Développement de méthodes et outils d'analyse transcriptomique par réseaux de co-expression de gènes pour la détection de gènes candidats dans le vieillissement de différents tissus humains

Thèse

Gwenaëlle Lemoine

Doctorat en médecine moléculaire
Philosophiæ doctor (Ph. D.)

Québec, Canada

Développement de méthodes et outils d'analyse transcriptomique par réseaux de co-expression de gènes pour la détection de gènes candidats dans le vieillissement de différents tissus humains

Thèse

Gwenaëlle Lemoine

Sous la direction de:

Arnaud Droit, directeur de recherche

Résumé

L'analyse par réseau de co-expression de gènes est un outil entré il y a 15 ans dans l'ensemble des outils disponibles pour l'analyse transcriptomique. En étudiant la variation de synchronisation de l'expression des gènes, cet outil permet de révéler de nouveaux gènes impliqués dans des maladies ou phénotypes dont l'expression seule n'est pas significativement différente. Il est également capable de détecter des groupes de gènes, ou modules, interagissant préférentiellement et sur lesquels il est possible d'effectuer une exploration étendue. Il est ainsi possible d'utiliser des méthodes avec injection de connaissance préalable comme l'enrichissement de gènes ou l'association phénotypique, ou des méthodes guidées par les données comme l'analyse topologique ou la co-expression différentielle. Pourtant, ce type d'analyse reste sous exploitée actuellement par rapport à son potentiel, et notamment dans certaines maladies ou phénotypes où l'altération est une désorganisation du système comme le vieillissement.

Afin de faciliter à tout chercheur l'emploi de cette méthode, un progiciel R disponible sur Bioconductor et nommé GWENA a été développé. Organisé comme un pipeline d'analyse simplifié et allant de la construction du réseau jusqu'à l'aide à l'interprétation des modules entre différentes conditions, c'est également le seul pipeline actuel à intégrer la co-expression différentielle. Pour assister l'utilisateur, il comprend de nombreux avertissements sur l'intégrité des données rentrées et sur la plausibilité des résultats. Afin d'éviter de devoir recourir à d'autres logiciels, il contient également un système de visualisation des réseaux. Enfin, GWENA est un outil dont l'architecture modulaire lui permettra d'évoluer avec le temps.

L'efficacité de GWENA a été démontrée dans une première étude du vieillissement du muscle squelettique humain où un sous ensemble de gènes a été priorisé pour l'étude de la sarcopénie. Il a également permis de préciser une topologie du réseau spécifique du vieillissement et observée auparavant : la perte de connectivité du réseau, ou déconnexion. En effet, parallèlement à la déconnexion, il a été constaté grâce à GWENA une reconnexion locale située au niveau des gènes pivots. Pour étudier cette topologie à large échelle, l'analyse a été répétée sur un ensemble élargi de tissus humains. Par un recoupement des modules différentiellement exprimés, des phénomènes communs du vieillissement entre tissus sont apparus ainsi que des phénomènes spécifiques à certains tissus. L'analyse topologique, notamment de la déconnexion, des

gènes inclus dans ces recoupements pour deux exemples, un phénomène commun et un phénomène spécifique, a à son tour permis la priorisation de gènes encore mal étudiés ou inconnus dans ces phénomènes.

En finalité, les travaux présentés au cours de cette thèse auront amené à la création d'un outil utile à la communauté de biologistes comme bio-informaticiens pour faciliter l'accès à une analyse à a haut potentiel dans l'analyse du vieillissement et toute autre condition, notamment celles axées sur la dérégulation de l'expression systémique.

Mots-clefs : co-expression, réseau, vieillissement, transcriptomique, progiciel R, Bioconductor, co-expression différentielle.

Abstract

Gene co-expression network analysis is a tool that entered the transcriptomics analysis toolbox 15 years ago. By studying the variation in the synchronization of gene expression, this tool can reveal new genes involved in diseases or phenotypes whose expression alone is not significantly different. It is also able to detect groups of genes, or modules, that interact preferentially and on which it is possible to carry out an extended exploration. It is therefore possible to use knowledge-driven methods such as gene enrichment or phenotypic association, or data-driven methods such as topological analysis or differential co-expression. Nevertheless, this type of analysis is currently under-exploited compared to its potential, especially in certain diseases or phenotypes where the alteration is a disorganization of the system such as aging.

In order to facilitate the use of this method by any researcher, an R software package available on Bioconductor and named GWENA has been developed. Organized as a simplified analysis pipeline from the construction of the network to the interpretation of the modules between different conditions, it is also the only current pipeline to integrate the differential co-expression. To assist the user, it includes numerous warnings about the integrity of the data entered and the plausibility of the results. In order to avoid having to use other software, it also contains a network visualization system. Finally, GWENA is a tool whose modular architecture allows it to evolve over time.

The effectiveness of GWENA has been demonstrated in a first study of human skeletal muscle aging, where a subset of genes was prioritized for the study of sarcopenia. It also allowed to clarify a network topology specific to aging and previously observed: the loss of network connectivity, or disconnection. Indeed, in parallel to the disconnection, a local reconnection located at the level of hub genes was observed thanks to GWENA. To study this topology on a large scale, the analysis was repeated on an extended set of human tissues. By cross-referencing differentially expressed modules, common aging phenomena between tissues were identified as well as tissue-specific phenomena. Topological analysis, including disconnection, of the genes included in these overlaps for two examples, a common and a specific phenomenon, in turn allowed the prioritization of genes still poorly studied or unknown in these phenomena.

Overall, the work presented in this thesis will have led to the creation of a useful tool for the

community of biologists as bioinformaticians to facilitate access to a high-potential analysis in the analysis of aging and any other condition, especially those focused on the deregulation of systemic expression.

Keywords: co-expression, network, aging, transcriptomics, R package, Bioconductor, differential co-expression.

Table des matières

Liste des tableaux

Liste des figures

Acronymes

ADN Acide DesoxyriboNucléique.

ADNc ADN complémentaire.

ARN Acide RiboNucléique.

ARNlnc ARN longs non codants (*lncRNA* en anglais).

ARNm ARN messager.

ARNmi Micro ARN (*miRNA* en anglais).

ARNpi ARN interferants de petite taille (*siRNA* en anglais).

ARNt ARN de transfert.

CORUM Comprehensive Resource of Mammalian protein complexes.

CP Composante principale.

CPM Count per million.

DCE Co-expression différentielle (*differential co-expression* en anglais).

DNA DesoxyriboNucleic Acid.

EST Expressed sequence tag.

FPKM Fragments par kilo base de transcript par million de fragments alignés.

GCN Gene Co-expression Network.

GEO Gene Expression Omnibus.

GMT (*Gene Matrix Transposed* en anglais.

GO Gene Ontology.

GO :BP GO Biological Process.

GO :CC GO Cellular Compartment.

GO :MF GO Molecular Function.

GWENA Gene Whole co-Expression Network Analysis.

HP Human Phenotype Ontology.

HPA Human Protein Atlas.

IFN Interféron.

IFN-I IFN de type I.

KEGG Kyoto Encyclopedia of Genes and Genomes.

LCQ Locus de Caractère Quantitatif.

MEC Matrice extra-cellulaire.

MIRNA mirTarBase.

MP Modérément préservé.

mRNA messenger RNA.

MSE Marqueurs de Séquence Exprimée (EST en anglais).

NC Non conclusive.

NP Non préservé.

PC Principal Component.

PCA Principal Component Analysis.

QTL Quantitative Trait Loci.

REAC Reactome.

RLE Relative log expression.

RNA RiboNucleic Acid.

RNA-seq Technique de séquençage ARN.

RPKM Lectures (*Reads* en anglais) par kilo base de transcript par million de fragments alignés.

RT-qPCR Reverse transcription quantitative polymerase chain reaction.

SAGE Serial Analysis of Gene Expression.

TF TRANSFAC.

TMM Trimmed Mean of M-values.

TPM Transcripts par million.

WP WikiPathway.

La vie est absurde, après si t'aime ce genre d'humour c'est plus facile à vivre.

Isabelle Stévant, chercheuse en bio-informatique et collègue du blog bioinfo-fr.net

īn nīz bogzarad - Cela aussi passera

Farīd ud-Dīn, l'Attār de Nishapur, poète perse racontant la fable d'un roi puissant qui demanda à ses sages de lui inventer une phrase, à graver dans un anneau. Elle devrait être vraie, appropriée en tout temps et en toute situation. Des lunes de refléxion plus tard, les sages présentairent alors ces mots au roi : "Cela aussi passera."

Remerciements

La recherche n'est pas un processus en ligne droite, ce n'est pas un processus où celui qui est en avance va systématiquement trouver la nouvelle réponse à une grande question. C'est une avancée collective ou la moindre découverte n'est en aucun cas l'œuvre d'une seule personne. Il aura au préalable fallut accumuler la connaissance sur laquelle repose ladite découverte, il aura fallu le soutien des proches qu'on oublie trop souvent pour leur contribution certes indirecte mais essentielle. La recherche c'est une avancée à tâtons qui devrait donner sa chance à chaque branche de l'arbre qu'est la recherche car il est bien impossible de prédire exactement quelle sera la prochaine à donner un fruit. Qui plus est un fruit intéressant vis-à-vis de notre utilitarisme contemporain.

Pour pouvoir profiter des rares moments de joie suite à un résultat, une découverte, qui nous font aimer ce métier de chercheur, il faut donc dans la recherche faire preuve d'abnégation ou de résilience. Je ne saurais dire laquelle de ces deux notions de persévérance évoquées sied le plus à la description de ces 4 ans de doctorat. Ce que je sais, c'est qu'il n'aurait pas été possible de les tenir sans la contribution de certains et sans le soutien inconditionnels d'autres. Cette page a donc pour but de remercier tous ces gens dans un ordre indépendamment de leur degré de contribution à cette thèse.

Je remercie donc mon directeur de thèse Arnaud Droit pour m'avoir accueillie dans son laboratoire.

Je remercie les collaborateurs de L'Oréal dans le cadre de la Chaire de Recherche en Biologie numérique pour m'avoir donné l'opportunité d'y contribuer.

Merci aux membres du jury, la Dre. Francine Durocher, le Dr. Simon Hardy, le Dr. Yohan Bossé, la Dre. Sarah Gagliano Taliun, pour avoir accepté de siéger et évaluer mes travaux.

Merci à Marie Pier Scott-Boyer, superviseure de mes débuts de doctorat puis collègue source de nombreuses discussions scientifiques. À nos désaccords de Recherche sincèrement bénéfiques pour ma formation, même si ce n'était pas de la façon que je l'aurai pensée.

Merci à Julien Prunier, professionnel de recherche qui a su jouer un avocat du diable parfois

plus vrai que nature à grands renforts de mauvaise foi assumée pour me forcer à apporter de la nuance dans mon entêtement constant.

Merci à Charles Joly-Beauparlant, professionnel de recherche et véritable rocher immuable dans la tempête des projets s'accumulant et toujours ravi de donner un coup de main. Tu resteras le grand R master à mes yeux et ce fut un plaisir d'échanger cuisine avec toi.

Merci à Éric Fournier, professionnel de recherche à qui je dois le nom de mon package Bioconductor GWENA suite à une blague en réunion de labo. On va me penser narcissique à tort, mais comme tu l'as dit "Ça fonctionne trop bien et sera peut-être qu'une fois dans ta vie".

Merci aux autres membres de l'ADLab qui ont contribué à cette expérience qu'est le doctorat d'une façon ou d'une autre. Nos parties de babyfoot me manqueront. Et merci à ces anciens du labo qui ont en leur temps contribué à cette thèse. Merci donc à Maxime Vallée, professionnel de recherche pour son honnêté dans nos échanges sur la vie professionnelle et québécoise. Merci à notre ancien homme de l'ombre, véritable Batman de l'administration système, Adrien Dessmond.

Merci à Alexandra Elbakyan sans qui nombreuses seraient les publications qui me seraient restées hors de portée.

Des mercis en masse à toutes ces rencontres en terre canadienne qu'une thèse en un autre lieu n'auraient su me donner. : Marine, Camille, Charlie, Clément, Thomas, Xavier, Antoine. Votre bonne humeur et nos sorties découvertes de la culture/nature allez me manquer.

À cette communauté qu'est celle de la vulgarisation scientifique et notamment celle du Vortex, un merci sincère. Partager la science est un plaisir tant pour vous que pour moi alors on va continuer comme ça. Un merci supplémentaire à Mel, Mallou, Piloy, Kimist, Penta, Flax, Clem, Pha, Tofu, Pollo, Viper, Ilwa, Jerry, Bubble, Egz, Hibou, Apo, Oldu et toutes ces autres personnes que je ne pourrai pas citer sans risquer de faire des remerciements plus longs que ma thèse elle-même.

Pour cette confiance que je ne m'accorde jamais mais qu'ils ont su me donner, merci à mes coéquipiers administrateurs de Bioinfo-fr.net Isabelle et Yoann de m'avoir recrutée et donné l'occasion de faire mes preuves en publication et gestion d'édition. Un merci un peu long également aux membres de son canal IRC car ils sont nombreux à avoir enrichi scientifiquement et personnellement mes réflexions durant ce long doctorat. Un merci en ordre alphabétique, pour qu'ils ne trouvent pas une n-ième raison de troller, à Aestra, aurelbzh, azerin, Billbis, bjonnh, Chopopope, eorn, lhtd, Lins‘, maxulysse, MoUsSoR, Natir, Nedgang, neolem, noctisLab, Norore, schneu, slybzh, YaknotiS, ZaZo0o.

Un merci tout particulier à ma famille pour son soutien de toujours et nos skypes dominicaux. À Isabelle et Florent, mes parents, car on a beau être adulte, partir au bout du monde, crier haut et fort son indépendance, on reste toujours l'enfant de ses parents. Je suis fière d'être votre fille.

À Alexandre et Elzia, parce qu'à 3 enfants terribles on a toujours su se soutenir à notre façon. À mes grands-parents, Georges et Geneviève, véritables forces de la nature et source d'un profond respect.

Vous l'avez déjà entendu mille fois mais une fois de plus n'est pas de trop : merci à Romain, Jacques, Gabriel, Lauriane, Camille, Margaux, Maxime, compagnons de jeux vidéos, de société, ou de rôle qui malgré les milliers de kilomètres nous séparant ont répondu présent à l'appel et ont permis de parler librement, de décompresser. À nos rires, nos sessions chant et notre sel qui ne saurait égaler celui de nos adversaires face à nos âneries.

Pour avoir répondu à l'appel nuit et jour malgré le décalage horaire, avoir été un soutien indéfectible en croyant bien plus à ma réussite que moi-même, avoir ramené de la raison où j'en manquais, avoir épanché mes douleurs de maladies chroniques physique et mentale que le doctorat n'aura clairement pas amélioré, avoir contribué au soin de cette dernière, avoir menacé tour à tour de prendre l'avion pour venir me botter le cul, un millier de mercis ne suffirait pas à Léopold et Amandine. Un câlin supplémentaire à Léopold qui pourra presque prétendre à l'HDR avec tous ces points scientifiques qu'on aura faits, et une promesse de session cuisine et canapé à ma toupine.

Un merci plus que particulier à Raijich (Régis) et Audrey, camarades doctorants devenus des amis chers à mon cœur. Je n'oublierai jamais à quel point je leur dois d'avoir tenu jusqu'au bout. Vous avez été là à chaque fois que j'en avais besoin consciemment ou non et on a pu progresser ensemble en se serrant les coudes. Je souhaite de tout mon cœur qu'on poursuive nos sessions détox régulières même après ce doctorat. Merci Audrey pour les rires, le mentorat, les sorties avec des diamants plein les yeux ou couronnée d'un poster trophée, la conversion au sous-estimé legging. Merci à Régis pour ce partage si subtile que sont l'humour absurde et la culture internet, pour ces papotages informatique au sens large ou cuisine, pour les grandes discussions en pleine crise existentielle tard (beaucoup trop tard) dans le labo, pour la reprise de l'escalade en bloc au grand dam de mes mains à présent couvertes de corne. À vous deux, à nos lavages de tasse à thé !

Enfin, à cette personne connue il y a bientôt 6 ans mais que j'aurai attendu 4 ans de plus avant de me rendre compte qu'il était la moitié que j'attendais pour me sentir complète, merci Clément. Ton soutien sans réserve pour m'aider à tenir jusqu'à la fin, ta tolérance face à mes râleries et excès, ton réconfort lors de mes doutes les plus viscéraux, font de toi l'homme que j'ai envie de chérir du plus profond de mon être. Je saurais te rendre la pareille pour ta (fausse) thèse ;D. À notre aventure à deux que j'espère la plus longue possible, nous Sain(t) Clément et le Graoully.

Avant-propos

Projets principaux

Cette thèse est réalisée avec l'insertion d'articles écrits durant mon doctorat. Elle présente l'état de mes travaux dont le but principal était le développement d'outils et méthodes pour la détection de gènes candidats au vieillissement humain par l'utilisation de réseaux de co-expression de gènes. Chaque chapitre est donc constitué d'un article publié ou visant à l'être.

Les articles insérés sont les suivants :

— *GWENA : gene co-expression networks analysis and extended modules characterization in a single Bioconductor package*, publié dans la revue *BMC Bioinformatics* le 25 mai 2021.

— *Analyse trans-tissus par réseau de co-expression de gènes pour la détection de fonctions physiologiques communes et spécifiques au vieillissement*, article en préparation.

Auteurs impliqués :

— Gwenaëlle G. Lemoine, Département de médecine moléculaire, Faculté de médecine, Université Laval, 2325 rue de l'Université, Québec G1V 0A6, Canada.

— Marie Pier Scott-Boyer, Centre de recherche du CHU de Québec-Université Laval, 2705 boulevard Laurier Québec, Québec G1V 4G2, Canada.

— Bathilde Ambroise, L'Oréal Research and Innovation, 15 rue Pierre Dreyfus, 92110 Clichy, France.

— Olivier Périn, L'Oréal Research and Innovation, 15 rue Pierre Dreyfus, 92110 Clichy, France.

— Arnaud Droit, Département de médecine moléculaire, Faculté de médecine, Université Laval, 2325 rue de l'Université, Québec G1V 0A6, Canada.

Contribution à l'article "GWENA"

Je suis responsable de la conception, développement et maintenance de l'outil GWENA ainsi que de l'écriture de l'article. Concernant la réalisation de l'analyse pour le cas d'utilisation sur le

muscle, je suis responsable du traitement des données ainsi que de leur analyse. Le choix de la méthodologie fut un travail conjoint de Marie Pier Scott-Boyer qui a également supervisé le projet. Elle a également avec Olivier Périn, Bathilde Ambroise et Arnaud Droit participé à la relecture de l'article. L'intégralité de l'article a été validé par tous les auteurs. Arnaud Droit s'est également chargé de la recherche de financement.

Modification à l'article "GWENA"

Des erreurs dans le texte ne modifiant pas les conclusions de l'article ont été trouvé à la relecture par le jury. Celles-ci ont été corrigées et sont listée ci-dessous :

- Chapitre 1, Figure 2.1 2.1, points ① et ② : les colonnes et lignes du tableau ont été transposés pour correspondre au texte "a table with genes as columns and samples as rows"
- Chapitre 1, Section *Single condition modules analysis* : une précision disparue lors des relectures avant publication et concernant la sélection des modules a été remise : "Modules 19, 21 and 25 were the top 3 enriched for terms related to muscle function" changé pour "Modules 19, 21 and 25 were the top 3 enriched for terms related to muscle function also associated with a phenotype impacting muscle"
- Chapitre 1, Section *Single condition modules analysis*, Figure 2.2, point C de la légende : correction du module 10 en module 19
- Chapitre 1, Section *Single condition modules analysis*, Figure 2.2 : inversion des textes des points D et E pour correspondre au bon panel de la figure
- Chapitre 1, Section *Background*, une référence oubliée à la base de donnée GTEx a été rajoutée

Contribution à l'article d'analyse trans-tissus du vieillissement via GWENA

Je suis responsable de la conception du projet, du traitement et de l'analyse des données, de l'interprétation des résultats, et de la rédaction de l'article. Marie Pier Scott-Boyer a assisté dans la consolidation de la méthodologie et sa validation. Arnaud Droit s'est chargé de la recherche de financement.

Projets annexes

Durant mon doctorat, j'ai également pu m'investir dans différents projets scientifiques :

— *Weighted gene co-expression network analysis identifies inflammaging biomarkers in aged skin of humans in vivo*. Projet de ré-analyse des données de transcriptomique de Kuehne et al. 2017 par le biais de GWENA. Il a permis de mettre en évidence des gènes impliqués dans le phénomène d'inflammation chronique de faible intensité dans des biopsies d'épiderme. Par respect envers la clause de confidentialité de la chaire de recherche et d'innovation L'Oréal en biologie numérique, ces travaux n'ont pas été soumis à publication.

— Étude à travers de multiples points temporels sur 28 jours de la reconstruction épidermique basée sur un modèle d'épiderme *in vitro* provenant de biopsies de circoncisions. Travaux effectués pour la chaire de recherche et d'innovation L'Oréal en biologie numérique et confidentiels.

D'autre travaux scientifiques non académiques ont également été réalisés et sont visibles en Annexe D.

Financements

Les travaux présentés dans cette thèse ont étés soutenus par la Chaire de recherche et d'innovation L'Oréal en biologie numérique.

Notes

— L'intégralité des figures a été réalisée par mes soins et est sous licence CC-BY-NC sauf mention contraire ou citation d'une figure d'une publication.

— L'article situé en Chapitre 1 est publié dans BMC Bioinformatics sous la licence CC-BY.

— Pour plus d'information sur les licences Creative Commons : https://creativecommons.org/about/cclicenses/.

— Toute figure ou table reprise d'un article est cité et traduite lorsque les licences du journal le permettent.

— Pour des raisons de cohérence avec l'article rédigé en anglais, de cohérence avec la littérature scientifique rédigée majoritairement en anglais, les acronymes de cette thèse utilisés dans la recherche et rarement trouvés dans le langage commun seront en anglais après avoir été explicité en français lors de leur première utilisation.

Introduction

La plasticité du transcriptome face aux nombreuses perturbations qu'il subit permet à un organisme de préserver ses fonctions vitales en tous temps dans une certaine mesure. Par un appel à de nombreux mécanismes de régulation de l'expression des gènes, les cellules vont alors contrer, compenser ou limiter l'impact de la perturbation. Si les mécanismes majeurs de la régulation ont pu être identifiés dès les premières quantifications de l'expression, le récent développement des technologies de transcriptomique a permis d'accélérer cette tendance. La recherche en médecine moléculaire s'est donc emparée de cet outil dans une démarche d'identification systématique de la fonction biologique de gènes encore mal connus ou de la cause de leur implication dans une maladie, un phénotype ou une condition.

La quantité de données faisant, le besoin de méthodes d'analyse biostatistique capables de gérer autant d'information pour en ressortir des biomarqueurs, caractéristiques spécifiques et mesurable d'une condition, s'est accru. C'est dans ce contexte qu'on a vu l'essor des méthodes d'analyse par réseau en biologie [1] et plus particulièrement celles dites de réseau de co-expression de gènes qui reposent sur la synchronisation de l'expression entre gènes. Par leur capacité à saisir l'ensemble des variations les plus subtiles dans le transcriptome et les variations conjointes entre gènes, les réseaux de co-expression de gènes ont su montrer leur intérêt dans la recherche de biomarqueurs de maladies complexes et encore mal comprises telles qu'Alzheimer [2] ou le cancer du sein [3]. Pourtant, leur emploi reste encore limité aux équipes possédant l'expertise d'un·e bio-informaticien·ne, retardant leur potentiel pour l'étude de nombreux phénomènes. Parmi eux, le vieillissement y gagnerait pourtant à utiliser des approches par réseau. De nature complexe, le vieillissement est un phénomène qu'on résume encore trop souvent à sa seule composante de sénescence. Il est pourtant le résultat de nombreux mécanismes entrecroisés que la co-expression est à même de décomposer grâce à son étude à l'échelle d'un système et pas seulement de ses acteurs individuels, les gènes.

Les travaux réalisés durant ce doctorat ont donc été divisé en deux parties avec un but chacune. Premièrement, tester la pertinence du développement d'un outil d'architecture pipeline modulaire, sous forme de paquet R, contenant une aide à l'interprétation et visant à pérenniser l'analyse de co-expression de gènes au vu de l'évolution des méthodes pour la réaliser. Deuxièmement,

évaluer le potentiel de découverte de l'analyse de co-expression de gènes à travers différents tissus et tranches d'âge pour une meilleure compréhension du vieillissement humain en priorisant des gènes.

Cette introduction vise à donner les clefs essentielles à la compréhension des différents concepts biostatistiques, bio-informatiques et médicaux employés tout au long de cette thèse. Elle est suivie par deux chapitres dont le premier est la présentation de l'outil GWENA dédié à l'analyse de réseaux de co-expression de gènes et de son application sur l'étude du vieillissement du muscle squelettique. Le second chapitre est une étude à plus large spectre du vieillissement analysant les fonctions communes et spécifiques du vieillissement à travers 10 tissus humains. S'ensuivent une conclusion résumant les apports de mes travaux, puis une discussion les replaçant dans le contexte des connaissances actuelles et les perspectives futures dans ce domaine.

1.1 La complexité du vivant

1.1.1 Un fonctionnement multiple issu d'un code unique

Le vivant est fait d'une diversité d'organismes à l'image de sa complexité. Chacun de ces organismes est composé d'une ou plusieurs **cellules** œuvrant ensemble pour la survie et la reproduction de celui-ci. Ces êtres multicellulaires sont bien souvent constitués d'une association de cellules à la fonction et la forme bien différentes de ses voisines. Ainsi, chez des organismes constitués d'un très faible nombre de cellules comme les Myxozoa constitués au maximum de 7 cellules [4] tout comme chez des organismes de taille bien plus importante comme l'humain fait d'entre 10^{12} et 10^{16} cellules [5]), on retrouve des cellules spécialisées pour une tâche spécifique [6] (Figure 1.1). Celles-ci servent différents objectifs tels que la protection, l'acheminement d'éléments nutritifs, ou encore la transmission de matériel génétique. Chez l'humain, ces cellules seront par exemple respectivement des kératinocytes [7], des cellules endothéliales de l'intima [8], des ovocytes [9]. En s'assemblant entre cellules de même type, puis en formant des complexes avec d'autres types cellulaires, les organismes parviennent à former alors des structures

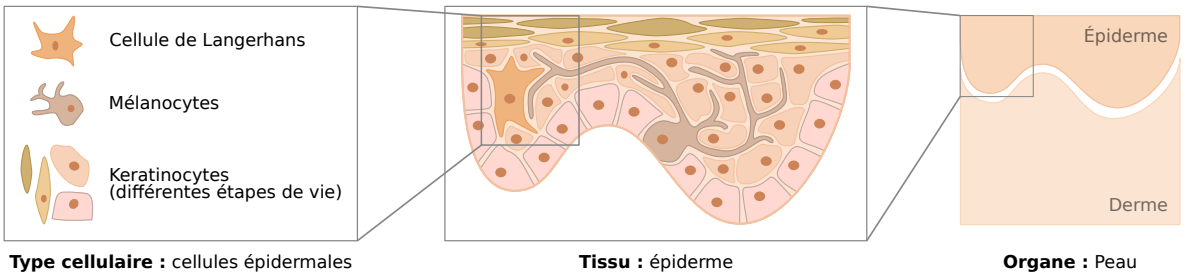

FIGURE 1.1 – Différentes échelles de visualisation de l'organisation cellulaire avec l'exemple de la peau. Les cellules majoritairement présentes sont les kératinocytes responsable de la fonction barrière, les mélanocytes responsables de la protection solaire notamment, et les cellules de Langerhans responsables d'une fonction immunitaire primaire

nommées **tissu** [10] qui elles-mêmes assemblées vont former un organe. La combinaison de kératinocytes, de mélanocytes, de cellules de Langerhans et les cellules de Merkel connectées par une matrice extra-cellulaire (MEC) forment ainsi le tissu épidermal, ou plus simplement l'épiderme [11].

