, 2018) and PCA screen (Tarassov et al., 2008): 5-FC (0.2 M), fluconazole (100 M), metformin (50 mM), and trifluoperazine (17.5 M). Two biological replicates were passaged for all libraries and all drug conditions. For each sample, 40 L of frozen cells ($\sim$ 4 x 105 cells) were inoculated into 760 L media in 96 deep-well plates. The cells were grown at 30°C for 96 hours and passaged every 24 hours at a ratio of 1:8 (100 L of cultured cells were added into 700 L of fresh media). We measured the OD600 of the cultures at the end of each passage to calculate the cell number. After each passage, the deep-well plates were centrifuged at $\sim$ 1,900 x g for 10 minutes to pellet the cells. We stored every passage at -20°C for the downstream analysis.

3.4.7. Genomic barcode extraction and deep sequencing

Yeast genomic DNA extraction

The 61 PPIs measured in 5 environmental conditions (Drug-Condition = noDrug, 5-FC, fluconazole, metformin, and trifluoperazine; multiplied by with and without MTX), two timepoints (T=0h and T=96h), and two replicates resulted in 1,302 biological samples in twenty-one 96-well plates for downstream NGS analysis. To extract the gDNA, the cell pellets were resuspended in 300 L of 50 mM EDTA containing 1 unit of Zymolyase. We centrifuged the mixtures at 1,900 x g for 10 minutes after incubating them for 30 to 60 minutes at 37°C. By aspirating the supernatant, we removed the excess liquid. Next, we added 200 L of lysis buffer (2% Triton X-100, 1% SDS, 100 mM NaCl, 10% Tris-HCl, 1 mM EDTA, pH 8), 2 L of RNaseA, and 2 L of protease K, and incubated for 10 minutes at 55 °C. The samples were vortexed, centrifuged at 1,900 x g for 10 minutes, and 200 L of the supernatant was then transferred to fresh deep-well plates with 600 L of 100% ethanol in each well. The samples were kept at room temperature for 15 minutes, then centrifuged at 1,900 x g for 10 minutes at 4°C, discarding the supernatant. We next obtained the isolated genome pellets, washed them using 150 liters of 70% ethanol, and centrifuged them at 1,900 x g for 10 minutes at 4°C. We removed the supernatant and let the pellet air dry for a few minutes at 60 °C. The pellets were rehydrated for an hour at 65°C after being dissolved in 40 L of TE. The extracted genomic DNA samples were stored at -20°C for the subsequent analysis.

Deep sequencing

Strain barcodes were amplified from the extracted genomic DNA in a two-step PCR. The first PCR was performed using PrimeSTAR GXL polymerase and 1:2000 SYBR Green I (20 L/reaction) with 2 L of extracted genomic DNA per reaction as a template. Because the concentrations of extracted genomic DNA varied, the number of PCR cycles was first determined by qPCR; the number of PCR cycles was chosen such that the amplification curve reached the mid to late log-linear phase (typically $\sim$ 24 cycles). The PCR amplification conditions were one cycle of 98°C for 3 min (initiation), followed by amplification cycles of 98°C for 10 sec, 55°C for 15 sec, and 68°C for 30 until the amplification curve reached the mid to late log-linear phase. Primers used for this reaction are of the following format:

- Forward: TCGTCGGCAGCGTCAGATGTGTATA AGAGACAG NNNNNNNN AGGCGGTTTGCGTATTGGGC
- Reverse: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG NNNNNNNN ACGGCACGCATCTTTGGCTG

The Ns corresponds to in-house sample indices, allowing the unique labeling of samples in each well of the 96-well plate during multiplexed Illumina NGS. The complete list of all NGS primers used in this study is on Github (https://github.com/ladyson1806/piQTL_mapping/blob/main/data/pipeline/PPI_reference_barcodes.csv). The PCR products were pooled per growth media condition and purified with DNA Clean Concentrator (Zymo). The second PCR was performed using PrimeSTAR GXL polymerase with 5 ng of purified product from the previous PCR was used as a template (50 L/reaction). Primers were unique pairs of index primers (i5 and i7) from the Nextera XT DNA library preparation kit (Illumina) (the complete sequence is in Supplemental Table). The PCR amplification conditions were one cycle of 98°C for 45 sec, then 12 cycles each of 98°C for 10 s, 55°C for 15 s, and 68°C for 30 s. The second PCR products were purified using Agencourt AMPure XP PCR purification kit (Beckman Coulter). We performed pair-end sequencing on Illumina NextSeq platforms with 30% balanced DNA (phiX).

Computational Methods

All computational analysis are available on Github (https://github.com/ladyson1806/piQTL_mapping).

3.4.8. FASTQ demultplexing and pre-processing

Our NGS run resulted in >350M reads with 89% at Phred score >= 30. Using AdapterRemoval (version 2.3.2) (Schubert et al., 2016), we split the demultiplexed the FASTQ file into 1,302 files corresponding to the biological samples. We allowed at most 2 mismatches within the inner-adapters (–barcode-mm-r1 = 2 and –barcode-mm-r2 = 2) to account for mutations that could arise from PCR or/and NGS. Since the in-house sample indices have a Hamming distance >2, $\sim$98% of the reads uniquely mapped to their biological samples.

Then, we merged the reads of the 1,302 demultiplexed samples with PEAR (Paired-End reAd merger) (version 0.9.6) (J. Zhang et al., 2014) resulting in $\sim$ 651 merged FASTQ files, which serve as input for strain barcode extraction.

3.4.9. Barcode extraction and mapping to inbred strains

We extracted the strains barcodes from the reads using the command bartender_extractor_com of Bartender (version 1.1) (Zhao et al., 2018) by providing the following regular expression: "TGGGC[5]CAGGTCTGAAGCTGTCGCAC[5]GAAAT", where TGGGC (preceding sequence) and GAAAT (succeeding sequence) that frame the construct double barcodes. The two 5-nucleotide barcode regions are linked by a constant spacer (CAGGTCTGAAGCTGTCGCAC). Additionally, we assigned a mismatch parameter of 4 (option -m 4), this parameter allows 2 mismatches for the proceeding and for the succeeding sequences to account for mutations during PCR or/and sequencing steps. The barcodes were scanned within the reads in forward and reverse complement direction (option -d both). This step resulted in 10-nucleotide barcode reads, which we mapped to the 361 unique strain barcodes (Table S2, Strains, available on request) using in-house Python scripts (https://github.com/ladyson1806/piQTL_mapping/blob/main/pipeline). Since the Hamming distance between our strain barcodes is >=3, we allowed for at most 1 mismatch between the barcode and strain IDs and successfully mapped 98% of the barcodes.

3.4.10. Fitness and PPI estimation

In each well containing the pooled strains but tagged PPI of a bait/prey pair π (Figure 3.1c(iii)), we estimated the relative fitness of a strain i as the log-fold-change (LFC) of its normalized barcode frequency between initial and final timepoints (0h and 96h, respectively):

$$Relative\ Fitness_i = \log_2 \left(\frac{n_{\pi,i,final}}{n_{\pi,i,initial}} \right) \tag{3.3}$$

We compared the fitness estimate between two replicates for all strains across all conditions (223,820 fitness values) using a density plot in Matplotlib Python library (Figure 3.1f). The density values are estimated using Gaussian kernels. To estimate the PPI strength for each bait/prey π in strain i, we subtracted the average fitness of the reference strains (Supplementary Table 2, Strains, available on request) that were untagged with DHFR (Equation (3.1)).

3.4.11. piQTL association mapping

Encoding of the genotype matrix and calculation of residual linkage disequilibrium

The inbred strains in this study is a subset (n=357, Suppementary Table 2) of the 1,125 F6 haploid progenies generated and genotyped by She and Jarosz (She and Jarosz, 2018). They provide the genotype matrix containing 12,054 SNPs, where 1 denotes the wine parental haplotype, -1 the clinical parental haplotype, and 0 is unphased haplotype which we consider as missing genotype. Using snpStats (Clayton, 2023), we calculate amount of linkage disequilibrium (LD) between all pairs of SNPs as the Pearson correlation between their genotype values across all 357 strains. LD blocks were defined as genetic regions with neighboring SNPs with LD > 0.75.

piQTL analyses

Our first piQTL approach relies on a simple linear regression model implemented in Matrix eQTL (Shabalin, 2012), where instead of providing a gene expression matrix as inputs for the phenotyping, we provide our PPI quantification matrix. The association between the changes of PPI and genotype is assumed to be linear (Equation (3.2)) where $\vec{y_{obs}}$ is the vector of observed PPI values across all strains, $\vec{s}$ is the strain locus genotype, $\hat{\alpha}$ is the intercept, $\hat{\beta}$ is the slope coefficient ("the QTL effect size") and ϵ is a random variable from independent and normal distribution. This is followed by calculation of a t-statistic test that provides a P-value. Due to the performance of multiple association's tests, Matrix eQTL correct the P-values by calculating false discovery rate (FDR).

Our second piQTL approach model the relationship between the genotype and the a phenotype (PPI) as a generalized linear model implemented in rMVP (Yin et al., 2021) as follows:

$$\vec{y} = Xb + \epsilon \tag{3.4}$$

where $\vec{y}$ is a vector of observed PPI values across all strains, X is a matrix of fixed effects and testing SNPs, b is an incidence matrix for X, and ϵ is a vector of residuals. A crucial distinction between the first approach and this one is that we add as a covariate in X the kinship matrix inferred from the genotypes. This means that the hypothesis testing accounts for residual LD that exists between the SNPs. We performed both approaches on all SNPs (12,054) and for all PPI under different environmental conditions. SNPs that were known be QTLs for methotrexate in the previous GWAS study were masked in the analysis. SNPs are considered piQTLs if they were significant (FDR < 0.05) in both approaches.

3.4.12. post-piQTL analyses

Heatmap of colocalized piQTLs across 61 PPIs

For each environmental condition (no Drug, 5-FC, fluconazole, metformin, trifluoroperazine), we built a matrix 62 PPI rows x 12,054 SNP columns where we assigned a value of 1 for SNPs defined as a piQTL under the presence of the drug (FDR<0.05 or P-value < 0.0001) where its FDR changed by an order of magnitude between the presence and absence of the drug (Figure 3.3a). For the visualization, SNPs positions with no piQTL across all the PPIs are not shown. The 61 PPIs are ordered based on their proximity in the PPI network defined from genome-wide *in vivo* yeast interactome (Tarassov et al., 2008). Proximity on the network is calculated as the shortest weighted path between nodes, where the weight is the *in vivo* PPI strength (Tarassov et al., 2008). This calculation is implemented in NetworkX (Hagberg et al., 2008).

Interactive visualization of piQTLs

We provided two Shiny/R applications to dynamically visualize the results of the piQTL mapping. Both webtools can be run locally after installing the required library from R (version 4.2.2). These are user-friendly web interfaces that allow 1) to visualize the piQTL graphs, such as Manhattan, quantile-quantile and volcano plots (see manhathan_interactive_plots.R), and 2) browse the piQTL results across the yeast annotated genome and several genome annotations (see igv_interactive_plots.R).
i) Manhattan, quantile-quantile (QQ) and volcano plots. Before visualization, the MTX-specific piQTLs were excluded. MTX-specific piQTLs are defined as piQTLs that appear under methotrexate both in the presence and absence of the four environmental drug conditions. These piQTLs correspond to 45 SNPs, all located in chromosome 15. The Manhattan plot display all the – log10-transformed P-values calculated from the association's tests for the identification of piQTL candidates. The quantile-quantile plot represents the distribution of theoretical versus experimental P-values. This graph shows the different trends of P-values distribution that can be used to determine the quality / reliability of the results we observed in their associated Manhattan plot. These graphs were generated by using Manhattanly, a R library (Bhatnagar, 2021). All the graphs are interactive.

ii) Interactive genome browser. This R/Shiny app is a custom wrapper to load the piQTL Manhattan plots (previously described in Manhattan, QQ and Volcano plots) on the yeast annotated genome available from the Broad Institute website that host the last annotated version of *S. cerevisia*e genome (sacCer3). Clicking on the hits will yield information on the SNP's annotations (SNP ID, chromosome ID, base-pair position, information of their

SNP class, locus ID, gene/genome feature ID, and SGD ID) and piQTL association statistics calculated from the rMVP analyses (P-value, effect size, standard error). Additional information are plotted, such as locations of non-coding RNAs (Dijk et al., 2011; Xu et al., 2009) CUT, SUT and XUT, as well as residual LD blocks (LD score > 0.75). These graphs were generated by using the R library igvR (Shannon, 2022).

3.4.13. Colocalization between yeast GWAS and piQTLs

We compare the Manhattan plots generated from the piQTL mapping and from previous GWAS results under the presence of Fluconazole (She and Jarosz, 2018). Overlapping peaks are considered as colocalized signals between the two strategies.

3.4.14. piQTL mapping on PPI networks biological networks

To observe the distribution of the piQTLs within different biological networks (in particular, the yeast protein-protein interaction network (Tarassov et al., 2008). For a given PPI, we extracted the subgraph of the PPI network with the proteins involved in the PPI and the proteins resulting from the ORFs that contain a SNP defined as a piQTL, their direct and their one-away neighbors. These networks were used to estimate the distances between the piQTLs by calculating the shortest path algorithm from NetworkX. To test if the distances between piQTLs are significantly shorter compared to the distances between nodes that could potentially be associated piQTLs, we performed 10,000 simulations to compare the distribution of distances between the piQTLs for a specific PPI under a given environmental condition and the distances between randomly selected proteins. We estimated the z-score of the mean distance shared between piQTLs to see where the value would fall in the distribution of distance between random nodes.

3.4.15. SAFE analyses

We performed SAFE (Spatial Analysis of Functional Enrichment) (Baryshnikova, 2018) analyses by mapping the SNPs associated with piQTLs on their associated ORFs, their promoter and/or their 3'-UTR regions under a given environmental condition (noDrug, 5-FC, Fluconazole, Metformin, Trifluoperazine). The piQTLs were mapped on the genetic interaction network generated by Costanzo and colleagues (Costanzo et al., 2016). SAFE is an automated network annotation algorithm that performs a local enrichment analysis to determine which regions of the network are over-represented for a given feature, in our case the different environmental conditions. SAFE visualizes the network and maps the detected enrichments onto the network.

Matériel Supplémentaire pour le second article

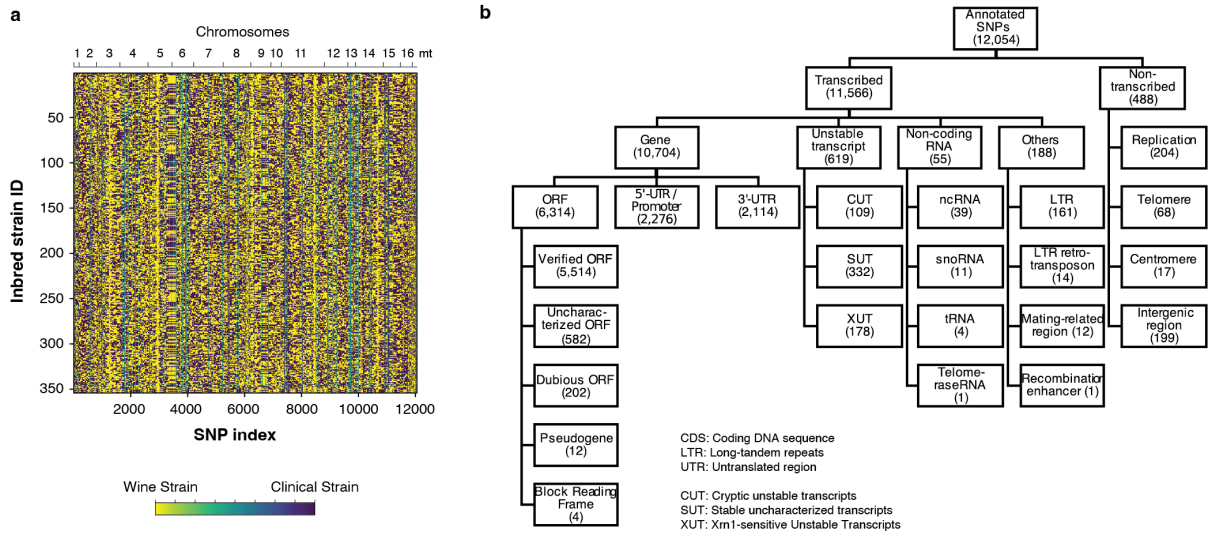

Fig. 3.5. Genotype map of inbred strains and genomic architecture of ∼ 12K mutations.

a), Genotyping of the 354 inbred strains from a previous study1 shows 12K SNPs across all yeast chromosomes and mitochondrial DNA. **b**, Distribution of the 12K SNPs across the functional and architectural features of the yeast genome. Recent annotation (Rossi et al., 2021) of *S. cerevisiae S288C* was used as a reference.

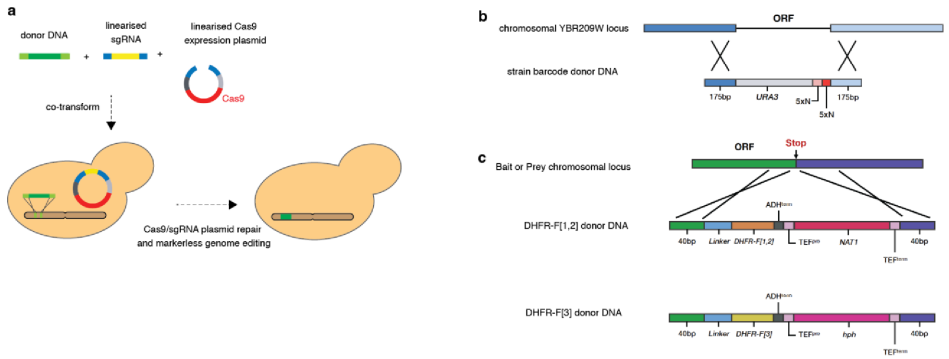

Fig. 3.6. Genome-editing strategy for strain barcoding.

a), High-efficiency yeast genome editing that couples CRISPR/Cas9 with homologous recombination. A "donor DNA" with homology arms to the target genomic locus, an sgRNA targeting the locus of integration, and a linearized Cas9 expression vector are co-transformed into yeast. Gap repair leads to the reconstitution of the Cas9 expression plasmid, followed by the cutting of the genomic target locus by Cas9/sgRNA, and subsequent repair by homologous recombination. **b**, To perform the chromosomal barcoding, the donor DNA contains two NNNNN sites (denoted as 5xN in the diagram) that uniquely label each of the 354 strains. This donor cassette is in tandem with a Ura3 gene selection marker that allows selection on minimal media without uracil. The sgRNA targets for barcode integration a neutral locus in the genome (YBR209W).

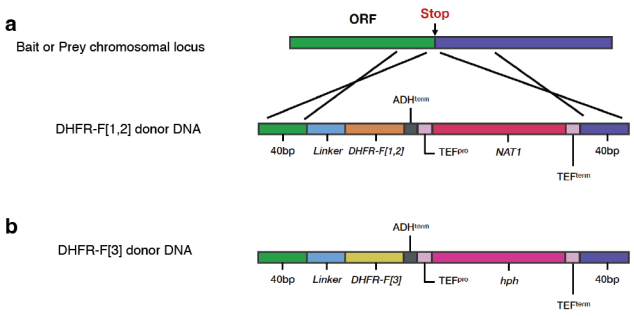

Fig. 3.7. Genome-editing strategy for DHFR fragment tagging.

The genome editing follows the design outlined in Extended Data Figure 3.6a. **a**, The donor cassette for DHFR-F[1,2] consists of a 40 bp homologous sequence on the left, linker DHFR-F[1,2] ADH_{term} followed by TEF promoter, nourseothricin N-acetyl transferase (NAT1) which confers resistance to nourseothricin, TEF terminator and finally a 40 bp homologous sequence on the right. Both fragments have a 10 amino acid (Gly-Gly-Gly-Gly-Ser)2x linker. **b**, The donor cassette for DHFR-F[3] consists of the same factors, excluding the selecting marker: it contains hygromycin B phosphotransferase, which confers resistance to hygromycin B, as a selection marker.

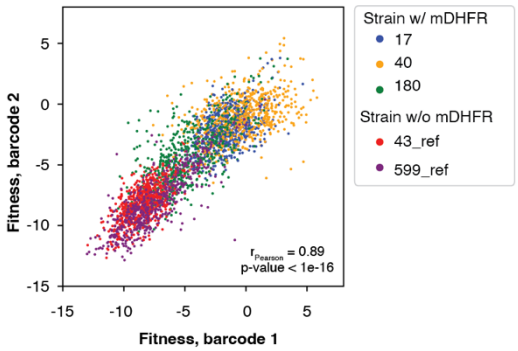

Fig. 3.8. Within-well replicability of fitness estimates.

Three strains (#17, #40, and #180) were independently labeled with two barcodes prior to being pooled and tagged with the mDHFR fragments. Two other strains (#43 and #599) were also labeled with two barcodes but were not tagged with mDHFR. The fitness of these strains served as reference for "no PPI" or fitness in the absence of mDHFR complementation. Shown correlation of the fitness estimates from the two unique barcodes under methotrexate. Expectedly, strains without mDHFR have the lowest fitness.