Pour effectuer chacune de ces tâches spécialisées, les différents types de cellules vont produire en partie des **protéines** ayant des fonctions biologiques différentes de leurs voisines. Bien qu'il existe une très large diversité de fonctions, 32 catégories de fonctions selon le projet Gene Ontology (GO) [12], on peut toutefois les regrouper en 3 familles de fonctions majeures [13] :

— **les fonctions catalytiques** : les protéines, dans ce cas appelées enzymes, présentent une activité d'augmentation du taux de réaction chimique en diminuant l'énergie d'activation nécessaire. Ces enzymes sont spécifiques d'une transformation d'un substrat vers un produit. Elles peuvent cependant voir leur fonction altérée par des activateurs (co-facteurs, co-enzymes) et des inhibiteurs (co-répresseurs). Par exemple, l'alcool déshydrogénase va, avec la liaison du coenzyme NAD+ et d'un cofacteur zinc, catalyser l'oxydation d'alcools en aldéhydes ou cétones.

— **les fonctions de signalisation ou liaison** : en se fixant à un récepteur ou en étant un récepteur à la fixation d'un ligand, les protéines permettent de faire transiter de l'information en intra ou extra cellulaire. La fixation entraîne la transmission d'un signal physique ou chimique via une cascade d'événements qui vont conduire à une réponse cellulaire. Par exemple, la production d'une protéine d'insuline par le pancréas va, en se fixant sur les récepteurs membranaires des cellules, déclencher des mécanismes de stockage du glucose. Ceci se traduit dans le foie par l'initiation de la transformation du glucose en glycogène.

— **les fonctions de structuration** : les protéines par leur forme et propriétés physico-chimiques hautement modulables permettent de donner une architecture ou des propriétés mécaniques aux cellules dans/autour desquelles elles se trouvent. Par exemple la combinaison de protéines de myosine II et d'actine permettent la contraction musculaire par des phénomènes de glissement de ces protéines.

Pour être produites, ces protéines suivent chacune un plan de construction défini grâce au code génétique comportant 4 unités, des nucléotides : adénine, thymine, guanine et cytosine. Ils s'enchainent dans une séquence qui va encoder l'information génétique d'un individu et forment un polymère de longueur variable nommée Acide DesoxyriboNucléique (ADN, en anglais DNA). Cette molécule d'ADN, identique dans chacune des cellules d'un organisme, est découpée en plusieurs chromosomes. Cette molécule d'ADN est présente à l'identique dans chacune des cellules d'un organisme et encode au long de sa séquence l'information relative aux protéines sous forme d'unité nommée **gène**. Chez l'humain sur qui on se concentrera dans ce manuscrit, la version 37 de l'annotation du génome humain du projet GENCODE [14] considère ainsi qu'il existe 60 651 gènes. Cependant, tous parmi eux ne produisent pas de protéine. On distingue en réalité les gènes dits **codants** (19951 toujours dans GENCODE 37) car produisant des protéines et les

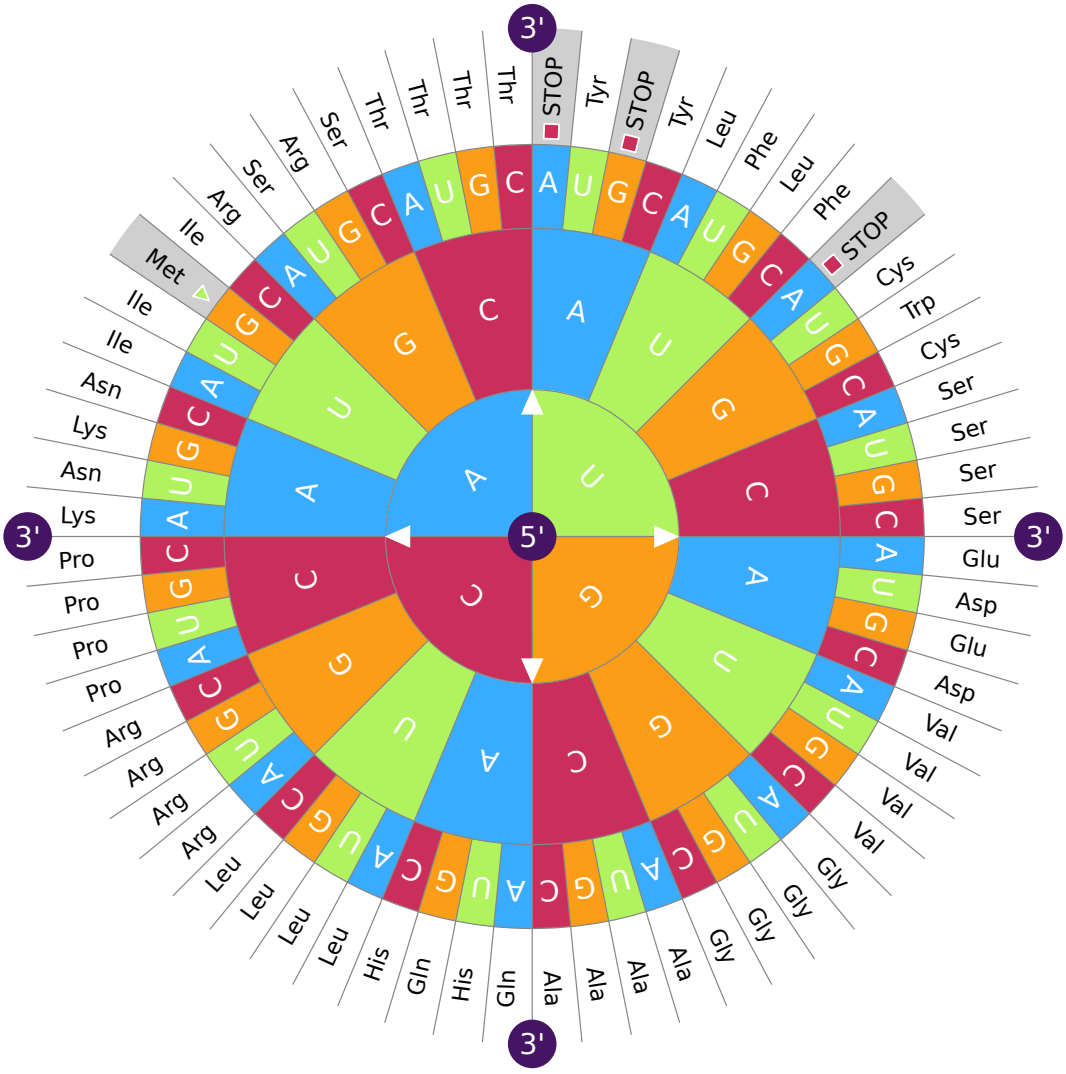

FIGURE 1.2 – Correspondance des codons avec l'acide aminé produit. La lecture s'effectue du centre du cercle vers l'extérieur pour respecter le sens 5' → 3' du RNA. Le codon de la méthionine qui est le codon de départ de toute traduction est indiqué par un triangle vert, et les trois codons stop sont indiqués avec un carré rouge.

gènes **non-codants** qui codent en fait d'autres entités moléculaires que des gènes, mais tout aussi utiles à la cellule.

Chez l'humain et plus généralement les eucaryotes, l'ADN est contenu dans le noyau de la cellule où il se trouve condensé après plusieurs étapes de repliement à l'aide de protéines appelées histones. La machinerie cellulaire de production des protéines étant située à l'extérieur du noyau, une étape intermédiaire de transcription est nécessaire afin de copier la séquence ADN d'un gène sous la forme d'une molécule capable de sortir du noyau : l'Acide RiboNucléique messager (ARNm, en anglais mRNA). Semblable à l'ADN, l'ARNm est cependant monobrin alors que l'ADN est double brin. Dans cette molécule, la thymine est remplacée par un autre nucléotide, l'uracile. La taille d'un brin d'mRNA est par ailleurs plus petite que celle d'un fragment d'ADN car il embarque l'information correspondant à un gène seulement. Chacun des triplets de nucléotides

sur sa séquence, appelés codons, encode pour un des 20 types d'acides aminés constituant la protéine associée (Figure 1.2).

Une fois copié depuis l'ADN, le mRNA immature, composé de sous unités appelées introns et exons, sort du noyau vers le cytoplasme pour subir une étape de maturation. Durant celle-ci, il se voit retirer ses introns et certains exons ou sections d'exons selon les besoins de sous-fonction de la protéine finale. On nomme **transcrits** ces différentes versions d'ARNm issues d'un même gène (Figure 1.3). Ainsi, on considère que la transcription d'un gène sous n'importe lequel de ses transcrits correspond à l'expression de son gène dans un échantillon. Par la suite, une étape de traduction de l'ARNm en une séquence d'acides aminés donnera une molécule qui sera la base de la protéine. Après le repliement de ce polymère et des modifications dites post-traductionnelles par la machinerie cellulaire, la protéine sera enfin opérationnelle pour réaliser la fonction qui lui incombe.

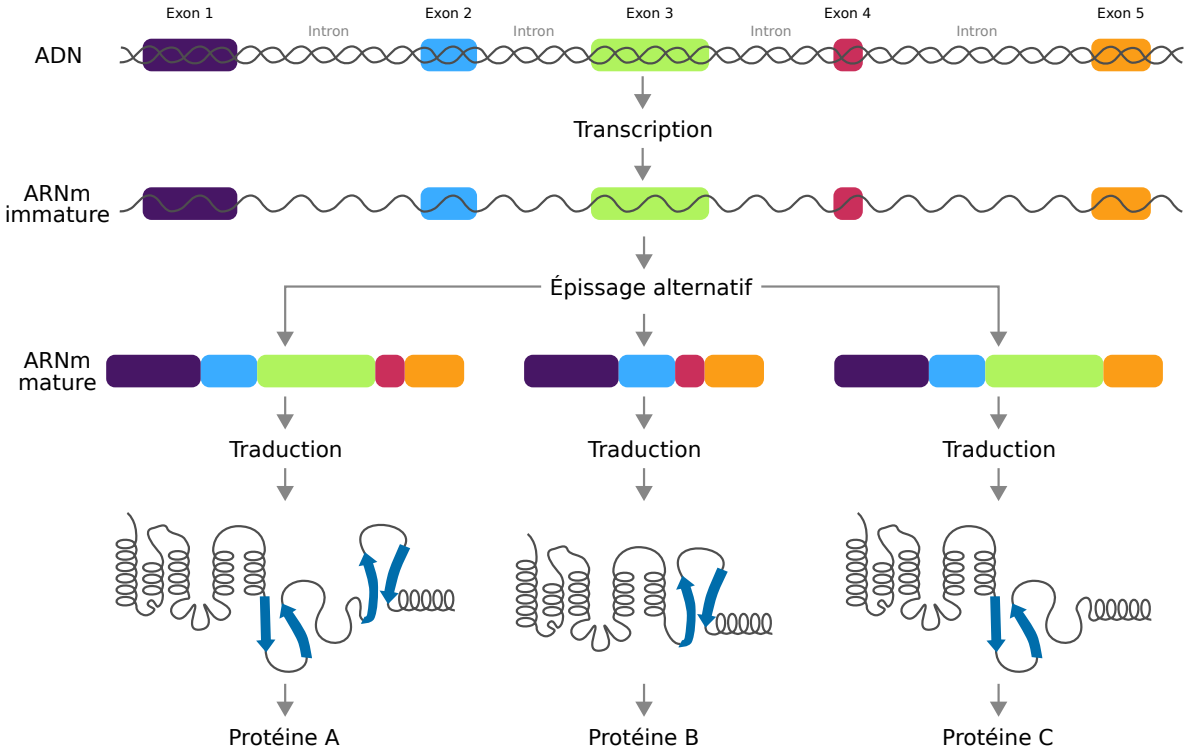

FIGURE 1.3 – Épissage alternatif avec un exemple sur 5 exons et 3 protéines différentes pouvant donc résulter du même ARNm.

1.1.2 La régulation de l'expression pour une spécification cellulaire

Comme mentionné précédemment, l'ADN est identique dans toutes les cellules d'un organisme. Chaque cellule pourtant est capable de n'exprimer, de ne transcrire, que les gènes spécifiques à sa fonction. Les myocytes seront par exemple les cellules majoritaires à produire de la myosine et les kératinocytes seront les cellules majoritaires à produire de la kératine. Cette capacité de

production spécialisée des cellules est permise par les différents mécanismes de régulation de l'expression des gènes (et par extension leur répression) qui sont établis dès les premières différenciations cellulaires au cours du développement d'un organisme. Tous ces mécanismes, dits de régulation, sont interconnectés pour ajuster en temps réel la quantité de protéines nécessaires au fonctionnement immédiat de la cellule [15]. On distingue les mécanismes suivants :

- **La régulation de l'initiation de la transcription** : lors de la transcription de l'ADN d'un gène en ARNm, l'accessibilité des régions à transcrire dépend de l'impact de plusieurs éléments de régulations parmi lesquels on retrouve les amplificateurs (en anglais *enhancers*) et les inactivateurs (en anglais *silencers*), des séquences riches en motifs de liaison de facteurs de transcription capables respectivement d'augmenter ou diminuer le niveau d'expression de gènes [16]. Très souvent présents dans des boucles de rétro-action, ces facteurs de transcription qui sont des protéines peuvent eux-mêmes voir leur expression augmentée ou diminuée par des acteurs variés tels que des ARN non-codants et des protéines [17].
- **La conformation de la chromatine** : pour rentrer dans l'espace que représente le noyau d'une cellule, l'ADN doit être replié sur lui-même à l'aide de protéines telles que des histones. Cette condensation de l'ADN implique de rendre spatialement disponibles certaines régions à la transcription et indisponibles d'autres [18]. La variation d'accessibilité de ces régions est notamment due à la localisation des boucles que forme la chromatine avec elle-même.
- **La modification post-transcriptionnelle** : les mRNA nécessitent l'ajout d'une 7-methylguanosine sur l'extrémité 5' (coiffe 5'), un épissage et une polyadénylation sur l'extrémité 3' (ajout d'une queue poly A) afin d'être traduits en protéine. En l'absence de ces modifications, les ARNm sont détruits par la cellule [19].
- **La méthylation** : les îlots CpG sont des structures de l'ADN où une cytosine est suivie d'une guanine dans le sens 5' $\rightarrow$ 3'. Très présents dans les régions d'initiation de la transcription, ils peuvent subir l'ajout d'un groupement méthyl empêchant la fixation de différents agents de la transcription [20].
- **La susceptibilité à la dégradation** : afin de perdurer plus de quelques minutes dans la cellule, certains RNA contiennent ou évitent des motifs de nucléotides influant la vitesse de dégradation (éléments riches en AU, codon stop prématuré, taille de queue poly-A) [21]. Des RNA non-codants tels que les RNA interférents de petite taille (ARNpi, ou en anglais piRNA), les ARN micro (ARNmi, en anglais miRNA) et les RNA longs non codants (ARNlnc, en anglais lncRNA) peuvent également favoriser la dégradation de certains RNA cibles [22].
- **La régulation de la traduction** : en perturbant l'initiation, l'élongation ou la terminaison de la traduction des ARNm, des acteurs tels que les fragments dérivés d'RNA de transfert (ARNt, en anglais tRNA) [23], le niveau de ribosomes disponibles [24] ou encore des ARNmi [25].

1.1.3 L'étude de l'expression des gènes pour la résolution de conditions cliniques

Qu'il s'agisse du génome entier, du génome de tissus ou du génome de cellule spécifiques, les études de profilage des transcrits produits, ou **transcriptome**, ont permis de mieux comprendre le fonctionnement basal et sain de ces entités [26, 27]. Aussi appelées approches systématiques, ces études servent en premier lieu à identifier de gènes impliqués dans des fonctions physiologiques (annotation) [28], à associer des régions de l'ADN à des traits quantitatifs (Locus de Caractère Quantitatif (LCQ), en anglais Quantitative Trait Loci (QTL)) [29], ou encore à identifier des mécanismes de régulation de la transcription [30] ou des mécanismes cellulaires tels que la différenciation [31], la réparation de l'ADN [32], etc. De plus en plus de ces études permettent par la suite une réutilisation de ces données en les déposant dans des répertoires en ligne tels que Gene Expression Omnibus (GEO) [33] et ArrayExpress [34]. Ils ont notamment été utilisées pour créer Expression Atlas [35], une carte globale d'expression des gènes avec une mise à jour récente pour inclure des données provenant du séquençage de cellules uniques [36]. Ces données, ainsi que la connaissance associée par les études dont elles sont issues, permettent alors d'avoir une vision générale du fonctionnement sain de l'humain pour mieux étudier par contraste les perturbations qui ont lieu.

Ces perturbations du fonctionnement cellulaire peuvent provenir de différentes conditions telles qu'une maladie, une mutation, un stress, ou encore l'âge. Avec chacune d'elle, la régulation de l'expression se retrouve altérée différemment par rapport au fonctionnement sain. Le transcriptome, ensemble des ARN ou transcrits produits, est alors un témoin direct des fonctions mises en défaut [37]. L'étude de la quantité de chaque transcrit exprimé donne ainsi un moyen direct de comprendre l'origine de manifestations macroscopique telles que la présence de tumeur ou nécrose dans les tissus [38, 39]. Ce type d'étude permet également une compréhension plus fine des variations moléculaires telles que des variations de pH ou la mauvaise absorption de nutriment dans l'organisme [40, 41]. En recherchant les transcrits dont la quantité change significativement entre une personne saine et une personne affectée par la condition considérée, il est alors possible d'isoler des biomarqueurs de la condition en question. Outre leur rôle d'aide au dépistage de certaines conditions, les biomarqueurs ont un intérêt en termes de soin car ils peuvent représenter ce qu'on nomme une cible thérapeutique [42]. En conceptualisant des molécules en fonction du biomarqueur, on peut être capable d'interférer avec le biomarqueur, ses précurseurs, ses co-acteurs ou ses produits, le tout en vue de limiter leurs potentiels impacts négatifs. La détection fiable de ces biomarqueurs est alors primordiale et dépend tant de la méthode de quantification des transcrits que de leur analyse.

1.2 Les technologies de transcriptomique pour la quantification de l'expression des gènes

1.2.1 Historique des technologies de transcriptomique

L'ARN est une molécule dont la stabilité est faible et la dégradation rapide en raison des nombreuses enzymes de dégradation la ciblant dans une cellule. Les premières technologies de transcriptomique ont donc rarement ciblé les ARN en eux-mêmes et ont plutôt construit des ADN complémentaires (ADNc, en anglais cDNA) de leurs séquences à l'aide de transcriptases inverses découvertes en 1970 [43]. Ces premières technologies furent basées sur les marqueurs de séquence exprimée (en anglais Expressed sequence tag (EST)), de courtes séquences d'ADNc identifiantes de transcrits qu'on détectait ensuite par migration sur gel. Développés au début des années 70, les EST ont servi de base à la méthode SAGE (Serial Analysis of Gene Expression) créée en 1995 [44] qui analyse les EST concaténés via le séquençage Sanger lui-même inventé en 1975 [45]. Cette technique de séquençage alors très populaire à l'époque [46] est dite rétrospectivement de bas débit et a permis les premiers séquençages de transcriptome partiels [47] ou complets pour des organismes à RNA de petite taille comme le bactériophage MS2 [48]. La quantification par réaction en chaine d'RNA retro transposé (qRC-RT, en anglais Reverse transcription quantitative polymerase chain reaction (RT-qPCR)) combinée à un buvardage de northern [1] (en anglais *northernblot*) fut également utilisée à partir de la fin des années 80 en raison de sa grande précision[49].

L'utilisation de SAGE et de RT-qPCR dans l'étude de l'expression des gènes a toutefois diminué au profit de nouvelles technologies capables de quantifier un plus grand nombre de transcrits simultanément [50]. Les puces à ADN par hybridation (en anglais *hybridization-based microarrays*) ont ainsi été prisés dès le milieu des années 90 [51] après la commercialisation de la première puce Affymetrix [52] en raison de leur faible rapport coût sur transcrits quantifiés. Cependant, face à la contrainte des puces à ADN de devoir connaître les séquences des transcrits à quantifier, la technologie du RNA-seq (séquençage de l'ARN, RNA-seq en anglais) est devenue un incontournable dans nombre d'études. Profitant de l'amélioration des technologies de séquençage génomique via toujours une transcription inverse de l'ARNm en ADNc, la technologie du RNA-seq va émerger dans les années 2000 et est annoncée comme révolutionnaire [53]. Toutefois, sa mention dans une publication n'arrivera qu'en 2008 [54] bien qu'elle soit utilisée dès 2006 [55].

Les puces à ADN et le RNA-seq sont encore aujourd'hui les deux technologies de quantification de l'expression des gènes utilisées, malgré le déclin annoncé des puces à ADN (Figure 1.4). Les bio-informaticien·ne·s sont donc amené·e·s à devoir traiter et analyser les données issues

1. Traduction issue de http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8392423

de ces deux approches en tenant compte des propriétés biologiques, techniques et statistiques respectives à chacune des deux technologies.

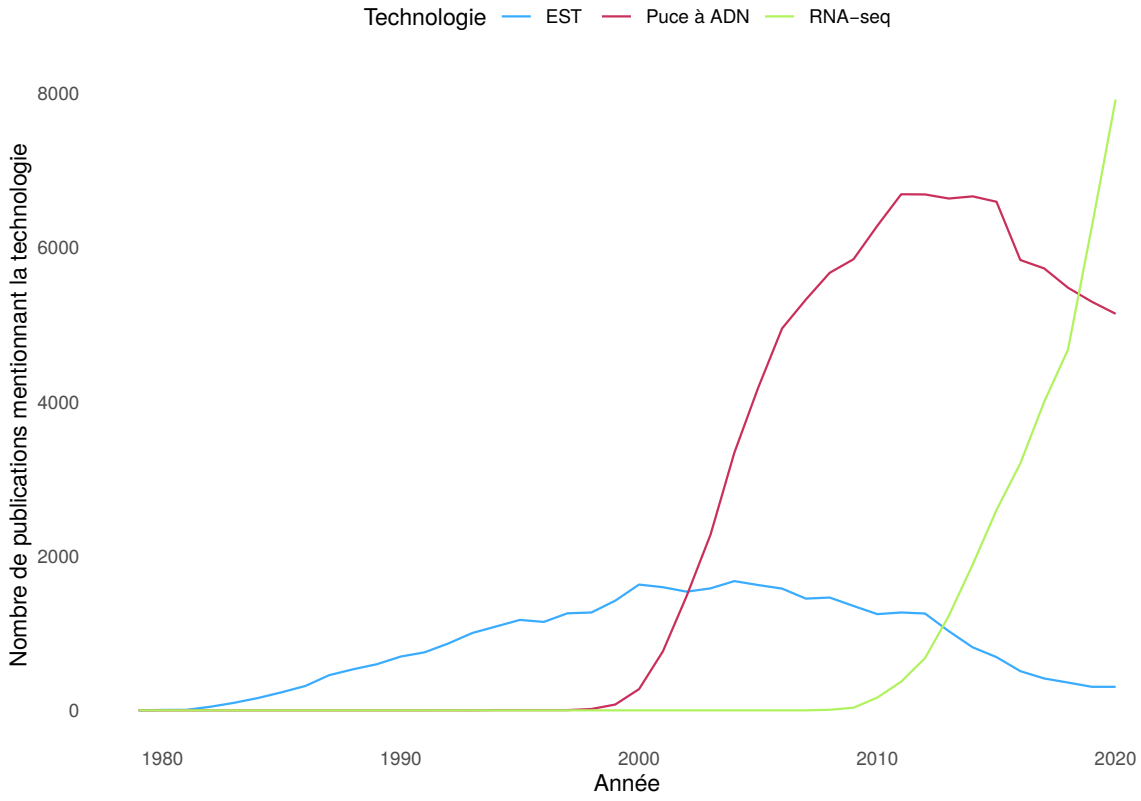

FIGURE 1.4 – Évolution de l'utilisation des différentes techniques de transcriptomique en se basant sur le nombre de publications les mentionnant. Les résultats proviennent du site PubMed (https://pubmed.ncbi.nlm.nih.gov) avec les requêtes suivantes : **EST** = "cDNA library" OR "cDNA libraries" OR "complementary DNA library" OR "complementary DNA libraries" OR "expressed sequence tag" OR "expressed sequence tags", **Puce à ADN** = "micro array" OR "microarray" NOT "chromosomal", **RNA-seq** = "RNA Seq" OR "RNA-Seq" OR "RNASeq".

1.2.2 Les puces à ADN

Principe

Une puce à ADN moderne consiste en une lame de verre sur laquelle est déposé un ensemble de fragments de ADN nommés amorces (en anglais *probes*) dans des puits qui correspondent aux RNA que l'on souhaite quantifier. Les constructeurs de puces à ADN tels que Affymetrix, Illumina ou Agilent mettent donc à disposition plusieurs modèles de puces contenant un ensemble d'amorces prédéfinies pour réaliser la quantification de l'expression de gènes chez un organisme [56]. Certains laboratoires universitaires disposent également de l'équipement pour construire leurs propres puces avec des amorces créées ou sélectionnées pour une étude [57]. Cette option est particulièrement pertinente lors de la quantification de l'expression de gènes d'organismes

dont l'annotation est récente, d'organismes non disponibles en puce chez les constructeurs ou bien lors d'étude de sous-ensembles de transcrits particuliers.

Pour venir réagir avec les amorces de la puce, les transcrits sont tout d'abord isolés de l'échantillon par purification. Une étape de transcription inverse donne ensuite l'ensemble des ADNc capables d'être liés avec les amorces. Afin de quantifier par fluorescence l'hybridation des ADNc aux amorces, les ADNc sont associés à des marqueurs fluorescents appelés fluorochromes. Certaines puces, prévues pour une utilisation avec un seul fluorochrome, sont dites à un seul canal, tandis que d'autres, dites à deux canaux, permettent l'hybridation de deux échantillons différents tels que deux conditions : sain/malade, sauvage/muté [58]. Chacun est reconnaissable sur la puce car marqué avec un fluorochrome différent : la Cyanine 3 qui émet à 570 nm (vert) et la Cyanine 5 qui émet à 670 nm (rouge). Dans les puits contenant plusieurs amorces visant un même transcrit, une hybridation compétitive a lieu et permet pour un même transcrit de quantifier relativement chacun des échantillons [59]. Après une étape de rinçage, les puces sont scannées par un laser qui va exciter les fluorochromes. Un détecteur équipé d'un capteur de fluorescence va mesurer l'intensité émanant de chaque puits et chaque longueur d'onde s'il s'agit d'une puce à deux canaux. Il en résulte une image, ou deux si il y a deux canaux, tel que visible en Figure 1.5.

En parallèle de cette méthode la plus courante, d'autres méthodes existent pour la construction des puces à ADN. On a présenté ici la version par dépôt d'amorces, mais il existe aussi une méthode par synthèse *in situ*. Cette méthode permet notamment l'utilisation d'amorces de plus grande taille (> 70 mers ou nucléotides) qui augmentent la spécificité [56]. Certaines puces à ADN utilisent également des billes de polystyrènes à la place de la lame de verre pour la fixation des amorces et se servent de ratios entre deux colorations n'interférant pas avec la fluorescence pour identifier l'amorce [60].

Afin de transformer les images obtenues, quelle que soit la technique, en une donnée exploitable, le signal détecté sur chacun doit ensuite être traité en fonction de la puce utilisée. Les puces à ADN sur lame de verre étant les plus courantes, on s'attardera par la suite sur leur traitement en particulier.

Pré-traitement des données

L'intensité de fluorescence, aussi appelée abondance, détectée par le capteur constitue la donnée retournée pour un transcrit donné sur une puce à ADN en théorie. Comme on peut le voir en Figure 1.5, cette abondance n'est toutefois pas uniforme au sein d'un même puits. L'image capturée va donc rarement être utilisée en tant que telle pour donner les abondances chiffrées finales et va passer par un premier ensemble d'ajustements.

En 2004, Petrov *et al.* [62] ont ainsi exploré en détail l'impact de différents paramètres expérimen-

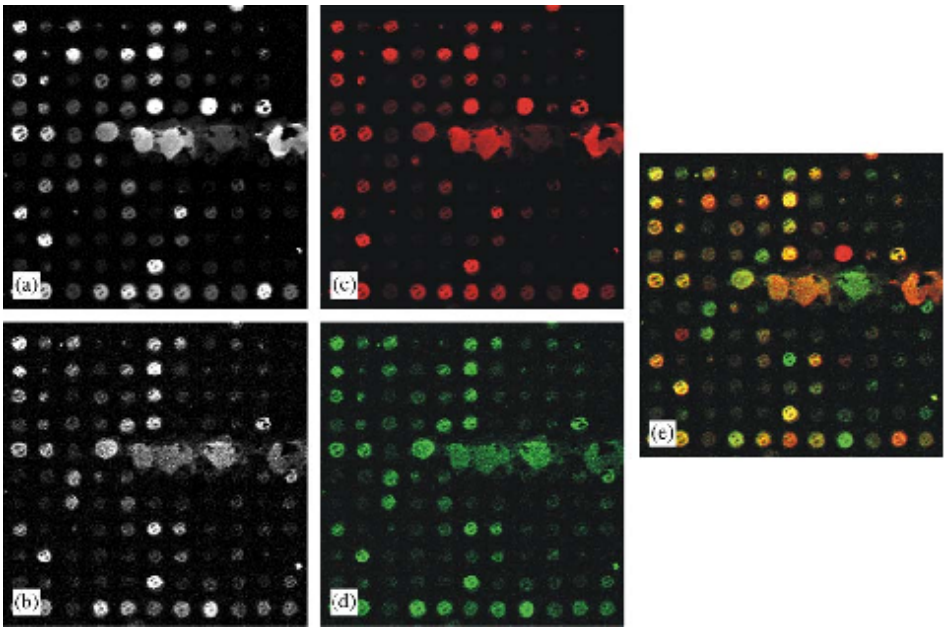

FIGURE 1.5 – Image d'une puce à ADN à deux canaux rouge et vert. (a) Canal rouge en échelle de gris, (b) Canal vert en échelle de gris, (c,d) Canaux respectifs colorés selon leur fluorescence, (e) Puce à ADN visualisée en coloration RGB (*Red Green Blue*). Des artefacts de fluorescence sont visibles en ligne 5. Issu de Lukac *et al.* 2005 [61].

taux et de différents choix de correction du signal en se basant sur un contrôle qualité approfondit des puces étudiées. Le pré-traitement de l'image prise d'une puce à ADN y est découpé selon 4 étapes, plus une s'il s'agît d'une puce à deux canaux : l'alignement d'images si deux canaux, le placement de grille, la détection de puits, la segmentation et la mesure de qualité (Figure 1.6). Ces étapes permettent en finalité d'assurer la reproductibilité de la valeur d'abondance détectée, ainsi qu'une robustesse face à des évènements tels que des artefacts de fluorescences, des contaminations de puits, ou encore des variations de fluorescence entre réplicats.

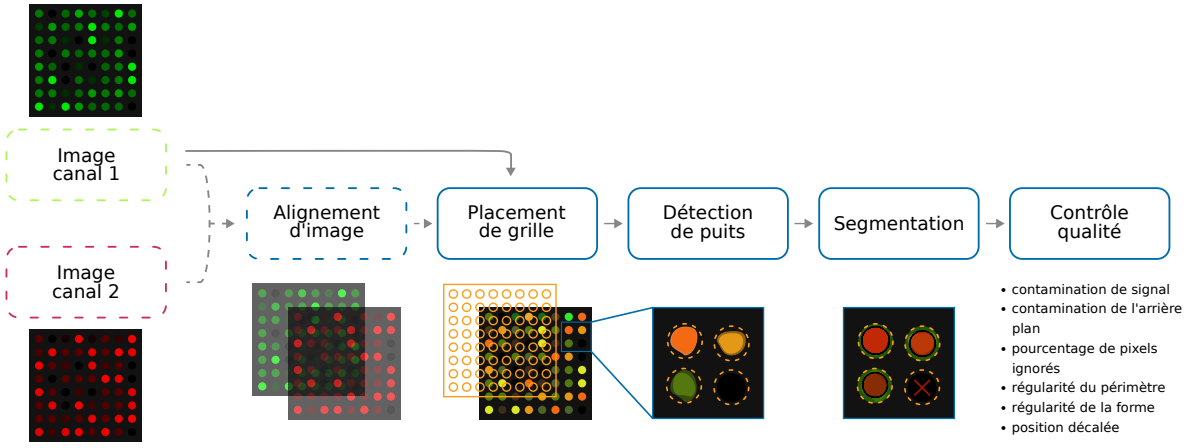

FIGURE 1.6 – Ordre des différentes étapes de pré-traitement d'une image de puce à ADN. Modifié d'après Petrov et al [62].

Une fois cette étape de traitement de l'image effectuée, de nombreux biais techniques peuvent encore impacter la valeur chiffrée retournée depuis l'image corrigée. On applique donc une étape dite de normalisation qui visera tant à palier les biais techniques contrôlables qu'incontrôlables afin de comparer au mieux différentes conditions. Bien que la variété de modèles de puces à ADN et de conditions d'utilisation puisse nécessiter l'utilisation de normalisations spécifiques, trois approches font consensus [63] (Figure 1.7) : la normalisation d'arrière-plan, la normalisation intra-puce, et la normalisation inter-puces.

La normalisation de l'arrière-plan va venir corriger le biais d'intensité parasite lié au phénomène d'hybridation non spécifique des transcrits sur les amorces dans les puits. Par une simple soustraction de l'intensité de l'arrière-plan à l'intensité détectée dans chaque puits on va permettre de compenser le biais. Lors de l'utilisation d'une puce à deux canaux, les intensités en fonction de la fluorescence utilisée tendent à ne pas couvrir le même intervalle d'intensité. Il est possible également que d'autres biais non-linéaires impactent les différents canaux de la puce. Ces biais sont corrigés par une normalisation dite intra-puce et n'est donc généralement pas utilisée dans le cas de puce à un canal. Plusieurs méthodes se sont succédé pour réaliser cette normalisation avec dans les débuts une utilisation d'un puits de contrôle, de la somme de toutes les intensités, ou de la déviation absolue médiane pour diviser les intensités de chaque transcrit. Plus récemment, l'approximation des données par une régression pondérée locale (en anglais *LOESS* pour *LOcally Estimated Scatterplot Smoothing*) inspirée du théorème de Taylor et aboutie par Cleveland et Devlin en 1988 [64] a prouvé être particulièrement efficace pour retirer ce biais intra-puce [63].

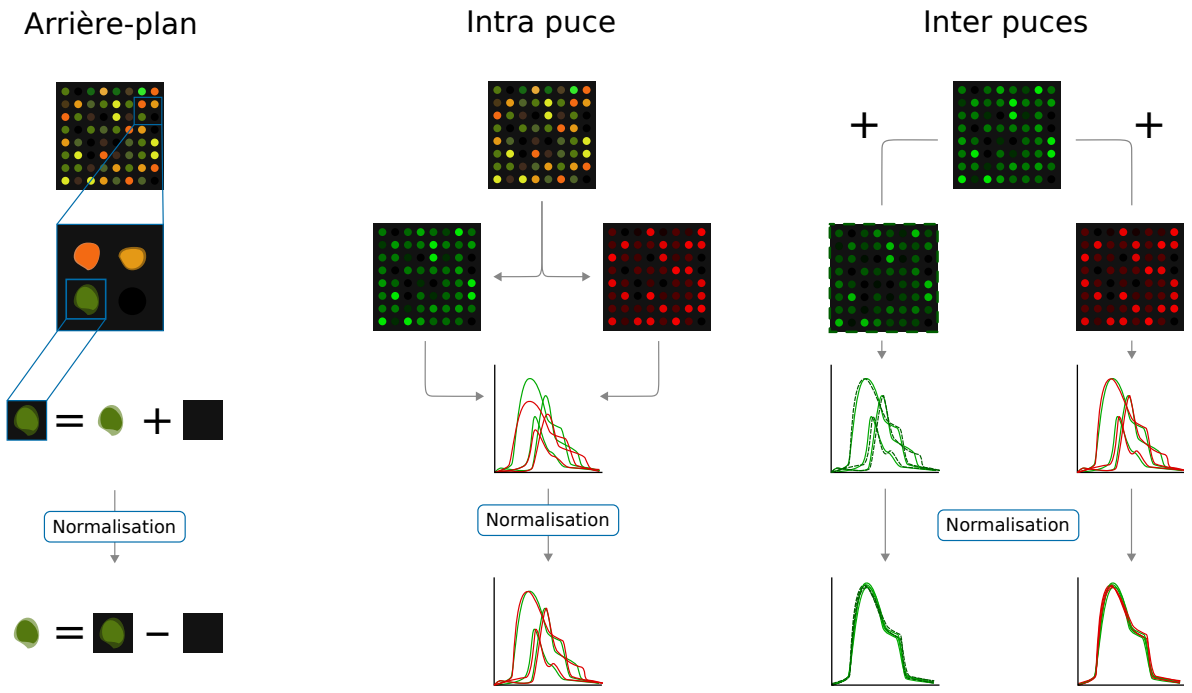

FIGURE 1.7 – Normalisations consensus de puces à ADN : la normalisation de l'arrière-plan, la normalisation intra puce, et la normalisation inter puces.

Enfin, la finalité des puces à ADN dans l'analyse d'expression de gène étant la comparaison de conditions, on va effectuer une étape de normalisation inter-puces pour s'assurer de leur comparabilité pour détecter au mieux les variations significatives entre elles. Cette correction consiste donc à rendre similaire la distribution des intensités. La méthode variera cependant selon le type et le nombre de canaux des puces bien que la plupart soient une adaptation d'une normalisation par quantiles [65].

À ces normalisation s'ajoutent d'autres transformations usuelles des données des puces à ADN telles que des filtrations [66]. On y retrouve la suppression des données d'intensité des transcrits dont la fluorescence du puits aurait saturé le capteur et engendré une quantification tronquée [67]. Lors de l'utilisation de réplicats biologiques, les transcrits présentant une faible intensité après normalisation de l'arrière-plan et une faible variation entre les réplicats peuvent être supprimées des données. Également, il est d'usage de transformer les données d'intensité en logarithme, le plus souvent un log2. Ce changement d'une expression en échelle additive en une échelle proportionnelle de facteur 2 vise à faciliter l'interprétation. En effet, dans cette disposition, une intensité augmentant ou diminuant de 1 unité signifiera respectivement un doublement ou une division par deux de l'intensité [63], ce qui est souvent considéré comme un changement de l'expression significatif en biologie.

Ces différentes normalisations et transformations peuvent mathématiquement se réaliser dans n'importe quel ordre. Toutefois, bio-informatiquement, il est important de s'attacher à l'ordre et au choix des méthodes employées car chacune possède des contraintes, un contexte d'utilisation, et doit tenir compte du type d'analyse réalisée sur les données par la suite. On parle alors de stratégie de normalisation. Parmi les précautions à prendre dans la stratégie, il est par exemple nécessaire lors d'analyses comparatives de conditions de tenir compte des transcrits supprimés d'une des deux puces à ADN uniquement. Dans le cas contraire, une différence significative artificielle pourrait être obtenue, faussant alors les résultats. De même, l'étape de normalisation intra-puce doit être faite avec parcimonie car elle peut entraîner dans certains cas une non-comparabilité ultérieure avec d'autres puces en fixant la distribution relativement à la puce considérée [68]. Enfin, plus généralement, une utilisation à mauvais escient de certaines normalisations peut aussi entraîner une suppression du signal biologique d'intérêt. Une fois totalement préparées, ces données peuvent finalement être analysées par diverses méthodes pour détecter des différences d'expression.

Mais face à la limitation en nombre de transcrits et à l'impossibilité de quantifier des transcrits inconnus, l'utilisation de puces à ADN tend à diminuer de plus en plus au profit du RNA-seq. Bolon-Canedo *et al.* soulignent ainsi en 2019 [69] qu'elles ne sont plus pertinentes à ce jour pour de la recherche exploratoire telle que l'étude du transcriptome de nouveaux organismes ou l'analyse comparative de l'expression dans des conditions mal comprises. L'utilisation des puces à ADN reste toutefois pertinente dans d'autres types d'analyses plus ciblées de transcrits. Elles

restent un outil au rapport qualité/prix imbattable dans des analyses de génotypage pour du diagnostic ou de l'étude de population.

1.2.3 RNA-Seq

Principe

Aujourd'hui, le séquençage NGS de l'ARN (RNA-seq) est devenu la technologie privilégiée pour les études transcriptomiques. Issue des techniques de séquençage nouvelle génération, le RNA-seq permet d'étudier le transcriptome d'un organisme sans connaissance préalable de sa séquence [53]. Le protocole de préparation des transcrits est le même que pour les puces à ADN sauf dans le cas d'un séquençage direct de l'ARN. Après prélèvement des échantillons, les transcrits sont isolés par purification et rétro-transcrits en ADNc. Ceux-ci sont alors amplifiés par PCR et fragmentés selon la taille requise par la technologie de séquençage. Le séquençage peut se faire actuellement par seconde ou troisième génération dépendant de l'objectif poursuivit. Les technologies de seconde génération, avec notamment celle d'Illumina (Figure 1.8), sont ainsi adaptées dans le cadre d'une comparaison de conditions. Elles produisent rapidement un séquençage au rapport qualité prix raisonnable avec un minimum de 30 millions de lectures (en anglais *reads*) alignées recommandées par le projet ENCODE [70] pour de l'ARN total chez l'humain. Les technologies de troisième génération telles que PacBio ou Nanopore (Figure 1.8) quant à elles prennent plus de temps mais peuvent séquencer des fragments de taille bien supérieure à ceux de seconde génération. Cette capacité les rend donc particulièrement adaptées à l'analyse de la variété de transcrits issus de l'épissage alternatif [71].

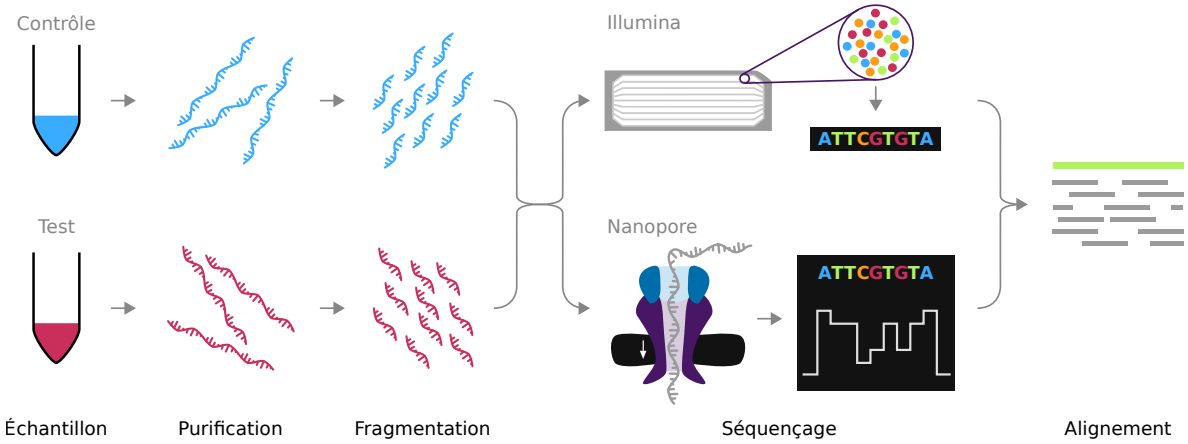

FIGURE 1.8 – Déroulé d'une opération de quantification d'expression par RNA-seq.

Les lectures de ces fragments sont ensuite alignées sur un transcriptome de référence ou bien sont l'objet d'un assemblage de transcriptome *de novo*, c'est-à-dire un nouvel assemblage sans connaissance préalable. Plusieurs méthodes partant de différents postulats existent pour réali-

ser cette étape d'alignement loin d'être triviale. Chacune possède ses avantages et limites qu'il faut toutefois prendre en compte en cela qu'elles impactent l'expression finale quantifiée [72, 73]. Dans la catégorie des associations aligneurs et quantifieurs utilisés pour la quantification, les aligneurs STAR [74] et minimap2 [75] associés au quantifieur RSEM [76] sont parmi les plus utilisés et donnent des quantifications extrêmement précises car ils donnent la position de chaque lecture de transcrits sur un génome de référence ou *de novo*. Ces aligneurs sont cependant assez lents et notamment dans le cas d'un alignement *de novo*. À l'opposé se trouvent les logiciels Salmon [77] et Kallisto [78], des quantifieurs basés sur un principe de d'alignement probabiliste nommé *selective-alignment* pour Salmon et *pseudo-alignment* pour Kallisto. Plus rapides que les aligneurs classiques, ils sont, en raison de leurs euristiques, plus sujets aux erreurs de quantification bien qu'elles restent rares. Ces aligneurs probabilistes sont également incapables de détecter de nouveaux gènes étant donné leur utilisation d'un index de transcriptome. Quelle que soit la méthode employée pour comptabiliser les lectures par transcrit, ces données doivent ensuite être traitées avant d'être analysées.

Pré-traitement des données

Comme les puces à ADN, le RNA-seq nécessite de normaliser et transformer les données pour les rendre exploitables en retirant les biais techniques et biologiques. Des contrôles qualité existent ainsi pour contrôler dans un premier temps d'éventuels artefacts de séquençage tels que la sur-représentation d'une lecture, ou encore des erreurs de formatage ou d'encodage des fichiers par exemple. Concernant la normalisation, une des approches est la division des comptes bruts par la profondeur de séquençage et 1 million, ce qui donne des comptes par million (en anglais Count per million ou CPM). Si cette normalisation est assez courante comme d'autres normalisations, elles ne peuvent toutefois être appliquées de façon systématique. La normalisation donnant des fragments par million de kilobases (en anglais *Fragments Per Kilobase Million* ou FPKM) pour le séquençage à lecture par paire (en anglais *paired-end*) ou des lectures par million de kilobases (en anglais *Reads Per Kilobase Million* ou RPKM) lorsqu'il s'agit de séquençage à lecture unique (en anglais *single-end*) sont ainsi des normalisations à éviter lors d'analyses comparatives [79]. En divisant directement les CPM par la longueur des gènes pour tenir compte du biais de comptage en faveur des gènes de plus grande taille, l'utilisation de la normalisation en FPKM/RPKM entraîne potentiellement une somme de lectures normalisée différente dans chaque échantillon. L'analyse comparative étant un des intérêts majeur de la transcriptomique en recherche, en médecine moléculaire et en biologie, d'autres mesures normalisées compatibles avec ce type de problématique ont été développées. Le RLE, pour *Relative Log Expression* [80] et le TMM pour Trimmed Mean of M-values [81] sont ainsi des normalisations recommandées par l'étude du French StatOmique Consortium [2] en 2013 [82]. Tous deux partent

2. Dans cette publication, le RLE est introduit par le biais du progiciel (en anglais *package*) DESeq [80] qui implémente cette mesure

du principe que l'expression de la majorité des gènes ne change pas entre deux conditions. Le RLE consiste tout d'abord à calculer une pseudo référence (Équation 1.1) via une moyenne géométrique sur tous les échantillons [83] qui contrairement au RPKM conserve bien une somme de lectures identique entre échantillons et est moins sensible aux valeurs extrêmes que la moyenne arithmétique. Un facteur de taille (Équation 1.2) est ensuite calculé comme la médiane des rapports entre les comptages de chaque échantillon avec ceux de la pseudo-référence. Celui-ci est alors utilisé pour normaliser les comptes bruts de chaque échantillon.

$$\text{pseudo-référence} = \left(\prod_{t=1}^{m} k_{it}\right)^{1/m} \tag{1.1}$$

$$\text{facteur de taille} = \hat{s}_j = \underset{i}{\text{median}} \frac{k_{ij}}{\left(\prod_{v=t}^{m} k_{it}\right)^{1/m}} \tag{1.2}$$

Où :

$i =$ gène $= 1, ..., n$

$j =$ échantillon $= 1, ..., m$

$k =$ table de comptes de lectures observés $n \times m$

Le TMM quant à lui effectue une normalisation par le nombre de fragments d'ARN pour chaque gène dans chaque échantillon et s'en sert pour calculer la valeur M qui est rapport du niveau moyen en base 2, soit un log2 entre les échantillons de deux conditions (Équation 1.3). Certains comptes pouvant être nuls ce que ne peut accepter un logarithme, il est d'usage d'employer des pseudo-comptes qui sont des comptes auxquels on a ajouté 1 [84]. Parallèlement, il calcule la valeur A qui est la valeur absolue du niveau d'expression (Équation 1.4). Ces deux valeurs sont respectivement tronquées de 30% et 5% et servent à estimer une moyenne pondérée d'après un échantillon référence qui est l'échantillon au 3ème quartile le plus proche de la moyenne des 3èmes quartiles (Équation 1.5) qui permet enfin de calculer le facteur de normalisation (Équation 1.6).

$$M_g = \log_2 \left(\frac{Y_{gk}/N_k}{Y_{gk'}/N'_k}\right) \tag{1.3}$$

$$A_g = \frac{1}{2}\left(\log_2 \left(\frac{Y_{gk}}{N_k}\right) + \log_2 \left(\frac{Y_{gk'}}{N'_k}\right)\right) \tag{1.4}$$

$$w^r_{gk} = \frac{N_k - Y_{gk}}{N_k \times Y_{gk}} + \frac{N_r - Y_{gr}}{N_r \times Y_{gr}} \tag{1.5}$$

$$\log_2 (TMM^r_k) = \frac{\sum_{g \in G^*} w^r_{gk} \times M^r_{gk}}{\sum_{g \in G^*} w^r_{gk}} \tag{1.6}$$

Où :

$g =$ gène

$k =$ échantillon

$N_{...} =$ nombre total de lectures pour . . .(soit k, soit r)

$Y_{g...} =$ comptes observés pour . . .(soit k, soit r)

$r =$ référence

$G^* =$ ensemble des gènes non tronqués lors du tronquage de M_g et A_g

Une étude détaillée des performances et propriétés mathématiques du RLE et du TMM a par ailleurs été publiée en 2016 en reprenant les différentes étapes de leur normalisation pour en tirer des équivalences [85] (Figure 1.9). À ces méthodes de normalisation s'en ajoutent d'autres se basant sur le contenu en GC qui va favoriser le séquençage de gènes par rapport à d'autres [86]. D'autres méthodes encore utilisent des gènes contrôle, le plus souvent des gènes dit de ménage (en anglais *housekeeping genes*) [87] à l'instar des puces à ADN. Chaque méthode de normalisation va donc modifier considérablement la distribution des données d'expression et la stratégie de normalisation mise en place va et doit donc dépendre de l'analyse ultérieure faite sur les données.

Step	Description	TMM (edgeR)	RLE (DESeq2)
I	Pre-normalization by library size	$Y_{gkr} = \frac{X_{gkr}}{N_{kr}}$	
II	Reference sample, or *pseudo-reference sample* (DESeq2)	$Y_g^{\mathrm{TMM}} = Y_{g11}$	$Y_g^{\mathrm{RLE}} = \sqrt[KR]{\prod_{k=1}^{K}\prod_{r=1}^{R} X_{gkr}}$
III	Relative sizes of transcriptomes and reference sample, or *relative scaling factors* (edgeR), or *size factors* (DESeq2)	$\tau_{kr}^{\mathrm{TMM}} = \frac{1}{\#\mathcal{G}_{kr}^*}\sum_{g\in\mathcal{G}_{kr}^*}\frac{Y_{gkr}}{Y_g^{\mathrm{TMM}}}$ *where* $\mathcal{G}_{kr}^*$ represents the set of not trimmed genes	$\tau_{kr}^{\mathrm{RLE}} = \underset{g}{\mathrm{median}}\left(\frac{X_{gkr}}{Y_g^{\mathrm{LRE}}}\right)$
IV	*Relative scaling factors* adjusted to multiply to 1 (edgeR)	$\tilde{\tau}_{kr}^{\mathrm{TMM}} = \frac{\tau_{kr}^{\mathrm{TMM}}}{\tilde{\tau}^{\mathrm{TMM}}}$ where $\tilde{\tau}^{\mathrm{TMM}} = \sqrt[KR]{\prod_{k=1}^{K}\prod_{r=1}^{R}\tau_{kr}^{\mathrm{TMM}}}$	
V	Taking into account both the relative size and the library size, or *effective library size* (edgeR)	$e_{kr}^{\mathrm{TMM}} = \tilde{\tau}_{kr}^{\mathrm{TMM}} N_{kr}$	
VI	Normalization factors, or *relative normalization factors* (edgeR), or *size factors* (DESeq2)	$f_{kr}^{\mathrm{TMM}} = \tilde{\tau}_{kr}^{\mathrm{TMM}}$	$f_{kr}^{\mathrm{RLE}} = \tau_{kr}^{\mathrm{RLE}}$
VII	Normalization of counts, or *counts-per-million* (edgeR)	$Z_{gkr}^{\mathrm{TMM}} = \frac{X_{gkr}}{e_{kr}^{\mathrm{TMM}}} 10^6$	$Z_{gkr}^{\mathrm{RLE}} = \frac{X_{gkr}}{f_{kr}^{\mathrm{RLE}}}$

FIGURE 1.9 – Table de comparaison des étapes de normalisation avec équivalences re-exprimées. Issu de la Table 2 (tronquée) de Maza 2016 [85].

1.3 L'analyse transcriptomique par réseaux de co-expression

L'explosion de la taille et quantité de jeux de données mis à disposition publiquement engendre une demande croissante en techniques d'analyse capable de les gérer mais également de profiter de la précision de séquençage apportée. L'analyse d'expression différentielle fut la première méthode dédiée spécifiquement à l'analyse des données de quantification d'expression contrairement à des méthodes de statistique exploratoire comme l'analyse par composantes principales (ACP, en anglais PCA) [88]. Elle consiste à comparer l'expression de chaque gène entre différentes conditions, le plus souvent un contrôle contre un test, pour déterminer avec un test statistique quels gènes voient leur expression significativement changer dans la condition test. Si plusieurs méthodes ont été développées pour s'assurer de la robustesse de l'identification des gènes différentiellement exprimés [89, 90], le principe reste le même : identifier les gènes sur-exprimés ou sous-exprimés. Dans la recherche permanente qu'est la priorisation de gènes candidats à une maladie, l'analyse d'expression différentielle a donc grandement contribué à l'approfondissement de la connaissance de certaines pathologies et a permis d'y associer des biomarqueurs pour réaliser des diagnostics [91]. Outre les problèmes de reproductibilité soulevés par plusieurs études [92], les analyses d'expression différentielle présentent également l'inconvénient d'être limitées à quelques gènes d'un ou plusieurs mécanismes qui impliquent pourtant bien plus de gènes. À titre d'exemple, nombre de gènes associés à des maladies Mendéliennes sont exprimés dans plusieurs tissus sans entraîner nécessairement de dysfonctionnement, suggérant plutôt une interaction délétaire qu'un gène problématique à lui seul [10]. De plus, l'analyse d'expression différentielle est mise en difficulté dans le cas de gènes à large variance à travers les échantillons étant donné la tendance des plages de valeurs d'expression à se chevaucher [92]. Elle en vient alors à manquer de sensibilité et va omettre des gènes dont l'implication est effective dans certaines condition [93]. L'analyse par réseaux de co-expression de gènes a notamment été développée dans l'optique de répondre à ce type de problématique. En ne considérant non pas seulement les gènes mais leur dynamique ou profil d'expression et leur synchronisation, les réseaux de co-expression visent à identifier les motifs de variation de la transcription qui indiquent des interactions fonctionnelles ou de régulation des relations entre gènes [94]. La découverte d'information est cependant rarement faite à partir du simple réseau construit. Pour discerner les gènes interagissant préférentiellement ensemble, une méthode de partitionnement est appliquée et va définir des groupes de gènes qu'on nomme modules. Ceux-ci sont ensuite interprétés à l'aide de différentes méthodes incluant de la connaissance a priori, ou de l'analyse topologique pour de la connaissance *ex nihilo*. Il en résulte un pipeline d'analyse (Figure 1.10) [95] qu'on va décrire en détail dans les parties qui suivent.