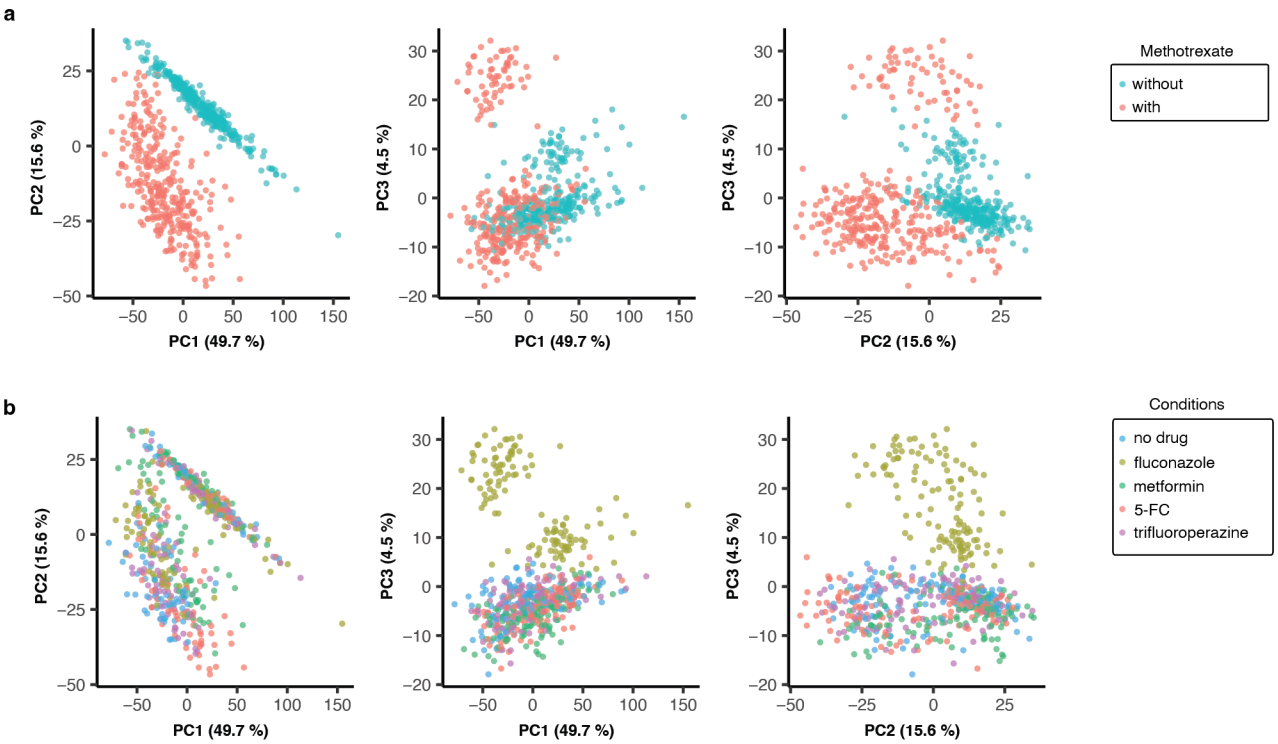

Fig. 3.9. Principal component analyses for the PPI measured across different conditions.

a), Projection of the 62 PPI (each as an average over 353 backgrounds) onto the first three principal components. Variances explained are PC1: 69.8%, PC2: 5.07%, and PC3: 3.39%. Colors correspond to with and without methotrexate. **b**, Similar to panel a, but the colors correspond to different drug conditions.

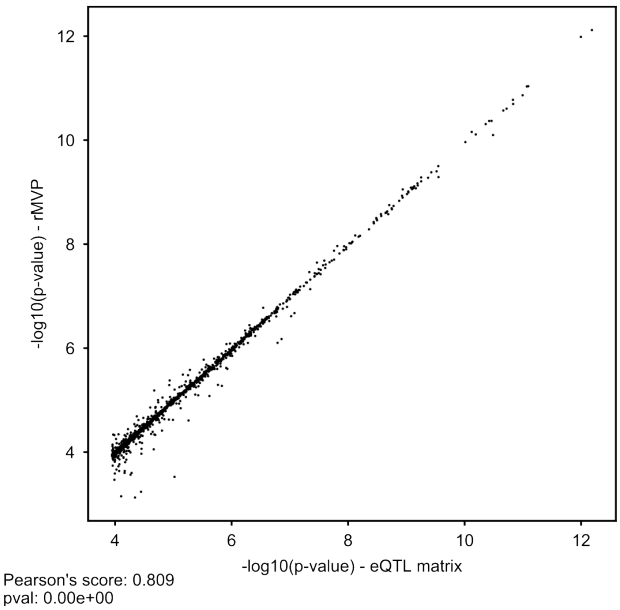

Fig. 3.10. Correlation between two fine-mapping approaches.

We perform piQTL mapping using matrix eQTL and rMVP that also accounts for the correlation in the genotypic matrix. Due to the minimal presence of LD, the association with and without LD correction are strongly correlated. We only select as piQTL those SNPs that have FDR<0.05 ($\sim$P-value=$10^{-3.9}$) in both approaches.

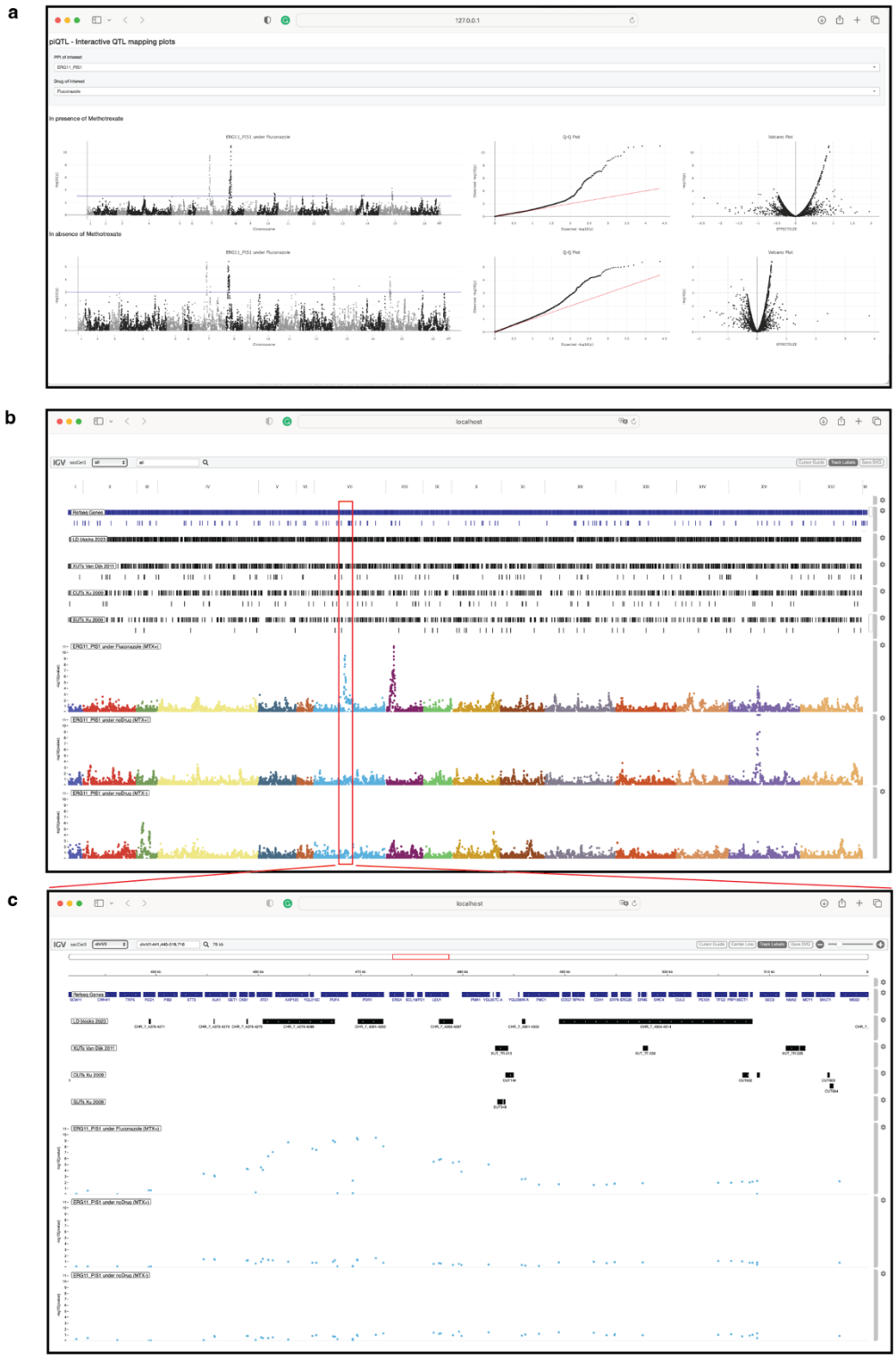

Fig. 3.11. Interactive piQTL webserver.

All PPI genome-wide association analysis results are available on in interactive server showing the statistical significance of the association (Q-Q and volcano plots in **panel a**)). A genome-browser is also provided showing the comparison between the associations under drug+methotrexate, no drug+methotrexate, and no methotrexate (**panel b**). Zoom-in showing details of the genome annotation, residual LD blocks (LD r2> 0.75) and annotations of non-coding RNA SUTs, XUTs, and CUTs (**panel c**).

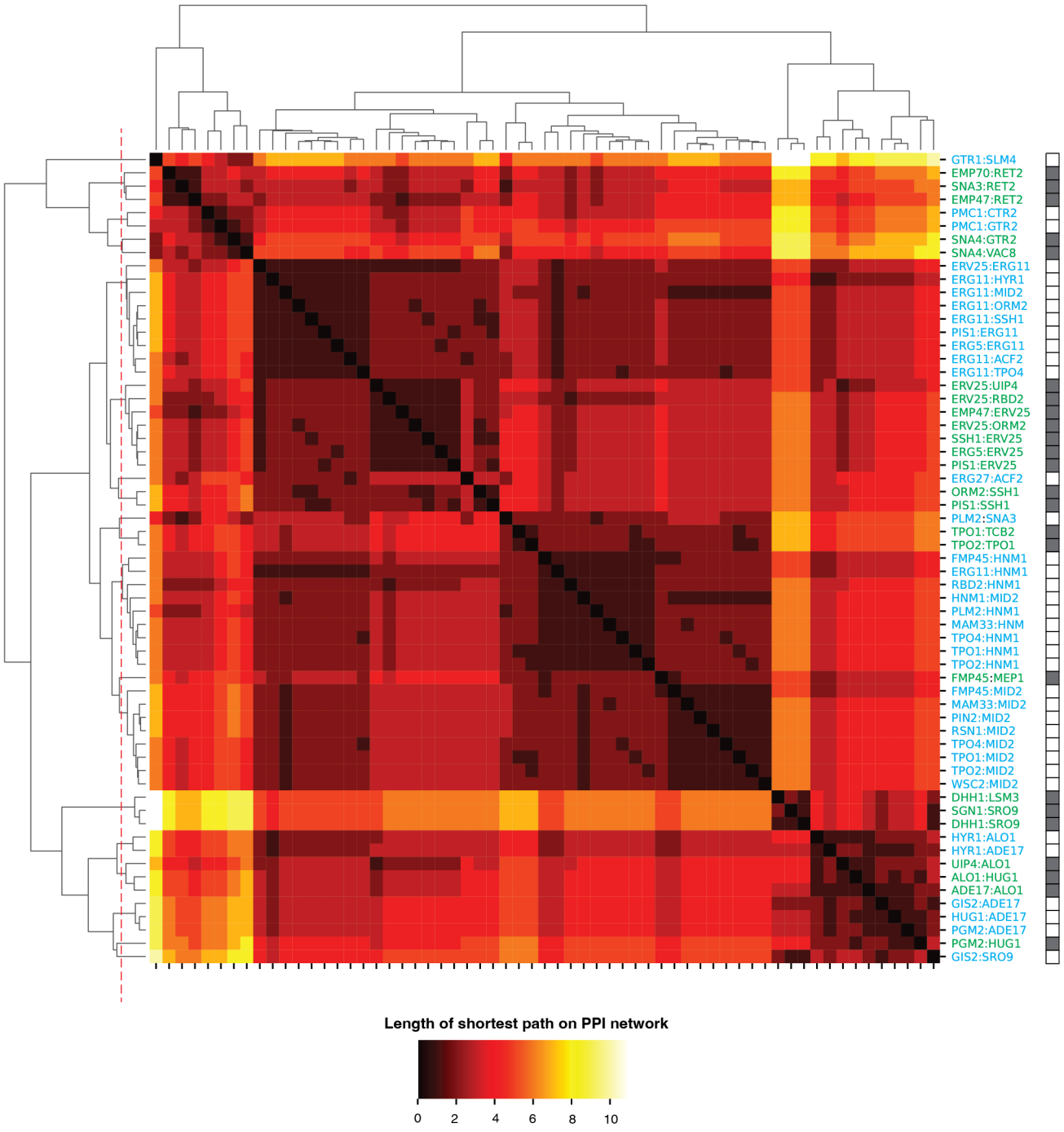

Fig. 3.12. Clustering of the 62 reporter bait/prey pairs on the PPI network.

Distance matrix is based on shortest path between PPIs. The hierarchical tree and the PPI groups (rightmost bar) define the y-axes of the heatmaps in Figures 3.3 and **??**.

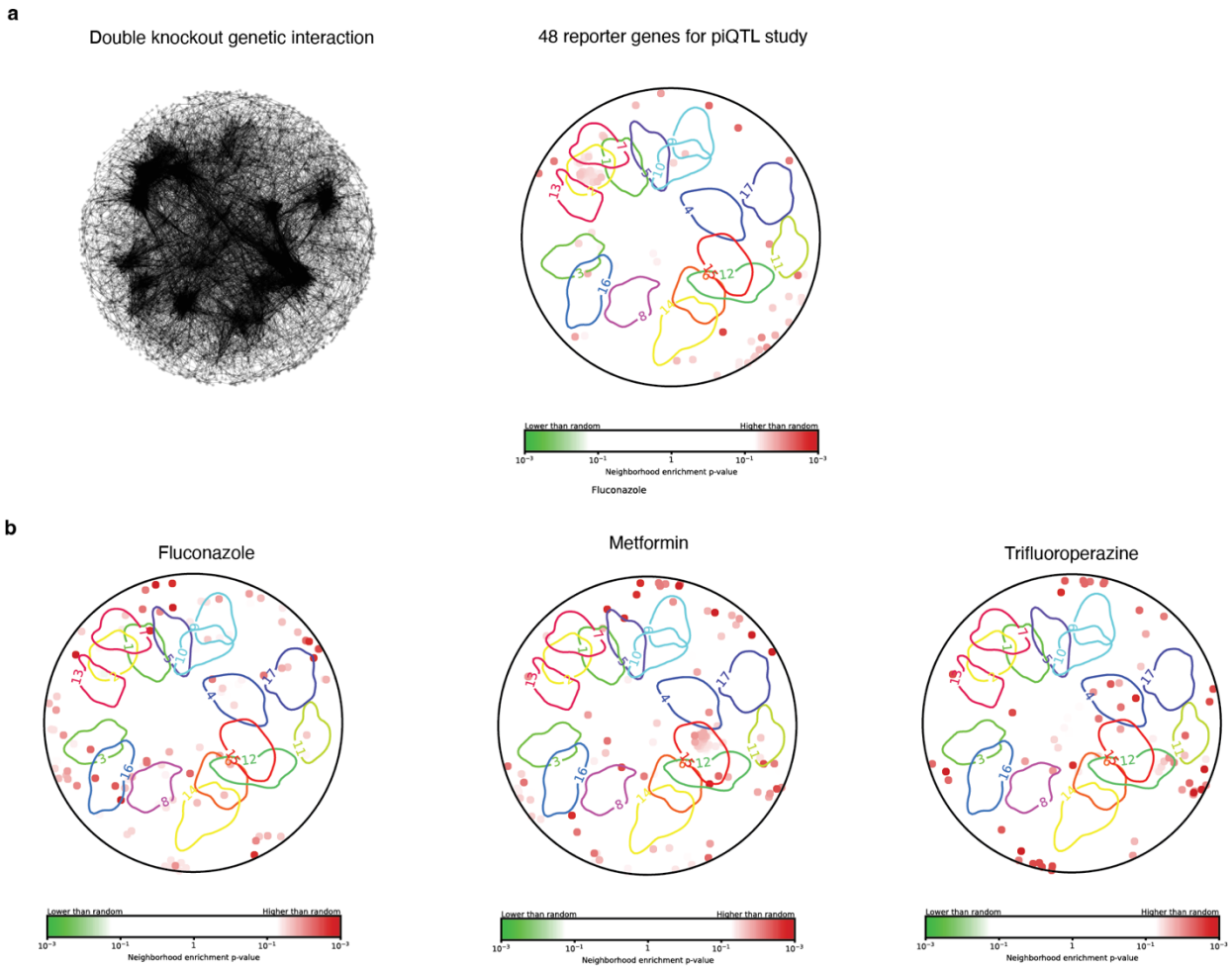

Fig. 3.13. Clustering of piQTL on the gene interaction network defined by double-gene knockouts.

Gene	ORF	Uniprot ID	Functional annotation in Uniprot
ACF2	YLR144C	Q12168	Intracellular beta-1,3-endoglucanase.
ADE16	YLR028C	P54113	Enzyme of 'de novo' purine biosynthesis; ADE16 has a paralog, ADE17.
ADE17	YMR120C	P38009	Enzyme of 'de novo' purine biosynthesis; ADE17 has a paralog, ADE16.
ALO1	YML086C	P54783	Can oxidize L-gulono-1,4-lactone as well as D-arabinono-1,4-lactone and L-galactono-1,4-lactone.
CTR2	YHR175W	P38865	Provides bioavailable copper via mobilization of vacuolar copper stores and export to the cytoplasm.
DED1	YOR204W	P06634	ATP-binding RNA helicase involved in translation initiation.
DHH1	YDL160C	P39517	ATP-dependent RNA helicase involved in mRNA turnover, and more specifically in mRNA decapping by activating the decapping enzyme DCP1.
EMP47	YFL048C	P43555	Involved in the secretion of glycoproteins and in nucleus architecture and gene silencing.
EMP70	YLR083C	P32802	With TMN2 and TMN3, plays a critical role in the late stages of a nutrient-controlled pathway notably regulating FLO11 gene expression.
ERG11	YHR007C	P10614	Lanosterol 14-alpha demethylase; part of the third module of ergosterol biosynthesis pathway that includes the late steps of the pathway.
ERG27	YLR100W	Q12452	3-keto-steroid reductase; part of the third module of ergosterol biosynthesis pathway that includes the late steps of the pathway.
ERG5	YMR015C	P54781	C-22 sterol desaturase; part of the third module of ergosterol biosynthesis pathway that includes the late steps of the pathway.
ERV25	YML012W	P54837	Constituent of COPII-coated endoplasmic reticulum-derived transport vesicles.
FMP45	YDL222C	Q07651	Involved in sporulation and affects the sphingolipid composition of the plasma membrane.
GIS2	YNL255C	P53849	May act in the sexual differentiation pathway.
GTR1	YML121W	Q00582	GTPase component of the GSE complex, a GTPase complex required for intracellular sorting of GAP1 out of the endosome.
GTR2	YGR163W	P53290	GTPase. Component of the GSE complex, a GTPase complex required for intracellular sorting of GAP1 out of the endosome.
HNM1	YGL077C	P19807	Sole choline transporter in yeast.
HUG1	YML058W-A	Q6Q5K6	Involved in the MEC1-mediated checkpoint response to DNA damage and replication arrest.

Fig. 3.14. Genes tagged in the piQTL study (1/3)

HYR1	YIR037W	P40581	Involved in oxidative stress response and redox homeostasis. Functions as a sensor and transducer of hydroperoxide stress.
LAT1	YNL071W	P12695	The pyruvate dehydrogenase complex catalyzes the overall conversion of pyruvate to acetyl-CoA and CO2.
LSM3	YLR438C-A	P57743	Component of LSm protein complexes, which are involved in RNA processing and may function in a chaperone-like manner.
MAM33	YIL070C	P40513	Not known. Binds to the sorting sequence of cytochrome b2.
MEP1	YGR121C	P40260	Transporter for ammonium (both charged and uncharged NH3 and NH4) to use as a nitrogen source.
MID2	YLR332W	P36027	Cell wall stress sensor. Involved in activation of a response that includes both stress-related increased chitin synthesis and the MPK1 mitogen-activated protein kinase cell integrity pathway.
MSC7	YHR039C	P38694	Not known. MSC7 mutants are defective in directing meiotic recombination events to homologous chromatids.
ORM2	YLR350W	Q06144	Component of the SPOTS complex that acts as a negative regulator of sphingolipid synthesis.
PGM2	YMR105C	P37012	Major phosphoglucomutase isozyme that catalyzes the reversible interconversion of glucose 1-phosphate and glucose 6-phosphate.
PIN2	YOR104W	Q12057	Not known. Seems to be able to provoque the non-Mendelian trait [PIN+] which is required for the de novo appearance of the [PSI+] prion.
PIS1	YPR113W	P06197	Phosphatidylinositol synthase; required for biosynthesis of phosphatidylinositol.
PLM2	YDR501W	Q04383	Binds to the promoters of genes with functions important for the G1/S (start) transition.
PMC1	YGL006W	P38929	This magnesium-dependent enzyme catalyzes the hydrolysis of ATP coupled with the transport of calcium.
RBD2	YPL246C	Q12270	Probable serine protease.
RET2	YFR051C	P43621	The coatomer is a cytosolic protein complex that binds to dilysine motifs and reversibly associates with Golgi non-clathrin-coated vesicles via the Golgi up to the trans Golgi network.
RPN2	YIL075C	P32565	Acts as a regulatory subunit of the 26S proteasome which is involved in the ATP-dependent degradation of ubiquitinated proteins.
RSN1	YMR266W	Q03516	Acts as an osmosensitive calcium-permeable cation channel.
SGN1	YIR001C	P40561	mRNA-binding protein that may play a role in modulating the expression of cytoplasmic mRNA.
SLM4	YBR077C	P38247	Component of the GSE complex, a GTPase complex required for intracellular sorting of GAP1 out of the endosome.