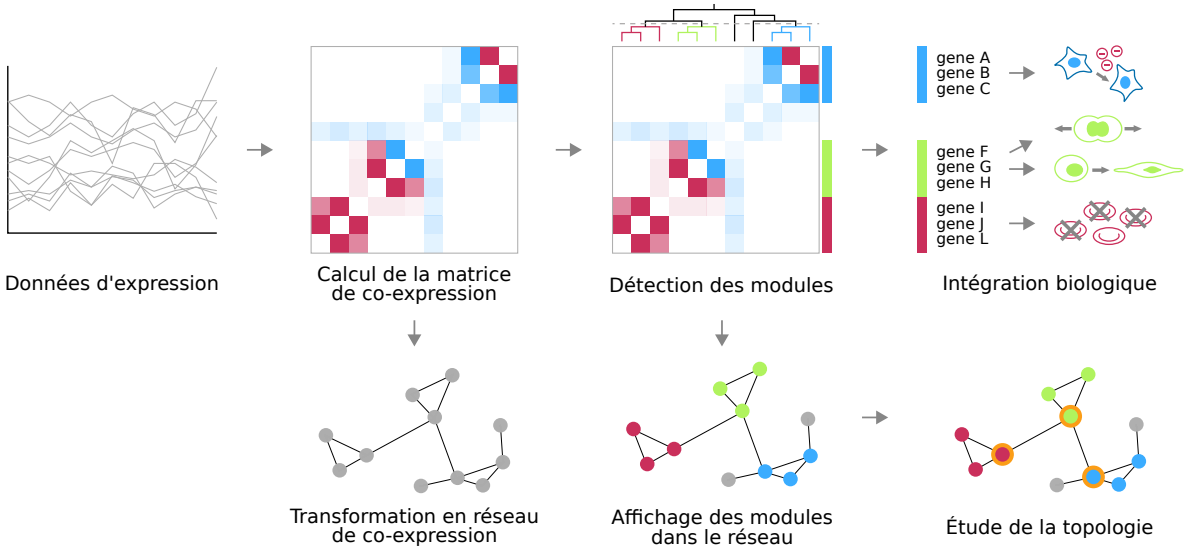

FIGURE 1.10 – Étapes de réalisation d'une analyse par réseaux de co-expression.

1.3.1 Définitions des réseaux et principe de la co-expression

L'analyse par réseaux de co-expression de gènes (en anglais Gene Co-expression Network (GCN)) a su bénéficier des avancées conceptuelles en théorie des réseaux depuis sa conception pour encoder et étudier au mieux les relations entre gènes [96]. La théorie des réseaux appartient elle-même à la théorie des graphes, structure mathématique derrière les réseaux [97]. Un **réseau** n'est en effet qu'une implémentation d'un **graphe**, un cas particulier, pour représenter des entités et leur connections dans un contexte donné. Ainsi, si les termes de graphe et réseau sont souvent utilisés de façon interchangeable hors de la recherche en informatique et mathématique, il est à noter qu'ils désignent des concepts différents bien qu'imbriqués, et au vocabulaire spécifique. Ainsi, les entités considérées, ici les gènes, sont appelés **nœuds** dans un réseau et **sommets** dans un graphe. De même, les connections reliant les entités seront nommées respectivement **lien** et **arrête**. Dans cette thèse, les termes de nœud et lien seront donc préférés car issus de la théorie des réseaux. Les liens représentent de façon binaire la présence (valeur du lien égale à 1) ou l'absence, (valeur égale à 0) de relation entre les nœuds. Les relations entre gènes en biologie ne sont toutefois pas aussi dichotomiques et nécessitent plus de granularité pour décrire la multiplicité et le niveau des interactions. Une même cellule échantillonnée à des temps différents aura un profil plus ou moins différent où les fonctions clefs auront peu changé mais celles plus variables auront varié bien plus. C'est pourquoi on utilise plus souvent des réseaux dit pondérés qui vont ajouter une valeur réelle $\mathbb{R}$, le plus souvent décimale $\mathbb{D}$ entre 0 et 1 qui vont représenter une force de liaison entre deux gènes. Si le réseau est **signé**, ces valeurs seront en plus positives si le lien représente une co-expression conjointe, et négative s'il représente une co-expression opposée (Figure 1.11).

Lorsque deux nœuds sont reliés par un lien, on dit qu'ils sont **voisins**. De cette unité de base

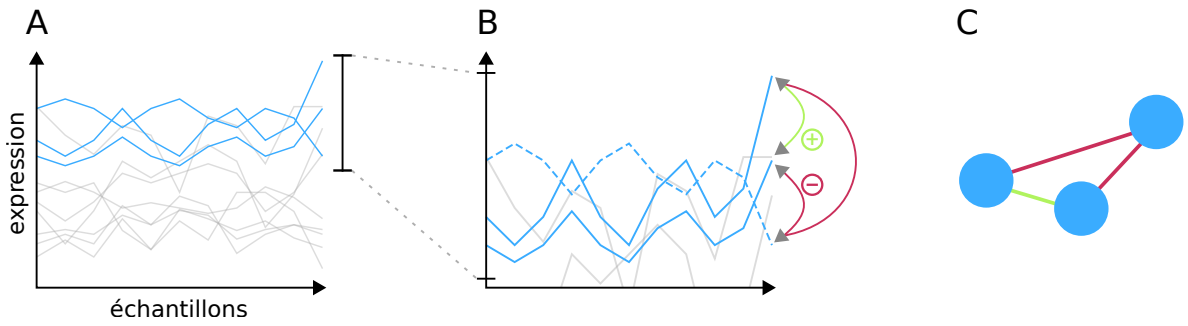

FIGURE 1.11 – Définition des liens du réseau pour des gènes se co-exprimant. A. Profils d'expression de gènes selon plusieurs échantillons avec en bleu 3 gènes co-exprimés. B. Zoom sur l'expression de ces 3 gènes avec mise en avant de deux gènes avec un profil dans un sens et un gène avec un profil dans le sens inverse ; les deux profils en sens identique ont donc un score de similarité positif entre eux, et négatif avec le troisième gène. C. Ces scores et leur signe sont utilisés pour visualiser les liens entre ces gènes sous forme de nœuds dans un réseau.

vont découler plusieurs mesures qui vont permettre de caractériser un réseau du point de vue de sa topologie. Le **degré** d'un nœud est ainsi le nombre de voisins dont il dispose tandis que la somme des valeurs des liens pondérés ou non sera appelée la **force** ou la **connectivite** d'un nœud selon le contexte. À l'échelle du réseau on va retrouver des mesures qui vont permettre de localiser des zones aux propriétés différentes du reste du réseau. Le cumul des **plus court chemin** pour se rendre d'un nœud à un autre est ainsi un moyen de mesurer la **centralité** d'une zone ou d'un nœud. Ces métriques peuvent également tenir compte de la direction des liens si une direction est précisée. Dans ce cas on parle de réseau dirigé en opposition avec les réseaux non dirigés où le sens du lien est inconnu ou inexistant. Habituellement cependant, les GCN sont non dirigés car leur construction ne permet pas nativement de déterminer une causalité.

Les GCN se construisent conventionnellement à partir d'une matrice d'expression $n \times m$ d'indice $i = 1 \dots n$ pour les échantillons et de variable $j = 1 \dots m$. Une métrique est alors calculée sur chaque paire de gènes pour quantifier leur relation. Plusieurs approches existent dans le choix d'une métrique comme des modèles de mélange gaussien ou des probabilités bayésiennes. Mais l'approche la plus commune et sur laquelle on s'est concentré durant cette thèse est une approche par score de similarité. Ce terme généraliste permet d'englober la quantité importante de méthodes employée pour quantifier la similarité des gènes deux à deux.

1.3.2 Considérations préalables à l'analyse par réseaux de co-expression de gènes

L'analyse de données d'expression de gènes par réseaux de co-expression nécessite des précautions et préparations en plus de celles effectuée sur chaque technologie de transcriptomique. Sur le plan de la normalisation, certaines normalisations de puce à ADN et de RNA-seq ne seront pas compatible avec l'analyse par co-expression telle que décrite dans la section suivante [95]. Pour des données provenant de RNA-seq, les normalisations FPKM/RPKM et TPM ou tout

autre normalisation visant à corriger pour le nombre total de lectures ou la taille des gènes sont à proscrire. En présence de gènes très fortement exprimés, le restant des gènes va décroître relativement ce qui va entraîner une co-expression artificielle entre ces gènes [98]. Les informations pointant vers une normalisation adaptée en particulier pour le RNA-seq en co-expression sont toutefois manquantes [3]. Les TMM et RLE sont conceptuellement compatibles mais les dernières publications tendent à utiliser la correction par composante principale (CP, en anglais Principal Component PC) [94] qui effectue en plus une correction de multiples biais et va retirer des facteurs confondants non identifiés. Ces facteurs confondants sont d'ailleurs une source majeure de co-expression erronée car ils favorisent la corrélation de gènes sur des critères indépendants de la question biologique étudiée. Dans le cas des puces à ADN, une étude de Reverter *et al.* trouve la normalisation par modèles mixtes comme idéale [100]. Cependant, une étude de Lim *et al.* trouve plus tard le MAS5 comme idéal mais sans comparaison avec la normalisation par modèle mixte [101].

Malgré une normalisation visant à donner des résultats les plus proches de la vérité possible, chacune des technologies de transcriptomique, même si effectuées sur des échantillons identiques, va donner des réseaux différents comme l'a étudié Giorgi *et al.* [102]. Les réseaux basés sur des puces à ADN tendant à être plus similaires aux réseaux biologiques déjà connus par rapport au RNA-seq. Un biais est toutefois présent dans cette conclusion car les réseaux biologiques en question influent en partie le design des puces à ADN en termes de transcrits mesurés. Ainsi, les réseaux basés sur RNA-seq sont moins similaires aux réseaux biologiques car ils prennent en compte la totalité des transcrits présents et permettent une meilleure découverte de nouveaux gènes d'intérêt. La nature de cette différence, signal réel ou bruit, est toutefois un cas particulier à chaque couple réseau/organisme. Ballouz *et al.* sont venu en 2015 [103] préciser les différences et ont constaté que la différence majeure repose sur la variation de topologie et plus particulièrement de connectivité. De nombreux gènes détectés comme pivot dans les réseaux basé sur puce à ADN ne sont pas retrouvés dans ceux à RNA-seq et inversement bien que leurs propriétés fonctionnelles soient les mêmes. En investiguant la contribution des transcrits aux gènes pivots, ils concluent que le RNA-seq tend à capturer plus de variation d'expression dans les données, mais qu'il est en contrepartie très sensible au bruit causé par les faibles intensités d'expression. Sur le plan de la compensation de données manquantes, on peut mentionner que les techniques actuelles d'inférence de l'expression des gènes va impacter la construction finale du réseau [104]. Les données de puces à ADN particulièrement nécessitent de l'inférence du fait du nombre restreint de transcripts quantifiés. Les méthodes pour inférer l'expression diffèrent toutefois entre puce à ADN et RNA-seq (voir scRNA-seq) du fait de leurs distributions différentes.

Le nombre d'échantillons disponibles pour construire le GCN va également impacter sa topologie finale et les informations pouvant en être extraites. Plus il a d'échantillons, plus la qualité des

3. À ce jour et lors de la réalisation des travaux présentés dans cette thèse. Une publication préliminaire (en anglais *preprint*) réalisé par A. Vandenbon en mars 2021 [99] a toutefois réalisé une comparaison étendue de 6 normalisations pour relever que le quartile supérieur est la méthode la plus adaptée globalement.

GCN tend à s'améliorer, cependant cette relation que Ballouz et al définissait comme linéaire, ne l'est pas. Liesecke *et al.* ont démontré en 2019 que celle-ci atteint un plateau qui dépend de la technologie de transcriptomique [105]. À taille d'échantillons équivalente, les réseaux de puces à ADN permettent une meilleure découverte de fonctions physiologiques. Ceci est toujours à mettre en perspective avec le fait que le RNA-seq comprends plus de transcrits et que de nombreuses fonctions physiologiques sont issues d'analyses sur puce à ADN, et que d'autres fonctions sont encore à découvrir avec l'avènement du RNA-seq. En finalité, il est recommandé d'utiliser au minimum 100 échantillons pour la construction d'un réseau, bien qu'un nombre inférieur soit utilisable à la condition d'assurer une qualité des données via une normalisation et correction des artéfacts rigoureuse.

Tous les transcrits ne sont également pas bon à prendre en compte dans une construction de GCN. Leur filtration cependant doit tenir compte des propriétés et postulats faits par les méthodes de construction du GCN, car une filtration mal menée peu significativement altérer le réseau. Une erreur courante est ainsi de n'utiliser que les gènes différentiellement exprimés pour construire un réseau, ce qui va changer la loi de distribution de l'expression sur laquelle de nombreuses méthodes se basent [95]. À l'inverse, certaines filtrations vont être essentielles afin de réduire le bruit dans l'expression. Les gènes ayant une très faible variation d'expression entre échantillons sont ainsi à retirer des jeux de données car ils vont avoir tendance à corréler entre eux. De même en RNA-seq, on retirera les gènes dont le nombre de lectures est inférieur à un seuil dépendant du séquenceur. Similairement dans les puces à ADN, on supprimera les gènes dont les intensités sont proches du seuil minimal de détection. En considérant l'intégralité de ces précautions avant la construction d'un GCN, il est possible d'assurer un réseau représentatif de la condition étudiée.

1.3.3 Construction

Score de base

La méthode de construction abordée dans cette thèse se focalise sur une approche par score de similarité entre gènes deux à deux. Si ce score était initialement basé sur une approche naïve par corrélation de Pearson [106] (Équation 1.7), d'autres scores plus robustes ont depuis été proposés. Ainsi, La corrélation de Spearman (Équation 1.8) est une méthode couramment utilisée en raison de sa faible sensibilité aux artefacts grâce à sa corrélation par rang [104, 107, 108]. L'utilisation de Pearson n'est toutefois pas à proscrire mais demande un plus grand nombre d'échantillons et une vigilance vis-à-vis des artéfacts dans les données. Toujours dans les scores basés sur des coefficients de corrélation, la corrélation médiane bi-pondérée, aussi appelée bicor [109] (Équation 1.9), se base sur une corrélation par médiane qui se veut plus robuste que Pearson. Certaines mesures de distance sont aussi utilisées en tant que score de similarité. C'est le cas de la mesure dite d'information mutuelle (Équation 1.10) qui mesure la quantité d'information ob-

tenable d'une variable en en utilisant une autre pour estimer la dépendance entre celles-ci [110]. En définitive, le choix de la méthode de base de calcul du score de similarité dépend de la nature des données (ex : linéaire ou non entre gènes) bien que la corrélation de Spearman et la bicor soient aujourd'hui les plus courantes [107]. Aucune de ces méthodes ne met toutefois à profit les outils de théorie des réseaux pour ajuster son score de similarité.

$$\text{Pearson } \mathrm{cor}_{jj'} = \frac{\mathrm{cov}(X_j, X'_j)}{\sigma_{X_j}\sigma_{X_{j'}}} \tag{1.7}$$

$$\text{Spearman } \mathrm{cor}_{jj'} = \frac{\mathrm{cov}(rg_{X_j}, rg_{X_{j'}})}{\sigma_{X_j}\sigma_{X_{j'}}} \tag{1.8}$$

$$\text{bicor } \mathrm{bicor}_{jj'} = \frac{\sum\limits_i \left(X_{ij} - \mathrm{med}(X_j)\right) w_i^{(X_j)} \left(X_{ij'} - \mathrm{med}(X_{j'})\right) w_i^{(X_{j'})}}{\sqrt{\sum\limits_i \left[\left(X_{ij} - \mathrm{med}(X_j)\right) w_i^{(X_j)}\right]^2} \sqrt{\sum\limits_i \left[\left(X_{ij'} - \mathrm{med}(X_{j'})\right) w_i^{(X_{j'})}\right]^2}} \tag{1.9}$$

$$\text{Info. mut. } \mathrm{mi}_{jj'} = \sum_{i \in X_j} \sum_{i \in X_{j'}} p(X_j, X_{j'}) \log\left(\frac{p(X_j, X_{j'})}{p(X_j)p(X_{j'})}\right) \tag{1.10}$$

Où :

$$rg_X = \text{rang des valeurs de } X$$

$$u_i = \frac{X_{ij} - \mathrm{med}(X_j)}{9\,\mathrm{mad}(X_j)} \qquad v_i = \frac{X_{ij'} - \mathrm{med}(X_{j'})}{9\,\mathrm{mad}(X_{j'})}$$

$$w_i^{(X)} = \left(1 - u_i^2\right)^2 I\left(1 - |u_i|\right) \qquad I\left(1 - |u_i|\right) = \begin{cases} 1 & \text{si } 1 - |u_i| > 0 \\ 0 & \text{sinon} \end{cases}$$

$$p(X_j) = \text{densité de la loi de } X_j$$

Adjacence

En 1999, les physiciens A. Barabasi et R. Albert découvrent une propriété topologique à laquelle de nombreux réseaux de très grande taille répondent ou par laquelle ils sont approximables [111] : l'**invariance d'échelle** (en anglais *scale-free*) [112]. Cette propriété, qui permet de classer ces réseaux comme réseaux **ultra petit monde** [113] (en anglais *ultra small-world*), indique que la probabilité $P(k)$ qu'un nœud du réseau interagisse avec k autres nœuds décroît selon une loi de puissance telle que $P(k) \sim k^{-\lambda}$. Communément, on dit qu'il existe un grand nombre de nœuds peu connectés, et peu de nœuds avec un grand nombre de connexions. En théorie des réseaux, on parle de **nœuds pivots** (en anglais *hub*) qui vont avoir une forte centralité [114]. C'est le cas des GCN qui possèdent un nombre limité de gènes contribuant à réguler l'expression par le biais de facteurs de transcription, d'éléments de méthylation ou encore de QTL [107]. Évolutivement

parlant, cela s'explique par la relation préférentielle de gènes naissants *de novo* avec des gènes déjà bien établis et centraux car préservés au long des mutations d'un organisme [1].

Six ans plus tard, les biostatisticiens S. Horvath et B. Zhang proposent de tirer parti de cette propriété d'invariance d'échelle pour améliorer le score de similarité afin qu'il puisse refléter la réalité biologique [95]. Pour cela, ils proposent d'appliquer une fonction d'adjacence à la matrice de similarité selon ce qu'ils nomment un seuillage souple. Il s'agit en fait d'une fonction qui se base sur une loi de puissance (Équation 1.11) pour déterminer l'adjacence, plutôt qu'un seuillage strict qui revient à binariser le score de similarité (Équation 1.12). Pour estimer le paramètre β de la loi de puissance à appliquer au score de similarité, il faut paramétrer plusieurs lois de puissance avec $\beta = 1, \ldots, n$ et mesurer laquelle suit au mieux les données. Une fois le β optimal trouvé, la loi de puissance ajustée est appliquée à la matrice de similarité qui devient une matrice d'adjacence.

$$a_{ij} = \text{power}\left(s_{ij}, \beta\right) \equiv |s_{ij}|^{\beta} \tag{1.11}$$

$$a_{ij} = \text{signum}\left(s_{ij}, \tau\right) \equiv \begin{cases} 1 & \text{if } s_{ij} \geq \tau \\ 0 & \text{if } s_{ij} < \tau \end{cases} \tag{1.12}$$

Où :

$m =$ taille max de la matrice (carrée) de similarité

$i, j = 1, \ldots, m$

$s =$ score de similarité

$a =$ adjacence

Topologie de modularité hiérarchique

La topologie d'invariance d'échelle des réseaux de gènes va être encore précisée par E. Ravasz *et al.* en 2002 [115]. Bien qu'appliqués aux réseaux métaboliques, leurs travaux ouvrent sur une généralisation à l'échelle de l'organisation cellulaire de différents acteurs dont les gènes. Ils démontrent que les réseaux purement invariants d'échelle ne suffisent pas à correctement modéliser la distribution des degrés des nœuds et la topologie des modules présents dans les réseaux biologiques. Pour le prouver, ils utilisent le coefficient de partitionnement C qui permet de mesurer cette modularité et est défini comme $C_i = 2n/k_i\left(k_i - 1\right)$. Ils observent alors que ce coefficient de partitionnement suit une loi d'échelle $C(k) \sim k^{-}1$, qui indique l'existence d'une modularité hiérarchique (Figure 1.12).

Ainsi, les gènes au sein d'un réseau vont avoir tendance à s'organiser sous forme de multiples groupes de gènes, des modules, formant eux-mêmes des modules plus importants. L'origine

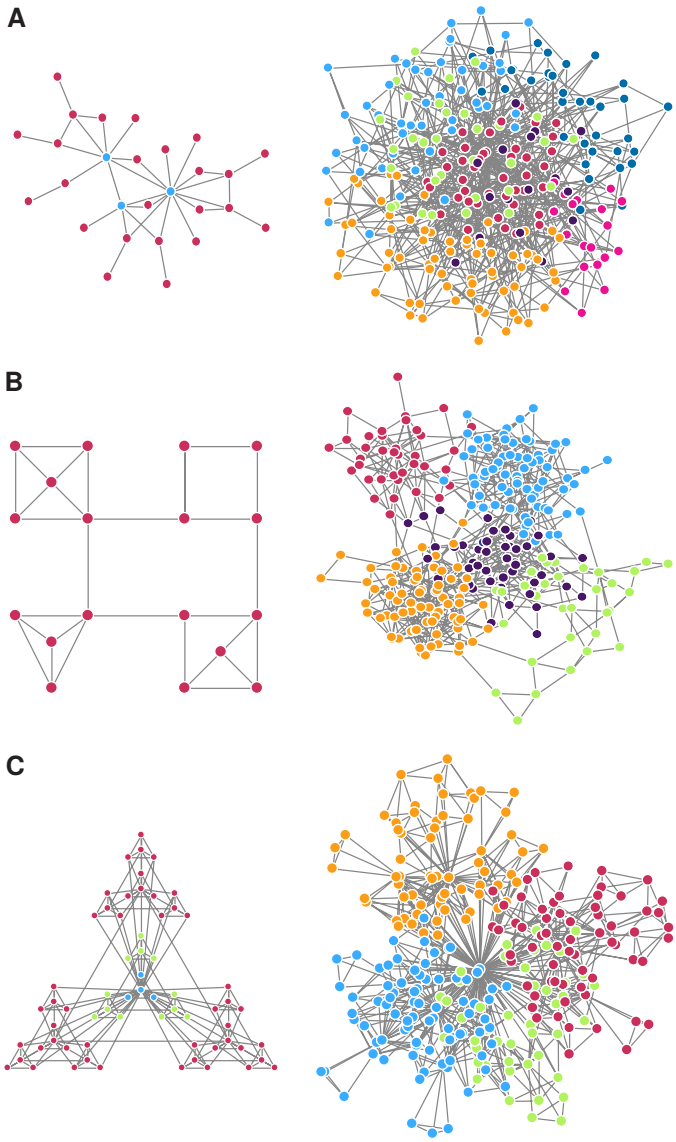

FIGURE 1.12 – Modèles de réseaux complexes, issus et traduits de Ravasz *et al.* [115] et adaptés pour être compatible au daltonisme. **(A)** Illustration schématique (à gauche) d'un réseau à invariance d'échelle, dont la distribution des degrés suit une loi de puissance. Dans un tel réseau, quelques nœuds fortement connectés, ou hubs (cercles bleus), jouent un rôle important pour maintenir la cohésion de l'ensemble du réseau. Une configuration typique (à droite) d'un réseau sans échelle de 256 nœuds est également illustrée, obtenue à l'aide du modèle à invariance d'échelle, qui requiert l'ajout d'un nouveau nœud à chaque fois de sorte que les nœuds existants ayant des degrés de connectivité plus élevés ont plus de chances d'être liés aux nouveaux nœuds [112]. Les nœuds sont disposés dans l'espace à l'aide d'un algorithme de partitionnement standard [116] pour illustrer l'absence de modularité sous-jacente. **(B)** Illustration schématique (à gauche) d'un réseau manifestement modulaire composé de quatre modules fortement interconnectés et reliés entre eux par quelques liens. Cette topologie intuitive n'a pas de distribution de degrés sans échelle, car la plupart de ses nœuds ont un nombre similaire de liens, et les hubs sont absents. Un algorithme de partitionnement standard révèle la modularité inhérente du réseau (à droite) en partitionnant un réseau modulaire de N = 256 nœuds en quatre structures isolées intégrées au système. **(C)** Le réseau hiérarchique (à gauche) présente une topologie d'invariance d'échelle avec une modularité intégrée. Les niveaux hiérarchiques sont représentés par ordre croissant de bleu à vert et rouge. Les algorithmes de partitionnement standard (à droite) ne parviennent pas à mettre en évidence la modularité sous-jacente du réseau. Une caractérisation quantitative détaillée des trois modèles de réseau est disponible dans www.nd.edu/ networks/cell/index.html [4].

biologique de ce type d'organisation est l'objet de nombreuses hypothèses sans qu'aucune n'ait pu être estimée plus plausible qu'une autre [117]. Une forte co-expression locale est présente et on y retrouve des **gènes pivot intra-modules** qui centralisent la majorité des plus courts chemins. Les modules sont également reliés entre eux par un faible nombre de **gènes pivot inter-modules** qui font office de passerelle.

Pour prendre en compte cette topologie en vue de faciliter la détection de gènes, S. Horvath et B. Zhang proposent alors d'ajouter une surcouche à l'adjacence calculée. En se basant sur les travaux de Ravasz *et al.*, ils implémentent le score de recouvrement topologique qui va mesurer l'inter-connectivité relative de deux nœuds (Équation 1.13, Figure 1.13) dans leur calcul du score de similarité. Ils vont cependant faire le choix de passer d'un score de similarité à un score de dissimilarité (Équation 1.14).

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}} \tag{1.13}$$

$$d_{ij}^{\omega} = 1 - \omega_{ij} \tag{1.14}$$

Où :

$$a_{ij} = \text{adjacence}$$
$$l_{ij} = \sum_u a_{iu} a_{uj}$$
$$k_i = \sum_u a_{iu} = \text{connectivité ou force du nœud}$$

Le score de similarité qui en résulte est donc une métrique de la concordance entre les voisins directs de deux nœuds. Par la suite, d'autres améliorations ont également été proposées comme l'utilisation d'un score de recouvrement topologique généralisé [119] ou encore une implémentation différente du recouvrement de topologie, wTO [120], pour pouvoir lui associer une **valeur p**. Chacun de ces scores ayant ses propres avantages et limites, le choix de l'un ou l'autre pour la construction d'un réseau reposera avant tout sur la question de recherche. wTO est ainsi meilleur pour étudier les réseaux dans lesquels il est important de savoir si une interaction est activatrice ou inhibitrice/répressive. Le score de S. Horvath et B. Zhang présente lui une comparabilité accrue avec d'autres méthodes de construction de réseaux de co-expression aux métriques non signées. Une fois le score adapté choisi, il est enfin possible de détecter plus facilement les modules de co-expression.

4. Lien originel inclut dans la publication. Renvoi aujourd'hui une erreur "404 - page inconnue"

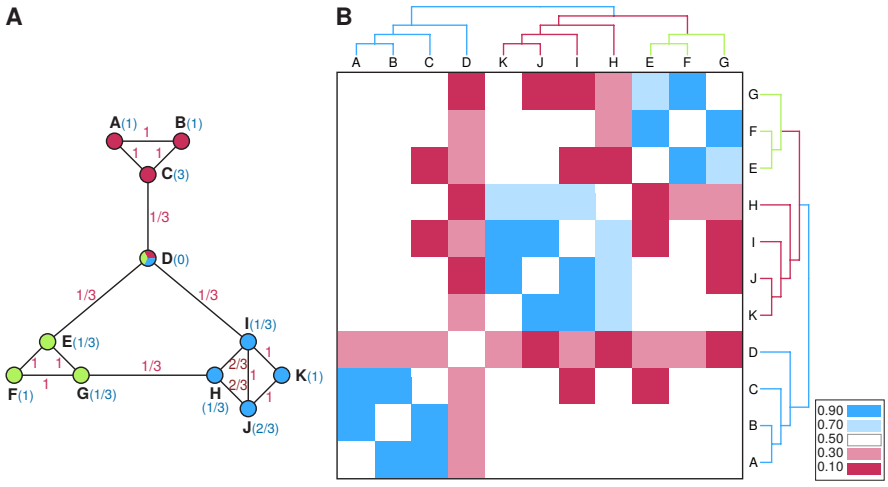

FIGURE 1.13 – Découverte de la modularité sous-jacente d'un réseau complexe, issu de Ravasz *et al.* [115] et adaptés pour être compatible avec le daltonisme. **(A)** Recouvrement topologique illustré sur un petit réseau hypothétique. Pour chaque paire de nœuds, i et j, nous définissons le recouvrement topologique $O_T(i,j) = J_n(i,j)/[min(k_i,k_j)]$, où $J_n(i,j)$ désigne le nombre de nœuds auxquels i et j sont liés (+ 1 s'il existe un lien direct entre i et j) et $[min(k_i,k_j)]$ est le plus petit des degrés k_i et k_j. Sur chaque lien, nous indiquons le recouvrement topologique pour les nœuds connectés, et entre parenthèses à côté de chaque nœud, nous indiquons le coefficient de partitionnement du nœud. **(B)** La matrice de recouvrement topologique correspondant au petit réseau présenté en (A). Les lignes et les colonnes de la matrice ont été réorganisées par l'application d'une méthode de partitionnement par liaison moyenne [118] à ses éléments, ce qui nous permet d'identifier et de placer à proximité les uns des autres les nœuds qui présentent un recouvrement topologique élevé. Le code couleur indique le degré de recouvrement topologique entre les nœuds. L'arbre associé reflète les trois modules distincts construits dans le modèle en (A), ainsi que le fait que les modules EFG et HIJK sont plus proches les uns des autres au sens topologique que le module ABC.

1.3.4 Détection de modules

Point sur les méthodes

Le module est l'unité majeure d'interprétation des GCN. En effet, il regroupe des gènes au profil d'expression similaire au sein de plusieurs échantillons car ceux-ci tendent à être impliqués dans des fonctions physiologiques communes ou des phénotypes définis. La détection des modules vise donc à grouper les gènes dans le réseau en tenant compte des scores de similarité des gènes. Face au défi qu'est l'identification d'unités fonctionnelles dans des réseaux de régulation biologique complexes et chevauchants, de nombreuses méthodes ont été utilisées sont classables en plusieurs catégories (Figure 1.14) [121] :

— Partitionnement : regroupement de gènes sur la base d'un score de similarité globale issu de la matrice d'expression des gènes.

— Décomposition : extraction des composants correspondant aux modules de co-expression en décomposant la matrice d'expression en un produit de matrices plus petites.

— Bi-partitionnement : regroupement simultané de gènes et d'échantillons en classification double sur la base d'un comportement local similaire en matière d'expression.

— Inférence de réseau itérative : optimisation itérative d'un réseau inféré et d'un ensemble de partitions.

— Inférence de réseau directe : inférence d'un réseau de régulation basé sur la similarité de l'expression génétique entre les régulateurs et les gènes cibles.

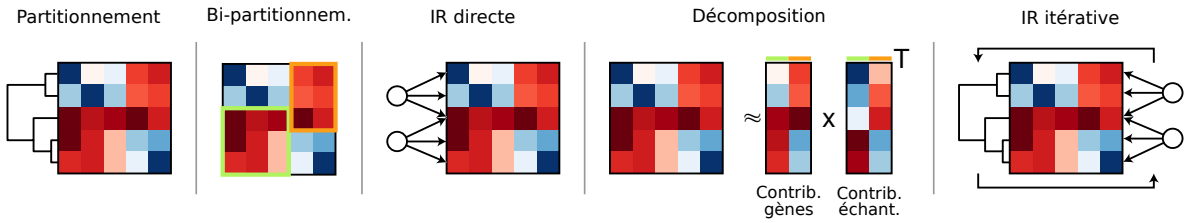

FIGURE 1.14 – Illustration des 5 catégories de méthodes de détection de modules. Extrait de Saelens *et al.* 2018 (CC-BY) [121] et adapté pour être compatible au daltonisme. IR = Inférence de réseau.

La détection de modules servant différents objectifs, certaines approches sont plus adaptées que d'autres dépendant de la question de recherche. Dans la majorité des cas, la détection de modules peut permettre d'avoir un aperçu global des données sans apport de connaissances extérieures comme des voies d'activations, d'identifier des fonctions biologiques ou des maladies et d'inférer des réseaux de régulation. La méthode par partitionnement permettant de répondre à tous ces besoins [122, 123, 124] et étant encore à ce jour la plus utilisée [121], c'est sur elle qu'on s'est concentré dans cette thèse.

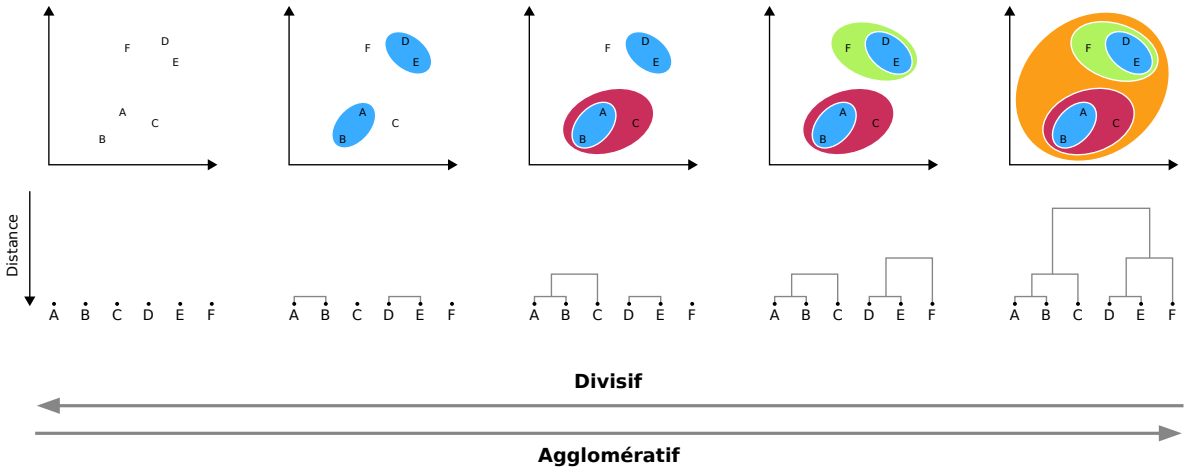

FIGURE 1.15 – Principe du partitionnement hiérarchique. Chacun des 6 points du nuage est associé à chaque étape avec le point le plus proche dans un groupe. L'arbre est ensuite coupé traditionnellement pour obtenir les groupes, dans le cas de la co-expression, les modules.

Partitionnement hiérarchique

Les méthodes de partitionnement sont variées mais seules trois semblent être utilisées en majorité dans la recherche en réseaux de co-expression : le partitionnement par connectivité, le partitionnement par centroïdes, et le partitionnement par distribution. Si ces deux dernières méthodes ont fait leurs preuves pour détecter des modules d'intérêt [125, 126, 127], le partitionnement par connectivité, aussi appelé partitionnement hiérarchique (Figure 1.15) reste une procédure largement utilisée [128, 129, 130]. Il est apprécié pour son compromis entre facilité d'implémentation, pertinence des modules détectés, et besoins en ressources tant en temps qu'en puissance de calcul [121]. Il est également particulièrement adapté aux réseaux biologiques qui suivent une topologie modulaire hiérarchique comme ceux de gènes.

La méthode initiale [131] consiste à construire un arbre de distances (un dendrogramme) en calculant un critère de liaison entre chaque point. Chaque point et ensuite soit regroupé itérativement selon les points les plus proches (processus agglomératif), soit divisé en deux groupes à chaque tour les points (processus divisif). L'arbre final est coupé selon un seuil qui va définir les points regroupés ensemble, dans notre cas les gènes regroupés dans un même module. Ce seuil ayant été plusieurs fois critiqué pour son côté trop statique dans la hauteur de découpe et son côté arbitraire, une amélioration consiste à rendre le seuil dynamique [132]. Pour cela, les branches du dendrogramme sont analysées à la recherche de structures de sous partitions récursivement jusqu'à une stabilisation des sous clusters. Cette approche permet donc une définition de clusters de hauteur différente dans le dendrogramme pour chaque branche.

Ces méthodes de construction et de détection des modules sont finalement celles implémentées dans le populaire progiciel (en anglais *package*) R WGCNA créé par P. Langfelder et S. Hortvath [133]. Développé en 2008, ce progiciel a permis de démocratiser l'analyse par réseaux de co-expression et de donner des modules prêts à être interprétés. Pour leur donner tout leur sens, il est cependant nécessaire d'effectuer des analyses supplémentaires.

1.3.5 Exploitation des modules de gènes

Approche guidée par la connaissance préalable : l'intégration biologique

Les modules tendent à regrouper des gènes aux fonctions physiologiques semblables ou co-régulés [96, 117]. Une façon de comprendre la raison du regroupement de ces gènes est donc de leur rattacher une fonction, une voie d'activation, un tissu ou une maladie dans laquelle ils interviennent. Cette approche est alors qualifiée de guidée par la connaissance préalable (en anglais *knowledge-driven*). Cependant, les gènes œuvrant rarement à une seule fonction, ils se retrouvent annoté de plusieurs fonctions [134] plus ou moins pertinentes dans le contexte du module. Pour estimer lesquelles sont réellement significatives dans chaque module, on effectue donc

des analyses d'enrichissement [135] par rapport à diverses sources d'annotations (Table 1.1). Appelée également par raccourci analyse de voies d'activation (en anglais *pathway analysis*), cette analyse peut être effectuée de trois façons dépendamment des méta-données disponibles : l'analyse de sur ou sous représentation, la cotation de classe fonctionnelle, et l'approche basée sur la topologie des voies d'activation.

Source de données	Acronyme	Description	Compatible PT ?
Gene Ontology	GO	Modèle informatique de systèmes biologiques sous forme d'ontologie décrivant des fonctions physiologiques, moléculaires et cellulaire de différents organismes	oui
Kyoto Encyclopedia of Genes and Genomes	KEGG	Ensemble de bases de données relatives aux génomes, aux voies métaboliques et aux composés biochimiques, associées à une représentation en réseau des intéractions moléculaires	oui
Reactome	Reactome	Processus biologiques humains annotés sous la forme d'une série d'événements moléculaires imbriqués	oui
WikiPathways	WP	Ressource communautaire pour la contribution et la maintenance de contenu dédié aux voies d'activation métaboliques	oui
MicroRNA-Target Interactions database	mirTarBase	Base de données revue des interactions micro ARN et séquence cible	non
TRANSFAC	TRANSFAC	Ressource de facteurs de transcriptions, sites de liaisons et éléments régulés	non
Comprehensive Resource of Mammalian protein complexes	CORUM	Collection de complexes protéiques chez les mammifères expérimentalement vérifiés	non
Human Protein Atlas	HPA	Cartographie de l'ensemble des protéines humaines par type cellulaire, tissu, organe	non
Human Phenotype Ontology	HPO	Modèle informatique de systèmes biologiques sous forme d'ontologie décrivant des phénotypes humains anormaux et/ou associés à des maladies	oui
Molecular Signatures Database	MSigDB	Ensemble de 9 collections d'annotations sur la localisation chromosomique, les voies métaboliques, la régulation, les signatures oncologique, l'immunologie, le type cellulaire, et différentes ontologies	non

TABLE 1.1 – Sources d'enrichissement communes en analyse d'enrichissement. PT = approche basée sur la topologie des voies d'activation

En l'absence d'autres données que les identifiants de gènes, on peut effectuer une analyse de sur ou sous représentation (en anglais *over-representation analysis* ou ORA) par rapport à un contexte et une source d'annotation donnée [135]. Par exemple, on pourra chercher si un ensemble d'identifiants de gènes issus d'un module est sur-représenté par rapport à la normale dans groupe de gènes annotant une voie métabolique sur l'ensemble des gènes chez l'humain (Figure 1.16). Pour estimer si la sur-représentation est significative, on utilise le plus souvent un test hyper-géométrique bien que d'autres distributions comme le chi^2 ou la binomiale soient également utilisés plus occasionnellement [136]. Le test étant répété pour de nombreuses annotations, une correction pour test multiple doit être réalisée. L'inconvénient majeur de l'enrichissement par sur-représentation réside dans sa pondération uniforme de l'impact de chaque gène, ce qui est loin de représenter la réalité biologique qui est hiérarchique [135].

Si en plus des identifiants de gènes on possède un score, par exemple un poids, un taux de

variation, un rang, une valeur p, on peut effectuer une cotation de classe fonctionnelle (en anglais *functional class scoring* ou FSC), aussi appelée de façon confondante analyse d'enrichissement de collection de gènes (en anglais *gene set enrichment analysis* ou GSEA). Le score à considérer pour les gènes va impacter fortement les enrichissements obtenus [137] et doit donc être choisi avec précaution. Pour obtenir l'enrichissement, la cotation de classe fonctionnelle va tout d'abord calculer une statistique au niveau des gènes (ex : valeur p, un rang, un taux de fausses découvertes), si le score fournit n'est pas déjà de cette nature. Ces statistiques de niveau de gènes sont ensuite agglomérées par source d'annotation donnée selon une méthode de résumé de l'information comme une médiane, une statistique de Kolmogorov-Smirnov, ou une somme de rang de Wilcoxon [135]. Finalement, cette statistique agglomérée est testée pour sa significativité par test de permutation ou de comparaison au hasard (Figure 1.16). Si cette méthode répond à la limitation des tests de sur-représentation évoqués plus haut, elle possède elle-même une limite majeure car elle considère les annotations données de façon indépendante. Les annotations se chevauchant régulièrement du fait de l'implication récurrente d'un gène dans plusieurs annotations, un biais est nécessairement présent.

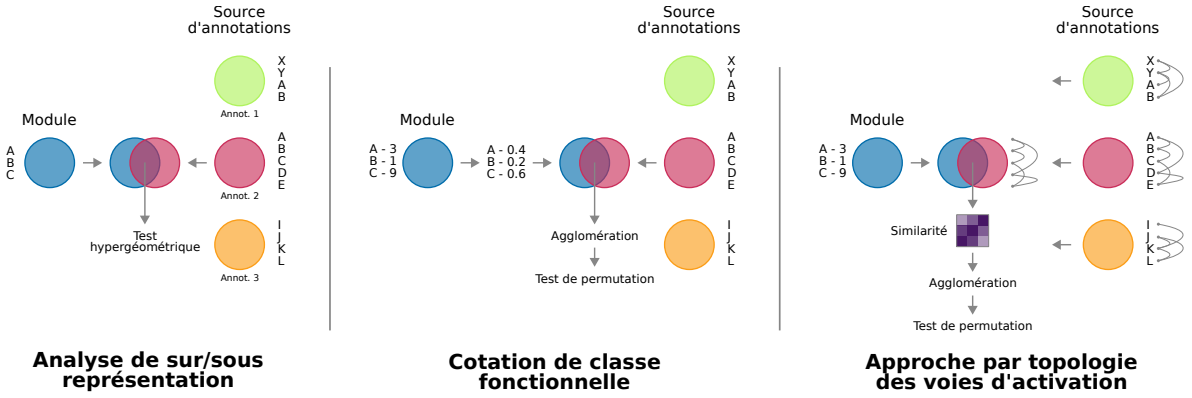

FIGURE 1.16 – Schématisation des différentes méthodes d'enrichissement de groupes de gènes. Le module composé des gènes A,B,C est dans chacun des 3 exemples enrichit selon une source d'annotation contenant trois annotations au nombre de gènes variables. Dans le cas de l'analyse de sur ou sous représentation, seuls les identifiants sont fournis coté module. Dans la cotation fonctionnelle de classe et l'approche par topologie de voie d'activation, des scores sont en plus associés aux gènes du module. Dans le cas de l'approche par topologie de voie d'activation, des informations sur les liens entre les différents gènes des annotations est en plus fournit.

La méthode d'enrichissement la plus récente est l'approche basée sur la topologie des voies d'activation (en anglais *pathway topology based approache* ou PT). Cette méthode prend en compte non pas juste une liste d'identifiants de gènes et leur score associé, mais les liens existants entre les gènes dans les sources d'annotation [135] (Table 1.1). Ainsi, si une voie métabolique se voit changer l'agencement des liens entre les différents gènes, sans pour autant changer la liste des gènes, l'approche par topologie des voies d'activation sera capable d'en tenir compte dans son résultat d'enrichissement contrairement aux deux méthodes précédentes. Son fonctionnement est similaire à celui de la cotation par classe fonctionnelle qui, à la différence que la statistique de niveau de gène telle que ScorePAGE [138] calcule un niveau de similarité entre chaque paire de gènes de l'annotation donnée et le divise par le nombre de réactions nécessaires pour connec-

ter les deux gènes (Figure 1.16). L'apport d'information supplémentaire permet donc de préciser encore plus les enrichissements obtenus. Toutefois, cette méthode requière des sources d'annotation possédant des liens entre les gènes donnés, et celles-ci ne représentent qu'un faible nombre de sources.

En plus des tests d'enrichissement qui apportent une information généraliste par le biais des sources d'annotations, on peut apporter une information plus spécifique si des phénotypes sont fournis pour les échantillons via les tests d'association phénotypique [133]. Afin de tester chaque module avec chaque phénotype, on cherche tout d'abord à représenter chaque module sous une forme moyenne. Pour ce faire, on effectue une décomposition en valeurs singulières de l'expression des gènes contenus dans un module, et on sélectionne la première composante calculée qu'on nommera gène propre (en anglais *eigengene*). Dans le cas de variables quantitatives comme l'âge, la taille, le poids, il est directement possible d'effectuer un test de corrélation avec le gène propre d'un module pour déterminer si celui-ci est significativement associé à la variable phénotypique considérée. Dans le cas d'une variable qualitative, il est nécessaire de la binariser pour pouvoir effectuer l'association avec le gène propre [139]. Cette technique consiste transformer chaque modalité de la variable en une variable à part entière dont la valeur sera 1 lorsque l'échantillon possède cette modalité, et 0 pour le reste de la variable ainsi que les autres variables modales (Figure 1.17).

Échantillon	Âge	Poids	Sexe
A	67	82	Masculin
B	42	75	Masculin
C	19	56	Féminin
D	29	52	Intersexe
E	56	64	Féminin

Échantillon	Âge	Poids	Masculin	Féminin	Intersexe
A	67	82	1	0	0
B	42	75	1	0	0
C	19	56	0	1	0
D	29	52	0	0	1
E	56	64	0	1	0

FIGURE 1.17 – Exemple de binarisation de variable catégorielle. La variable du sexe comportant 3 modalités, homme, femme, et intersexe est transformée en 3 variables distinctes qui codent pour les catégories initiales par des valeurs binaires.

Ces différentes approches guidées par la connaissance permettent finalement d'effectuer des prédictions fonctionnelles *in silico* quand associées à l'analyse de GCN. Le paradigme de l'association par culpabilité (en anglais *guilt by association*) présume en effet que des gènes fortement co-exprimés, comme ceux présents dans les modules, contribuent ou partagent de mêmes fonctions physiologiques[103]. Cette association est donc particulièrement utile pour l'annotation de gènes qui sont peu ou pas étudiés en se basant sur les enrichissements et associations phénotypiques obtenus [140]. Il faut toutefois noter que si l'ajout de connaissances établies permet d'acquérir de nouvelles connaissances et de comprendre le contexte biologique dans lequel se situent des modules, il apporte aussi avec lui un risque de biais. Les sources d'enrichissement sont par exemple disproportionnées concernant certains gènes ou familles de gènes [141]. L'utilisation conjointe de méthodes dites guidées par les données (en anglais *data-driven*) est donc un moyen de compléter de façon moins biaisée l'interprétation des modules.

Approche guidée par les données : l'étude topologique

Les motifs d'organisation et les structures présentes dans un réseau sont depuis longtemps étudiés dans une volonté de rattacher ensemble topologie et fonction biologique. Les gènes concentrant les liens dans un réseau les gènes pivot, étant des structures facilement notables, elles ont été étudiées très tôt. Les gènes pivot ont ainsi été trouvés associés à des gènes essentiels à la survie d'un organisme [142]. Une analyse de ces gènes pivot croisée aux gènes responsables de maladies Mendéliennes a cependant écarté l'hypothèse intuitive que les gènes pivot étaient la clef de ces maladies mono-géniques[143]. À la place, les gènes associés à ces maladies et aux maladies non-Mendéliennes ont plutôt été retrouvés en périphérie ou à des positions neutres du réseau [142]. La détection des gènes pivot reste donc intéressante pour favoriser la détection de gènes aux fonctions essentielles ou bien celle de gènes potentiellement liés à une maladie, dans le cas de l'étude de gènes voisins [144].

La méthode de détection des gènes pivot a toutefois souffert d'une définition assez vague de ce que sont les gènes pivot. On retrouve ainsi des définitions telles que sélectionner les dix gènes les plus connectés [145], un pourcentage de gènes les plus connectés et une façon de mesurer cette connectivité qui varie selon l'étude [123, 128] ou encore un métrique de théorie des réseaux seuillée arbitrairement [146]. Deux mesures parmi celles existantes ressortent pour leur définition de gènes pivots basé uniquement sur la topologie et un test statistique, sans seuil extérieur :

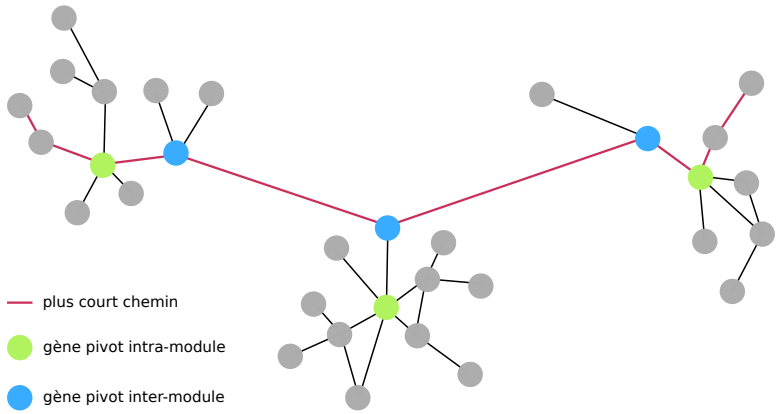

FIGURE 1.18 – Réseau illustrant la topologie présentée par les gènes pivot intra et inter-modules. Les gènes pivot intra-module (en vert) sont centraux au sein des modules. Les gènes pivot inter-modules (en bleu) sont le lieu de passage d'une grande quantité de plus courts chemins (en rouge).

— La sur-connectivité [147] : la connectivité, c'est-à-dire la somme des poids des liens d'un nœud, et testée statistiquement par rapport à la connectivité moyenne du réseau pour estimer si la différence est significative. Le degré est aussi parfois utilité comme mesure initiale.

— L'adhésion au module [148] : la corrélation entre l'expression d'un gène et le gène propre d'un module est testée pour estimer sa significativité.

Plus précisément encore, ce type de mesure va permettre de détecter des gènes pivots dit intra-module (Figure 1.18). D'autres gènes possèdent une structure de pivot mais se placent plutôt en passerelle entre les modules ou sous-modules du réseau lié à la topologie hiérarchique. Ceux-ci sont détectables par une analyse des nœuds les plus présents dans les plus courts chemins pour parcourir le réseau d'un nœud à un autre (Figure 1.18). Leur implication et caractérisation biologique reste toutefois encore peu étudiée.

Comparaison de réseaux entre conditions : la co-expression différentielle

Les comparaisons sain contre malade, sauvage contre muté, ou toute autres conditions contrôle contre test ont toujours été une approche de choix pour comprendre les altérations d'une condition en biologie. Les GCN n'échappent pas à la règle et des méthodes de comparaison de réseau sont rapidement apparues [151]. L'analyse de co-expression différentielle (en anglais *differential co-expression* ou DCE) est complémentaire de l'analyse classique d'expression différentielle en détectant les gènes dont les motifs de régulation changent sans que l'expression des gènes varie elle-même significativement. Certains gènes associés à des maladies ont en effet un niveau d'expression constant mais avec une modification du transcrit telle qu'une mutation, un épissage alternatif, ou une altération post-transcriptionnelle qui vont venir affecter sa fonction [93]. Ce phénomène est observable dans le cas de perturbation d'éléments de régulation [130], dans le cas

Name	Description
Density	It measures compactness of a module. Density is defined as the average connection strength of all pairs of genes present in a module.
Clustering coefficient	Clustering co-efficient of a node quantifies the interconnectedness of its neighboring nodes. The average of clustering coefficients of all nodes measures the modularity of a module.
Maximum adjacency ratio	It generally deals with a weighted network and can be used to check whether a hub gene is connected to a large number of genes with moderate connection strengths or a few genes with strong connection strengths [148]. By computing the average MAR, we can distinguish among the connectivity patterns. More the correlation between MAR of nodes in two modules across conditions, then the preservation is higher.
Sign preserving mean correlation	This module preservation strategy takes element-wise multiplication of the correlation matrices for the same module in two different samples groups and then takes the mean. Higher mean means more preservation.
Eigengene based measure	Eigengene is an arbitrary gene represents a module very well. To calculate preservation, one can take the correlation between the eigengenes of the corresponding modules of different phenotypes. Higher correlation means more preservation.
Intra-modular connectivity	The sum of connection strengths of neighbors of a node. To investigate preservation, correlation of intra-modular connectivity of nodes present in the same module of different phenotypes can be used. More correlation implies higher preservation.
Fitting coefficient (R2) [95]	Scale-free topology criteria of a network can be quantified using the fitting coefficient of linear regression.
Global network connectivity [149]	This characteristic path length is the average of the shortest paths between all pairs of nodes in a network.
Zsummary [150]	It quantifies the amount of preservation of a module of reference in a test network by considering density and connectivity of that module in both networks.
MedianRank [150]	It gives the ranking of a reference network modules in a test network. The most preserved module gets the top rank.

TABLE 1.2 – Statistics for Module Preservation and Disruption Measurement. Statistiques pour la préservation des modules et la mesure des perturbations. Issu de Chowdhury *et al.* [104] ©2019 IEE. Traduction non autorisée.

de comparaisons de tissus [152], ou entre espèces similaires voir organismes [104].

Réaliser une analyse DCE demande cependant un regard critique sur les données utilisées en entrée tout comme sur la méthode de co-expression différentielle employée. Des différences individuelles de liens ou du poids dans le réseau ne sont ainsi pas suffisantes à elles seules pour indiquer une divergence de fonction et pourraient être le fruit de biais techniques, par exemple de la méthode de séquençage [153]. De la même manière que pour l'expression différentielle classique, il faut s'assurer de la significativité des changements de motif observés. Face à la diversité d'observations possibles, plusieurs méthodes ont vue le jour avec l'influence tant de la théorie des réseaux que de la biostatistique (Table 1.2). Initialement partis d'une simple comparaison de l'expression moyenne par nœuds ou jusqu'à un nombre défini de nœuds voisins, les méthodes de co-expression différentielle ont évolué pour s'orienter sur de la détection de modifications de motifs complexes dans un réseau [93].