Fig. 3.15. Genes tagged in the piQTL study (2/3)

SLY1	YDR189W	P22213	Able to suppress the functional loss of YPT1. SLY1 is essential for cell viability.
SNA3	YJL151C	P14359	Protein involved in efficient MVB sorting of proteins to the vacuole; may function as an RSP5 adapter protein for MVB cargos.
SNA4	YDL123W	Q07549	Protein of unknown function; localized to the vacuolar outer membrane; predicted to be palmitoylated.
SNQ2	YDR011W	P32568	Could be an ATP-dependent permease. Confers hyper-resistance to the mutagens 4-nitroquinoline-N-oxide (4-NQO) and triaziquone.
SRO9	YCL037C	P25567	May overlap in function with tropomyosin and may be involved in organization of actin filaments. Acts as a multicopy suppressor of RHO3 mutation.
SSH1	YDR086C	P35179	Part of the Sec61 complex, which is the major component of channel-forming translocon complex that mediates protein translocation across the endoplasmic reticulum (ER).
TCB2	YNL087W	P48231	May play a role in membrane trafficking.
TPO1	YLL028W	Q07824	Cell membrane polyamine/proton antiporter, involved in the detoxification of excess polyamines in the cytoplasm.
TPO2	YGR138C	P53283	Cell membrane polyamine/proton antiporter, involved in the detoxification of excess polyamines in the cytoplasm.
TPO4	YOR273C	Q12256	Cell membrane polyamine/proton antiporter, involved in the detoxification of excess polyamines in the cytoplasm.
TSC10	YBR265W	P38342	Catalyzes the reduction of 3-ketodihydrosphingosine (KDS) to dihydrosphingosine (DHS).
UIP4	YPL186C	Q08926	Protein required for nuclear envelope integrity; involved in distribution of nuclear pore complexes.
VAC8	YEL013W	P39968	Functions in both vacuole inheritance and protein targeting from the cytoplasm to vacuole.
VPH1	YOR270C	P32563	Subunit of the V0 complex of vacuolar(H+)-ATPase (V-ATPase), a multisubunit enzyme composed of a peripheral complex (V1) and a membrane integral complex (V0).
WSC2	YNL283C	P53832	Sensor-transducer of the stress-activated PKC1-MPK1 signaling pathway.

Fig. 3.16. Genes tagged in the piQTL study (3/3)

Chapitre 4

Discussion & Perspectives

Dans cette discussion, je résumerai de manière indépendante les résultats, limites, approches alternatives et perspectives associés à chaque article.

Par la suite, je proposerai une discussion générale qui mettra en avant la complémentarité des deux approches protéome-centrées proposées, et qui explicitera le choix des stratégies mises en place pour l'étude des phénotypes complexes par le biais de méthodes bio-informatiques.

Pour conclure, je mettrai en avant les différentes contributions de ma recherche dans le domaine de la bio-informatique et établirai le bilan de mes expertises à l'issue de mes 5 années de doctorat.

4.1. Étude comparative de la propension d'agrégation et de la tolérance de mutation des protéines chez le rat-taupe nu et la souris

4.1.1. Originalité de l'étude

Dans le chapitre 2, nous avons proposé une stratégie de génomique comparée pour étudier différentes propriétés intrinsèques du protéome, à savoir la propension d'agrégation et la stabilité des propriétés d'agrégation de la protéine aux mutations de sa séquence nucléotidique. Cette stratégie visait à comprendre les différences d'espérance de vie maximale entre le rat taupe nu et la souris. Elle a permis d'identifier des ensembles de protéines avec des différences significatives de propension d'agrégation et de tolérance de mutation entre le rat taupe nu et la souris. Certaines des protéines identifiées chez le rat-taupe nu sont impliquées dans des

processus biologiques précédemment identifiés comme liés aux manifestations du vieillissement. On peut aussi noter que certaines de ces protéines ont été précédemment identifiées comme bio-marqueurs pour des maladies associées à l'âge chez l'homme (ex : l'acide ceramidase ASAH1 pour la maladie d'Alzheimer (Parveen et al., 2019)). Grâce à ces résultats concordant avec la littérature, nous sommes confiants de proposer des protéines candidates d'intérêt pour caractériser le vieillissement dans des études moléculaires plus ciblées.

4.1.2. Limites de l'étude et stratégies alternatives

Choix stratrégiques dans l'étude des protéines orthologues partagées entre le rat-taupe nu et la souris

A l'issue de notre étude, nous avons pu réaliser des analyses comparatives pour les protéines orthologues partagées entre le rat-taupe nu et la souris. Nous avons restreint nos analyses à 9,522 paires de protéines orthologues bien que nous ayions initialement identifié un total de 13,806 protéines. La réduction de notre jeu de données passe d'abord par une première étape de vérification de la qualité de nos paires de protéines orthologues, où nous avons exclu toute paire dont le taux de similarité est inférieur à 60% et contenant plus de 20% de "gaps". Les gaps représentent un saut dans un alignement de protéines. Ces sauts d'alignements peuvent correspondre à des événements d'insertion-délétion dans l'une des séquences de l'alignement. Ici, nous nous sommes donc concentrés uniquement sur les protéines qui sont communes au deux espèces afin d'étudier les mécanismes communs régissant la modulation de leur espérance de vie maximale. Cela signifie qu'il n'est pas possible dans le cadre de cette étude d'identifier des protéines qui sont spécifiques au rat-taupe nu qui pourraient expliquer sa plus grande espérance de vie.

Ensuite, nous avons exclu les protéines membranaires et transmembranaires de notre analyse car le logiciel, Tango prédit uniquement la propension d'agrégation pour des protéines globulaires (Fernandez-Escamilla et al., 2004). Cela se justifie par le fait que les tests de performance du logiciel ont été réalisé sur un jeu de données contrôlé de protéines globulaires amylogéniques. Les protéines amylogéniques sont des protéines contenant des segments peptidiques connus pour être impliqués dans des phénomènes d'agrégation protéique (Fernandez-Escamilla et al., 2004), et il est donc difficile de généraliser la performance de Tango au delà de ce type de protéine.

Stratégies d'optimisation pour la prédiction de propension d'agrégation dans des jeux de données massifs

Dans notre étude, nous avons choisi de calculer la propension d'agrégation à l'aide du logiciel Tango, qui permet de prédire la structure secondaire la plus probable d'une séquence peptidique, dans des conditions environnementales pré-déterminées. Pour chaque acide aminé, l'algorithme de Tango estime la probabilité de formation d'hélices α ou feuillets β, en prenant en compte ses propriétés physico-chimiques et celles de ses voisins proches. L'hypothèse sous-jacente de l'algorithme de Tango est qu'au sein de la séquence primaire d'une protéine, si certains segments sont impliqués dans la formation de feuillets β, alors ces régions de la protéine peuvent potentiellement s'agréger. Le logiciel estime alors pour chaque résidu un pourcentage qui représente sa probabilité à participer à la formation d'un feuillet β et donc de participer à la propension d'agrégation globale de la protéine. Pour calculer la propension d'agrégation au niveau de sa séquence entière, ou au niveau de ses domaines fonctionnels, il a été nécessaire de créer des métriques spécifiques, permettant de caractériser la probabilité de formation d'agrégats protéiques pour une séquence codante donnée, de manière globale et locale. Nous avons calculé ce score en faisant la moyenne des scores de propension d'agrégation par résidu, normalisé en fonction deu type de séquence prise en compte (la séquence entière ou le domaine).

Tango est un logiciel publié en 2004, il fait partie des premiers outils permettant de caractériser la propension à l'agrégation des protéines et il a été utilisé pour répondre à des questions portant sur les variations de propension d'agrégation des protéines suite à l'introduction de mutations ponctuelles dans la séquence codant la protéine (plus particulièrement des substitutions non synonymes) (Martinez-Rivas et al., 2022; Melnik et al., 2022; Törner et al., 2022). Les mutations testées avec le logiciel Tango peuvent avoir été préalablement identifiées dans des études de conservation de séquences entre espèces, ou bien elles peuvent provenir d'expériences de mutagenèse dirigées qui permettent d'étudier la robustesse du maintien de la conformation protéique suite à de tels changements. Parmi les limites de ce logiciel, on peut constater que les paramètres environnementaux (température, pH et concentration, fixés par l'utilisateur) ne sont pas dynamiques, ce qui signifie qu'on ne peut étudier la propension d'agrégation que de manière statique. Pourtant, l'état conformationnel d'une protéine dépend de son contexte environnemental. Par exemple, au sein d'une cellule, les conditions de pH et de concentration moléculaire peuvent changer en fonction du compartiment cellulaire où les protéines se trouvent (Balut et al., 2008).

Comme souligné dans la section 1.1.4, il existe d'autres alternatives d'algorithmes pour l'estimation de propension d'agrégation des protéines. Dans l'étude de Tsolis et collègues

(Tsolis et al., 2013), les auteurs ont réalisé une analyse comparative des logiciels permettant la prédiction de la propension d'agrégation. On constate que les valeurs moyennes entre sensibilité et spécificité (score Q dans le Tableau 2) sont relativement proches (Q=57.32% pour Aggrescan et Q=54% pour Tango). Cette étude met en avant qu'Aggrescan et Tango sont deux méthodes aux performances équivalentes. Pour proposer des métriques plus robustes de prédiction de la propension d'agrégation, il serait donc pertinent d'utiliser une méthodologie qui repose sur l'utilisation d'algorithmes où différents modèles de formations d'agrégats protéiques sont pris en compte.

Le logiciel AmylPred2 propose cette approche en calculant des scores de propension d'agrégation par résidu, à l'aide de différents logiciels (dont Aggrescan et Tango). On constate que son score Q est plus haut que ceux d'Aggrescan et Tango (Q=62%). Ceci démontre que cette méthode, proposant des scores de propension d'agrégation et s'appuyant sur plusieurs méthodes, permet des prédictions plus robustes que si on utilisait les méthodes de manière indépendante. Cet outil est proposé sur un serveur web, ce qui permet à l'utilisateur de ne pas avoir besoin de faire par lui-même les étapes d'installation de tous les logiciels nécessaires pour l'utiliser. Cependant, ce choix de distribution n'est pas adéquat pour une utilisation à large échelle, avec un nombre important de protéines à étudier. Par exemple, dans notre étude, notre approche pan-génomique repose sur l'identification d'un grand nombre de protéines orthologues entre la souris et le rat-taupe nu ($\sim$ 10,000 protéines). Les données d'entrée pour AmylPred2 (les séquences primaires des protéines au forma FASTA) doivent actuellement être copiées-collées sur le serveur-web, sans possibilité d'envoyer les fichiers d'entrée de manière automatisée. Aussi, même si AmylPred2 propose 11 méthodes différentes pour calculer des scores de propension d'agrégation robuste, c'est à l'utilisateur que revient le choix de prendre en compte ou non toutes les méthodes proposées. Si tous les logiciels étaient sélectionnés, cela ajouterait un temps de calcul conséquent dans l'obtention des scores finaux de propension d'agrégation pour l'ensemble des protéines qu'on souhaite étudier.

En prenant en compte ces considérations, nous avons préféré choisir un logiciel performant pouvant être utilisé localement, dont l'utilisation par ligne de commande permet la création de scripts automatisables pour un calcul rapide de scores de propension d'agrégation. Ceci a permis l'étude globale (au niveau de la séquence entière) et locale (au niveau des domaines fonctionnels) de la tendance d'agrégation d'une importante quantité de protéines en commun entre le rat-taupe nu et la souris. De plus, ces scores de propension d'agrégation sont nécessaires pour estimer la tolérance de mutation qui est calculé à partir des séquences générées par la mutagenèse computationnelle à large échelle (plusieurs dizaines de millions de séquences). Grâce à l'utilisation de Tango directement avec un script shell, il est possible

de paralléliser les calculs via l'utilisation de noeuds de calcul. Dans notre cas, nous avons utilisé les ressources de l'Alliance de recherche Numérique du Canada pour calculer l'ensemble des scores de propension d'agrégation pour toutes les protéines et leurs séquences mutantes pour identifier des différences de tolérance de mutation entre le rat-taupe nu et la souris.

Cette étude est un exemple de méthodologie robuste et reproductible pour faire le traitement massif de données -omiques.

Commentaires sur le score caractérisant la tolérance de mutation

Dans notre étude, l'intégrité fonctionnelle d'une protéine se traduit par sa capacité à garder une conformation native stable qui la "protège" de changements au niveau de sa séquence en acides aminés pouvant conduire à des défauts de repliement et à la formation d'agrégats protéiques. Dans un premier temps, notre hypothèse de travail s'appuie principalement sur les mutations qui ne changent pas les propriétés de propension d'agrégation initiales de la protéine.

C'est pourquoi la définition de notre score de tolérance de mutation repose sur le calcul des différences de tolérance de mutation entre :

- les séquences mutées suite à un événement de substitution unique apparu au niveau des gènes ou transcrits codant pour les protéines étudiées.
- les séquences initiales de ces protéines.

Nous avons proposé deux types de score de tolérance de mutation :

- la tolérance de mutation "stricte" (voir Equation (2.6)), où seules les mutations qui n'ont aucun impact sur la mesure initiale de propension d'agrégation pour la séquence originale de la protéine, sont comptées,
- la tolérance de mutation "clémente" (voir Equation (2.8)), où en plus des mutations décrites plus hauts, on comptabilise également les mutations qui contribuent à une diminution de la propension d'agrégation. Ces mutations sont considérées comme bénéfiques puisqu'elles contribueraient à la stabilité de la protéine.

L'utilisation de ces deux définitions a permis de mettre en avant que l'étude de la tolérance de mutation stricte était suffisante pour observer des différences au niveau de la tolérance de mutation entre le rat-taupe nu et la souris.

Enfin, dans une étude ultérieure, il serait intéressant d'essayer de complexifier l'équation modélisant la tolérance de mutation des protéines. En effet, ici on ne considère que l'impact des substitutions sur la propension d'agrégation. Il serait également intéressant d'étudier les impacts d'événements d'insertion-délétion de résidus dans le contexte de propension d'agrégation. Par exemple, dans notre étude, nous avons mis en évidence qu'il existait des événements évolutifs spécifiques chez le rat-taupe nu. Nous avons observé que

l'insertion de nucléotides au sein d'un domaine fonctionnel permettait de contribuer à la stabilisation de ses propriétés de propension d'agrégation.

Stratégies permettant l'étude des différences d'espérance de vie maximale au sein de différents taxa phylogénétiques

Notre approche de génomique comparée prend le parti de comparer les protéomes de deux organismes phylogénétiquement proches dont la différence d'espérance de vie est importante. Cette stratégie a apporté des résultats intéressants, mais il est possible qu'ils ne se généralisent pas aux autres espèces, ceci est donc une avenue à explorer dans l'avenir. Dans un premier temps, nous pourrions proposer la mise à l'échelle de la méthode implémentée à l'ensemble des espèces eucaryotes pour lesquelles on dispose d'un génome de bonne qualité couplé à leurs données d'espérance de vie maximale. Cette approche aurait pour but de ne pas choisir par "picorage" les espèces à comparer. Pour différencier les espèces avec une espérance de vie maximale "courte" (SL, pour *Short-Lived species*) et celles avec une espérance de vie maximale "longue" (LL, pour *Long-Lived species*), il serait idéal de proposer un seuil permettant de différencier les deux groupes. Par exemple, on pourrait utiliser le quotient de longévité (Austad and Fischer, 1991), qui représente simplement le ratio entre l'espérance de vie maximale d'une espèce normalisé par son poids de corps. Des études comme celles de Farré et collègues Farré et al., 2021 ont utilisé cette métrique pour différencier deux groupes d'espèces avec des espérance de vie extrême, et identifié environ 2,000 gènes spécifique à la longévité. Il faut cependant constater que l'étude de Farré et collègues propose leur étude sur un nombre restreint d'organismes (N=12, 6 par groupe de longévité extrême). Pour faire une étude la plus exhaustive possible, il serait donc intéressant d'étudier l'ensemble des espèces ayant des données génomiques vérifiées, ainsi que des données d'espérance de vie maximale, afin de pouvoir proposer des groupes SL et LL de taille plus importante.

Quelques résultats préliminaires ont été générés sur cette question. Après vérification de la qualité des génomes pour l'ensemble des espèces pour lesquels on possèdes des informations sur l'espérance de vie maximale (provenant de la base de données AnAge (Tacutu et al., 2018), 185 espèces ont été sélectionnées. A l'aide du logiciel TimeTree (Kumar et al., 2017), j'ai pu utilisé l'arbre phylogénétique entre ces espèces (voir Figure 4.1). On constate que les données d'espérance de vie maximale ont été collectées préférentiellement dans 5 groupes taxonomiques particuliers : mammifères (94), des oiseaux (51), des téléostéens (19) des reptiles (11) et des insectes (4). À partir de cet ensemble d'espèces, j'ai souhaité savoir s'il était possible d'établir deux groupes avec des différences d'espérance de vie maximale significativement différentes : les groupes SL et LL.

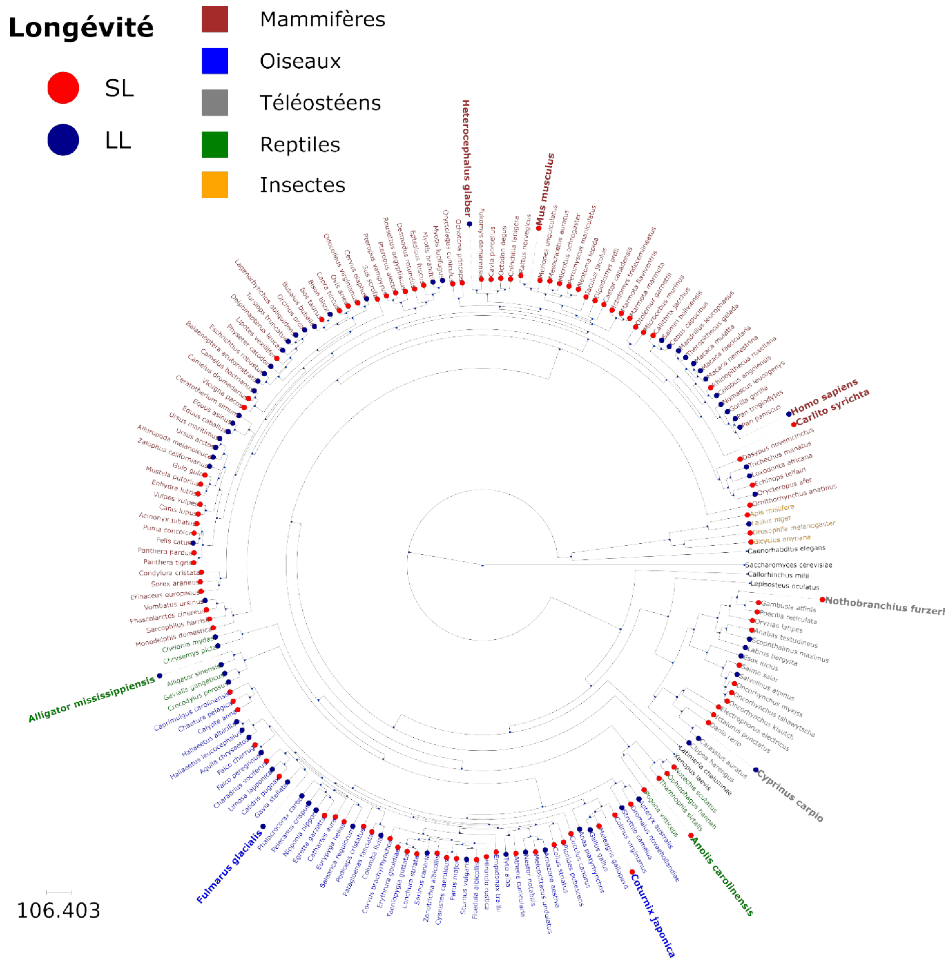

Fig. 4.1. Arbre phylogénétique des espèces eucaryotiques associées à leurs informations d'espérance de vie maximale.

(Créée par Savandara Besse)
L'arbre phylogénétique a été généré à partir de la ressource web TimeTree en donnant la liste spécifique des espèces qui nous intéressaient. Les mesures de longévité ont été recueillies sur la base de données AnAge et reposent sur la durée de vie maximale des espèces à l'état sauvage ou captif.

Dans une première approche, nous avons tenté de faire la distinction entre ces deux groupes sans pour autant faire de distinction entre les groupes taxonomiques. À l'aide d'une valeur seuil représentée par la valeur moyenne d'espérance de vie maximale dans l'ensemble de notre jeu de données (Figure 4.2A), nous avons obtenu un groupe SL de 114 espèces et un groupe LL de 77 espèces. Cependant la comparaison des groupes SL et LL n'est pas possible, car ces derniers n'ont pas une valeur de distribution de distances évolutives similaires (Figure 4.2B) (test de Mann-Whitney : p-valeur < 0.05). Ainsi, il n'est pas possible d'étudier ces deux groupes pour identifier des différences associées en fonction d'une valeur seuil dépendant uniquement de l'espérance de vie maximale. Les écarts conséquents des distances phylogénétiques entre les deux groupes peuvent s'expliquer par la présence

d'espèces ayant une évolution trop différente par rapport au reste du groupe, notamment la levure *S. cerevisiae*, seul représentant unicellulaire de ce jeu de données. Les espèces ayant peu de représentants de leurs groupes taxonomiques (en noir dans Figure 4.1) doivent donc être exclues de l'analyse.

Dans un second temps, pour établir des groupes SL et LL comparables, j'ai tenté de créer des sous-groupes qui possèdent des distances évolutives similaires. C'est pourquoi j'ai d'abord séparé notre jeu de données en fonction des annotations taxonomiques pour réduire les distances liées à l'évolution dans chacun de ces groupes. J'ai créé les groupes SL et LL à l'aide de la valeur moyenne d'espérance de vie maximale respective à chaque groupe (voir Tableau 1). De manière similaire à la première approche, j'ai comparé les distributions de distances phylogénétiques pour chaque groupe taxonomique. Deux groupes taxonomiques demeurent dominants dans notre jeu de données, les oiseaux et les mammifères. Pour le groupe des oiseaux (Figure 4.2C), les groupes SL et LL ont une similarité de distributions des distances évolutives (test de Mann-Whitney : p-valeur > 0.05), ce qui suppose qu'on peut les comparer. Pour le groupe des mammifères (Figure 4.2D), il a été nécessaire de réaliser une étape de sous-échantillonnage pour obtenir des groupes plus similaires. Les autres groupes taxonomiques n'ont pas un nombre suffisant d'espèces dans leurs groupes SL et LL pour permettre une analyse comparée de leurs protéines orthologues.