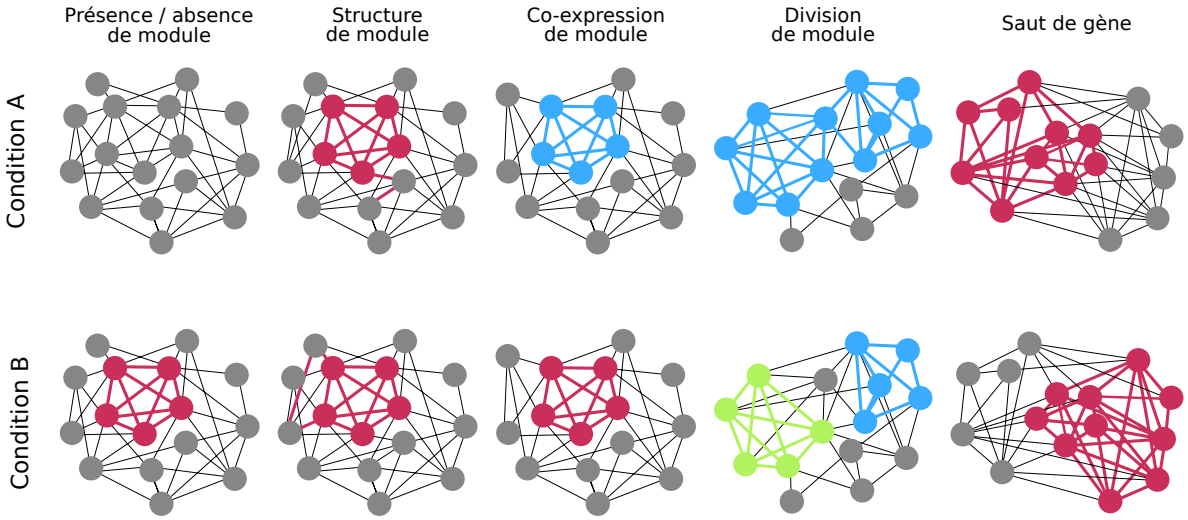

FIGURE 1.19 – Motifs de changements de co-expression entrainant une non-préservation des modules dans une analyse de co-expression différentielle. D'après Van Dam *et al.* [114] et adaptés pour être compatible au daltonisme. Les motifs peuvent consister en une apparition ou disparition de module en raison des changement d'intensité de corrélation, une modification partielle des gènes inclus dans le module, une variation globale de l'intensité de co-expression sans changer la conformation du module comme une diminution de tous les liens, une séparation en deux d'un module initial ou une jointure de deux modules initiaux, un saut de la co-expression d'un sous groupe de gène vers un autre groupe en raison notamment de la présence de facteurs de transcription.

Les variations singulières de gènes en DCE sont généralement le reflet de maladies Mendéliennes ou de traits phénotypiques monogéniques. Pour investiguer des conditions polygéniques, on utilise plus souvent l'échelle du module pour investiguer les différences de co-expression entre deux conditions. On parle alors d'analyse de préservation ou non préservation des modules de co-expression entre des conditions. De façon similaire à l'interprétation des modules qui peut se faire via des détections de voies d'activation ou via de la topologie, la co-expression différentielle peut reconnaître les modules préservés ou non préservés en étudiant la conservation de voies d'activation ou de motifs de liens au sein du module [104]. Face à la diversité de changement de topologie, l'analyse par modification des liens requière toutefois plus de détail qu'une analyse

présence/absence des comparaisons d'enrichissement. On peut ainsi observer des disparitions totales de modules, des variations de leur composition en gène, des séparations d'un module en plusieurs ou inversement, des changements de co-expression par grappe (Figure 1.19).

En 2005 lors de la publication de leur méthode d'analyse par co-expression, B. Zhang et S. Horvath présentent également sept métriques de caractérisation de la topologie d'un réseau (Annexe A.1) [95]. En évaluant sur plusieurs aspects du réseau, elles permettent d'avoir une vision globale des perturbations que peut subir l'organisation d'un réseau lorsqu'il est étudié dans d'autres conditions. Ils s'en serviront d'ailleurs pour développer une méthode d'estimation de préservation des modules entre deux conditions. Cependant, face au large nombre de modules testés par analyses et en l'absence de correction de test multiple, de nombreux faux positifs ont pu être relevés [154]. Une méthode développée par S. Ritchie en 2016 viendra toutefois contrer ce défaut dans la méthode de B. Zhang et S. Horvath en intégrant la mesure des sept métriques au sein d'un test de permutation des identifiants de nœuds au sein du réseau (Figure 1.20).

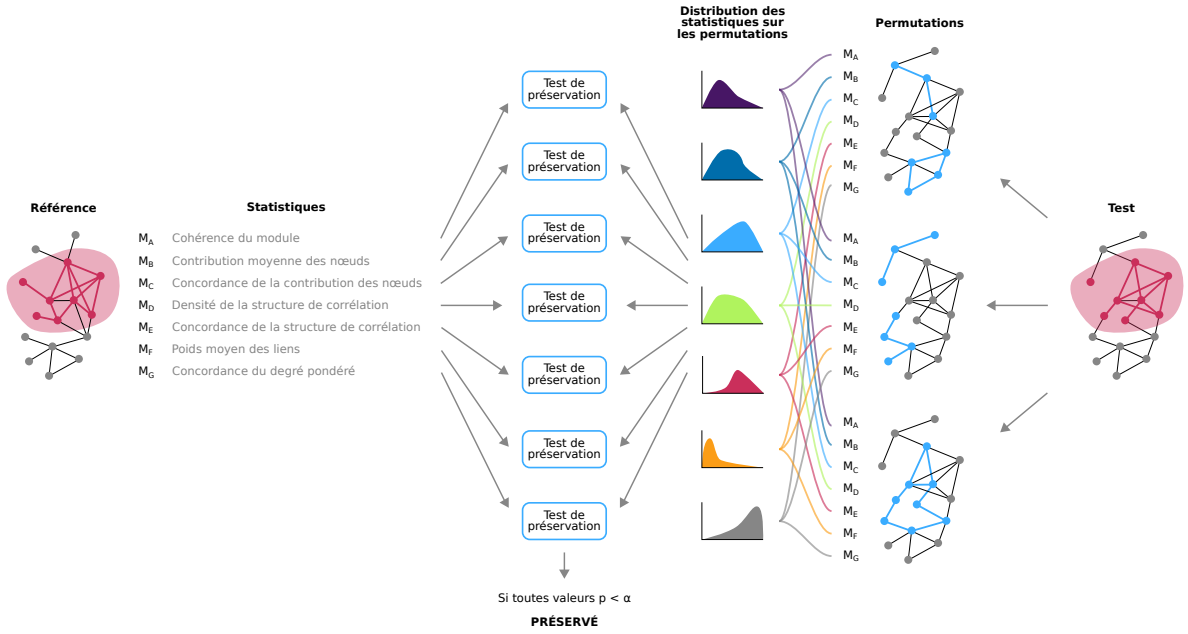

FIGURE 1.20 – Fonctionnement du test de permutation implémenté par S. Ritchie [154]. Les réseaux construits dans chaque condition sont assignés d'un rôle : référence pour le réseau qui servira de contrôle, et test pour celui dans lequel sera vérifié la préservation du module. Le réseau de test est permuté plusieurs fois (minimum 1000 fois) et les sept statistiques sont calculées dans chacun pour finalement obtenir la distribution nulle du réseau. Un test de préservation est ensuite réalisé à l'aide des 7 statistiques calculées dans le réseau de référence et des distributions nulles. Si la totalité des sept valeurs p est significative, alors le module est dit préservé.

1.4 Le vieillissement, une imbrication complexe de dérèglements

Source multi-factorielle de changements dans l'organisme, le vieillissement est, chez l'humain, un continuum de dérèglements conduisant une personne en bonne santé à une réduction de sa

qualité de vie en raison d'une dégradation progressive des fonctions de base du corps [155]. Ses différents aspects ne forment pas un chemin linéaire mais plutôt une arborescence où chaque individu va tendre à accumuler plus ou moins rapidement certaines marques du vieillissement (Figure 1.21). Si certaines de ces marques n'impactent que légèrement la dégradation de la personne vieillissante ou âgée, des événements traumatiques ou pathologiques tels qu'un accident vasculaire cérébral peuvent, par un effet de cascade, entraîner une accélération brutale dans la vitesse de dégradation. En augmentant la longévité des personnes, les avancées en santé publique et qualité de vie ont alors retardé la dégradation fonctionnelle. Paradoxalement, elles ont cependant augmenté l'âge maximal et donc l'accumulation des marques du vieillissement, rendant cette population plus susceptible à des maladies chroniques telles que le cancer ou les maladies cardiovasculaires [156].

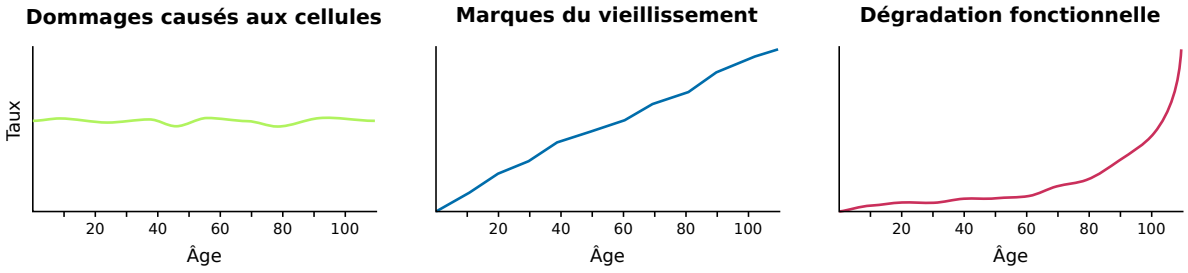

FIGURE 1.21 – Lois de distribution régissant le vieillissement sous un modèle simplifié : l'exposition aux facteurs de vieillissement, l'accumulation cellulaire de dommages, l'augmentation exponentielle de la dégradation fonctionnelle et du risque de mortalité. Cette augmentation s'explique par le niveau constant de dommages causé aux cellules couplé à la réduction progressive des capacités de réparation qui est elle-même affectée par les dommages dans une boucle délétère [157].

On distingue trois types de vieillissements : le vieillissement réussi ou en bonne santé, le vieillissement usuel, et le vieillissement pathologique [158]. Le vieillissement réussi est décrit comme associé à l'absence de facteurs de risque pour la santé et avec une perte minime de condition physiologique : pas d'anomalies sanguines, pas de problèmes à l'effort, seulement un ralentissement de l'activité en raison d'une fatigue plus rapide [155]. Le vieillissement usuel est défini par une diminution des capacités fonctionnelles globales telles que visible en moyenne chez des personnes du même âge. Il implique les différents changements de régulation cellulaire et moléculaire qui vont se manifester à l'échelle de la personne par des déséquilibres métaboliques, des facteurs de risques augmentés, et des pathologies chroniques sans décompensation c'est-à-dire sans dégradation brutale suite à une rupture des mécanismes de compensation d'un dérèglement [155]. Enfin, le vieillissement pathologique est un terme ne faisant pas l'unanimité car étant un processus physiologique intrinsèque, il est discutable de le considérer comme une maladie. De ce fait, il est le plus souvent décrit comme recoupant deux notions : le vieillissement prématuré et le vieillissement en continuité avec le grand âge [159]. Tous ces vieillissements entretiennent évidemment des liens étroits d'un point de vue physiologique, mais on a souhaité dans cette thèse se concentrer sur le vieillissement usuel.

1.4.1 Définition moléculaire

Le vieillissement usuel, qu'on nommera simplement par la suite vieillissement, est inévitable et ne reflète pas nécessairement l'âge chronologique d'un individu mais plutôt un âge biologique [160]. C'est un phénomène influencé par la génétique d'un individu [156, 161] ainsi que son environnement [162]. Des gènes associés à la longévité ont ainsi pu être identifiés mais sont loin d'expliquer la totalité du vieillissement au vu de sa très grande complexité moléculaire. Pour aider à la communication sur le sujet et pour tenter de catégoriser les différentes composantes du vieillissement, López-Otín *et al.* proposent en 2013 un découpage en neuf **marques principales** du vieillissement (*hallmarks of aging* en anglais) qui sont représentatives des causes, conséquences et résultats du vieillissement (Figure 1.22) [163] :

— **Instabilité génomique** : l'intégrité de l'ADN est régulièrement mise à l'épreuve lors de la vie cellulaire, qu'il s'agisse de contraintes physiques, de conditions chimiques ou d'agents biologiques [164]. Ils entraînent des mutations ponctuelles, des translocations, des duplications et des augmentation ou réduction de taille de chromosome, notamment au niveau des télomères. Pour conserver sa stabilité, des mécanismes de stabilité et réparation de l'ADN sont présents dans la cellule mais tendent à dysfonctionner avec l'âge, permettant aux dommages de perdurer et se transmettre [165].

— **Attrition des télomères** : les télomères, extrémités des chromosomes, jouent un rôle de protection vis-à-vis de ceux-ci. Lors de la réplication cellulaire et en l'absence de télomérase dans la majorité des cellules, ils tendent toutefois à réduire en taille à chaque fois [163]. Leur taille initiale limite donc le nombre maximal de réplication qu'une cellule pourra avoir, nombre qui est appelé la limite de Hayflick [166]. Les télomères jouent donc un rôle dans la longévité globale mais peuvent être affectés par l'instabilité génomique ou des pathologies ponctuelles telles que des infections [167].

— **Altérations épigénétiques** : l'épigénétique désigne tout mécanisme de modification du génome n'incluant pas d'altération de sa séquence. Lors du vieillissement, plusieurs de ces mécanismes, comme la modification des histones chargés du repliement de la chromatine, vont entraîner des modifications de la régulation de la transcription [168]. La perturbation de certaines voies comme la voie métabolique de l'insuline vont alors jouer sur la longévité des individus [169].

— **Perte de protéostasie** : la protéostasie regroupe l'intégralité des mécanismes visant à permettre aux protéines d'acquérir et conserver leur repliement prévu, ainsi que les mécanismes de lyse des protéines pour éviter leur malfonctionnement en cas de dégâts [170]. Les liens précis entre les marques principales précédentes et la perte de protéostasie restent encore à comprendre malgré une association avérée de l'augmentation de mauvais repliement et de perte de proteolyse avec le vieillissement [170].

— **Dérégulation de la perception des nutriments** : la perception des fluctuations des niveaux de nutriments dans l'organisme telles que les sucres, acides aminés, et lipides est

essentiel pour son fonctionnement et requière différentes voies d'activation spécifiques à chacun [171]. Face à la déficience de régulation de nutriments chez les personnes âgées, il a été observé que plusieurs voies tendent à être perturbées, par exemple mTOR dans la perception des hautes concentrations en acides aminés [172].

— **Dysfonction mitochondriale** : les mitochondries, organelles chargées entre autre de la respiration cellulaire, diminuent en activité avec le vieillissement ce qui cause un ralentissement de croissance [173]. La production d'adénosine triphosphate (ATP) nécessaire aux nombreuses réactions cellulaires décroit donc et avec elle le métabolisme global de la cellule. Des dérivés réactifs de l'oxygène (en anglais *reactive oxygen species* ou ROS) sont également mesurés en plus grande quantité en conséquence, bien que leur rôle soit ambigu. Il a été observé que leur présence tend dans certains contextes à stimuler des réponses de compensations du vieillissement qui augmentent la longévité, et dans d'autres contextes à diminuer la longévité en raison du stress cellulaire induit [174].

— **Sénescence cellulaire** : la duplication cellulaire fait partie de la vie normale d'une cellule et se retrouve arrêté de façon stable dans un processus qu'on nomme sénescence. Plusieurs mécanismes peuvent mener la cellule à ce stade tel que le raccourcissement des télomères précédemment cité. D'autres formes de sénescence existent toutefois sous la forme de sécurités mises en place par la cellule pour éviter la multiplication de cellules atteintes par des dommages conséquents à l'ADN [156]. Ces cellules sénescentes sont également connues pour contribuer à une inflammation chronique de faible intensité (en anglais *inflammaging*) [175, 176] et leur signalisation pro-sénescence à travers les différents tissus [156].

— **Épuisement des cellules souche** : la réduction des capacités de différenciation et régé-

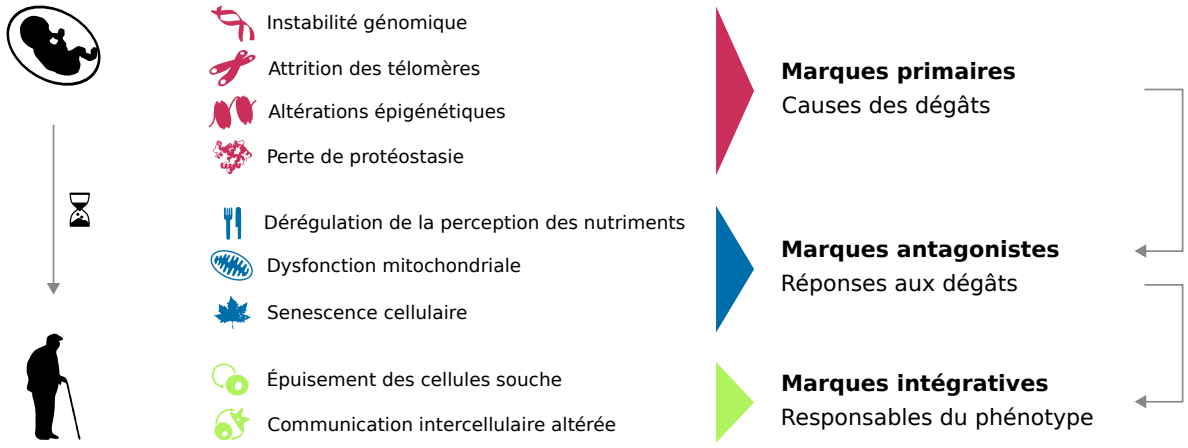

FIGURE 1.22 – Marques principales du vieillissement et interconnections fonctionnelles. Adapté et traduit d'après López-Otín 2013 [163][5]. Les neuf marques principales proposées du vieillissement sont regroupées en trois catégories. (En haut) Les signes distinctifs considérés comme les causes primaires des dommages cellulaires. (Au milieu) Ceux qui sont considérés comme faisant partie des réponses compensatoires ou antagonistes aux dommages. Dans un premier temps, ces réponses atténuent les dommages, mais à terme, si elles sont chroniques ou exacerbées, elles deviennent elles-mêmes délétères. (En bas) Signes distinctifs intégratifs qui sont le résultat final des deux groupes de signes distinctifs précédents et sont finalement responsables du déclin fonctionnel associé au vieillissement.

nération des cellules souches sont la conséquence des altérations des dommages à l'ADN et des mécanismes de sénescence en réponse pour limiter leur prolifération [177]. En atteignant par exemple les niches de cellules souches telles que celle des cellules hématopoïétiques, les dommages à l'ADN et à la régulation vont causer une diminution des capacités immunitaires (aussi appelée immuno-sénescence) ainsi qu'une anémie [163, 178].

— **Communication intercellulaire altérée** : les dérégulations de l'expression des gènes couplés à des stimulations de faible intensité du système immunitaire entraînent un brouillage de la communication entre cellules par diverses sécrétions dans la MEC [163]. Le système endocrinien comme paracrinien vont voir leurs capacités de coordination du métabolisme entre cellules réduites, ce qui va à son tour mener à des dommages dans l'ADN des cellules [156]. Ce phénomène peut s'accentuer localement au point d'engendrer un vieillissement accéléré d'un tissu qui peut promouvoir la sénescence dans d'autres tissus au point qu'on le nomme vieillissement métastatique [179].

Comme visible dans ces définitions ainsi que dans la Figure 1.22, plusieurs de ces marques principales du vieillissement tendent à s'entretenir les unes les autres dans un cercle vicieux s'accentuant à chaque cycle. Des mécanismes de compensations visent toutefois à contrer ce cycle en stoppant les cellules atteintes par la senescence, ou en re-stimulant leur régénération comme cela a pu être vu avec les ROS. Ces mécanismes sont d'ailleurs présents avant l'apparition du vieillissement qui n'est finalement qu'une illustration de la mise en échec progressive de la robustesse des mécanismes de compensation [180].

1.4.2 Enjeux

Si la quête pour l'immortalité a su inspirer nombre de personnes et de récits, elle n'est pas souhaitable pour plusieurs raisons tant éthiques qu'économiques [181]. Il n'est donc pas question de guérir ou empêcher le vieillissement dans le cadre de la recherche sur celui-ci mais de le comprendre. L'âge entraîne la multiplication des facteurs de risque et est associées à des maladies lui étant quasi spécifiques comme la maladie d'Alzheimer, l'athérosclérose, le cancer, l'insuffisance rénale chronique, l'obstruction pulmonaire chronique, l'insuffisance cardiaque, l'ostéoporose, la maladie de Parkinson, la sarcopénie, et le diabète de type 2 [182]. Leur fréquence dans la population a d'ailleurs augmenté depuis plusieurs années et va continuer d'augmenter puisque la part d'humains dépassant les 65 est supérieure aujourd'hui à celle des moins de 5 ans et qu'elle croît encore [183].

De la fragilité qu'engendrent ces maladies du vieillissement découlent de nombreuses contraintes physiques, économiques et sociales [184]. Face à la fatigue et la réduction de tonus corporel, les

5. Texte requis par l'éditeur : Cette traduction n'est pas officielle et n'a pas été approuvée par Elsevier

personnes âgées vont perdre progressivement en autonomie et nécessiter un accompagnement de plus en plus systématique par des structures d'aide à la personne et de soin chronique. L'augmentation de l'affluence de personnes à prendre en charge dans les années à venir va donc représenter un défi pour le système de santé publique [183]. Si l'option d'augmenter les capacités hospitalières et sociale existe, il serait moins coûteux de chercher à prévenir le vieillissement. La recherche actuelle vise donc à comprendre le vieillissement, non pas pour augmenter la longévité, mais plutôt pour contrer à la racine ses effets néfastes pour finalement vieillir en bonne santé [180].

1.4.3 L'étude du vieillissement à travers les réseaux de co-expression

Le vieillissement va directement impacter l'expression des gènes en raison des dérégulations causées. Les différentes marques principales du vieillissement seront donc perceptibles, chacune dans une mesure variable, à travers l'étude du transcriptome [185]. Ce profilage transcriptomique a permis par le passé l'identification de nombreux biomarqueurs uniques ou combinés associés à une marque principale du vieillissement ou à la longévité [186]. Les bases de données *GenAge* [187] et *Digital Aging Atlas* [188] répertorient ainsi respectivement 307 gènes humains avérés associés au vieillissement et 2599 gènes humains impliqués dans des variations lors du vieillissement ou de la longévité. Une étude réalisée par de Magalhães *et al.* identifie toutefois seulement 73 gènes constamment associés avec le vieillissement [189]. Les autres gènes présents dans ces différentes bases de données présentent des gènes variant spécifiquement dans un tissu, un sexe, ou en présence d'une maladie spécifique du vieillissement. Ils vont par ailleurs, pour une même marque principale du vieillissement comme le raccourcissement des télomères, avoir une manifestation différente entre tissus en raison de leur fonction ou leur taux de renouvellement (Table 1.3).

Le vieillissement est une imbrication simultanée et complexe de plusieurs altérations de l'intégrité des cellules. Ces gènes associés au vieillissement et agissant comme biomarqueurs ne permettent pas à eux seuls de comprendre quelles sont les voies métaboliques qui sont impactée et comment ces gènes vont contribuer à d'autres altérations. Les GCN sont alors particulièrement adaptés pour répondre à ce besoin grâce à l'étude des voisins de ces biomarqueurs [93]. Par l'étude des modules détectés dans leurs réseaux, il est également possible de détecter de nouveaux biomarqueurs, soit en analysant la topologie de modules associés au phénotype de l'âge, soit en effectuant une analyse de co-expression différentielle entre plusieurs tranches d'âge et en isolant les gènes responsables de la non-préservation des modules [191]. Il serait également possible d'effectuer cette co-expression différentielle entre tissus, espèces, populations pour estimer la part de spécificité et la part commune de vieillissement.

En 2009, les GCN vont même permettre de venir compléter l'idée de la désorganisation de la transcription dans le vieillissement. En observant différents tissus entre des souris de 16 et 24

Type de tissu	Nom du tissu	Manifestations pathologiques chez l'humain avec syndromes télomériques
Tissus à fort renouvellement	Peau	• Blanchissement prématuré des cheveux
	Moelle osseuse	• Anémie aplastique
	Immune	• Infections opportuniste • Immunodéficience des cellules B, T et NK
	Épithélium intestinal	• Entérocolite • Émoussement villositaire
Tissus à faible renouvellement	Poumon	• Emphysème prématuré • Fibrose pulmonaire idiopathique
	Foie	• Fibrose-cirrhose du foie cryptogénétique
	Os	• Ostéoporose • Nécrose avasculaire
Cancer	Multi-tissus	• Cancers épithéliaux • Hémopathies malignes

TABLE 1.3 – Phénotypes de maladies spécifiques aux organes chez les humains ayant des télomères courts. Modifié d'après la Table 2 de Armanios 2012, *The telomere syndromes* [190]

mois, Southworth *et al.* vont mettre à jour un phénomène de perte de densité de co-expression avec l'âge [153]. Plus précisément encore, ils vont démontrer que cette perte n'est pas uniforme, mais se fait de façon modulaire, probablement en raison d'une colocalisation sur un chromosome des gènes impliqués. Ce phénomène sera d'ailleurs retrouvé dans de nombreuses études sur les cancers, ensemble de maladies connu pour partager plusieurs marques principales du vieillissement [192]. Peu d'études sur ce phénomène ont suivi dans le cadre du vieillissement et une exploration plus détaillée de cette perte chez l'homme et sur plusieurs points dans l'âge d'un organisme sont à envisager.

1.5 Motivations et hypothèses de recherche

Face à la sous-exploitation des GCN dans l'analyse transcriptomique et à leur potentiel dans l'étude du vieillissement, j'ai souhaité faciliter l'accès à cette méthode et démontrer plusieurs aspects sur lesquels elle est à même de permettre de nouvelles découvertes. Les outils dédiés aux GCN existant au début de cette thèse présentaient de nombreux freins à l'utilisation par des utilisateurs sans formation en bio-informatique et biostatistique ou par des utilisateurs souhaitant faire une analyse rapide clef en main. Ces outils n'utilisaient également pas le potentiel d'un assemblage de logiciels façon pipeline, et si certains outils s'en inspirant ont été publiés par d'autres équipes durant cette thèse, aucun d'entre eux ne considérait la pérennité de l'outil final dans leur architecture implémentée.

Cette thèse a donc été le siège de deux hypothèses de recherche. La première consistait à s'interroger sur la faisabilité d'un outil pipeline sous la forme d'un progiciel R déposé sur Bioconductor qui tirerait parti des outils existants pour chaque étape de l'analyse de co-expression et

prévoirait la possibilité de les remplacer dans le futur par d'autres grâce à une architecture logiciel modulaire. La seconde supposait que l'utilisation d'un tel outil pourrait permettre de faciliter la découverte de nouveaux biomarqueurs spécifiques et communs du vieillissement entre différents tissus humain via des réseaux construits chacun. Chacune de ces hypothèses a donc été testée au cours de cette thèse dans un chapitre qui lui est dédié.

Chapitre 1 - GWENA : gene co-expression networks analysis and extended modules characterization in a single Bioconductor package

Gwenaëlle G. Lemoine[1], Marie Pier Scott-Boyer[2], Bathilde Ambroise[3], Olivier Périn[3], Arnaud Droit[1,2]

1. Département de médecine moléculaire, Faculté de médecine, Université Laval, 2325 rue de l'Université, Québec G1V 0A6, Canada.
2. Centre de recherche du CHU de Québec-Université Laval, 2705 boulevard Laurier Québec, Québec G1V 4G2, Canada.
3. L'Oréal Research and Innovation, 15 rue Pierre Dreyfus, 92110 Clichy, France.

2.1 Résumé

L'analyse de l'expression des gènes par le biais de réseaux de co-expression peut être utilisée pour étudier les relations modulaires entre des gènes remplissant différentes fonctions biologiques. À ce jour, aucun outil ne combine en un pipeline pérenne les analyses usuelles ainsi que la co-expression différentielle et la visualisation de réseau. Nous présentons ici GWENA, un nouveau progiciel R disponible sur Bioconductor qui répond à ce besoin. Pour démontrer ses performances, nous avons appliqué GWENA sur deux ensembles de données de muscle squelettique provenant de patients jeunes et âgés de l'étude GTEx. De façon remarquable, nous avons priorisé plusieurs gènes pour l'étude du vieillissement ainsi que précisé le phénomène connu de perte de connectivité. En effet, parallèlement à cette déconnexion se déroule une reconnexion de

divers gènes sur les gènes pivots du réseau, gènes codant pour des mécanismes de compensation.

2.2 Abstract

2.2.1 Background

Network-based analysis of gene expression through co-expression networks can be used to investigate modular relationships occurring between genes performing different biological functions. An extended description of each of the network modules is therefore a critical step to understand the underlying processes contributing to a disease or a phenotype. Biological integration, topology study and conditions comparison (e.g. wild vs mutant) are the main methods to do so, but to date no tool combines them all into a single pipeline.

2.2.2 Results

Here we present GWENA, a new R package that integrates gene co-expression network construction and whole characterization of the detected modules through gene set enrichment, phenotypic association, hub genes detection, topological metric computation, and differential co-expression. To demonstrate its performance, we applied GWENA on two skeletal muscle datasets from young and old patients of GTEx study. Remarkably, we prioritized a gene whose involvement was unknown in the muscle development and growth. Moreover, new insights on the variations in patterns of co-expression were identified. The known phenomena of connectivity loss associated with aging was found coupled to a global reorganization of the relationships leading to expression of known aging related functions.

2.2.3 Conclusion

GWENA is an R package available through Bioconductor (https://bioconductor.org/packages/release/bioc/html/GWENA.html) that has been developed to perform extended analysis of gene co-expression networks. Thanks to biological and topological information as well as differential co-expression, the package helps to dissect the role of genes relationships in diseases conditions or targeted phenotypes. GWENA goes beyond existing packages that perform co-expression analysis by including new tools to fully characterize modules, such as differential co-expression, additional enrichment databases, and network visualization.

2.2.4 Keywords

co-expression network, differential co-expression, R package, pipeline, aging, skeletal muscle

2.3 Background

The study of biological functions through discrete genes analysis methods has allowed the elucidation of numerous pathways and the understanding of gene-disease associations [1]. The full comprehension of the complex interactions taking place in cellular processes requires methods that are able to grasp the connections between the genes involved [2]. To address this issue, biological networks have been used as a framework to represent and study relationships between genes. In a gene network, a node represents a gene and an edge joining two nodes represents their relationship. Among the measures of relationship, weighted co-expression is one of the most widely used thanks to the popularity of the WGCNA R package [3] where the relationships are quantified (weight) instead of only a presence/absence information. The use of gene co-expression networks thus led to important discoveries such as the characterization of functional elements in *Arabidopsis* [4], help with prognosis in breast cancer [5], and more generally identification and prioritization of disease candidate genes [6].

When constructing gene co-expression networks, existing tools usually follow the same methodology. Using either microarray or RNA-seq gene expression, a co-expression score based on correlation is computed between each pair of genes in the samples. A clustering method is then selected to detect groups of strongly co-expressed genes called modules. The search for meaning in the co-expression relations classically involves the integration of biological information, as well as the study of topology [6]. Biological integration usually involves two methods, namely gene set enrichment and phenotypic association [6, 3]. A phenotypic association is based on the correlation between the eigengene (a representative of gene expression profile) of the module and a phenotype measured on the samples. Despite typically having a low yet significant correlation [7], phenotypic associations are used as a surrogate to study the molecular changes related to a condition. By looking for the genes responsible for the correlation, this method serves as a means of causal genes discovery or a way to find the effect of the condition on the phenotype [8]. As for the gene set enrichment, the most common enrichment test is based on the over-representation analysis (ORA) of a group of genes (in this case modules) compared to a reference of biological annotations such as Gene Ontology (GO) [9] or Reactome [10]. This approach, based on the guilt-by-association approach, allows the identification of new gene functions. The consideration of the scale-free topology property of gene co-expression networks also allow the use of graph theory metrics and methods to analyze the networks from a new perspective. The highly-connected genes also known as hub genes are often relevant for the functionality of the module, either being a regulator [11] or a gene coding for an essential function [12]. Their detection and the

investigation of the neighboring gene is therefore an opportunity to understand the mechanisms at work.

Like differential expression analysis, co-expression analysis can be used in a differential way to compare conditions (e.g. wild vs. mutant). This method aims to isolate dissimilarities [13] that would not be found by solely studying the GCN of a condition of interest (e.g. disease, phenotype). Variations in gene co-expression between multiple conditions can translate into appearance/disappearance of modules, changes in gene composition of a module, or rearrangement of genes within a module potentially leading to separation into several other modules [6]. These modifications of patterns reveal insights on the biological alterations in modules of interest and can suggest possible regulatory events linked to the studied condition (e.g. : transcription factors, miRNA). Such concepts were used successfully in recent publications to detect specific gene modules involved in ovarian or breast cancer [14, 15] or in recovery from water stress in *Cleistogenes* [16].

To date, multiple tools exist that perform one or few of the functionalities described previously but none combine them all into a single pipeline. Moreover, no available tool includes differential co-expression, exploits the potential of other topological metrics such as connectivity, or enables analysis to be carried out with other R packages or software as easily. In order to meet all these needs, we developed an R package for Gene Whole co-Expression Network Analysis (GWENA) available on Bioconductor (https://bioconductor.org/packages/release/bioc/html/GWENA.html). Based on a modified version of WGCNA for the network construction and module detection, GWENA is a modular pipeline that provides ORA enrichment on 9 biological sources, phenotypic association, hub genes detection, and differential co-expression between multiple conditions. These come with a set of descriptive visualizations that help the user understand and interpret complex results of gene co-expression network analysis.

In order to demonstrate the capabilities of our tool, we applied it to investigate skeletal muscle aging using publicly available gene expression data from donors spanning different age ranges from the GTEx database [17]. Skeletal muscle aging is indeed a major source of mobility loss in the elderly, resulting in a high fall ratio, depression, and therefore an increased mortality [18]. This decrease in the regenerative capacity of skeletal muscles and their progressive atrophy (sarcopenia) [19] gradually leads to a reduction of the contractile force and thus a loss of autonomy of the individuals [20]. Recent studies have made progress in finding factors associated to evolution of sarcopenia [21, 18], such as body weight [22], but the understanding of their intricate molecular mechanisms is still lacking.

In this article, we will therefore provide details on the implementation of our new R package GWENA. A presentation of its application will be done with the study of gene co-expression in young muscle, and then in the context of skeletal muscle aging by comparing samples from younger and older donors. Finally, a qualitative comparison will be made with other existing tools.

2.4 Implementation

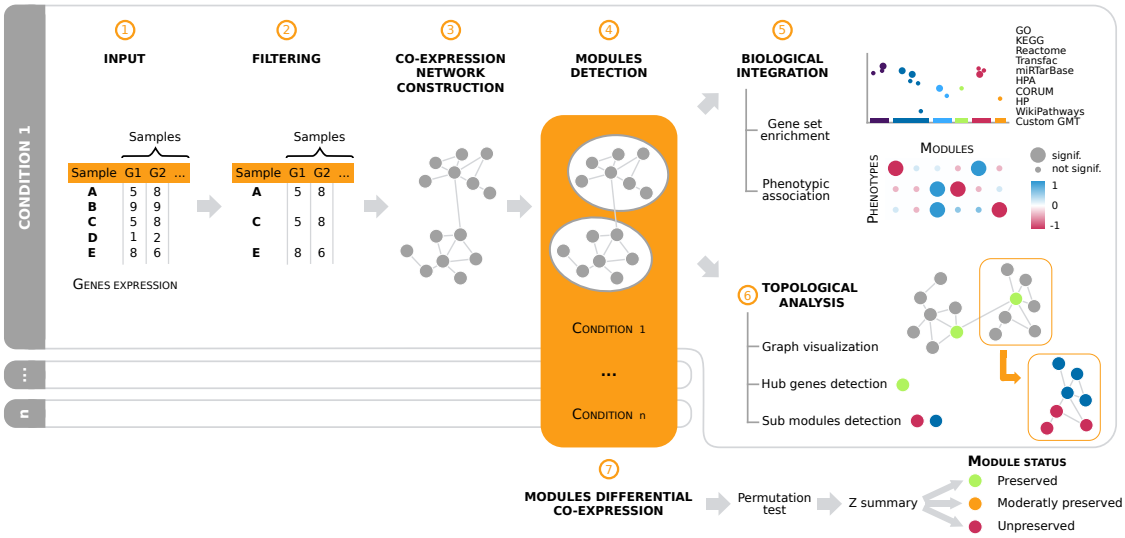

FIGURE 2.1 – Detailed steps of analysis performed in GWENA's pipeline, from expression data to characterization of the modules and comparison of conditions. ① Input : expression matrix pre-normalized and aggregated to gene level if it is a transcript matrix. ② Filtering : optional genes filtration according to transcriptomic input technology. ③ Co-expression network construction : computation through modified WGCNA function of a correlation matrix on the gene expression matrix, then transformation into an adjacency matrix, and finally into a topological overlap matrix (TOM). ④ Modules detection : genes clusterization over the TOM with another modified WGCNA function. ⑤ Biological integration : gene set enrichment of each module using g :Profiler services, and phenotypic association if a phenotype matrix is provided to describe the samples. ⑥ Graph analysis : transformation of the TOM in a graph to compute different topological metrics, detect the hub genes, and detect sub-modules of one module. ⑦ Modules differential co-expression over N conditions : permutation test using NetRep combined with a Z summary to detect preserved or unpreserved modules.

Designed as an R Bioconductor package, GWENA is a modular pipeline intended to ease the construction, interpretation and comparison of GCN. It reproduces a classical GCN analysis reinforced by complementary tools (Fig. 2.1).

2.4.1 Input

Both microarray and RNA-seq normalized expression can be used as input. The choice of normalization method is left to the user as it is highly dependent on the technology used to produce the raw data and the experimental design. Data must be stored in a table with genes as columns and samples as rows, or in a SummarizedExperiment object [23]. The minimal number of samples recommended is about 20 samples [24] with 100 samples ensuring a more robust networks [25].

Transcript-level data (probes or transcript) need to be aggregated to the gene level for the next steps (i.e. probes measurement summarized to their corresponding gene) [26]. Its execution is left to the user as the transcriptomic technology impacts the aggregation method to choose. However, it is recommended to use the highest mean probes expression for microarray data, and the counts

sum for RNA-seq. This can be achieved with the collapsing R function as described by Miller et al. [26].

2.4.2 Filtering

Genes are not always informative for modules detection as genes not always vary and their expression can be linked to technical biases. An additional filtering step can thus be applied to avoid noise and speed up the pipeline analysis. This operation must be carried out with caution as it may impact the network construction. Over-filtering may result in loss of informative signal and changing the data distribution could break the scale-free topology [27, 24]. In addition, co-expression network analysis is a method designed to handle larger amount of data than differential expression analyses and can capture more subtle significant gene expression variation [28, 8].

Two filters meeting these criteria are available in GWENA :

1. Low count filter : removes genes having a lower count than a pre-defined threshold (default is 5). It prevents confusing the true expression of a gene with an expression due to technical background noise.
2. Low variation filter : removes genes which expression is too similar across samples. As co-expression modules detection relies on the discrimination of similarity between gene expression profiles across samples, genes that so not vary sufficiently across samples may be randomly clustered in the same (or in different) modules which would not reflect the biological reality.

2.4.3 Co-expression network construction

The well-known R package WGCNA [3] was modified in order to be integrated it in our modular pipeline : the co-expression network construction step which computes the genes pairwise co-expression score has been isolated in its own R function. The first step of the co-expression score computation is the calculation of a correlation matrix based on the gene expression matrix. The Spearman correlation was added to the automated version of network construction in WGCNA as it ensures a better representation of genes monotonic relationships [29]. A power law distribution is then fitted on the correlation matrix and the “correlation matrix is then raised to the estimated power, resulting in an adjacency matrix [30]. According to the hierarchical organization of gene co-expression networks [31], a topological overlap matrix (TOM) [30] is then computed using the adjacency matrix which represents the final gene co-expression score matrix. Finally, the function return this matrix along with metadata information regarding the computation to ensure a good tracking of the performed operations.

2.4.4 Modules detection

The modules detection part from WGCNA was isolated in a new R function using the previously calculated gene co-expression score matrix as input. A hierarchical clustering is performed on the matrix which is then cut according to the dynamic cut tree method [32] in order to define the modules and the genes they contain. The first component of the principal component analysis of each module is used as a representative of their respective gene expression profile and is called an eigengene. In addition to its summarizing function, the eigengene is used to merge the highly-correlated modules. The gene co-expression profile of each module is visible using a dedicated function, with the eigengene highlighted. The function finally returns a detailed object with the detected modules as lists of genes identifiers, the dendrogram of the clustering, and the modules before merge.

2.4.5 Biological integration

Biological integration consists of two different analyses, namely gene set enrichment and phenotypic association.

The gene set enrichment (or functional enrichment) analysis is performed using g :Profiler [33] through their gprofiler2 R package. Their enrichment function covers 9 biological functional databases : Gene Ontology (GO) [9], Kyoto Encyclopedia of Genes and Genomes (KEGG) [34], Reactome [10], Transfac [35], miRTarBase [36], Human Protein Atlas (HPA) [37], CORUM [38], Human Phenotype ontology (HP) [39], WikiPathways [40]. Realizing a custom enrichment file through a Gene Matrix Transposed (GMT) format in gprofiler2 requires the use of additional functions. Also, gprofiler2 does not provide a merging function between the output of classical and custom enrichment to return all the enrichments in a single output. GWENA therefore provides a wrapper of these functions to have an all-in-one function.

The phenotypic association uses the eigengene returned in the output of the module detection function to perform a correlation test on a matrix of given phenotypes. If a phenotype is qualitative instead of quantitative, the variable encoding the phenotype is transformed into a binary variable (also known as dummy variable).

2.4.6 Graph analysis

To analyze the topology of the graph and allow its visualization, GWENA imports the igraph [41] R package. A wrapping function including integrity checks use the gene co-expression score matrix to build a graph object on which all igraph topological metrics can be computed (e.g. degree, connectivity, strength). Among the multiple metrics computable on a network, hub genes remain the most studied structure. As they can be defined according to different methods, the

three most popular ones were implemented : highest connectivity [42], highest degree [8], and Kleinberg's score [43]. GWENA visualization function simplifies the native plotting function of igraph and adapts it to GCN to assist in their interpretation (e.g. the native implementation of an edge filter parameter, as these are complete graphs). The layout selection was also favored towards scale-free topology compatible layouts as they are a main property of GCN. Sub-modules within previously detected modules can provide valuable information about the distribution and communication between biological functions within a module. GWENA allows their detection by performing a partitioning around medoids (PAM) clustering method [44, 45] with an automatic estimation of the number of cluster through a silhouette coefficient. These sub-modules can also be passed to the graph plot function to display them and see their organization.

2.4.7 Modules differential co-expression

Analysis of module preservation or non-preservation can be performed between different conditions such as treatments or phenotype. To isolate modules whose topology changes between conditions, GWENA first performs a permutation test using the NetRep R package [46]. Seven topological metrics are computed on each module in each condition. A permutation is then applied to the selected control condition where each node label of the modules is randomly reassigned without replacement to another and the seven metrics are then recomputed on it. Using these permutations as a null distribution [47], modules are considered preserved if all seven topological metrics are significant for the alternative hypothesis (one or two-sided) with the chosen alpha error.

As the unpreservation of a module cannot be assumed from the non significant modules, a second step of preservation evaluation is carried out using a Z summary score [48, 49]. The final score returned by GWENA is the combination of these two steps (Additional file : Figure A.1).

2.5 Results and discussion

To present GWENA's use and its capability to isolate genes groups or co-expression patterns of interest in a single condition or multiple conditions, we analyzed RNA-seq skeletal muscle data from GTEx (v8)[17] (Additional file : Table A.1). This data set contains 19 312 genes from 803 samples representing ages ranging from 20 to 70 years old. Low read counts and the low variation genes were discarded using the filtering function of GWENA to decrease the noise, resulting in 18 870 genes.

As GTEx data is known to be subject to multiple confounding factors (batch effect, experimental bias, read contamination, etc.) [50, 51, 27], a partial PC-correction[27] was applied to correct the data (Additional file : Figure A.2 and A.3). To investigate the aging process two subsets represen-

ting contrasting age classes were selected from the corrected data set : 73 samples between 20 and 30 years old (referred as young in this report), and 292 samples between 60 and 70 years old (referred as old in this report). Both datasets were analyzed using GWENA's pipeline with default parameters, except for the correlation method parameter which was selected to be "spearman" instead of the default "pearson" as it is less prone to outliers.

2.5.1 Single condition modules analysis

To illustrate the process of analyzing a single condition with GWENA, we initially focused on studying the muscle gene co-expression computed in the young sub-population. The 95 modules detected on the co-expression score matrix with GWENA were merged according to their similarity indices, which resulted in a total of 35 modules (Fig 2.2a). Each module was then tested for its association with a selected set of phenotypes related to muscle aging (i.e. age, sex, ethnicity, body weight and BMI) to isolate modules of interest. As shown in Fig. 2.2b, 15 of these modules were significantly associated with at least one of the phenotypes.

These modules were provided to GWENA enrichment analysis (p value <0.05 with g :SCS multiple testing correction) to identify their biological functions and assess their potential involvement in muscle function (Table 2.1). All modules were at least enriched in one term and 8 obtained enrichment terms related to muscle activity or metabolism (Table 2.1). Modules 19, 21 and 25 were the top 3 enriched for terms related to muscle function also associated with a phenotype impacting muscle. However, modules 21 and 25 terms were mostly coming from Human Protein Atlas and were also related to a wide range of additional tissues such as the pancreas, the cervix, the bladder, the stomach, or the skin and were thus deemed less specific for muscle aging than module 19.

Briefly, the remaining module 19 presented 77% of genes positively correlated to its eigengene (therefore 23 negatively, Fig. 2.2d), and the muscle enriched terms involved muscle adaptation and negative regulation of hypertrophy (Table 2.2, Fig. 2.2c). The detection of hub genes by GWENA returned 12 hub genes, some of which are known as transcription factors. Among them, ARID5B (ENSG00000150347) is a transcription factor strongly co-expressed with KLF15 (ENSG00000163884) and TRIM63 (ENSG00000158022) (Fig. 2.2e). These two genes are present in the GO term GO :0014888 (striated muscle adaptation) to which ARID5B is not associated. The function of ARID5B is well known in adipocytes and hepatocytes but is still rarely studied in skeletal muscle metabolism. However, the knockout of this gene in mice has shown structural defects in the sarcomere structure [52].

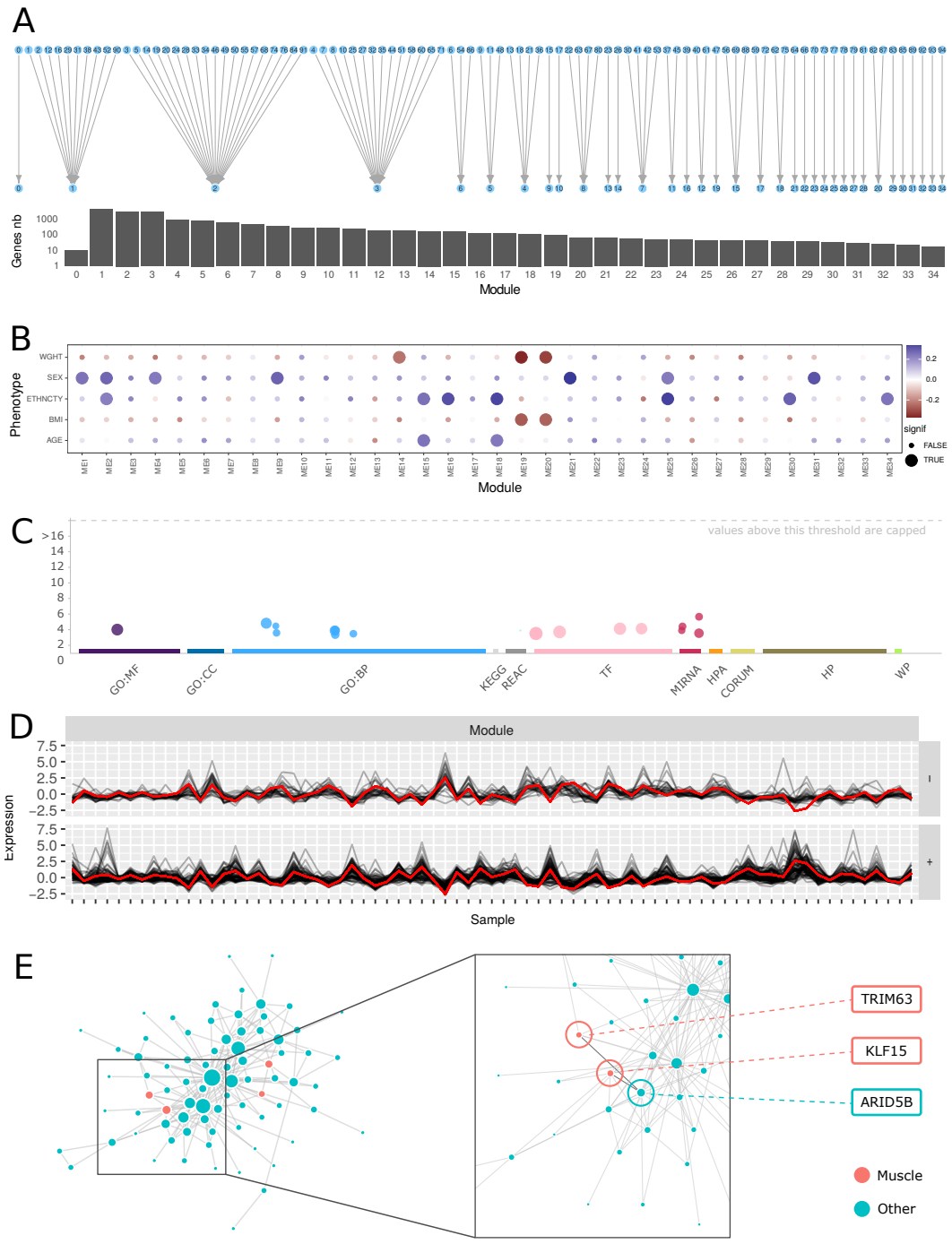

FIGURE 2.2 – Available visualizations in GWENA along the pipeline applied to the aging study on the whole age range. A : modules merge as a bipartite graph from plot_modules_merge function and the genes distribution inside each of them (log scale). B : phenotypic association between the 35 modules and age, sex, BMI, ethnicity, weight. C : Manhattan-like enrichment plot (interactive in GWENA) of module 19 on GO, KEGG, Reactome, Transfac, miRTarBase, Human Protein Atlas, CORUM, Human Phenotype ontology, WikiPathways. D : expression profile of module 19 split depending on the correlation sign to the eigengene. E : module 19's network visualization as a graph with muscle enrichment genes colored in red and others in blue. The zoom focus on ENSG00000158022 / ENSG00000107372 / ENSG00000265972 and related hub genes.

module	# genes	# pheno. asso.	# enrichment	% muscle enrich.
0	11	NA	NA	NA
1	5335	1	3288	0.6
2	3661	2	1098	0.4
3	3355	0	1620	0.3
4	1001	1	1883	1.0
5	987	0	1626	0.0
6	699	0	457	0.6
7	546	0	428	1.1
8	409	0	561	0.7
9	310	1	729	0.3
10	308	0	58	5.2
11	261	0	857	0.0
12	214	0	847	15.0
13	207	0	452	1.4
14	197	1	767	0.5
15	175	2	233	0.0
16	137	1	6	0.0
17	136	0	1	0.0
18	129	2	20	0.0
19	108	2	18	3.7
20	77	2	52	0.0
21	72	1	233	2.8
22	63	0	24	0.0
23	57	0	10	0.0
24	55	0	8	0.0
25	47	2	82	8.5
26	47	0	32	0.0
27	46	0	12	0.0
28	43	0	147	4.7
29	40	0	17	0.0
30	35	1	2	0.0
31	31	1	12	0.0
32	27	0	1	0.0
33	24	0	23	0.0
34	20	1	3	0.0

TABLE 2.1 – Summary of detected modules and related biological integration. The number (#) of genes is indicated for each module (module 0 being a false module containing the unassigned genes. The number of phenotypic associations with the variables of interest (weight, sex, ethnicity, bmi, age) are counted for each one. The number of enrichments corresponds to the count of significant terms on each cumulative biological database. The ratio (%) of enriched terms associated with muscle is then established as the number of terms containing one or more elements of the following corpus : "muscle", "sarco*", "*", "muscul*", "actin*", "myosin*" (where * denotes a completion by any other character string)

Coupled with the results of GWENA, this may corroborate the involvement of ARID5B in the adaptation of striated muscle in response to a stimulus. Moreover, it has recently been shown that ARID5B knockout in mice was associated with increased glucose metabolism via an increased

translocation of SLC2A4 (ENSG00000181856) [53]. Since SLC2A4 is a gene that is also regulated by KLF15 [54, 55], this supports the idea that ARID5B has implications in skeletal muscle function and more precisely in glucose metabolism. GWENA thus allowed the identification of a gene that may give new insight in the muscle development and growth which needs to be confirmed by further experiments.

source	term name	p val.
GO :BP	response to hormone	0.0015
GO :BP	negative regulation of muscle hypertrophy	0.0033
GO :BP	muscle adaptation	0.0118
GO :BP	response to peptide hormone	0.0129
GO :BP	striated muscle adaptation	0.0255
GO :BP	platelet-derived growth factor receptor signaling pathway	0.0328
GO :BP	regulation of muscle adaptation	0.0434
GO :MF	enzyme binding	0.0097
MIRNA	hsa-miR-6882-5p	0.0002
MIRNA	hsa-miR-197-5p	0.0039
MIRNA	hsa-miR-152-5p	0.0125
MIRNA	hsa-miR-6878-5p	0.0282
REAC	Regulation of FOXO transcriptional activity by acetylation	0.0126
TF	Factor : Zbtb37 ; motif : NYACCGCRNTCACCGCR ; match class : 1	0.0073
TF	Factor : RNF96 ; motif : BCCCGCRGCC ; match class : 1	0.0074
TF	Factor : ETF ; motif : GVGGMGG ; match class : 1	0.0193
TF	Factor : AP-2 ; motif : SNNNCCNCAGGCN	0.0306
TF	Factor : AP-2 ; motif : SNNNCCNCAGGCN ; match class : 0	0.0306

TABLE 2.2 – Module 19 young enriched terms table. Multiple enrichment are linked to muscle development and growth

2.5.2 Multiple conditions modules comparison and analysis

Differential expression analysis allowed the detection of genes involved in aging in the last years (GenAge [56], Digital Ageing Atlas [57]). Such discriminant analysis is limited in helping to understand aging as this phenomenon is composed of concomitant mechanisms [58]. Understanding the relationships between the genes is therefore crucial to determine the altered functions and the changes involved. Differential GCN between conditions overcomes this problem by detecting the subtle pattern modifications. Using our previously defined young (20 to 30 years old) and old (50 to 60 years old) skeletal muscle modules, we ran GWENA's GCN differential co-expression func-

tionality to compare the modules between these age ranges. The GCN of each module detected in the young sub-population were taken as a reference and tested against the ones detected in the old sub-population.

Comparison status	# modules	Modules id
preserved	11	1, 2, 3, 4, 6, 8, 9, 11, 12, 14, 19
moderately preserved	17	7, 10, 13, 17, 18, 20, 21, 22, 23, 24, 25, 27, 28, 30, 31, 33, 34
unpreserved	2	16, 32
inconclusive	4	5, 15, 26, 29

TABLE 2.3 – Modules comparison between young and old age range and their comparison status

From the 35 modules detected in the previously described single condition analysis of young muscle data, GWENA's differential GCN of young versus old age range returned 2 modules that were unpreserved, 17 modules that were moderately preserved, 11 modules that were preserved, and 4 that were inconclusive (Table 2.3, Additional file : Figure A.1). Unpreserved and moderately preserved modules are the most promising for identifying groups of genes differently expressed with age. Few and heterogeneous significant enrichment terms were associated to unpreserved modules while several moderately preserved modules had enrichment terms known to be linked to aging [59, 60, 61, 58] such as transcription regulation (module 21), cellular stress (modules 20 and 27), immune response (modules 7 and 28), cell proliferation (module 13).