Pour générer des groupes de sous-échantillons pour le groupe des mammifères, j'ai implémenté une méthode de Monte-Carlo qui permet de créer différents sous-groupes SL dont on diminue progressivement le nombre d'espèces afin de trouver le sous-échantillonnage idéal afin de le comparer au groupe LL sans biais phylogénétiques. La similarité des distributions de leurs distances phylogénétiques est évaluée grâce à un test statistique non paramétrique de Mann-Whitney. Si le résultat de ce test est supérieur au seuil de significativité, on considère que les deux groupes sont similaires. Sur plus de 10,000 simulations, un seul résultat a permis d'identifier un sous-groupe SL qui peut être comparable au groupe LL. On constate que la création de ce sous-groupe SL permet de diminuer les écarts de distances liée à l'évolution entre ces deux groupes. On a ici obtenu des groupes SL et LL comparables pour les groupes des oiseaux et des mammifères. Toutefois, on constate qu'il reste des différences visibles entre leurs distributions de distances phylogénétiques (FIGURE 4.2C et 4.2E,). Ces différences peuvent s'expliquer par la prédominance de l'ordre des rongeurs dans le groupe SL et celle de l'ordre des primates dans le groupe LL. Nous concluons que pour le groupe des mammifères, il faudrait étudier les ordres des rongeurs et des primates de manière séparée afin que la disparité de leurs distances phylogénétiques n'impacte pas la comparaison des groupes SL et LL des espèces restantes.

Ainsi, la détermination des groupes SL et LL reposant sur la valeur moyenne de longévité dans chaque groupe taxonomique n'apparait pas comme la solution idéale pour créer des groupes comparables avec des distances phylogénétiques proches. Pour créer des groupes SL et LL possédant des distributions de distances évolutives strictement identiques entre les espèces, il sera nécessaire de modifier la stratégie de sous-échantillonnage pour induire la création de sous-groupes SL et LL ayant strictement la même distribution de distances phylogénétiques.

Dans l'article présenté dans le chapitre 2, nous avons opté pour une stratégie simplifiée où l'on étudie une paire d'organismes au sein d'un groupe taxonomique avec des longévités totalement opposées. Cette stratégie nous permet de considérer que nos deux espèces sont indépendantes. Cette méthodologie et les analyses proposées pour étudier les différences d'espérance de vie pour deux espèces appartenant aux rongeurs pourraient donc être facilement étendue pour d'autres paires d'organismes provenant de différents taxa phylogénétiques : des exemples de paires d'organismes sont indiqués en gras dans l'arbre phylogénétique des espèces (Figure 4.1). Grâce à ces paires d'organismes, on pourrait identifier des adaptations spécifiques associées à leur longévité pour chaque groupe taxonomique dominant identifié.

4.1.3. Perspectives

Validation des protéines candidates associées à l'espérance de vie maximale

Parmi les perspectives envisagées, différentes approches peuvent être adoptées pour aller plus loin dans l'étude des protéines identifiées dans l'analyse décrite dans le chapitre 2.

Les approches expérimentales. Il serait intéressant de déterminer si les gènes codant pour ces protéines avec des différences de propension d'agrégation pourraient être utilisés comme bio-marqueurs du vieillissement ou de la longévité. Pour cela, on peut réaliser des études moléculaires ciblées comme des expériences de "knock-out" des gènes responsables de la synthèse de ces protéines candidates afin d'observer les conséquences de l'inactivation de ces gènes sur la longévité de l'organisme modèle dans lequel ces inactivations sont réalisées. Comme organisme modèle, il serait pertinent dans un premier temps de réaliser ces expériences soit directement chez le rat-taupe nu (mais relativement peu de laboratoires expérimentaux travaillent sur cet organisme modèle), ou bien chez la souris.

Les approches basées sur l'étude des réseaux. Pour ces analyses il sera nécessaire d'inférer les réseaux de PPIs chez le rat-taupe nu et la souris (par exemple avec la base de données STRING). Dans un premier temps, il sera intéressant de comparer les réseaux de chaque espèce pour vérifier si ces réseaux ont des topologies similaires ou non. Il sera

également intéressant d'étudier le réseau d'interaction protéine-protéine en commun pour les deux espèces et d'étudier si les protéines candidates que nous avons identifiées participent activement ou non à la topologie de ce réseau et à sa résilience. La création de ce réseau de PPI serait une manière de représenter le phénotype du vieillissement, et pourrait nous renseigner sur le contexte fonctionnel ou biologique auquel les protéines candidates sont associées.

Les études de cartographie génétique au sein de cohortes humaines centenaires ou ayant des maladies neuro-dégénératives. Enfin, il serait aussi intéressant d'étudier le rôle des gènes associés à ces protéines chez l'humain. La population humaine a une espérance de vie maximale d'environ 120 ans, ce qui la catégorise comme une espèce avec une espérance de vie "longue". En génétique des populations, des GWAS ont été réalisées sur des cohortes humaines centenaires afin d'identifier les loci associés à leur longévité (Franceschi et al., 1995). D'autres études GWAS ont également été conduites sur des cohortes présentant des maladies neuro-dégénératives, comme les maladies d'Alzheimer et de Parkinson pour identifier les loci associés à ces maladies (Ramanan and Saykin, 2013). Ainsi, il serait intéressant de voir si les gènes orthologues chez l'humain de nos gènes candidats ont été précédemment identifiés au sein de ces cohortes populationnelles. Si tel est le cas, il serait pertinent d'identifier les SNPs humains, qu'ils soient au sein de ces gènes ou à leur proximité, et de comprendre leur rôle fonctionnel à l'aide d'études eQTL ou pQTL dans ces cohortes.

On pourrait également s'intéresser à la présence de pressions de sélection entre différentes populations ethniques centenaires via des statistiques mesurant les différences entre population comme la statistique F_{ST}. Il existe différentes cohortes centenaires identifiées dans différentes régions du monde, comme en Italie (Caselli et al., 2018) ou au Japon (Arai et al., 2017; Willcox et al., 2017) et d'autres régions du monde (Robine and Cubaynes, 2017). Dans cette approche, il faudra bien faire attention à exclure les variants génétiques qui seraient spécifiques aux ethnies étudiées pour se concentrer exclusivement sur les variants génétiques qui seraient associés à l'âge. Pour cela, une étude comparative des MAFs des variants génétiques dans ces deux populations pourront permettre d'identifier les SNPs communs aux deux populations.

Enfin, il serait aussi intéressant d'étudier s'il existe des différences dans les manifestations du vieillissement entre les deux sexes. En effet, l'espérance de vie des femmes est, en général, plus haute que celle des hommes (Waldron and Johnston, 1976), il serait donc pertinent d'identifier les signatures génétiques spécifiques à chaque sexe. Pour cela, il serait important de sélectionner des cohortes d'individus avec un ratio homme/femme équilibré

pour proposer des études stratifiées robustes. On pourrait, par exemple, utiliser la cohorte centenaire ELCV (pour étude longitudinale Canadienne sur le vieillissement) pour mener ce type d'études. Une revue récente montre l'intérêt d'identifier les facteurs génétiques liées au sexe pour comprendre la plasticité des phénotypes des maladies associées à l'âge (Hägg and Jylhävä, 2021), il est cependant important de souligner l'importance des composantes environnementales permettant d'expliquer les différences de longévité liées au sexe (Austad and Fischer, 2016; Lemaître et al., 2020).

4.2. La cartographie fine de piQTL chez la levure

4.2.1. Originalité de l'étude

Dans le chapitre 3 , nous avons proposé une nouvelle stratégie pour faire de la cartographie fine de variants génétiques qui s'appuient exclusivement sur la mesure *in vivo* d'interaction protéine-protéine, que nous avons appelé piQTL. Cette approche permet d'étudier la relation génotype-phénotype dans différents contextes environnementaux chez la levure, permettant de mettre en avant l'influence de l'environnement sur le génotype dans ce modèle.

Cette approche propose des innovations majeures :

(1) la mise en place d'une stratégie expérimentale s'appuyant sur des techniques de séquençage à haut débit pour mesurer *in vivo* des interactions protéine-protéine à l'aide de la technique de PCA au sein d'une population génétiquement diversifiée chez *S. cerevisiae*

(2) la mise en place d'un protocole bio-informatique robuste permettant le pré-traitement des données NGS et leur analyse en vue d'estimer les valeurs de *fitness* des différentes souches au sein de notre population

(3) la proposition d'une métrique quantitative robuste pour estimer les changements de force d'interactions d'une PPI (via un proxy dépendant du nombre de complexes d'interaction formés sous environnement de sélection)

(4) la mise en place d'une stratégie d'analyse statistique permettant d'identifier les variants génétiques influençant les changements de force d'interaction des PPI dans un contexte environnemental donné

(5) l'élaboration et l'implémentation de différentes analyses bio-informatiques subséquentes aux résultats issus des analyses piQTL

4.2.2. Limites de l'étude et stratégies alternatives

Choix de l'organisme modèle

L'implémentation des méthodologies expérimentales et bio-informatiques du piQTL s'est réalisée sur l'organisme modèle *S. cerevisiae*. L'utilisation de cet organisme modèle permet

d'étudier à moindre coût des mécanismes biologiques spécifiques à des organismes eucaryotiques, dont certains peuvent être transposés dans des organismes eucaryotiques plus complexes, à l'aide d'approche de génomique comparée.

Cependant, *S. cerevisiae* n'est pas un modèle permettant d'expliquer toute la complexité des organismes eucaryotiques, et notamment celle de l'humain. Il sera donc important de proposer des méthodes expérimentales spécifiques pour la quantification à large échelle des PPIs chez l'humain pour pouvoir transposer l'analyse piQTL. L'utilisation des méthodes CRISPR/Cas9 pourrait permettre la mise en place de la technique PCA dans d'autres organismes modèles.

L'étude d'un nombre restreint de PPIs

Dans notre stratégie expérimentale, on peut considérer que l'étude de 62 PPIs représente une portion assez restreinte de l'interactome de la levure. Il serait donc important de pouvoir augmenter le nombre de PPIs à étudier, mais pour cela il sera nécessaire d'avoir des ressources financières et technologiques importantes pour permettre l'obtention des données de NGS pour l'ensemble de la collection de levures, dans les différents contextes environnementaux et pour chaque PPI. En théorie, pour l'étude de l'interactome entier de la levure, dans toutes nos conditions requises (les 4 médicaments d'intérêt, la présence ou non de la drogue de sélection, pour les 361 souches de notre collection de levures) pour l'approche piQTL, nous aurons besoin de séquencer plus des dizaines de milliers de conditions différentes. L'étape suivante serait de pouvoir mesurer l'interactome entier de cette population (plusieurs milliers de PPIs), pour proposer des méthodologies statistiques similaires à celles des GWAS, mais étudiant ici l'interactome, ce serait le pré-requis nécessaire pour introduire le concept de piWAS (protein interaction Wide Associations Study).

Mesures de la quantification des PPIs pour une mesure fixe de MTX

La quantification des PPIs repose principalement sur l'habileté des cellules à former les complexes de PPIs sous une pression de sélection associée au methotrexate, qui est la drogue nécessaire à métaboliser à l'aide de la restauration de l'activité de l'enzyme DHRF, utilisée comme une enzyme rapportrice. Ici, les mesures de quantification de PPIs sont mesurées à une concentration de MTX fixe (10μM), cette concentration ne permet pas forcément de pouvoir mesurer de manière équivalente tous les PPIs étudiés. Par exemple, certains PPIs peuvent avoir besoin de concentration en MTX plus élevé pour pouvoir activer l'enzyme rapportrice. L'utilisation de différentes concentrations de MTX pourrait donc permettre la quantification de PPIs ayant différentes forces d'interactions, elles pourraient par exemple permettre de distinguer des interactions spécifiques fortes des interactions spécifiques faibles.

Applicabilité et faisabilité d'une étude piQTL à large échelle en dehors du système PCA

Dans notre approche piQTL, la quantification des interactions protéine-protéine est grandement dépendante du succès d'incorporation de la construction génétique qui permet d'estimer le nombre de complexes PPIs formés grâce à la reconstitution de l'enzyme DHFR qui peut métaboliser la molécule de méthotrexate. Pour chaque souche préalablement taguée avec un code-barre spécifique, si un grand nombre de complexes PPIs est formé au sein de la cellule, ceci conférera une grande résistance à la drogue de sélection, se traduisant par un fitness important de la souche, et donc par la génération d'un grand nombre de code-barres. Au contraire, si peu de complexes PPIs sont formés, la souche sera plus sensible à la drogue de sélection, se traduisant par un fitness plus faible de la souche, et donc un nombre de code-barres plus restreint. Ce protocole a été mis en place pour l'étude de 61 PPIs sous 5 conditions environnementales différentes. Dans cette stratégie expérimentale, l'enjeu est de pouvoir identifier impartialement les différentes souches contenues dans la population de levures et d'associer à chaque souche la quantification de complexes PPIs formés pour réaliser par la suite les études de mapping piQTL.

L'étape clef de l'approche est donc la quantification des interactions protéine-protéine. Précédemment, Picotti et collègues ont précédemment proposé une étude pQTL où ils ont pu mesurer l'abondance protéique de presque la totalité du protéome de *S. cerevisiae* (Picotti et al., 2013) grâce à la spectrométrie de masse, ce qui met en avant la faisaibilité d'une étude piQTL utilisant cette technologie à condition d'adapter les protocoles expérimentaux pour isoler les complexes d'interactions protéine-protéine et permettre leur quantification. Pou cela, il sera nécessaire d'effectuer des étapes de purification par affinité où les conditions doivent être adéquates pour maintenir la formation des complexes PPIs. Cependant, la mise en place d'un tel protocole expérimental ne permettrait pas de quantifier ces interactions protéine-protéine de manière *in vivo*.

Pour mesurer la quantification des PPIs de manière *in vivo*, on peut substituer la méthode PCA par la méthode FRET (pour *Fluorescence resonance energy transfer*, en anglais) où la quantification des PPIs pourra être réalisée par une mesure de fluorescence (Truong and Ikura, 2001) qui sera émise seulement s'il y a formation des complexes PPIs (les fluorochromes permettant l'observation du signal FRET étant préalablement greffés aux différents interacteurs impliqués dans le PPI). Les mesures de quantification des PPIs pourraient être obtenues par cytométrie en flux qui permettrait d'identifier les différentes souches de notre population et de leur attribuer leur quantification respective d'interation protéine-protéine

(Lim et al., 2022). Dans ces deux cas de figure, la mise en place de ces stratégies expérimentales nécessite l'acquisition d'appareil de mesure onéreux, ainsi que la création de librairies moléculaires conséquentes pour l'identification d'un grand nombre de PPIs.

4.2.3. Perspectives

Les résultats présentés dans le chapitre 3 mettent en avant les observations globales qu'on peut établir grâce aux piQTLs.

Les analyses reposant sur les approches réseaux

Nous avons projeté nos résultats piQTL sur différents réseaux biologiques, que ce soit des réseaux d'interactions protéine-protéine, mais aussi sur des réseaux d'interaction géniques. A l'heure actuelle, nos analyses permettent de générer l'étude de ces réseaux de manière PPI- et environnement-spécifique. Pour étendre notre étude à une analyse plus globale reposant sur la topologie des réseaux, il serait nécessaire d'inférer les réseaux biologiques qui contiennent l'ensemble des variants génétiques associés à des ORFs étudiés dans notre cohorte populationnelle. Une fois ce réseau établi, nous pourrons projeter les différentes métriques inférées par les analyses statistiques des piQTL, comme les p-valeurs, les tailles d'effet et le score h^2 et établir leur relation potentielle avec la topologie de ces réseaux. Pour cela, on peut par exemple établir la corrélation entre ces métriques et le degré des gènes / protéines, soit le nombre d'interactions associés à chaque gène / protéine. On peut également calculer les chemins les plus courts entre les différents piQTLs et leurs PPIs associés pour établir les distances les séparant, et ainsi définir les concepts de cis- et trans-piQTL. Dans le contexte d'un réseau biologique, un cis-piQTL serait associé à un variant génétique qui serait à une distance proche du PPI influencé (voisin direct ou voisin à un noeud de distance d'un des gène/protéine impliqué dans le PPI). Un trans-piQTL serait associé à un variant génétique qui serait à une distance plus éloignée du PPI influencé (voisin à plus d'un noeud de distance d'un des gènes/protéines impliqué dans le PPI)). Ces concepts de cis- et trans-piQTL permettrait de mettre en évidence des phénomènes d'épistasie qui permettraient la mise en contexte biologique ou fonctionnelle des piQTLs et des gènes qui sont à l'origine des protéines permettant la création du PPI.

Il est à noter que les représentations actuelles de nos réseaux biologiques nous permettent de projeter seulement les métriques statistiques des piQTL qui sont situés directement sur des ORFs. Il serait intéressant d'enrichir nos réseaux avec les piQTLs qui sont situés à proximité de ces ORFs. Graphiquement, cela pourrait être réalisé à l'aide d'un code couleur ou de formes permettant de distinguer les différents types de piQTLs dans le réseau. Les réseaux d'interaction génique et d'interaction protéine-protéine ne permettent pas d'étudier les piQTLs présents sur des régions non codantes du génome (comme les transcrits instables,

les ARN non codants et les régions intergéniques etc.). Pour étudier spécifiquement ces piQTLs, on pourrait proposer un nouveau type de réseau où chaque noeud représenterait une annotation génomique (voir Figure Supplémentaire 3.5 pour voir les différentes annotations génomiques possibles), et où les liens entre ces noeuds pourraient représenter leur localisation spatiale (par exemple : on définira un lien entre les annotations génomiques situées sur le même chromosome). Cela constituerait un réseau de 12,054 noeuds, ce qui est en fait un réseau de données massif ("lourd") à analyser, pour lequel il faudra proposer des approches d'analyses computationnellement "légères" (en termes de temps de calcul ou d'espaces d'allocation de ressources) pour assurer la faisabilité. Les analyses par réseau nous permettront de valider notre hypothèse qui est que l'abstraction du modèle omnigénique à l'aide des réseaux des interactions protéine-protéine est une solution prometteuse pour décrire la variance phénotypique associée à des traits complexes.

Croisement des résultats piQTLs de la levure et données GWAS chez l'humain

Certains phénotypes complexes évoqués dans nos travaux, tels que la schizophrénie ou le diabète de type II, ont été étudiés avec des approches GWAS. Ces phénotypes complexes ont été étudiés avec des approches GWAS et ont permis d'identifier un certain nombre de gènes associés à ces maladies. Il serait donc intéressant de voir si les piQTLs identifiés chez la levure ont des homologues chez l'humain pour lesquels des variants génétiques ont été identifiés comme associé à ces phénotypes. Cela permettrait de proposer une validation fonctionnelle indirecte de ces variants par une approche de génomique comparée. Pour réaliser cette analyse, il sera nécessaire de collecter dans la littérature les loci associés aux SNPs identifiés dans les analyses GWAS sur ces deux phénotypes, afin d'étudier les SNPs associés à leur loci homologues chez la levure. Pour cela, il est possible de réaliser des études de chevauchement entre les gènes associés aux piQTLs de la levure et les gènes de la levure dont les homologues chez l'humain ont été précédemment identifiées comme impliqués au diabète de type II ou à la schizophrénie. Ces deux jeux de données seront comparés aux gènes contenant des variants génétiques de la levure, qu'ils soient des piQTLs ou non. A partir de ces études de chevauchement, des tests stastistiques hypergéométriques, ou des méthodes de sous-échantillonage (bootstrapping) pourraient être utilisés afin de valider que le sous-ensemble de gènes identifiés par le chevauchement des différents jeux de données ne soit pas dû au hasard. Ainsi les piQTLs permettraient de prioriser certains SNPs à étudier chez l'humain pour des analyses plus approfondies d'eQTL, pQTL ou encore de piQTL pour les deux maladies d'intérêt.

Utilisation du modèle linéaire mixte pour une meilleure représentation de la relation génotype-phénotype par la quantification des PPIs

Dans cette étude, les différents modèles proposés pour l'étude de la relation phénotype et génotype sont des modèles s'appuyant sur des modélisations linéaires simples. Nous n'avons pas inclus, dans nos analyses actuelles, de correction pour des effets génétiques qui ne seraient pas uniquement spécifiques à la réponse au médicament étudié. En particulier, parmi les conditions que nous avons mesuré, nous pourrions utiliser les quantifications des PPIs qui sont mesurées dans un contexte sans aucune drogue de sélection, afin de pouvoir identifier les piQTLs qui sont uniquement influencés par le bagage génétique des levures avant sélection artificielle. Dans des analyses subséquentes, il serait également intéressant de modéliser l'intervention de l'environnement dans la relation phénotype et génotype, par exemple par l'utilisation de modèles linéaires mixtes.

4.3. Les approches protéome-centrées et leur apport pour l'étude des phénotypes complexes

Avec cette thèse, nous avons mis en avant deux différentes stratégies bio-informatiques pour étudier les phénotype complexes avec une approche protéo-centrée. Dans le chapitre 2, nous avons étudié le phénotype du vieillissement à l'aide d'une approche de génomique comparée qui s'est concentrée exclusivement sur les propriétés intrinsèques et prédites du protéome orthologue partagées entre le rat taupe nu et la souris. Dans le chapitre 3, nous avons étudié la réponse aux médicaments chez la levure, dans une population consanguine génétiquement diverse à l'aide d'une quantification *in vivo* des PPIs.

Les sciences -omiques sont de toutes evidences des méthodologies en plein essor en sciences du vivant, avec un focus plus grand sur l'ADN et l'ARN. Pourtant, la protéomique, refletant la dimension la plus dynamique du dogme central de la biologie moléculaire. Ces approches sont plus compliquées à implémenter à large échelle, mais ont le potentiel de donner une vision plus globale de la complexité des systèmes moléculaires du vivant. Les deux études présentées dans cette thèse mettent en avant l'importance d'étudier les protéines et les interactions protéine-protéine pour caractériser deux exemples de phénotypes complexes. A travers la discussion séparée des deux études, il peut être constaté que les approches proposées respectivement dans chaque chapitre peuvent également servir à apporter différents angles de complexité dans la description d'un phénotype complexe Avec cette thèse, nous avons mis en évidence que l'abstraction des phénotypes par le biais des protéines et de leurs interactions est une approche robuste et révélatrice pour caractériser des phénotypes complexes. En complémentarité avec les modèles reposant sur l'expression des gènes ou les changements

épigénétiques associés à ces phénotypes complexes, elles mettent en avant l'importance de proposer un modèle plus systémique pour étudier la relation génotype-phénotype.

Il est très probable que l'avenir de ce domaine scientifique de la biologie des systèmes necessitera de proposer d'étudier des phénotypes complexes à l'aide de réseaux dit multi-échelles. Différents approches multi-échelles peuvent être définies, chacune ayant des stratégies bio-informatiques spécifiques à mettre en place :

- l'étude des réseaux avec une composante d'intégration multi-omiques (Hasin et al., 2017; Kreitmaier et al., 2023) (expression de gènes, abondances protéiques, données métaboliques, QTLs etc.) Pour cette approche il sera important de mettre en place des méthodes permettant de proposer différents types d'harmonisation de données afin qu'elles puissent être comparées, et des approches robustes pour établir des relations entre ces différentes données.
- les comparaisons de réseaux biologiques représentant différents types cellulaires et/ou tissulaires pour intégrer la diversité phénotypique à travers les tissus (Wang et al., 2022). En effet, l'expression des gènes au sein des tissus n'est pas uniforme. En fonction des traits ou maladies complexes étudiés, il est essentiel de mettre en avant les tissus qui montrent des différences de variance phénotypique. Par exemple, dans l'étude de la maladie d'Alzheimer, il a été démontré que leurs biomarqueurs génétiques se trouvaient principalement dans les tissus nerveux (Mathys et al., 2019).
- les comparaisons de réseaux biologiques représentant des cohortes populationnelles permettant d'inférer les gènes candidats en lien avec des phénotypes complexes. Il serait intéressant de pouvoir inférer des réseaux biologiques se basant sur les gènes identifiées par des études GWAS ou bien des QTLs afin de proposer des modèles d'abstraction des phénotypes complexes qui permettent de proposer une meilleure contextualisation fonctionnelles ou biologiques des gènes associées à des variants génétiques influençant la diversité phénotypique.
- les comparaisons de réseaux biologiques communes entre différences espèces, afin d'étudier les phénotypes complexes dans un contexte évolutifs.

Cette thèse met également en évidence l'importance du partage approprié des methodologies et des données pour implémenter des approches multi-échelle en biologie des systèmes. En effet, l'accès libre à ce type de ressources permet de s'affranchir des limites rencontrées quand les méthodes proposées sont propriétaires où dont l'accès est limité, ce qui rend difficile de proposer des stratégie de mise à l'échelle, comme nous l'avons démontré dans le chapitre 2. Aussi le partage de ressources expérimentales telle que la cohorte de levures génétiquement diversifiée, va permettre à la communauté scientifique de pouvoir réaliser des études puissantes comme celle proposée dans le chapitre 3.

4.4. Conclusion

En conclusion, ma thèse démontre l'intérêt de proposer des stratégies bio-informatiques à l'intersection de divers disciplines (ici, la génomique comparée, la biologie évolutive, la science des données appliquée aux données -omiques, la génétique des populations, la biologie des sytèmes et la biologie des réseaux) pour contribuer à une meilleure compréhension de la relation génotype-phénotype pour les phénotypes complexes. En effet, mes travaux confirment les apports significatifs du domaine de la bio-informatique pour l'implémentation d'approches protéome-centrées.

Cette thèse m'a permise de travailler sur différents organismes modèles (le rat-taupe nu et la souris, la levure) pour proposer des hypothèses à tester chez l'humain, me permettant ainsi d'explorer différents degrés de diversité du vivant. Elle m'a permise de mettre en avant l'importance de proposer des approches bio-informatiques robustes et reproductibles pour étudier de manière systémique différents phénotypes complexes qui sont des enjeux de société importants.

Les méthodes et idées proposées à travers ces différents chapitres sont le point de départ pour des études similaires dans d'autres contextes, et benificieront d'être partagées à la communauté scientifique dans le domaine bio-informatique, la biologie évolutive et la génétique statistique.

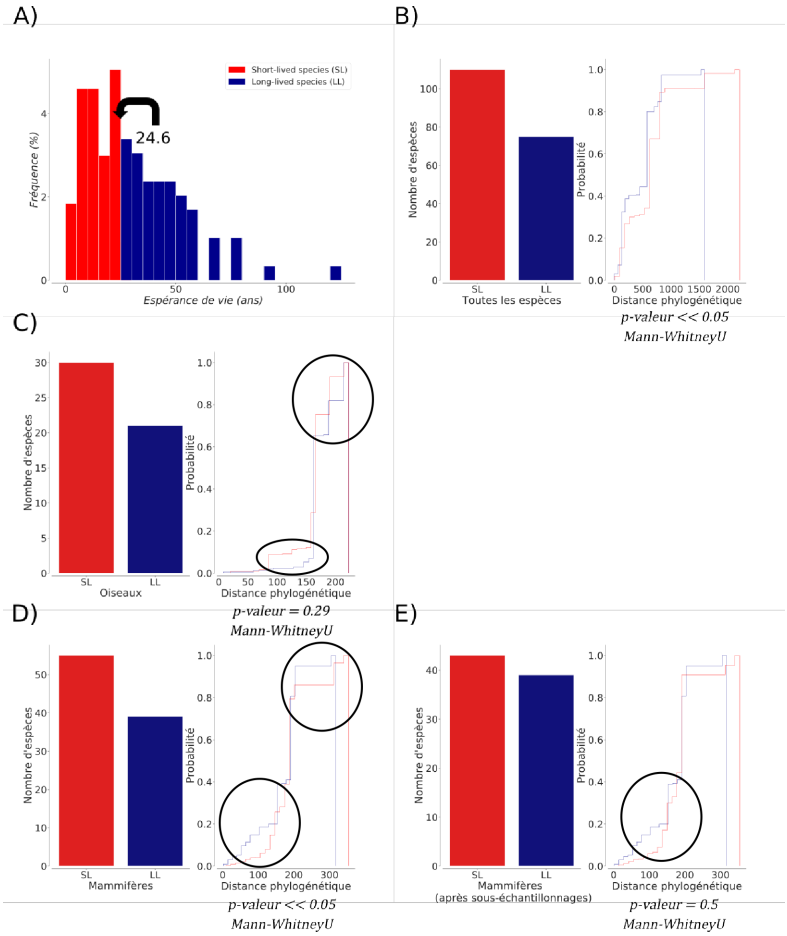

Fig. 4.2. Tentative de distinction des groupes SL et LL avec et sans distinction des relations phylogénétiques.

(Créée par Savandara Besse)
A), **Distribution de l'espérance de vie maximale chez les eucaryotes sans distinction phylogénétique** : Le seuil de distinction entre les groupes SL et LL est la valeur moyenne d'espérance de vie maximale définie à partir de l'ensemble des données d'espérance de vie maximale de toutes les espèces présentes dans Figure 4.1.
B), **Comparaison de la distribution des distances phylogénétiques entre les groupes SL et LL sans distinction taxonomique** : Il existe une différence significative entre les distributions des distances phylogénétiques des groupes définis sans distinction taxonomique d'après le test non paramétrique de Mann-Whitney. Cette comparaison démontre que la définition des deux groupes de longévité se basant sur la métrique de moyenne d'espérance de vie sur l'ensemble des espèces ne permet d'obtenir deux groupes similaires que l'on peut comparer pour identifier des différences associées à la longévité.
C), **Comparaison de la distribution des distances phylogénétiques entre les groupes SL et LL dans le groupe des oiseaux** : Ici, les distributions des distances phylogénétiques sont significativement similaires d'après le test non paramétrique de Mann-Whitney. Toutefois on peut constater qu'il semble persister des écarts de distances phylogénétiques entre les deux groupes (zones entourées).
D) & **E)**, **Diminution des différences de distribution des distances phylogénétiques entre les groupes SL et LL dans le groupe de mammifères** : Il a été nécessaire d'effectuer une étape de sous-échantillonnages pour réduire l'écart entre les distributions des distances phylogénétiques des deux groupes. Toutefois, bien que le second test non paramétrique de Mann-Whitney sur les sous-groupes créés soit non significatif, ce qui permet de spéculer sur la similarité des deux groupes. Ici aussi, on observe encore des écarts dans les deux distributions (zones entourées).

Matériel Supplémentaire pour la discussion

Classe	Moyenne d'espérance de vie maximale (ans)	Nombre d'espèces total	Nombre d'espèces dans le groupe SL	Nombre d'espèces dans le groupe LL
Mammifères	23	94	55	39
Téléostéens	30	51	30	21
Reptiles	17	19	12	7
Insectes	9	4	3	1

Tableau 1. Répartition des espèces dans les groupes SL et LL par classe taxonomique.

Méthode	VP	VN	FP	FN	Sensitivité (%)	Spécificité (%)	Q (%)
Aggrescan (Conchillo-Solé et al., 2007)	445	5,210	1,363	813	35.37	79.26	57.32
Tango (Fernandez-Escamilla et al., 2004)	12	6282	291	1086	13.67	95.57	54.62
AMYLPRED2 (Tsolis et al., 2013)	494	5553	1020	764	39.27	84.48	61.88

Tableau 2. Etude comparative des logiciels Aggrescan, Tango et Amylpred2

(Version modifiée et traduite provenant de Tsolis et al., 2013)
Tsolis et collègues ont évalué la performance de ces trois logiciels par le décompte de valeurs Vrai/Faux Positif (VP, FP) et Vrai/Faux Négatif (VN, FN) qui permettent de calculer les pourcentage de sensitivité (calculé par l'Equation (4.1)) et de sensitivité (calculé par l'Equation (4.2)). Les valeurs VP, FP, VN et FN représentent les compte de résidus qui sont correctement prédits dans régions qui peuvent être impliqué dans la formation d'agrégats protéiques dans 33 protéines amylogéniques. Le score Q représente la moyenne entre sensitivité et spécificité (Equation (4.3)). Le logiciel Pasta (Walsh et al., 2014) n'est pas présent dans ce tableau car ce logiciel est postérieur à l'étude de Tsolis et collègues.

$$\frac{VP}{VP+FN} \tag{4.1}$$

$$\frac{VN}{VN+FP} \tag{4.2}$$

$$\frac{Sensitivité + Spécificité}{2} \tag{4.3}$$

Références bibliographiques

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Abraham, A.-L. (2008, December). *Caractérisation et analyse évolutive des répétitions intragéniques : Une étude au niveau des gènes, des séquences protéiques et des structures tridimensionnelles* (Theses). Université Pierre et Marie Curie - Paris VI. https://theses.hal.science/tel-00482373

Abramovs, N., Brass, A., & Tassabehji, M. (2020). Hardy-weinberg equilibrium in the large scale genomic sequencing era. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00210

Adams, D. (2022). *Genotype* [NIH]. Retrieved December 27, 2022, from https://www.genome.gov/genetics-glossary/genotype

Aguzzi, A., & O'Connor, T. (2010). Protein aggregation diseases: Pathogenicity and therapeutic perspectives. *Nature Reviews. Drug Discovery*, *9*(3), 237–248. https://doi.org/10.1038/nrd3050

Al Shweiki, M. R., Mönchgesang, S., Majovsky, P., Thieme, D., Trutschel, D., & Hoehenwarter, W. (2017). Assessment of label-free quantification in discovery proteomics and impact of technological factors and natural variability of protein abundance. *Journal of Proteome Research*, *16*(4), 1410–1424. https://doi.org/10.1021/acs.jproteome.6b00645

Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M., & Gerstein, M. B. (2010). Annotating non-coding regions of the genome. *Nat Rev Genet*, *11*(8), 559–571. https://doi.org/10.1038/nrg2814

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Anand, S., Samuel, M., Ang, C.-S., Keerthikumar, S., & Mathivanan, S. (2017). Label-based and label-free strategies for protein quantitation [Series Title: Methods in Molecular

Biology]. In *Proteome bioinformatics* (pp. 31–43). Springer New York. https://doi.org/10.1007/978-1-4939-6740-7_4

Andziak, B., O'Connor, T. P., Qi, W., DeWaal, E. M., Pierce, A., Chaudhuri, A. R., Van Remmen, H., & Buffenstein, R. (2006). High oxidative damage levels in the longest-living rodent, the naked mole-rat. *Aging Cell*, *5*(6), 463–471. https://doi.org/10.1111/j.1474-9726.2006.00237.x

Anisimov, V., Zabezhinski, M., Popovich, I., Piskunova, T., Semenchenko, A., Tyndyk, M., & Blagosklonny, M. (2011). Rapamycin increases lifespan and inhibits spontaneous tumorigenesis in inbred female mice. *Cell cycle*, *10*(24), 4230–4236.

Arai, Y., Sasaki, T., & Hirose, N. (2017). Demographic, phenotypic, and genetic characteristics of centenarians in okinawa and honshu, japan: Part 2 honshu, japan. *Mechanisms of Ageing and Development*, *165*, 80–85. https://doi.org/10.1016/j.mad.2017.02.005

Aubert, G., & Lansdorp, P. M. (2008). Telomeres and aging. *Physiological Reviews*, *88*(2), 557–579. https://doi.org/10.1152/physrev.00026.2007

Austad, S. N., & Fischer, K. E. (1991). Mammalian aging, metabolism, and ecology: Evidence from the bats and marsupials. *Journal of Gerontology*, *46*(2), B47–B53. https://doi.org/10.1093/geronj/46.2.B47

Austad, S. N., & Fischer, K. E. (2016). Sex differences in lifespan. *Cell Metabolism*, *23*(6), 1022–1033. https://doi.org/https://doi.org/10.1016/j.cmet.2016.05.019

Aydin, S. (2015). A short history, principles, and types of ELISA, and our laboratory experience with peptide/protein analyses using ELISA. *Peptides*, *72*, 4–15. https://doi.org/10.1016/j.peptides.2015.04.012

Balch, W. E., Morimoto, R. I., Dillin, A., & Kelly, J. W. (2008). Adapting proteostasis for disease intervention. *Science (New York, N.Y.)*, *319*(5865), 916–919. https://doi.org/10.1126/science.1141448

Balut, C., vandeVen, M., Despa, S., Lambrichts, I., Ameloot, M., Steels, P., & Smets, I. (2008). Measurement of cytosolic and mitochondrial pH in living cells during reversible metabolic inhibition. *Kidney International*, *73*(2), 226–232. https://doi.org/10.1038/sj.ki.5002632

Balzi, E., Chen, W., Ulaszewski, S., Capieaux, E., & Goffeau, A. (1987). The multidrug resistance gene PDR1 from saccharomyces cerevisiae. *Journal of Biological Chemistry*, *262*(35), 16871–16879.

Bampi, G. B., Bisso-Machado, R., Hünemeier, T., Gheno, T. C., Furtado, G. V., Veliz-Otani, D., Cornejo-Olivas, M., Mazzeti, P., Bortolini, M. C., Jardim, L. B., Saraiva-Pereira, M. L., & Rede Neurogenetica. (2017). Haplotype study in SCA10 families provides further evidence for a common ancestral origin of the mutation. *Neuromolecular Medicine*, *19*(4), 501–509. https://doi.org/10.1007/s12017-017-8464-8

Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(2), 101–113. https://doi.org/10.1038/nrg1272

Barton, N. H., Etheridge, A. M., & Véber, A. (2017). The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, *118*, 50–73. https://doi.org/10.1016/j.tpb.2017.06.001

Baryshnikova, A. (2018). Spatial analysis of functional enrichment (SAFE) in large biological networks. *Methods in Molecular Biology (Clifton, N.J.)*, *1819*, 249–268. https://doi.org/10.1007/978-1-4939-8618-7_12

Bell, C. G., Lowe, R., Adams, P. D., Baccarelli, A. A., Beck, S., Bell, J. T., Christensen, B. C., Gladyshev, V. N., Heijmans, B. T., Horvath, S., Ideker, T., Issa, J.-P. J., Kelsey, K. T., Marioni, R. E., Reik, W., Relton, C. L., Schalkwyk, L. C., Teschendorff, A. E., Wagner, W., ... Rakyan, V. K. (2019). DNA methylation aging clocks: Challenges and recommendations. *Genome Biology*, *20*(1), 249. https://doi.org/10.1186/s13059-019-1824-y

Benayoun, B. A., Pollina, E. A., Singh, P. P., Mahmoudi, S., Harel, I., Casey, K. M., Dulken, B. W., Kundaje, A., & Brunet, A. (2019). Remodeling of epigenome and transcriptome landscapes with aging in mice reveals widespread induction of inflammatory responses. *Genome Research*, *29*(4), 697–709. https://doi.org/10.1101/gr.240093.118

Bennett, B., Farber, C., Orozco, L., Kang, H., Ghazalpour, A., Siemers, N., & Lusis, A. (2010). A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome research*, *20*(2), 281–290.

Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., & Schwede, T. (2017). Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Scientific Reports*, *7*(1), 10480. https://doi.org/10.1038/s41598-017-09654-8

Beynon, R. J., Doherty, M. K., Pratt, J. M., & Gaskell, S. J. (2005). Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nature Methods*, *2*(8), 587–589. https://doi.org/10.1038/nmeth774

Bhatnagar, S. (2021). *Manhattanly: Interactive q-q and manhattan plots using 'plotly.js'*. https://CRAN.R-project.org/package=manhattanly

Biesecker, L. (2022, December 27). *Haplotype* [NIH]. Retrieved December 30, 2022, from https://www.genome.gov/genetics-glossary/haplotype

Blackburn, A. N., Blondell, L., Kos, M. Z., Blackburn, N. B., Peralta, J. M., Stevens, P. T., Lehman, D. M., Blangero, J., & Göring, H. H. H. (2020). Genotype phasing in pedigrees using whole-genome sequence data. *European journal of human genetics: EJHG*, *28*(6), 790–803. https://doi.org/10.1038/s41431-020-0574-3

Blackburn, E. H., Greider, C. W., & Szostak, J. W. (2006). Telomeres and telomerase: The path from maize, tetrahymena and yeast to human cancer and aging [Publisher: Nature Publishing Group]. *Nature medicine*, *12*(10), 1133–1138.

Blasco, M. A. (2007). Telomere length, stem cells and aging [Publisher: Nature Publishing Group]. *Nature chemical biology*, *3*(10), 640–649.

Boomsma, D., Busjahn, A., & Peltonen, L. (2002). Classical twin studies and beyond [Publisher: Nature Publishing Group]. *Nature reviews genetics*, *3*(11), 872–882.

Bosch, J. A., Chen, C.-L., & Perrimon, N. (2021). Proximity-dependent labeling methods for proteomic profiling in living cells: An update. *Wiley Interdisciplinary Reviews. Developmental Biology*, *10*(1), e392. https://doi.org/10.1002/wdev.392

Boyle, E., Li, Y., & Pritchard, J. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell*, *169*(7), 1177–1186.

Brem, R., Yvert, G., Clinton, R., & Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science*, *296*, 752–755.

Bringans, S., Kendrick, T. S., Lui, J., & Lipscombe, R. (2008). A comparative study of the accuracy of several de novo sequencing software packages for datasets derived by matrix-assisted laser desorption/ionisation and electrospray. *Rapid communications in mass spectrometry: RCM*, *22*(21), 3450–3454. https://doi.org/10.1002/rcm.3752

Brosch, M., Yu, L., Hubbard, T., & Choudhary, J. (2009). Accurate and sensitive peptide identification with mascot percolator. *Journal of proteome research*, *8*(6), 3176–3181. https://doi.org/10.1021/pr800982s

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: Existing methods and new developments. *Nature Reviews. Genetics*, *12*(10), 703–714. https://doi.org/10.1038/nrg3054

Brückner, A., Polge, C., Lentze, N., Auerbach, D., & Schlattner, U. (2009). Yeast two-hybrid, a powerful tool for systems biology. *International Journal of Molecular Sciences*, *10*(6), 2763–2788. https://doi.org/10.3390/ijms10062763

Buccitelli, C., & Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, *21*(10), 630–644.

Buffenstein, R. (2005). The naked mole-rat: A new long-living model for human aging research. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *60*(11), 1369–1377. https://doi.org/10.1093/gerona/60.11.1369

Buffenstein, R., & Ruby, J. G. (2021). Opportunities for new insight into aging from the naked mole-rat and other non-traditional models. *Nature Aging*, *1*(1), 3–4. https://doi.org/10.1038/s43587-020-00012-4

Buntru, A., Trepte, P., Klockmeier, K., Schnoegl, S., & Wanker, E. E. (2016). Current approaches toward quantitative mapping of the interactome. *Frontiers in Genetics*, *7*. https://doi.org/10.3389/fgene.2016.00074

Burnett, B. G., & Pittman, R. N. (2005). The polyglutamine neurodegenerative protein ataxin 3 regulates aggresome formation. *Proceedings of the National Academy of Sciences*, *102*(12), 4330–4335. https://doi.org/10.1073/pnas.0407252102

Butler, J. M. (2015). Chapter 10 - STR population data analysis. In J. M. Butler (Ed.), *Advanced topics in forensic DNA typing: Interpretation* (pp. 239–279). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-405213-0.00010-5

Calder, P. C., Bosco, N., Bourdet-Sicard, R., Capuron, L., Delzenne, N., Doré, J., Franceschi, C., Lehtinen, M. J., Recker, T., Salvioli, S., & Visioli, F. (2017). Health relevance of the modification of low grade inflammation in ageing (inflammageing) and the role of nutrition. *Ageing Research Reviews*, *40*, 95–119. https://doi.org/10.1016/j.arr.2017.09.001

Campisi, J. (2013). Aging, cellular senescence, and cancer [Publisher: NIH Public Access]. *Annual review of physiology*, *75*, 685.

Candia, J., Daya, G. N., Tanaka, T., Ferrucci, L., & Walker, K. A. (2022). Assessment of variability in the plasma 7k SomaScan proteomics assay. *Scientific Reports*, *12*(1), 17147. https://doi.org/10.1038/s41598-022-22116-0

Cannavò, E., Koelling, N., Harnett, D., Garfield, D., Casale, F., Ciglar, L., & Furlong, E. (2017). Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature*, *541*(7637), 402–406.

Carpenter, E. P., Beis, K., Cameron, A. D., & Iwata, S. (2008). Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology*, *18*(5), 581–586. https://doi.org/10.1016/j.sbi.2008.07.001

Caselli, G., Battaglini, M., & Capacci, G. (2018). Beyond one hundred: A cohort analysis of italian centenarians and semisupercentenarians. *The Journals of Gerontology: Series B*. https://doi.org/10.1093/geronb/gby033

Ceballos, F. C., Hazelhurst, S., & Ramsay, M. (2018). Assessing runs of homozygosity: A comparison of SNP array and whole genome sequence low coverage data. *BMC Genomics*, *19*(1), 106. https://doi.org/10.1186/s12864-018-4489-0

Chakravarti, A., & Turner, T. N. (2016). Revealing rate-limiting steps in complex disease biology: The crucial importance of studying rare, extreme-phenotype families. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *38*(6), 578–586. https://doi.org/10.1002/bies.201500203

Chanock, S. J., & Ostrander, E. A. (2014). 22 - discovery and characterization of cancer genetic susceptibility alleles. In *Abeloff's clinical oncology (fifth edition)* (Fifth Edition, 309–321.e3). Churchill Livingstone. https://doi.org/https://doi.org/10.1016/B978-1-4557-2865-7.00022-9

Chatterjee, N., & Walker, G. C. (2017). Mechanisms of DNA damage, repair, and mutagenesis. *Environmental and Molecular Mutagenesis*, *58*(5), 235–263. https://doi.org/10.1002/em.22087

Chen, B., Liu, Q., Ge, Q., Xie, J., & Wang, Z.-W. (2007). UNC-1 regulates gap junctions important to locomotion in c. elegans. *Current Biology*, *17*(15), 1334–1339. https://doi.org/10.1016/j.cub.2007.06.060

Chen, J., Carter, M. B., Edwards, B. S., Cai, H., & Sklar, L. A. (2012). High throughput flow cytometry based yeast two-hybrid array approach for large-scale analysis of protein-protein interactions. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology*, *81*(1), 90–98. https://doi.org/10.1002/cyto.a.21144

Cheng, J., Yuan, Z., Yang, W., Xu, C., Cong, W., Lin, L., Zhao, S., Sun, W., Bai, X., & Cui, S. (2017). Comparative study of macrophages in naked mole rats and ICR mice. *Oncotarget*, *8*(57), 96924–96934. https://doi.org/10.18632/oncotarget.19661

Chiti, F., & Dobson, C. M. (2017). Protein misfolding, amyloid formation, and human disease: A summary of progress over the last decade. *Annual Review of Biochemistry*, *86*(1), 27–68. https://doi.org/10.1146/annurev-biochem-061516-045115

Chouraki, V., & Seshadri, S. (2014, January 1). Chapter five - genetics of alzheimer's disease. In T. Friedmann, J. C. Dunlap, & S. F. Goodwin (Eds.), *Advances in genetics* (pp. 245–294). Academic Press. https://doi.org/10.1016/B978-0-12-800149-3.00005-6

Clayton, D. (2023). *snpStats: SnpMatrix and XSnpMatrix classes and methods*. https://doi.org/10.18129/B9.bioc.snpStats

Clouard, C., Ausmees, K., & Nettelblad, C. (2022). A joint use of pooling and imputation for genotyping SNPs. *BMC Bioinformatics*, *23*(1), 421. https://doi.org/10.1186/s12859-022-04974-7

Collaborative Cross Consortium. (2012). The genome architecture of the collaborative cross mouse genetic reference population. *Genetics*, *190*(2), 389–401. https://doi.org/10.1534/genetics.111.132639

Conchillo-Solé, O., de Groot, N. S., Avilés, F. X., Vendrell, J., Daura, X., & Ventura, S. (2007). AGGRESCAN: A server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC bioinformatics*, *8*, 65. https://doi.org/10.1186/1471-2105-8-65

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., & Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, *10*(3), 184–194. https://doi.org/10.1038/nrg2537

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., Pelechano, V., Styles, E. B., Billmann, M., van Leeuwen, J., van Dyk, N., Lin, Z.-Y., Kuzmin, E., Nelson, J., Piotrowski, J. S., ... Boone, C. (2016). A global genetic interaction network maps a wiring diagram

of cellular function. *Science*, *353*(6306), aaf1420–aaf1420. https://doi.org/10.1126/science.aaf1420

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, *26*(12), 1367–1372. https://doi.org/10.1038/nbt.1511

Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*(5258), 561–563. https://doi.org/10.1038/227561a0

D'Amico, D., Sorrentino, V., & Auwerx, J. (2017). Cytosolic proteostasis networks of the mitochondrial stress response. *Trends in Biochemical Sciences*, *42*(9), 712–725. https://doi.org/10.1016/j.tibs.2017.05.002

Dantuma, N. P., Hoppe, T., & Herzog, L. K. (2020). The price of longevity. *Aging*, *12*(22), 22350–22351. https://doi.org/10.18632/aging.104215

David, D. C. (2012). Aging and the aggregating proteome. *Frontiers in Genetics*, *3*. https://doi.org/10.3389/fgene.2012.00247

Davidovic, M., Sevo, G., Svorcan, P., Milosevic, D. P., Despotovic, N., & Erceg, P. (2010). Old age as a privilege of the "selfish ones". *Aging and Disease*, *1*(2), 139–146.

De, S., & Klenerman, D. (2019). Imaging individual protein aggregates to follow aggregation and determine the role of aggregates in neurodegenerative disease. *Biochimica Et Biophysica Acta. Proteins and Proteomics*, *1867*(10), 870–878. https://doi.org/10.1016/j.bbapap.2018.12.010

de Magalhães, J. P., Curado, J., & Church, G. M. (2009). Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, *25*(7), 875–881. https://doi.org/10.1093/bioinformatics/btp073

Devlin, B., & Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping [Publisher: Elsevier]. *Genomics*, *29*(2), 311–322.

Diaz-Papkovich, A., Anderson-Trocmé, L., & Gravel, S. (2021). A review of UMAP in population genetics. *Journal of Human Genetics*, *66*(1), 85–91. https://doi.org/10.1038/s10038-020-00851-4

Dijk, E., Chen, C., d'Aubenton-Carafa, Y., Gourvennec, S., Kwapisz, M., Roche, V., & Morillon, A. (2011). XUTs are a class of xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature*, *475*(7354), 114–117.

Draceni, Y., & Pechmann, S. (2019). Pervasive convergent evolution and extreme phenotypes define chaperone requirements of protein homeostasis. *Proceedings of the National Academy of Sciences*, *116*(40), 20009–20014. https://doi.org/10.1073/pnas.1904611116

Dudbridge, F., & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, *32*(3), 227–234. https://doi.org/10.1002/gepi.20297

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Eichler, E., Flint, J., Gibson, G., Kong, A., Leal, S., Moore, J., Nadeau, J., & E., E. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, *11*(6), 446–450.

Elmore, S. (2007). Apoptosis: A review of programmed cell death. *Toxicologic Pathology*, *35*(4), 495–516. https://doi.org/10.1080/01926230701320337

El-Samad, H., Kurata, H., Doyle, J. C., Gross, C. A., & Khammash, M. (2005). Surviving heat shock: Control strategies for robustness and performance. *Proceedings of the National Academy of Sciences*, *102*(8), 2736–2741. https://doi.org/10.1073/pnas.0403510102

Eng, J. K., McCormack, A. L., & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, *5*(11), 976–989. https://doi.org/10.1016/1044-0305(94)80016-2

Ermolaeva, M., Neri, F., Ori, A., & Rudolph, K. L. (2018). Cellular and epigenetic drivers of stem cell ageing. *Nature Reviews Molecular Cell Biology*, *19*(9), 594–610. https://doi.org/10.1038/s41580-018-0020-3

Esposito, D., Weile, J., Shendure, J., Starita, L. M., Papenfuss, A. T., Roth, F. P., Fowler, D. M., & Rubin, A. F. (2019). MaveDB: An open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology*, *20*(1), 223. https://doi.org/10.1186/s13059-019-1845-6

Ewens, W. J. (2004). *Mathematical population genetics: Theoretical introduction* (Vol. 1). Springer.

Fagiolo, U., Cossarizza, A., Scala, E., Fanales-Belasio, E., Ortolani, C., Cozzi, E., Monti, D., Franceschi, C., & Paganelli, R. (1993). Increased cytokine production in mononuclear cells of healthy elderly people. *European Journal of Immunology*, *23*(9), 2375–2378. https://doi.org/10.1002/eji.1830230950

Falconer, D. S. (1996). *Introduction to quantitative genetics*. Pearson Education India.

Farré, X., Molina, R., Barteri, F., Timmers, P. R. H. J., Joshi, P. K., Oliva, B., Acosta, S., Esteve-Altava, B., Navarro, A., & Muntané, G. (2021). Comparative analysis of mammal genomes unveils key genomic variability for human life span (K. Nowick, Ed.). *Molecular Biology and Evolution*, *38*(11), 4948–4961. https://doi.org/10.1093/molbev/msab219

Faure, A. J., Schmiedel, J. M., Baeza-Centurion, P., & Lehner, B. (2020). DiMSum: An error model and pipeline for analyzing deep mutational scanning data and diagnosing

common experimental pathologies. *Genome Biology*, *21*(1), 207. https://doi.org/10.1186/s13059-020-02091-3

Fernández, M. E., Goszczynski, D. E., Lirón, J. P., Villegas-Castagnasso, E. E., Carino, M. H., Ripoli, M. V., Rogberg-Muñoz, A., Posik, D. M., Peral-García, P., & Giovambattista, G. (2013). Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, traceability and assessment of parentage in an inbred angus herd [Publisher: SciELO Brasil]. *Genetics and molecular biology*, *36*, 185–191.

Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J., & Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *NATURE BIOTECHNOLOGY*, *22*(10), 5.

Ferrucci, L., Schrack, J. A., Knuth, N. D., & Simonsick, E. M. (2012). Aging and the energetic cost of life. *Journal of the American Geriatrics Society*, *60*(9), 1768–1769. https://doi.org/10.1111/j.1532-5415.2012.04102.x

Fields, S., & Song, O.-k. (1989). A novel genetic system to detect protein–protein interactions. *Nature*, *340*(6230), 245–246. https://doi.org/10.1038/340245a0

Fink, A. L. (1998). Protein aggregation: Folding aggregates, inclusion bodies and amyloid. *Folding and Design*, *3*(1), R9–R23. https://doi.org/https://doi.org/10.1016/S1359-0278(98)00002-9

Fontana, L., Partridge, L., & Longo, V. D. (2010). Extending healthy life span—from yeast to humans. *Science*, *328*(5976), 321–326. https://doi.org/10.1126/science.1172539

Franceschi, C., Monti, D., Sansoni, P., & Cossarizza, A. (1995). The immunology of exceptional individuals: The lesson of centenarians. *Immunology Today*, *16*(1), 12–16. https://doi.org/10.1016/0167-5699(95)80064-6

Frankel, D., Davies, M., Bhushan, B., Kulaberoglu, Y., Urriola-Munoz, P., Bertrand-Michel, J., Pergande, M. R., Smith, A. A., Preet, S., Park, T. J., Vendruscolo, M., Rankin, K. S., Cologna, S. M., Kumita, J. R., Cenac, N., & St John Smith, E. (2020). Cholesterol-rich naked mole-rat brain lipid membranes are susceptible to amyloid beta-induced damage in vitro. *Aging*, *12*(21), 22266–22290. https://doi.org/10.18632/aging.202138

French, J. D., & Edwards, S. L. (2020). The role of noncoding variants in heritable disease. *Trends in Genetics*, *36*(11), 880–891. https://doi.org/https://doi.org/10.1016/j.tig.2020.07.004

Frenk, S., & Houseley, J. (2018). Gene expression hallmarks of cellular ageing. *Biogerontology*, *19*(6), 547–566. https://doi.org/10.1007/s10522-018-9750-z

Frizzell, M. A. (2013). Incomplete dominance. In *Brenner's encyclopedia of genetics (second edition)* (Second Edition, pp. 58–60). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-374984-0.00784-1

Fromont-Racine, M., Rain, J. C., & Legrain, P. (1997). Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nature Genetics*, *16*(3), 277–282. https://doi.org/10.1038/ng0797-277

Furlong, L. I. (2013). Human diseases through the lens of network biology. *Trends in Genetics*, *29*(3), 150–159. https://doi.org/https://doi.org/10.1016/j.tig.2012.11.004

Fushan, A. A., Turanov, A. A., Lee, S.-G., Kim, E. B., Lobanov, A. V., Yim, S. H., Buffenstein, R., Lee, S.-R., Chang, K.-T., Rhee, H., Kim, J.-S., Yang, K.-S., & Gladyshev, V. N. (2015). Gene expression defines natural changes in mammalian lifespan. *Aging Cell*, *14*(3), 352–365. https://doi.org/10.1111/acel.12283

Gamache, I., Legault, M.-A., Grenier, J.-C., Sanchez, R., Rhéaume, E., Asgari, S., Barhdadi, A., Zada, Y. F., Trochet, H., Luo, Y., Lecca, L., Murray, M., Raychaudhuri, S., Tardif, J.-C., Dubé, M.-P., & Hussin, J. (2021). A sex-specific evolutionary interaction between ADCY9 and CETP. *eLife*, *10*, e69198. https://doi.org/10.7554/eLife.69198

Gao, R., Liu, Y., Silva-Fernandes, A., Fang, X., Paulucci-Holthauzen, A., Chatterjee, A., Zhang, H. L., Matsuura, T., Choudhary, S., Ashizawa, T., Koeppen, A. H., Maciel, P., Hazra, T. K., & Sarkar, P. S. (2015). Inactivation of PNKP by mutant ATXN3 triggers apoptosis by activating the DNA damage-response pathway in SCA3 (C. E. Pearson, Ed.). *PLOS Genetics*, *11*(1), e1004834. https://doi.org/10.1371/journal.pgen.1004834

Gaspar, H. A., & Breen, G. (2019). Probabilistic ancestry maps: A method to assess and visualize population substructures in genetics. *BMC Bioinformatics*, *20*(1), 116. https://doi.org/10.1186/s12859-019-2680-1

Gauthier, L., Stynen, B., Serohijos, A., & Michnick, S. (2020). Genetics' piece of the PI: Inferring the origin of complex traits and diseases from proteome-wide protein-protein interaction dynamics. *Bioessays*, *42*, 1900169.

Gietz, R. D., & Schiestl, R. H. (2007). Frozen competent yeast cells that can be transformed with high efficiency using the LiAc/SS carrier DNA/PEG method. *Nature Protocols*, *2*(1), 1–4. https://doi.org/10.1038/nprot.2007.17

Gilad, Y., Rifkin, S. A., & Pritchard, J. K. (2008). Revealing the architecture of gene regulation: The promise of eQTL studies [Publisher: Elsevier]. *Trends in genetics*, *24*(8), 408–415.

Goede, O., Nachun, D., Ferraro, N., Gloudemans, M., Rao, A., Smail, C., & Li, Q. (2021). Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. *Cell*, *184*(10), 2633–2648.

Gold, L., Ayers, D., Bertino, J., Bock, C., Bock, A., Brody, E. N., Carter, J., Dalby, A. B., Eaton, B. E., Fitzwater, T., Flather, D., Forbes, A., Foreman, T., Fowler, C., Gawande, B., Goss, M., Gunn, M., Gupta, S., Halladay, D., ... Zichi, D. (2010). Aptamer-based

multiplexed proteomic technology for biomarker discovery [Publisher: Public Library of Science]. *PLOS ONE*, *5*(12), 1–17. https://doi.org/10.1371/journal.pone.0015004

Goldberg, D., & Roth, F. (2003). Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences*, *100*(8), 4372–4376.

Golubev, A., Hanson, A. D., & Gladyshev, V. N. (2017). Non-enzymatic molecular damage as a prototypic driver of aging. *Journal of Biological Chemistry*, *292*(15), 6029–6038. https://doi.org/10.1074/jbc.R116.751164

Goodell, M. A., & Rando, T. A. (2015). Stem cells and healthy aging. *Science*, *350*(6265), 1199–1204. https://doi.org/10.1126/science.aab3388

Gray, A. (1860). Darwin on the origin of species [Publisher: Taylor & Francis]. *Annals and Magazine of Natural History*, *6*(35), 373–386.

Green, E. (2022). *Shotgun sequencing* [NIH]. Retrieved December 27, 2022, from https://www.genome.gov/genetics-glossary/Shotgun-Sequencing

Greenbaum, D., Colangelo, C., Williams, K., & Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology*, *4*(9), 117. https://doi.org/10.1186/gb-2003-4-9-117

Griffin, N. M., Yu, J., Long, F., Oh, P., Shore, S., Li, Y., Koziol, J. A., & Schnitzer, J. E. (2010). Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nature Biotechnology*, *28*(1), 83–89. https://doi.org/10.1038/nbt.1592

Groh, N., Bühler, A., Huang, C., Li, K. W., van Nierop, P., Smit, A. B., Fändrich, M., Baumann, F., & David, D. C. (2017). Age-dependent protein aggregation initiates amyloid- aggregation. *Frontiers in Aging Neuroscience*, *9*, 138. https://doi.org/10.3389/fnagi.2017.00138

Guo, H. H., Choe, J., & Loeb, L. A. (2004). Protein tolerance to random amino acid change. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(25), 9205–9210. https://doi.org/10.1073/pnas.0403255101

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., & Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, *17*(10), 994–999. https://doi.org/10.1038/13690

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th python in science conference* (pp. 11–15). http://conference.scipy.org/proceedings/SciPy2008/paper_2/

Hägg, S., & Jylhävä, J. (2021). Sex differences in biological aging with a focus on human studies (Y. Suh & J. K. Tyler, Eds.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *10*, e63425. https://doi.org/10.7554/eLife.63425

Hanke, S., Besir, H., Oesterhelt, D., & Mann, M. (2008). Absolute SILAC for accurate quantitation of proteins in complex mixtures down to the attomole level. *Journal of Proteome Research*, *7*(3), 1118–1130. https://doi.org/10.1021/pr7007175

Harraan, D. (1955). Aging: A theory based on free radical and radiation chemistry.

Harrison, D. E., Strong, R., Sharp, Z. D., Nelson, J. F., Astle, C. M., Flurkey, K., Nadon, N. L., Wilkinson, J. E., Frenkel, K., Carter, C. S., Pahor, M., Javors, M. A., Fernandez, E., & Miller, R. A. (2009). Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature*, *460*(7253), 392–395. https://doi.org/10.1038/nature08221

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, *18*(1), 83. https://doi.org/10.1186/s13059-017-1215-1

Hein, M., Hubner, N., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., & Mann, M. (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell*, *163*(3), 712–723.

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, *38*(6), 226–231. https://doi.org/10.1007/BF01245622

Hilton, H. G., Rubinstein, N. D., Janki, P., Ireland, A. T., Bernstein, N., Fong, N. L., Wright, K. M., Smith, M., Finkle, D., Martin-McNulty, B., Roy, M., Imai, D. M., Jojic, V., & Buffenstein, R. (2019). Single-cell transcriptomics of the naked mole-rat reveals unexpected features of mammalian immunity (A. Bhandoola, Ed.). *PLOS Biology*, *17*(11), e3000528. https://doi.org/10.1371/journal.pbio.3000528

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits [Publisher: National Acad Sciences]. *Proceedings of the National Academy of Sciences*, *106*(23), 9362–9367.

Hipp, M. S., Kasturi, P., & Hartl, F. U. (2019). The proteostasis network and its decline in ageing. *Nature Reviews Molecular Cell Biology*, *20*(7), 421–435. https://doi.org/10.1038/s41580-019-0101-y

Hoeijmakers, J. H. (2009). DNA damage, aging, and cancer [Publisher: Mass Medical Soc]. *New England Journal of Medicine*, *361*(15), 1475–1485.

Høie, M. H., Cagiada, M., Frederiksen, A. H. B., Stein, A., & Lindorff-Larsen, K. (2022). Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Reports*, *38*(2), 110207. https://doi.org/https://doi.org/10.1016/j.celrep.2021.110207

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., & Zamir, A. (1965). Structure of a ribonucleic acid. *Science*, *147*(3664), 1462–1465. https://doi.org/10.1126/science.147.3664.1462

Horvath, S., & Raj, K. (2018). DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature Reviews Genetics*, *19*(6), 371–384. https://doi.org/10.1038/s41576-018-0004-3

Horwitz, A. (2015). Efficient multiplexed integration of synergistic alleles and metabolic pathways in yeasts via CRISPR-cas. *Cell Syst*, *1*, 88–96.

Huang, J., Zhu, H., Haggarty, S. J., Spring, D. R., Hwang, H., Jin, F., Snyder, M., & Schreiber, S. L. (2004). Finding new components of the target of rapamycin (TOR) signaling network through chemical genetics and proteome chips. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(47), 16594–16599. https://doi.org/10.1073/pnas.0407117101

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A., & Mackay, T. (2014). Natural variation in genome architecture among 205 drosophila melanogaster genetic reference panel lines. *Genome research*, *24*(7), 1193–1208.

Hudry, B., Viala, S., Graba, Y., & Merabet, S. (2011). Visualization of protein interactions in living drosophila embryos by the bimolecular fluorescence complementation assay. *BMC Biology*, *9*(1), 5. https://doi.org/10.1186/1741-7007-9-5

Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, *132*(2), 583–589. https://doi.org/10.1093/genetics/132.2.583

Hudson, R. (2004). Linkage disequilibrium and recombination [Publisher: Wiley Online Library]. *Handbook of statistical genetics*.

Irvine, G. B., El-Agnaf, O. M., Shankar, G. M., & Walsh, D. M. (2008). Protein aggregation in the brain: The molecular basis for alzheimer's and parkinson's diseases. *Molecular Medicine*, *14*(7), 451–464. https://doi.org/10.2119/2007-00100.Irvine

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., & Mann, M. (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Molecular & cellular proteomics: MCP*, *4*(9), 1265–1272. https://doi.org/10.1074/mcp.M500061-MCP200

Jackson, M. P., & Hewitt, E. W. (2016). Cellular proteostasis: Degradation of misfolded proteins by lysosomes. *Essays in Biochemistry*, *60*(2), 173–180. https://doi.org/10.1042/EBC20160005

Jackson, R. J., Hellen, C. U. T., & Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews Molecular Cell Biology*, *11*(2), 113–127. https://doi.org/10.1038/nrm2838

Jakobson, C. M., & Jarosz, D. F. (2019). Molecular origins of complex heritability in natural genotype-to-phenotype relationships. *Cell Systems*, *8*(5), 363–379.e3. https://doi.org/10.1016/j.cels.2019.04.002

Jakociunas, T., Jensen, M., & Keasling, J. (2016). CRISPR/cas9 advances engineering of microbial cell factories. *Metab Eng*, *34*, 44–59.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., ... D'Eustachio, P. (2019). The reactome pathway knowledgebase. *Nucleic Acids Research*, gkz1031. https://doi.org/10.1093/nar/gkz1031

Jin, C., Li, J., Green, C. D., Yu, X., Tang, X., Han, D., Xian, B., Wang, D., Huang, X., Cao, X., et al. (2011). Histone demethylase UTX-1 regulates c. elegans life span by targeting the insulin/IGF-1 signaling pathway [Publisher: Elsevier]. *Cell metabolism*, *14*(2), 161–172.

Joo, Y., Ficarro, S., Soares, L., Chun, Y., Marto, J., & Buratowski, S. (2017). Downstream promoter interactions of TFIID TAFs facilitate transcription reinitiation. *Genes & Development*, *31*(21), 2162–2174.

Jordá, T., & Puig, S. (2020). Regulation of ergosterol biosynthesis in saccharomyces cerevisiae. *Genes*, *11*(7), 795. https://doi.org/10.3390/genes11070795

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. https://doi.org/10.1002/bip.360221211

Kachroo, A., Laurent, J., Yellman, C., Meyer, A., Wilke, C., & Marcotte, E. (2015). Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science*, *348*(6237), 921–925.

Kaiser, S. E., Riley, B. E., Shaler, T. A., Trevino, R. S., Becker, C. H., Schulman, H., & Kopito, R. R. (2011). Protein standard absolute quantification (PSAQ) method for the measurement of cellular ubiquitin pools. *Nature Methods*, *8*(8), 691–696. https://doi.org/10.1038/nmeth.1649

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., & Yamanishi, Y. (2007). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, *36*, D480–D484. https://doi.org/10.1093/nar/gkm882

Kao, K., & Sherlock, G. (2008). Molecular characterization of clonal interference during adaptive evolution in asexual populations of saccharomyces cerevisiae. *Nature genetics*, *40*(12), 1499–1504.

Kennedy, B. K., Berger, S. L., Brunet, A., Campisi, J., Cuervo, A. M., Epel, E. S., Franceschi, C., Lithgow, G. J., Morimoto, R. I., Pessin, J. E., Rando, T. A., Richardson, A., Schadt, E. E., Wyss-Coray, T., & Sierra, F. (2014). Geroscience: Linking aging to chronic disease. *Cell*, *159*(4), 709–713. https://doi.org/10.1016/j.cell.2014.10.039

Kennedy, S. R., Loeb, L. A., & Herr, A. J. (2012). Somatic mutations in aging, cancer and neurodegeneration. *Mechanisms of Ageing and Development*, *133*(4), 118–126. https://doi.org/10.1016/j.mad.2011.10.009

Kirkpatrick, D. S., Gerber, S. A., & Gygi, S. P. (2005). The absolute quantification strategy: A general procedure for the quantification of proteins and post-translational modifications. *Methods (San Diego, Calif.)*, *35*(3), 265–273. https://doi.org/10.1016/j.ymeth.2004.08.018

Kirkwood, T. B. (1977). Evolution of ageing [Publisher: Nature Publishing Group]. *Nature*, *270*(5635), 301–304.

Klug, W. S., & Ward, S. M. (Eds.). (2013). *Essentials of genetics* (8th ed). Pearson.

Koga, H., Kaushik, S., & Cuervo, A. M. (2011). Protein homeostasis and aging: The importance of exquisite quality control. *Ageing Research Reviews*, *10*(2), 205–215. https://doi.org/10.1016/j.arr.2010.02.001

Kraft, P., Zeggini, E., & Ioannidis, J. P. A. (2009). Replication in genome-wide association studies. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *24*(4), 561–573. https://doi.org/10.1214/09-STS290

Kreitmaier, P., Katsoula, G., & Zeggini, E. (2023). Insights from multi-omics integration in complex disease primary tissues. *Trends in Genetics*, *39*(1), 46–58. https://doi.org/10.1016/j.tig.2022.08.005

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019). OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, *47*, D807–D811. https://doi.org/10.1093/nar/gky1053

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes11edited by f. cohen. *Journal of Molecular Biology*, *305*(3), 567–580. https://doi.org/10.1006/jmbi.2000.4315

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution*, *34*(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

Labbadia, J., & Morimoto, R. I. (2015). The biology of proteostasis in aging and disease. *Annual Review of Biochemistry*, *84*(1), 435–464. https://doi.org/10.1146/annurev-biochem-060614-033955

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Research*, *37*(13), 4181–4193. https://doi.org/10.1093/nar/gkp552

Lee, C.-M., Adamchek, C., Feke, A., Nusinow, D. A., & Gendron, J. M. (2017). Mapping protein–protein interactions using affinity purification and mass spectrometry. In W. Busch (Ed.), *Plant genomics: Methods and protocols* (pp. 231–249). Springer. https://doi.org/10.1007/978-1-4939-7003-2_15

Lee, M. B., Dowsett, I. T., Carr, D. T., Wasko, B. M., Stanton, S. G., Chung, M. S., Ghodsian, N., Bode, A., Kiflezghi, M. G., Uppal, P. A., Grayden, K. A., Elala, Y. C., Tang, T. T., Tran, N. H. B., Tran, T. H. B., Diep, A. B., Hope, M., Promislow, D. E. L., Kennedy, S. R., ... Herr, A. J. (2019). Defining the impact of mutation accumulation on replicative lifespan in yeast using cancer-associated mutator phenotypes. *Proceedings of the National Academy of Sciences*, *116*(8), 3062–3071. https://doi.org/10.1073/pnas.1815966116

Lemaître, J.-F., Ronget, V., Tidière, M., Allainé, D., Berger, V., Cohas, A., Colchero, F., Conde, D. A., Garratt, M., Liker, A., Marais, G. A. B., Scheuerlein, A., Székely, T., & Gaillard, J.-M. (2020). Sex differences in adult lifespan and aging rates of mortality across wild mammals. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(15), 8546–8553. https://doi.org/10.1073/pnas.1911999117

Levy, S., Blundell, J., Venkataram, S., Petrov, D., Fisher, D., & Sherlock, G. (2015). Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*, *519*(7542), 181–186.

Lewontin, R. C. (1964). THE INTERACTION OF SELECTION AND LINKAGE. i. GENERAL CONSIDERATIONS; HETEROTIC MODELS. *Genetics*, *49*(1), 49–67. https://doi.org/10.1093/genetics/49.1.49

Li, F., Salit, M. L., & Levy, S. F. (2018). Unbiased fitness estimation of pooled barcode or amplicon sequencing studies. *Cell Systems*, *7*(5), 521–525.e4. https://doi.org/10.1016/j.cels.2018.09.004

Li, J., Zhang, D., Wiersma, M., & Brundel, B. J. (2018). Role of autophagy in proteostasis: Friend and foe in cardiac diseases. *Cells*, *7*(12), 279. https://doi.org/10.3390/cells7120279

Li, Y. (2016). RNA splicing is a primary link between genetic variation and disease. *Science*, *352*, 600–604.

Lim, J., Petersen, M., Bunz, M., Simon, C., & Schindler, M. (2022). Flow cytometry based-FRET: Basics, novel developments and future perspectives. *Cellular and Molecular Life Sciences*, *79*(4), 217. https://doi.org/10.1007/s00018-022-04232-2

Linding, R., Schymkowitz, J., Rousseau, F., Diella, F., & Serrano, L. (2004). A comparative study of the relationship between protein structure and -aggregation in globular and intrinsically disordered proteins. *Journal of Molecular Biology*, *342*(1), 345–353. https://doi.org/10.1016/j.jmb.2004.06.088

Liu, X., Li, Y. I., & Pritchard, J. K. (2019). Trans effects on gene expression can drive omnigenic inheritance. *Cell*, *177*(4), 1022–1034.e6. https://doi.org/10.1016/j.cell.2019.04.014

Liu, Y., Wang, X., & Liu, B. (2019). A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Briefings in Bioinformatics*, *20*(1), 330–346. https://doi.org/10.1093/bib/bbx126

Lodato, M. A., Rodin, R. E., Bohrson, C. L., Coulter, M. E., Barton, A. R., Kwon, M., Sherman, M. A., Vitzthum, C. M., Luquette, L. J., Yandava, C. N., Yang, P., Chittenden, T. W., Hatem, N. E., Ryu, S. C., Woodworth, M. B., Park, P. J., & Walsh, C. A. (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science*, *359*(6375), 555–559. https://doi.org/10.1126/science.aao4426

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2013). The hallmarks of aging. *Cell*, *153*(6), 1194–1217. https://doi.org/10.1016/j.cell.2013.05.039

Lu, P., Vogel, C., Wang, R., Yao, X., & Marcotte, E. M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotechnology*, *25*(1), 117–124. https://doi.org/10.1038/nbt1270

Luheshi, L. M., Tartaglia, G. G., Brorsson, A.-C., Pawar, A. P., Watson, I. E., Chiti, F., Vendruscolo, M., Lomas, D. A., Dobson, C. M., & Crowther, D. C. (2007). Systematic in vivo analysis of the intrinsic determinants of amyloid beta pathogenicity. *PLoS biology*, *5*(11), e290. https://doi.org/10.1371/journal.pbio.0050290

Lundberg, M., Eriksson, A., Tran, B., Assarsson, E., & Fredriksson, S. (2011). Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood. *Nucleic Acids Research*, *39*(15), e102. https://doi.org/10.1093/nar/gkr424

Ma, S., & Gladyshev, V. N. (2017). Molecular signatures of longevity: Insights from cross-species comparative studies. *Seminars in Cell & Developmental Biology*, *70*, 190–203. https://doi.org/10.1016/j.semcdb.2017.08.007

Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M., Richardson, M. F., Anholt, R. R. H., Barrón, M., Bess, C., Blankenburg, K. P., Carbone, M. A., Castellano, D., Chaboub, L., Duncan, L., … Gibbs, R. A. (2012). The drosophila melanogaster genetic reference panel. *Nature*, *482*(7384), 173–178. https://doi.org/10.1038/nature10811

Magalhães, J. P. d., Costa, J., & Church, G. M. (2007). An analysis of the relationship between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts. *The Journals of Gerontology: Series A*, *62*(2), 149–160. https://doi.org/10.1093/gerona/62.2.149

Manolio, T., Collins, F., Cox, N., Goldstein, D., Hindorff, L., Hunter, D., & Visscher, P. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511. https://doi.org/10.1038/nrg2796

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R., Kjaer, T. R., Fine, R. S., Lu, Y., Schurmann, C., Highland, H. M., Rüeger, S., Thorleifsson, G., Justice, A. E., Lamparter, D., Stirrups, K. E., Turcot, V., Young, K. L., Winkler, T. W., Esko, T., ... Lettre, G. (2017). Rare and low-frequency coding variants alter human adult height. *Nature*, *542*(7640), 186–190. https://doi.org/10.1038/nature21039

Marquardt, S., Hazelbaker, D. Z., & Buratowski, S. (2011). Distinct RNA degradation pathways and 3' extensions of yeast non-coding RNA species. *Transcription*, *2*(3), 145–154. https://doi.org/10.4161/trns.2.3.16298

Marquioni, V., Nunes, F. M. F., & Novo-Mansur, M. T. M. (2021). Protein identification by database searching of mass spectrometry data in the teaching of proteomics [Publisher: American Chemical Society]. *Journal of Chemical Education*, *98*(3), 812–823. https://doi.org/10.1021/acs.jchemed.0c00853

Martinez-Rivas, G., Bender, S., & Sirac, C. (2022). Understanding AL amyloidosis with a little help from in vivo models. *Frontiers in Immunology*, *13*, 1008449. https://doi.org/10.3389/fimmu.2022.1008449

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., Menon, M., He, L., Abdurrob, F., Jiang, X., Martorell, A. J., Ransohoff, R. M., Hafler, B. P., Bennett, D. A., Kellis, M., & Tsai, L.-H. (2019). Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, *570*(7761), 332–337. https://doi.org/10.1038/s41586-019-1195-2

Maynard, S., Fang, E. F., Scheibye-Knudsen, M., Croteau, D. L., & Bohr, V. A. (2015). DNA damage, DNA repair, aging, and neurodegeneration. *Cold Spring Harbor Perspectives in Medicine*, *5*(10), a025130. https://doi.org/10.1101/cshperspect.a025130

McCarroll, S. A., Murphy, C. T., Zou, S., Pletcher, S. D., Chin, C.-S., Jan, Y. N., Kenyon, C., Bargmann, C. I., & Li, H. (2004). Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nature Genetics*, *36*(2), 197–204. https://doi.org/10.1038/ng1291

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), 356–369. https://doi.org/10.1038/nrg2344

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS genetics*, *5*(10), e1000686. https://doi.org/10.1371/journal.pgen.1000686

Medawar, P. B. (1952). An unsolved problem of biology [Publisher: College].

Melnik, B. S., Katina, N. S., Ryabova, N. A., Marchenkov, V. V., Melnik, T. N., Karuzina, N. E., & Nemtseva, E. V. (2022). Relationship between changes in the protein folding pathway and the process of amyloid formation: The case of bovine carbonic anhydrase II. *International Journal of Molecular Sciences*, *23*(23), 14645. https://doi.org/10.3390/ijms232314645

Members of the Complex Trait Consortium. (2003). The nature and identification of quantitative trait loci: A community's view. *Nature Reviews Genetics*, *4*(11), 911–916. https://doi.org/10.1038/nrg1206

Mendel, G. (1865). Versuche uber pflanzen-hybriden. *Vorgelegt in den Sitzungen*.

Messier, V., Zenklusen, D., & Michnick, S. (2013). A nutrient-responsive pathway that determines m phase timing through control of b-cyclin mRNA stability. *Cell*, *153*, 1080–1093.

Messner, C., Demichev, V., Muenzner, J., Aulakh, S., Barthel, N., Röhl, A., & Ralser, M. (2023). The proteomic landscape of genome-wide genetic perturbations. *Cell*, *186*(9), 2018–2034.

Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., & Thomas, P. D. (2019). Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nature Protocols*, *14*(3), 703–721. https://doi.org/10.1038/s41596-019-0128-8

Michnick, S. W. (2003). Protein fragment complementation strategies for biochemical network mapping. *Current Opinion in Biotechnology*, *14*(6), 610–617. https://doi.org/10.1016/j.copbio.2003.10.014

Michnick, S., Ear, P., Manderson, E., Remy, I., & Stefan, E. (2007). Universal strategies in research and drug discovery based on protein-fragment complementation assays. *Nat Rev Drug Discov*, *6*, 569–582.

Mitchell, S. J., Scheibye-Knudsen, M., Longo, D. L., & de Cabo, R. (2015). Animal models of aging research: Implications for human aging and age-related diseases. *Annual Review of Animal Biosciences*, *3*(1), 283–303. https://doi.org/10.1146/annurev-animal-022114-110829

Molendijk, J., & Parker, B. L. (2021). Proteome-wide systems genetics to identify functional regulators of complex traits. *Cell Systems*, *12*(1), 5–22. https://doi.org/10.1016/j.cels.2020.10.005

Moore, R., Casale, F. P., Jan Bonder, M., Horta, D., BIOS Consortium, Franke, L., Barroso, I., & Stegle, O. (2019). A linear mixed-model approach to study multivariate gene-environment interactions. *Nature Genetics*, *51*(1), 180–186. https://doi.org/10.1038/s41588-018-0271-0

Morimoto, R. I., & Cuervo, A. M. (2009). Protein homeostasis and aging: Taking care of proteins from the cradle to the grave. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *64A*(2), 167–170. https://doi.org/10.1093/gerona/gln071

Morimoto, R. I., & Cuervo, A. M. (2014). Proteostasis and the aging proteome in health and disease. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, *69 Suppl 1*, S33–38. https://doi.org/10.1093/gerona/glu049

Moskalev, A. A., Shaposhnikov, M. V., Plyusnina, E. N., Zhavoronkov, A., Budovsky, A., Yanai, H., & Fraifeld, V. E. (2013). The role of DNA damage and repair in aging through the prism of koch-like criteria. *Ageing Research Reviews*, *12*(2), 661–684. https://doi.org/10.1016/j.arr.2012.02.001

Ng, P. C., & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research*, *11*(5), 863–874. https://doi.org/10.1101/gr.176601

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: Present and future. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*(1620), 20120362. https://doi.org/10.1098/rstb.2012.0362

Nicastro, G., Masino, L., Esposito, V., Menon, R. P., De Simone, A., Fraternali, F., & Pastore, A. (2009). Josephin domain of ataxin-3 contains two distinct ubiquitin-binding sites. *Biopolymers*, *91*(12), 1203–1214. https://doi.org/10.1002/bip.21210

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., & Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics: MCP*, *1*(5), 376–386. https://doi.org/10.1074/mcp.m200025-mcp200

Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., Zhang, F., Dolma, S., Willems, A., Coulombe-Huntington, J., Chatr-aryamontri, A., Dolinski, K., & Tyers, M. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, *47*, D529–D541. https://doi.org/10.1093/nar/gky1079

Paik, H., Kim, J., Lee, S., Heo, H.-S., Hur, C.-G., & Lee, D. (2012). Prioritization of SNPs for genome-wide association studies using an interaction model of genetic variation, gene expression, and trait variation. *Molecules and Cells*, *33*(4), 351–361. https://doi.org/10.1007/s10059-012-2264-7

Papadopoli, D., Boulay, K., Kazak, L., Pollak, M., Mallette, F., Topisirovic, I., & Hulea, L. (2019). mTOR as a central regulator of lifespan and aging. *F1000Research*, *8*, F1000 Faculty Rev–998. https://doi.org/10.12688/f1000research.17196.1

Park, C. B., & Larsson, N.-G. (2011). Mitochondrial DNA mutations in disease and aging [Publisher: The Rockefeller University Press]. *Journal of cell biology*, *193*(5), 809–818.

Parker, S., Fraczek, M., Wu, J., Shamsah, S., Manousaki, A., Dungrattanalert, K., & O'Keefe, R. (2018). Large-scale profiling of noncoding RNA function in yeast. *PLoS genetics*, *14*(3), 1007253.

Partridge, L., & Gems, D. (2007). Benchmarks for ageing studies. *Nature*, *450*(7167), 165–167. https://doi.org/10.1038/450165a

Parveen, F., Bender, D., Law, S.-H., Mishra, V. K., Chen, C.-C., & Ke, L.-Y. (2019). Role of ceramidases in sphingolipid metabolism and human diseases. *Cells*, *8*(12), E1573. https://doi.org/10.3390/cells8121573

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Pérez, V. I., Buffenstein, R., Masamsetti, V., Leonard, S., Salmon, A. B., Mele, J., Andziak, B., Yang, T., Edrey, Y., Friguet, B., Ward, W., Richardson, A., & Chaudhuri, A. (2009). Protein stability and resistance to oxidative stress are determinants of longevity in the longest-living rodent, the naked mole-rat. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(9), 3059–3064. https://doi.org/10.1073/pnas.0809620106

Perls, T. T. (2007). Centenarians. In *Encyclopedia of gerontology (second edition)* (Second Edition, pp. 269–275). Elsevier. https://doi.org/https://doi.org/10.1016/B0-12-370870-2/00035-4

Pevtsov, S., Fedulova, I., Mirzaei, H., Buck, C., & Zhang, X. (2006). Performance evaluation of existing de novo sequencing algorithms. *Journal of Proteome Research*, *5*(11), 3018–3028. https://doi.org/10.1021/pr060222h

Pickering, A. M., & Davies, K. J. (2012). Degradation of damaged proteins: The main function of the 20s proteasome [Publisher: Elsevier]. *Progress in molecular biology and translational science*, *109*, 227–248.

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, *94*, 559–573.

Picotti, P., Clément-Ziza, M., Lam, H., Campbell, D. S., Schmidt, A., Deutsch, E. W., Röst, H., Sun, Z., Rinner, O., Reiter, L., Shen, Q., Michaelson, J. J., Frei, A., Alberti, S., Kusebauch, U., Wollscheid, B., Moritz, R. L., Beyer, A., & Aebersold, R. (2013). A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature*, *494*(7436), 266–270. https://doi.org/10.1038/nature11835

Piper, M. D., Selman, C., McElwee, J. J., & Partridge, L. (2005). Models of insulin signalling and longevity. *Drug Discovery Today: Disease Models*, *2*(4), 249–256. https://doi.org/10.1016/j.ddmod.2005.11.001

Powers, E. T., Morimoto, R. I., Dillin, A., Kelly, J. W., & Balch, W. E. (2009). Biological and chemical approaches to diseases of proteostasis deficiency. *Annual Review of Biochemistry*, *78*(1), 959–991. https://doi.org/10.1146/annurev.biochem.052308.114844

Pritchard, J. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*, *69*, 124–137.

Pritchard, J. K., & Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data [Publisher: Elsevier]. *The American Journal of Human Genetics*, *69*(1), 1–14.

Proctor, C. J., & Lorimer, I. A. J. (2011). Modelling the role of the hsp70/hsp90 system in the maintenance of protein homeostasis (C. Chan, Ed.). *PLoS ONE*, *6*(7), e22038. https://doi.org/10.1371/journal.pone.0022038

Qi, X., Yang, G., & Liu, L. (2020). Robustness analysis of the networks in cascading failures with controllable parameters. *Physica A: Statistical Mechanics and its Applications*, *539*, 122870. https://doi.org/10.1016/j.physa.2019.122870

Qu, J., Zou, T., & Lin, Z. (2021). The roles of the ubiquitin–proteasome system in the endoplasmic reticulum stress pathway. *International Journal of Molecular Sciences*, *22*(4), 1526. https://doi.org/10.3390/ijms22041526

Quinn, J., & Chang, H. (2016). Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics*, *17*(1), 47–62.

Ramanan, V. K., & Saykin, A. J. (2013). Pathways to neurodegeneration: Mechanistic insights from GWAS in alzheimer's disease, parkinson's disease, and related disorders. *American Journal of Neurodegenerative Disease*, *2*(3), 145–175.

Rappsilber, J., Ryder, U., Lamond, A. I., & Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. *Genome Research*, *12*(8), 1231–1245. https://doi.org/10.1101/gr.473902

Reeb, J., & Rost, B. (2019). Secondary structure prediction. In *Encyclopedia of bioinformatics and computational biology* (pp. 488–496). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20267-7

Remm, M., Storm, C. E., & Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, *314*(5), 1041–1052. https://doi.org/10.1006/jmbi.2000.5197

Richards, A. L., Eckhardt, M., & Krogan, N. J. (2021). Mass spectrometry-based protein-protein interaction networks for the study of human diseases. *Molecular Systems Biology*, *17*(1), e8792. https://doi.org/10.15252/msb.20188792

Robine, J.-M., & Cubaynes, S. (2017). Worldwide demography of centenarians. *Mechanisms of Ageing and Development*, *165*, 59–67. https://doi.org/10.1016/j.mad.2017.03.004

Rodriguez, K. A., Edrey, Y. H., Osmulski, P., Gaczynska, M., & Buffenstein, R. (2012). Altered composition of liver proteasome assemblies contributes to enhanced proteasome activity in the exceptionally long-lived naked mole-rat (J. L. Brodsky, Ed.). *PLoS ONE*, *7*(5), e35890. https://doi.org/10.1371/journal.pone.0035890

Rodriguez, K. A., Valentine, J. M., Kramer, D. A., Gelfond, J. A., Kristan, D. M., Nevo, E., & Buffenstein, R. (2016). Determinants of rodent longevity in the chaperone-protein

degradation network. *Cell Stress and Chaperones*, *21*(3), 453–466. https://doi.org/10.1007/s12192-016-0672-x

Roepstorff, P., & Fohlman, J. (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical Mass Spectrometry*, *11*(11), 601. https://doi.org/10.1002/bms.1200111109

Romanov, N., Kuhn, M., Aebersold, R., Ori, A., Beck, M., & Bork, P. (2019). Disentangling genetic and environmental effects on the proteotypes of individuals. *Cell*, *177*(5), 1308–1318.

Ross, C. A., & Poirier, M. A. (2004). Protein aggregation and neurodegenerative disease. *Nature Medicine*, *10*, S10–S17. https://doi.org/10.1038/nm1066

Rossi, M. J., Kuntala, P. K., Lai, W. K. M., Yamada, N., Badjatia, N., Mittal, C., Kuzu, G., Bocklund, K., Farrell, N. P., Blanda, T. R., Mairose, J. D., Basting, A. V., Mistretta, K. S., Rocco, D. J., Perkinson, E. S., Kellogg, G. D., Mahony, S., & Pugh, B. F. (2021). A high-resolution protein architecture of the budding yeast genome. *Nature*, *592*(7853), 309–314. https://doi.org/10.1038/s41586-021-03314-8

Ruby, J. G., Smith, M., & Buffenstein, R. (2018). Naked mole-rat mortality rates defy gompertzian laws by not increasing with age. *eLife*, *7*, e31157. https://doi.org/10.7554/eLife.31157

Ruggeri, F. S., Šneideris, T., Vendruscolo, M., & Knowles, T. P. (2019). Atomic force microscopy for single molecule characterisation of protein aggregation [Publisher: Elsevier]. *Archives of biochemistry and biophysics*, *664*, 134–148.

Ryan, O. (2014). Selection of chromosomal DNA libraries using a multiplex CRISPR system. *Elife*, *3*.

Sánchez, K., & Maguire-Zeiss, K. (2020). MMP13 expression is increased following mutant -synuclein exposure and promotes inflammatory responses in microglia. *Frontiers in Neuroscience*, *14*, 585544. https://doi.org/10.3389/fnins.2020.585544

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage x174 DNA. *Nature*, *265*(5596), 687–695. https://doi.org/10.1038/265687a0

Sanger, F., & Thompson, E. O. P. (1953). The amino-acid sequence in the glycyl chain of insulin. 1. the identification of lower peptides from partial hydrolysates. *Biochemical Journal*, *53*(3), 353–366. https://doi.org/10.1042/bj0530353

Santos, J., Pujols, J., Pallarès, I., Iglesias, V., & Ventura, S. (2020). Computational prediction of protein aggregation: Advances in proteomics, conformation-specific algorithms and biotechnological applications. *Computational and Structural Biotechnology Journal*, *18*, 1403–1413. https://doi.org/10.1016/j.csbj.2020.05.026

Santra, M., Dill, K. A., & de Graff, A. M. R. (2019). Proteostasis collapse is a driver of cell aging and death. *Proceedings of the National Academy of Sciences*, *116*(44), 22173–22178. https://doi.org/10.1073/pnas.1906592116

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC research notes*, *9*(1), 1–7.

Schwersensky, M., Rooman, M., & Pucci, F. (2020). Large-scale in silico mutagenesis experiments reveal optimization of genetic code and codon usage for protein mutational robustness. *BMC biology*, *18*(1), 146. https://doi.org/10.1186/s12915-020-00870-9

Serohijos, A. W. R., Rimas, Z., & Shakhnovich, E. I. (2012). Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Reports*, *2*(2), 249–256. https://doi.org/10.1016/j.celrep.2012.06.022

Shabalin, A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, *28*(10), 1353–1358.

Shannon, P. (2022). *igvR: igvR: Integrative genomics viewer*. https://paul-shannon.github.io/igvR/

She, R., & Jarosz, D. (2018). Mapping causal variants with single-nucleotide resolution reveals biochemical drivers of phenotypic change. *Cell*, *172*, 478–490 415.

Singh, P. P., Demmitt, B. A., Nath, R. D., & Brunet, A. (2019). The genetics of aging: A vertebrate perspective. *Cell*, *177*(1), 200–220. https://doi.org/10.1016/j.cell.2019.02.038

Sionkowska, A., Skrzyński, S., Śmiechowski, K., & Kołodziejczak, A. (2017). The review of versatile application of collagen: Versatile application of collagen. *Polymers for Advanced Technologies*, *28*(1), 4–9. https://doi.org/10.1002/pat.3842

Skinnider, M., Scott, N., Prudova, A., Kerr, C., Stoynov, N., Stacey, R., & Foster, L. (2021). An atlas of protein-protein interactions across mouse tissues. *Cell*, *184*(15), 4073–4089.

Smith, E. D., Kennedy, B. K., & Kaeberlein, M. (2007). Genome-wide identification of conserved longevity genes in yeast and worms. *Mechanisms of Ageing and Development*, *128*(1), 106–111. https://doi.org/10.1016/j.mad.2006.11.017

Smith, K., & Rennie, M. J. (2007). New approaches and recent results concerning human-tissue collagen synthesis: *Current Opinion in Clinical Nutrition and Metabolic Care*, *10*(5), 582–590. https://doi.org/10.1097/MCO.0b013e328285d858

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., & Harris, L. (2023). The NHGRI-EBI GWAS catalog: Knowledgebase and deposition resource. *Nucleic Acids Research*, *51*, 977–985.

Soriano-Tárraga, C., Lazcano, U., Jiménez-Conde, J., Ois, A., Cuadrado-Godia, E., Giralt-Steinhauer, E., Rodríguez-Campello, A., Gomez-Gonzalez, A., Avellaneda-Gómez, C.,

Vivanco-Hidalgo, R. M., et al. (2021). Biological age is a novel biomarker to predict stroke recurrence [Publisher: Springer]. *Journal of Neurology*, *268*(1), 285–292.

Soriano-Tárraga, C., Giralt-Steinhauer, E., Mola-Caminal, M., Vivanco-Hidalgo, R. M., Ois, A., Rodríguez-Campello, A., Cuadrado-Godia, E., Sayols-Baixeras, S., Elosua, R., Roquer, J., et al. (2016). Ischemic stroke patients are biologically older than their chronological age [Publisher: Impact Journals, LLC]. *Aging (Albany NY)*, *8*(11), 2655.

Squier, T. C. (2001). Oxidative stress and protein aggregation during biological aging [Publisher: Elsevier]. *Experimental gerontology*, *36*(9), 1539–1550.

Stefani, M., & Dobson, C. M. (2003). Protein aggregation and aggregate toxicity: New insights into protein folding, misfolding diseases and biological evolution. *Journal of Molecular Medicine*, *81*(11), 678–699. https://doi.org/10.1007/s00109-003-0464-5

Stevens, T. J., & Arkin, I. T. (2000). Do more complex organisms have a greater proportion of membrane proteins in their genomes? *Proteins*, *39*(4), 417–420. https://doi.org/10.1002/(sici)1097-0134(20000601)39:4<417::aid-prot140>3.0.co;2-y

Stynen, B., Abd-Rabbo, D., Kowarzyk, J., Miller-Fleming, L., Aulakh, S., Garneau, P., & Michnick, S. (2018). Changes of cell biochemical states are revealed in protein homomeric complex dynamics. *Cell*, *175*(5), 1418–1429.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Sun, N., Youle, R. J., & Finkel, T. (2016). The mitochondrial basis of aging. *Molecular Cell*, *61*(5), 654–666. https://doi.org/10.1016/j.molcel.2016.01.028

Tacutu, R., Thornton, D., Johnson, E., Budovsky, A., Barardo, D., Craig, T., Diana, E., Lehmann, G., Toren, D., Wang, J., Fraifeld, V. E., & de Magalhães, J. P. (2018). Human ageing genomic resources: New and updated databases. *Nucleic Acids Research*, *46*, D1083–D1090. https://doi.org/10.1093/nar/gkx1042

Takasugi, M., Firsanov, D., Tombline, G., Ning, H., Ablaeva, J., Seluanov, A., & Gorbunova, V. (2020). Naked mole-rat very-high-molecular-mass hyaluronan exhibits superior cytoprotective properties. *Nature Communications*, *11*(1), 2376. https://doi.org/10.1038/s41467-020-16050-w

Talens, R. P., Christensen, K., Putter, H., Willemsen, G., Christiansen, L., Kremer, D., Suchiman, H. E. D., Slagboom, P. E., Boomsma, D. I., & Heijmans, B. T. (2012). Epigenetic variation during the adult lifespan: Cross-sectional and longitudinal data on monozygotic twin pairs [Publisher: Wiley Online Library]. *Aging cell*, *11*(4), 694–703.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467–484. https://doi.org/10.1038/s41576-019-0127-1

Taormina, G., Ferrante, F., Vieni, S., Grassi, N., Russo, A., & Mirisola, M. G. (2019). Longevity: Lesson from model organisms. *Genes*, *10*(7), 518. https://doi.org/10.3390/genes10070518

Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Molina, M. M. S., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., & Michnick, S. W. (2008). An in vivo map of the yeast protein interactome. *Science*, *320*(5882), 1465–1470. https://doi.org/10.1126/science.1153878

Taylor, R. C., & Dillin, A. (2011). Aging as an event of proteostasis collapse. *Cold Spring Harbor Perspectives in Biology*, *3*(5), a004440–a004440. https://doi.org/10.1101/cshperspect.a004440

The GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213. https://doi.org/10.1038/nature24277

The GTEx Consortium, Aguet, F., Anand, S., Ardlie, K. G., Gabriel, S., Getz, G. A., Graubert, A., Hadley, K., Handsaker, R. E., Huang, K. H., Kashin, S., Li, X., MacArthur, D. G., Meier, S. R., Nedzel, J. L., Nguyen, D. T., Segrè, A. V., Todres, E., Balliu, B., ... Volpi, S. (2020). The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, *369*(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776

The International Schizophrenia, C. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*, 748.

Thebault, P. (2004, July). *Formalisme CSP (constraint satisfaction problem) et localisation de motifs structurés dans les textes génomiques* (Theses). Université Paul Sabatier - Toulouse III. https://theses.hal.science/tel-00011452

Thiffault, I., Dicaire, M. J., Tetreault, M., Huang, K. N., Demers-Lamarche, J., Bernard, G., Duquette, A., Larivière, R., Gehring, K., Montpetit, A., McPherson, P. S., Richter, A., Montermini, L., Mercier, J., Mitchell, G. A., Dupré, N., Prévost, C., Bouchard, J. P., Mathieu, J., & Brais, B. (2013). Diversity of ARSACS mutations in french-canadians. *The Canadian Journal of Neurological Sciences. Le Journal Canadien Des Sciences Neurologiques*, *40*(1), 61–66. https://doi.org/10.1017/s0317167100012968

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A. K. A., & Hamon, C. (2003). Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, *75*(8), 1895–1904. https://doi.org/10.1021/ac0262560

Tian, X., Azpurua, J., Hine, C., Vaidya, A., Myakishev-Rempel, M., Ablaeva, J., Mao, Z., Nevo, E., Gorbunova, V., & Seluanov, A. (2013). High-molecular-mass hyaluronan

mediates the cancer resistance of the naked mole rat. *Nature*, *499*(7458), 346–349. https://doi.org/10.1038/nature12234

Tian, X., Seluanov, A., & Gorbunova, V. (2017). Molecular mechanisms determining lifespan in short- and long-lived species. *Trends in Endocrinology & Metabolism*, *28*(10), 722–734. https://doi.org/10.1016/j.tem.2017.07.004

Tiessen, A., Pérez-Rodríguez, P., & Delaye-Arredondo, L. J. (2012). Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC research notes*, *5*, 85. https://doi.org/10.1186/1756-0500-5-85

Törner, R., Kupreichyk, T., Hoyer, W., & Boisbouvier, J. (2022). The role of heat shock proteins in preventing amyloid toxicity. *Frontiers in Molecular Biosciences*, *9*, 1045616. https://doi.org/10.3389/fmolb.2022.1045616

Toto, A., Malagrinò, F., Visconti, L., Troilo, F., Pagano, L., Brunori, M., Jemth, P., & Gianni, S. (2020). Templated folding of intrinsically disordered proteins. *Journal of Biological Chemistry*, *295*(19), 6586–6593. https://doi.org/10.1074/jbc.REV120.012413

Tower, J. (2011). Heat shock proteins and drosophila aging. *Experimental Gerontology*, *46*(5), 355–362. https://doi.org/10.1016/j.exger.2010.09.002

Truong, K., & Ikura, M. (2001). The use of FRET imaging microscopy to detect protein-protein interactions and protein conformational changes in vivo. *Current Opinion in Structural Biology*, *11*(5), 573–578. https://doi.org/10.1016/s0959-440x(00)00249-9

Tsai, H., Krol, A., Sarti, K., & Bennett, J. (2006). Candida glabrata PDR1, a transcriptional regulator of a pleiotropic drug resistance network, mediates azole resistance in clinical isolates and petite mutants. *Antimicrobial agents and chemotherapy*, *50*(4), 1384–1392.

Tsolis, A. C., Papandreou, N. C., Iconomidou, V. A., & Hamodrakas, S. J. (2013). A consensus method for the prediction of 'aggregation- prone' peptides in globular proteins. *PLOS ONE*, *8*(1), 6.

Tuerk, C., & Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage t4 DNA polymerase. *Science (New York, N.Y.)*, *249*(4968), 505–510. https://doi.org/10.1126/science.2200121

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies [Number: 1 Publisher: Nature Publishing Group]. *Nature Reviews Methods Primers*, *1*(1), 1–21. https://doi.org/10.1038/s43586-021-00056-9

Venkataram, S. (2016). Development of a comprehensive genotype-to-fitness map of adaptation-driving mutations in yeast. *Cell*, *166*, 1585–1596 1522.

Victor, M. P., Acharya, D., Chakraborty, S., & Ghosh, T. C. (2020). Chaperone client proteins evolve slower than non-client proteins. *Functional & Integrative Genomics*, *20*(5), 621–631. https://doi.org/10.1007/s10142-020-00740-1

Vijg, J. (2014). Somatic mutations, genome mosaicism, cancer and aging. *Current Opinion in Genetics & Development*, *26*, 141–149. https://doi.org/10.1016/j.gde.2014.04.002

Visscher, P. M., Hill, W. G., & Wray, N. R. (2008). Heritability in the genomics era–concepts and misconceptions. *Nature Reviews. Genetics*, *9*(4), 255–266. https://doi.org/10.1038/nrg2322

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

von Mering, C. (2004). STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, *33*, D433–D437. https://doi.org/10.1093/nar/gki005

Voorman, A., Lumley, T., McKnight, B., & Rice, K. (2011). Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PloS One*, *6*(5), e19416. https://doi.org/10.1371/journal.pone.0019416

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., Brugge, H., Oelen, R., de Vries, D. H., van der Wijst, M. G. P., Kasela, S., Pervjakova, N., Alves, I., Favé, M.-J., Agbessi, M., … Franke, L. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics*, *53*(9), 1300–1310. https://doi.org/10.1038/s41588-021-00913-z

Wainberg, M. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nat Genet*, *51*, 592–599.

Wainschtein, P., Jain, D., Zheng, Z., TOPMed Anthropometry Working Group, Aslibekyan, S., Becker, D., Bi, W., Brody, J., Carlson, J. C., Correa, A., Du, M. M., Fernandez-Rhodes, L., Ferrier, K. R., Graff, M., Guo, X., He, J., Heard-Costa, N. L., Highland, H. M., Hirschhorn, J. N., … Visscher, P. M. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics*, *54*(3), 263–273. https://doi.org/10.1038/s41588-021-00997-7

Waldron, I., & Johnston, S. (1976). Why do women live longer than men? *Journal of Human Stress*, *2*(2), 19–30. https://doi.org/10.1080/0097840X.1976.9936063

Walker, G. A., & Lithgow, G. J. (2003). Lifespan extension in c. elegans by a molecular chaperone dependent upon insulin-like signals [Publisher: Wiley Online Library]. *Aging cell*, *2*(2), 131–139.

Walsh, I., Seno, F., Tosatto, S. C. E., & Trovato, A. (2014). PASTA 2.0: An improved server for protein aggregation prediction. *Nucleic Acids Research*, *42*, W301–307. https://doi.org/10.1093/nar/gku399

Wang, C., Lue, W., Kaalia, R., Kumar, P., & Rajapakse, J. C. (2022). Network-based integration of multi-omics data for clinical outcome prediction in neuroblastoma. *Scientific Reports*, *12*(1), 15425. https://doi.org/10.1038/s41598-022-19019-5

Washburn, M. P., Wolters, D., & Yates, J. R. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, *19*(3), 242–247. https://doi.org/10.1038/85686

Wery, M., Descrimes, M., Vogt, N., Dallongeville, A., Gautheret, D., & Morillon, A. (2016). Nonsense-mediated decay restricts LncRNA levels in yeast unless blocked by double-stranded RNA structure. *Molecular cell*, *61*(3), 379–392.

Wik, L., Nordberg, N., Broberg, J., Björkesten, J., Assarsson, E., Henriksson, S., Grundberg, I., Pettersson, E., Westerberg, C., Liljeroth, E., Falck, A., & Lundberg, M. (2021). Proximity extension assay in combination with next-generation sequencing for high-throughput proteome-wide analysis. *Molecular & cellular proteomics: MCP*, *20*, 100168. https://doi.org/10.1016/j.mcpro.2021.100168

Willcox, B. J., Willcox, D. C., & Suzuki, M. (2017). Demographic, phenotypic, and genetic characteristics of centenarians in okinawa and japan: Part 1—centenarians in okinawa. *Mechanisms of Ageing and Development*, *165*, 75–79. https://doi.org/10.1016/j.mad.2016.11.001

Williams, G. C. (2001). Pleiotropy, natural selection, and the evolution of senescence: Evolution 11, 398-411 (1957). [Publisher: American Association for the Advancement of Science]. *Science of Aging Knowledge Environment*, *2001*(1), cp13–cp13.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, *15*(4), 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

Xia, X. (2013). Codominance. In *Brenner's encyclopedia of genetics (second edition)* (Second Edition, pp. 63–64). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-374984-0.00278-3

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., & Steinmetz, L. (2009). Bidirectional promoters generate pervasive transcription in yeast. *Nature*, *457*(7232), 1033–1037.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. https://doi.org/10.1038/ng.608

Yao, D. W., O'Connor, L. J., Price, A. L., & Gusev, A. (2020). Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nature Genetics*, *52*(6), 626–633. https://doi.org/10.1038/s41588-020-0625-2

Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliasen, A. U., Jiang, Y., Raghavan, S., Miao, J., Arias, J. D., Graham, S. E., Mukamel, R. E., Spracklen, C. N., Yin, X., Chen, S.-H., Ferreira, T., Highland, H. H., ... Hirschhorn, J. N. (2022). A saturated map of common genetic variants associated with human height [Number: 7933 Publisher: Nature Publishing Group]. *Nature*, *610*(7933), 704–712. https://doi.org/10.1038/s41586-022-05275-y

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., & Liu, X. (2021). rMVP: A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics, Proteomics & Bioinformatics*, *19*(4), 619–628. https://doi.org/10.1016/j.gpb.2020.10.007

Ytournel, F. (2008, January). *Déséquilibre de liaison et cartographie de QTL en population sélectionnée* (Theses) [Issue: 2008AGPT0004]. AgroParisTech. https://pastel.archives-ouvertes.fr/pastel-00003789

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate illumina paired-end reAd mergeR. *Bioinformatics*, *30*(5), 614–620.

Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C., & Yates, J. R. (2013). Protein analysis by shotgun/bottom-up proteomics. *Chemical Reviews*, *113*(4), 2343–2394. https://doi.org/10.1021/cr3003533

Zhao, L., Zhao, J., Zhong, K., Tong, A., & Jia, D. (2022). Targeted protein degradation: Mechanisms, strategies and application. *Signal Transduction and Targeted Therapy*, *7*(1), 113. https://doi.org/10.1038/s41392-022-00966-4

Zhao, L., Liu, Z., Levy, S. F., & Wu, S. (2018). Bartender: A fast and accurate clustering algorithm to count barcode reads (B. Berger, Ed.). *Bioinformatics*, *34*(5), 739–747. https://doi.org/10.1093/bioinformatics/btx655

Zhu, B.-L., Long, Y., Luo, W., Yan, Z., Lai, Y.-J., Zhao, L.-G., Zhou, W.-H., Wang, Y.-J., Shen, L.-L., Liu, L., Deng, X.-J., Wang, X.-F., Sun, F., & Chen, G.-J. (2019). MMP13 inhibition rescues cognitive decline in alzheimer transgenic mice via BACE1 regulation. *Brain*, *142*(1), 176–192. https://doi.org/10.1093/brain/awy305

Zügel, U., & Kaufmann, S. H. E. (1999). Role of heat shock proteins in protection from and pathogenesis of infectious diseases. *Clinical Microbiology Reviews*, *12*(1), 19–39. https://doi.org/10.1128/CMR.12.1.19