In addition to this biological information, the topological comparison of these modules allows to grasp the nature of the variations in the relationships between genes (nodes in the network) and their co-expression score (edge weight in the network). Connectivity, as defined by J. Dong and S. Horwath [62], is a common topological metric computed in GCN as it is representative of the network robustness and is known to be linked to network deregulation [63, 64]. Over all modules, the connectivity of the genes in module 7 was noticeably dropping between young and old age range (Fig. 2.3, Additional file : Figure A.2.1). Using a co-expression score filter of 0.95, this loss of connectivity materialized in the network through a disconnection (edge loss) of peripheral genes (genes with low degree) such as in sub-module 4 between the young and old age range (Fig. 2.4a, b). Several other genes of the module 7 from the young age range also showed an increased connectivity when observed in the old age range, which therefore reflects a reconnection (edge gain) to other genes. These results confirm the observations from previous studies of a connectivity loss in the network of modules linked to aging [65, 64]. Overall, they support an alteration of the transcription regulation.

GWENA sub-module detection method on the module 7 revealed an impact of this reorganization of the gene connections by detecting 5 optimal sub-modules for the young age range, and 6

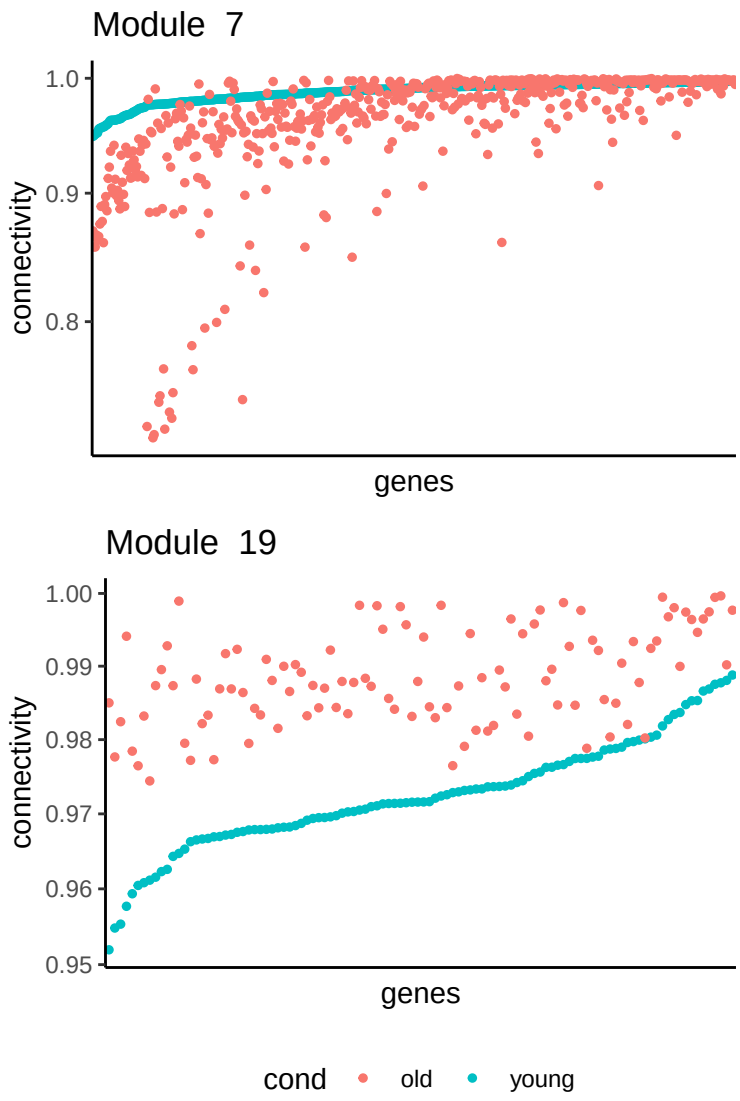

FIGURE 2.3 – Modules 7 and 19 genes (nodes) connectivity distribution between young and old age ranges. Young age range is used as reference for sorting genes by increasing connectivity. A comparison over all modules can be found in Additional file : Figure A.2.1

.

optimal sub-modules for the old age range. The gene composition of the sub-modules was highly similar between the condition (at least 81% common genes). Most of the differences in the gene composition are due to the disconnection of peripheral genes as previously spotted, and a small portion of the differences are due to the reconnection of genes or their attribution to another sub-module (Fig. 2.4a, b). This rewiring of the network is in line with known compensatory processes occurring during aging [60]. By triggering molecular processes involved in limiting or repairing cellular stress damage, these adaptive modifications aim to restore a homeostatic state.

To support this information, the new sub-module (sub-module 6 in Fig. 2.4b) appearing in the old

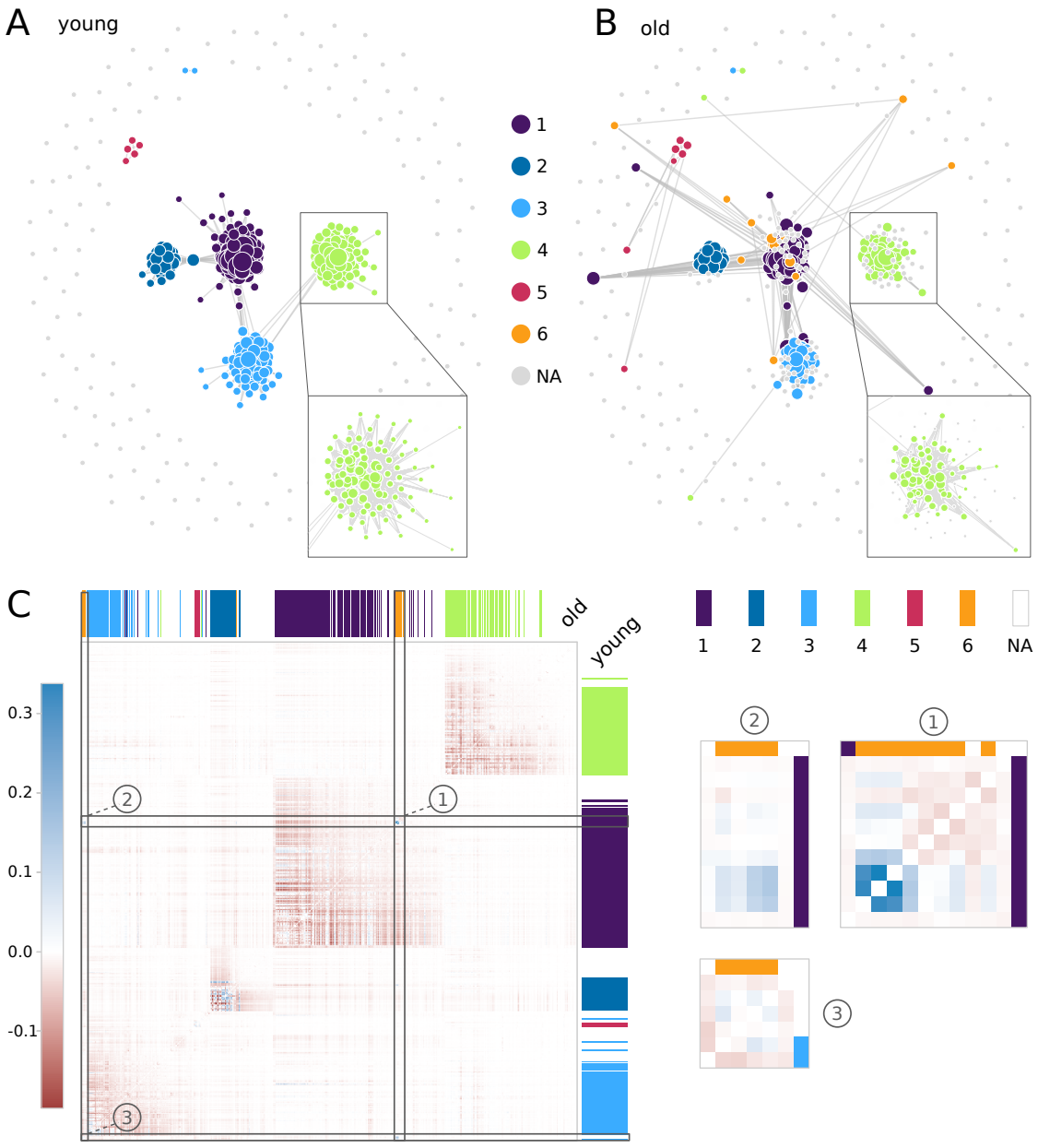

FIGURE 2.4 – Module 7 network comparison between young and old. A : module 7 GCN graph plotted with GWENA (0.95 co-expression score filter) for young age range with sub-clusters detected. B : same as A but for the old age range. A zoom is made on sub-module 4 to show the peripheral genes disconnection. The new sub-module 6 is visible in purple in the old graph. C : difference network heatmap (old - young) ordered according to young age range network dendrogram. Sub-modules from old age range are visible on the top of the heatmap in columns, and sub-modules from young age range on the right in rows. Three zooms are made on the heatmap on the areas corresponding to sub-module 6 genes. Zoom ① contains the genes reconnecting in the old age range.

age range was investigated further. Its creation is at the expense of the sub-module 1 of the young age range and of the 13 genes composing it, 8 of the genes are from the sub-module 1, 3 are reconnecting genes, and 2 are from sub-modules 2 and 3. A gene set enrichment analysis with GWENA of this sub-module 6 revealed significant enrichment in functions related to wound healing, coagulation, vessel diameter, platelet degranulation, and plasminogen activation (Additional

file : Table A.2). This is coherent with known morphological alterations of the vascular system in the aged skeletal muscle [66, 67], and the global immune/inflammatory response increased in aging [61]. Also, 51 enrichments from young sub-module 1 were not found significant in any of the sub-modules in the old age range. These enrichments involve antibacterial humoral response and negative regulation of endopeptidase activity. These terms are known to be associated with satellite cells (muscle stem cells responsible for muscle regeneration) regulators released by the vasculature in higher quantity in young skeletal muscle [67].

To complement these analyses, we investigated the variations of co-expression scores leading to the appearance of sub-module 6 in the old age range. Using the network co-expression matrix δ returned by GWENA for each condition, a co-expression difference matrix (Fig. 2.4c) was computed such as $\delta_{old} - \delta_{young}$. In this matrix, gene pairs with a negative score indicates a decrease in the co-expression over aging while a positive score indicates an increase in the co-expression. Among the variations, 3 genes showed a significant increase in co-expression between them but also towards other genes of sub-module 6. The pattern visible in Fig. 2.4c ① and ② suggest that these genes may be driving the co-expression changes occurring in this sub-module. These genes are FGG (ENSG00000171557), FGA (ENSG00000171560), and FGB (ENSG00000171564), the three fibrinogen chain coding genes involved in the polymerization of a fibrin matrix. This finding is consistent with previous studies about the increasing fibrinogen content in the elderly skeletal muscle leading to persistent fibrin deposition preventing myofiber repair [68, 69]. They also support the hypothesis of an inflammatory response triggered by a fibrin accumulation. All these results allowed by GWENA's capacity to study the topology and the biological context easily tend to support the idea of not only a global loss in connectivity in aging but also of a gene co-expression reorganization. Our tools also highlighted the biological translation of this reorganization as a potential compensatory response from antibacterial humoral response and endopeptidase activity towards coagulation and wound response.

2.5.3 GWENA's contribution and comparison with existing tools

Weighted GCN can be computed from existing tools such as WGCNA [3], wTO [70], CEMiTool [71]. As both GWENA and CEMiTool use elements from WGCNA, they share notable functionalities. They use similar network construction and modules detection functions from WGCNA but offer their own filter on the datasets. GWENA has been enhanced with additional checks on the network construction (such as aberrant power check) compared to WGCNA and CEMiTool. On its side, wTO used a different version of a topological score to construct the network as they don't perform a power law conversion on the correlation matrix and don't use the same definition of topological transformation. Therefore, the main differences between WGCNA, wTO, CEMiTool and GWENA lie in the added functionalities for module analysis.

Regarding biological integration, wTO provides neither phenotypic association nor gene set en-

Functionalities		GWENA	WGCNA	CEMiTool	wTO
Gene set enrichment	- Gene ontology	yes	yes	no	no
	- Pathways (KEGG/Reactome)	yes	no	no	no
	- Regulation actors (TRANSFAC/miRTarBase)	yes	no	no	no
	- Protein databases (Human Protein Atlas/CORUM)	yes	no	no	no
	- Custom GMT import	yes	no	yes	no
Native network visualization		yes	no	no [1]	yes
Phenotype association		yes	yes	yes	no
Hub gene detection		yes	yes [2]	yes [3]	no
Igraph compatibility for extended topology metrics calculation		yes	no	no	no
Sub-module detections inside module & graph coloration accordingly		yes	no	no	no
Modules differential co-expression		yes	yes [4]	no	no [5]

TABLE 2.4 – Key features of GWENA compared to similar tools such as WGCNA, CEMiTool and wTO. As some differences remain under the same labels, details are provided about their content. 1) CEMiTool allows network visualization only if a protein-protein interaction network file is provided, 2) WGCNA only provides a single hub gene selection by module, 3) CEMiTool persistently provides the top 10 hub genes independently of the module size or connectivity, 4) WGCNA's differential co-expression does not correct for multiple testing, 5) wTO have no differential co-expression method available but provides a consensus network method.

richment. The other three tools allow phenotypic association but differ on gene set enrichment analysis. While CEMiTool only allows enrichment on imported GMTs, WGCNA and GWENA allow enrichment on gene ontology. GWENA is the only one allowing enrichment on other databases of pathways, regulatory agents, and proteins (in addition to imported GMTs).

Additional topological analysis functions are also available in several of these tools. The most common, hub gene detection, is present in WGCNA, CEMiTool, and GWENA in different forms. CEMitool and WGCNA offer respectively as hub gene the top 10 most connected genes and genes with a top kME score (membership module based on eigengene). However, methods based on a fixed number of hub genes tend to bias the information since the number of hub genes can vary according to the number of genes present in the module. GWENA therefore proposes several methods (highest connectivity, superior degree, Kleinberg's score) based on a selection of genes with a hub score above a threshold. Another addition specific to GWENA is the ability to re-detect sub-modules into a defined module in order to further investigate the co-expression reconnection organization, and then identify the relations between enrichments associated to each sub-modules by visualizing them on the graph plot.

GWENA includes a differential co-expression analysis in the analysis pipeline as opposed to packages dedicated solely to it (DiffCoEx [72], CoDiNA [73], CoXpress [74]) or packages like wTO or CEMiTool that do not contain this analysis. The method in GWENA differs from the one present in WGCNA in that it includes a permutation test to prevent the problem of multi-testing. With the addition of the Z-summary score to detect unpreserved modules, GWENA is therefore the only pipeline including a differential co-expression analysis with high confidence in modules found unpreserved. The Table 2.4 finally provide a head-to-head comparison for all functionalities made available by GWENA.

CEMiTool and wTO are built as stand-alone tools with little or no eased interfacing with other tools. WGCNA is similar to them except for exporting networks to Cytoscape [75] or VisANT [76]. Conversely, GWENA has been developed according to a modular architecture in order to facilitate the realization with an external tool of one of the stages of the analysis pipeline defined in Fig. 2.1. GWENA will thus be more easily adaptable to follow future developments in co-expression network analysis methodology.

GWENA, as other GCN analysis tool has limitations. A first one common to all GCN construction method is that the quality of input data (e.g. filtration and/or proper normalization) will inevitably bias the results, especially if it breaks the scale-free property. A second limitation is the design of the permutation test that prevents reporting a significant unpreservation. The non-rejection of the null hypothesis of unpreservation can only state a lack of evidence of preservation. Therefore, unpreserved modules are determined among these modules lacking evidence (the non-significant modules) by the calculation of Z summary which only provide a tendency in the unpreservation [48]. The present application of GWENA to skeletal muscle aging also presents its own limitation. All analyses were performed on skeletal muscle sample and results were commented regarding this context. However, to be sure of the specificity of the findings, an additional differential co-expression of the modules should be performed on samples from other tissues from subjects with similar age range. As single-cell technologies are becoming common, the differential co-expression could also be used to target the cell-to-cell specific aging variation inside a tissue. Finally, as co-expression networks were unsigned and aging is a complex phenomenon involving actors beyond gene expression, causal effect of any finding need to be experimentally verified.

2.6 Conclusion

In this paper, we introduced GWENA, an R package on Bioconductor to construct and analyze GCN in a single pipeline through a whole range of tools from biological integration, topological analysis, and differential co-expression. The package reduces complexity of the GCN analysis through simple input and output functions combined to a set of visualizations to explore the results. The separation of each step of the analysis in one function also allows quick and easy replacement if users wish to use another method for this block.

GWENA demonstrated its performances on both single and multiple condition analysis through an exploration of variations of skeletal muscle function and processes in aging. The single condition analysis showed it is possible to find new genes potentially involved in an existing GO annotation using hub genes, network neighboring genes and gene sets enrichments. The differential co-expression analysis between young and old samples isolated modules specifically linked to aging and detected the rearrangement in connectivity related to aging. Additional analysis supported the observed genes co-expression reorganization beyond simple connectivity loss. This resulted

in a reinforcement of previous supposition on inflammatory response to fibrin increases in skeletal muscle aging.

2.7 References

[1] Albert László Barabási and Zoltán N. Oltvai. Network biology : Understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2) :101–113, 2004.

[2] Leland H. Hartwell, John J. Hopfield, Stanislas Leibler, and Andrew W. Murray. From molecular to modular cell biology. *Nature*, 402(6761 SUPPL. 1) :47–52, 1999.

[3] Peter Langfelder and Steve Horvath. WGCNA : An R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1) :559, 2008.

[4] Linyong Mao, John L. Van Hemert, Sudhansu Dash, and Julie A. Dickerson. Arabidopsis gene co-expression network and its functional modules. *BMC Bioinformatics*, 10 :1–24, 2009.

[5] Jianing Tang, Deguang Kong, Qiuxia Cui, Kun Wang, Dan Zhang, Yan Gong, and Gaosong Wu. Prognostic genes of breast cancer identified by gene co-expression network analysis. *Frontiers in Oncology*, 8(SEP) :1–13, 2018.

[6] Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, (December 2016) :bbw139, 2017.

[7] Bin Zhang, Chris Gaiteri, Liviu Gabriel Bodea, Zhi Wang, Joshua McElwee, Alexei A. Podtelezhnikov, Chunsheng Zhang, Tao Xie, Linh Tran, Radu Dobrin, Eugene Fluder, Bruce Clurman, Stacey Melquist, Manikandan Narayanan, Christine Suver, Hardik Shah, Milind Mahajan, Tammy Gillis, Jayalakshmi Mysore, Marcy E. MacDonald, John R. Lamb, David A. Bennett, Cliona Molony, David J. Stone, Vilmundur Gudnason, Amanda J. Myers, Eric E. Schadt, Harald Neumann, Jun Zhu, and Valur Emilsson. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell*, 153(3) :707–720, 2013.

[8] G. C. Tseng, E. Sibille, C. Gaiteri, Y. Ding, and B. French. Beyond modules and hubs : the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. *Genes, Brain and Behavior*, 13(1) :13–24, 2013.

[9] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology : tool for the unification of biology. *Nature Genetics*, 25(1) :25–29, May 2000.

[10] Antonio Fabregat, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, Steven Jupe, Florian Korninger, Sheldon McKay, Lisa Matthews, Bruce May, Marija Milacic, Karen Rothfels, Veronica Shamovsky, Marissa Webber, Joel Weiser, Mark Williams, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 44(D1) :D481–D487, 2016.

[11] Emma Pierson, Daphne Koller, Alexis Battle, and Sara Mostafavi. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLoS Computational Biology*, 11(5) :1–19, 2015.

[12] Matthew W. Hahn and Andrew D. Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22(4) :803–806, 2005.

[13] Hussain Ahmed Chowdhury, Dhruba Kumar Bhattacharyya, and Jugal K. Kalita. (Differential) Co-Expression Analysis of Gene Expression : A Survey of Best Practices. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(c) :1–1, 2019.

[14] Esra Gov and Kazim Yalcin Arga. Differential co-expression analysis reveals a novel prognostic gene module in ovarian cancer. *Scientific Reports*, 7(1) :1–10, 2017.

[15] Dharmesh D. Bhuva, Joseph Cursons, Gordon K. Smyth, and Melissa J. Davis. Differential co-expression-based detection of conditional relationships in transcriptional data : Comparative analysis and application to breast cancer. *Genome Biology*, 20(1) :1–21, 2019.

[16] Qi Yan, Fan Wu, Zhuanzhuan Yan, Jie Li, Tiantian Ma, Yufei Zhang, Yufeng Zhao, Yanrong Wang, and Jiyu Zhang. Differential co-expression networks of long non-coding RNAs and mRNAs in Cleistogenes songorica under water stress and during recovery. *BMC Plant Biology*, 19(1) :1–19, 2019.

[17] Kristin G. Ardlie, David S. DeLuca, Ayellet V. Segrè, Timothy J. Sullivan, Taylor R. Young, Ellen T. Gelfand, Casandra A. Trowbridge, Julian B. Maller, Taru Tukiainen, Monkol Lek, Lucas D. Ward, Pouya Kheradpour, Benjamin Iriarte, Yan Meng, Cameron D. Palmer, Tõnu Esko, Wendy Winckler, Joel N. Hirschhorn, Manolis Kellis, Daniel G. MacArthur, Gad Getz, Andrey A. Shabalin, Gen Li, Yi Hui Zhou, Andrew B. Nobel, Ivan Rusyn, Fred A. Wright, Tuuli Lappalainen, Pedro G. Ferreira, Halit Ongen, Manuel A. Rivas, Alexis Battle, Sara Mostafavi, Jean Monlong, Michael Sammeth, Marta Melé, Ferran Reverter, Jakob M. Goldmann, Daphne Koller, Roderic Guigó, Mark I. McCarthy, Emmanouil T. Dermitzakis, Eric R. Gamazon, Hae Kyung Im, Anuar Konkashbaev, Dan L. Nicolae, Nancy J. Cox, Timothée Flutre, Xiaoquan Wen, Matthew Stephens, Jonathan K. Pritchard, Zhidong Tu, Bin Zhang, Tao Huang, Quan Long, Luan Lin, Jialiang Yang, Jun Zhu, Jun Liu, Amanda Brown, Bernadette Mestichelli, Denee Tidwell, Edmund Lo, Michael Salvatore, Saboor Shad, Jeffrey A. Thomas, John T. Lonsdale, Michael T. Moser, Bryan M. Gillard, Ellen Karasik, Kimberly Ramsey, Christopher

Choi, Barbara A. Foster, John Syron, Johnell Fleming, Harold Magazine, Rick Hasz, Gary D. Walters, Jason P. Bridge, Mark Miklos, Susan Sullivan, Laura K. Barker, Heather M. Traino, Maghboeba Mosavel, Laura A. Siminoff, Dana R. Valley, Daniel C. Rohrer, Scott D. Jewell, Philip A. Branton, Leslie H. Sobin, Mary Barcus, Liqun Qi, Jeffrey McLean, Pushpa Hariharan, Ki Sung Um, Shenpei Wu, David Tabor, Charles Shive, Anna M. Smith, Stephen A. Buia, Anita H. Undale, Karna L. Robinson, Nancy Roche, Kimberly M. Valentino, Angela Britton, Robin Burges, Debra Bradbury, Kenneth W. Hambright, John Seleski, Greg E. Korzeniewski, Kenyon Erickson, Yvonne Marcus, Jorge Tejada, Mehran Taherian, Chunrong Lu, Margaret Basile, Deborah C. Mash, Simona Volpi, Jeffery P. Struewing, Gary F. Temple, Joy Boyer, Deborah Colantuoni, Roger Little, Susan Koester, Latarsha J. Carithers, Helen M. Moore, Ping Guan, Carolyn Compton, Sherilyn J. Sawyer, Joanne P. Demchok, Jimmie B. Vaught, Chana A. Rabiner, and Lockhart. The Genotype-Tissue Expression (GTEx) pilot analysis : Multitissue gene regulation in humans. *Science*, 348(6235) :648–660, 2015.

[18] Esra Ates Bulut, Pinar Soysal, Ali Ekrem Aydin, Ozge Dokuzlar, Suleyman Emre Kocyigit, and Ahmet Turan Isik. Vitamin B12 deficiency might be related to sarcopenia in older adults. *Experimental Gerontology*, 95 :136–140, 2017.

[19] Valter Santilli, Andrea Bernetti, Massimiliano Mangone, and Marco Paoloni. Clinical definition of sarcopenia. *Clinical Cases in Mineral and Bone Metabolism*, 11(3) :177–180, 2014.

[20] Ian Janssen, Steven B. Heymsfield, and Robert Ross. Low relative skeletal muscle mass (sarcopenia) in older persons is associated with functional impairment and physical disability. *Journal of the American Geriatrics Society*, 50(5) :889–896, 2002.

[21] Kunihiro Sakuma, Wataru Aoi, and Akihiko Yamaguchi. Molecular mechanism of sarcopenia and cachexia : recent research advances. *Pflugers Archiv European Journal of Physiology*, 469(5-6) :573–591, 2017.

[22] Jianqin Jiao and Fabio Demontis. Skeletal muscle autophagy and its role in sarcopenia and organismal aging. *Current Opinion in Pharmacology*, 34 :1–6, 2017.

[23] Martin Morgan, Valerie Obenchain, Jim Hester, and Hervé Pagès. Summarizedexperiment : Summarized-experiment container, 2018, 2018.

[24] Horvath S. Langfelder P. Frequently asked questions, 2014. Accessed : 2020-08-26.

[25] Franziska Liesecke, Johan Owen De Craene, Sébastien Besseau, Vincent Courdavault, Marc Clastre, Valentin Vergès, Nicolas Papon, Nathalie Giglioli-Guivarc'h, Gaëlle Glévarec, Olivier Pichon, and Thomas Dugé de Bernonville. Improved gene co-expression network quality through expression dataset down-sampling and network aggregation. *Scientific Reports*, 9(1) :1–16, 2019.

[26] Jeremy A Miller, Chaochao Cai, Peter Langfelder, Daniel H Geschwind, Sunil M Kurian, Daniel R Salomon, and Steve Horvath. Strategies for aggregating gene expression data : the collapserows r function. *BMC bioinformatics*, 12(1) :1–13, 2011.

[27] Princy Parsana, Claire Ruberman, Andrew E. Jaffe, Michael C. Schatz, Alexis Battle, and Jeffrey T. Leek. Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biology*, 20(1) :94, 2019.

[28] Nicholas J. Hudson, Antonio Reverter, and Brian P. Dalrymple. A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Computational Biology*, 5(5), 2009.

[29] Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures : Mutual information, correlation, and model based indices. *BMC Bioinformatics*, 13(1), 2012.

[30] Andy M. Yip and Steve Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, 8 :1–14, 2007.

[31] Erzsébet Ravasz and Albert László Barabási. Hierarchical organization in complex networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 67(2) :7, 2003.

[32] Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree : The Dynamic Tree Cut package for R. *Bioinformatics*, 24(5) :719–720, 2008.

[33] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g :Profiler : a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1) :W191–W198, 2019.

[34] Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, 47(D1) :D590–D595, 2019.

[35] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel : transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(Database issue) :108–110, 2006.

[36] Chih Hung Chou, Sirjana Shrestha, Chi Dung Yang, Nai Wen Chang, Yu Ling Lin, Kuang Wen Liao, Wei Chi Huang, Ting Hsuan Sun, Siang Jyun Tu, Wei Hsiang Lee, Men Yee Chiew, Chun San Tai, Ting Yen Wei, Tzi Ren Tsai, Hsin Tzu Huang, Chung Yu Wang, Hsin Yi Wu, Shu Yi Ho, Pin Rong Chen, Cheng Hsun Chuang, Pei Jung Hsieh, Yi Shin Wu, Wen Liang Chen, Meng Ju Li, Yu Chun Wu, Xin Yi Huang, Fung Ling Ng, Waradee Buddhakosai, Pei Chun Huang, Kuan Chun Lan, Chia Yen Huang, Shun Long Weng, Yeong Nan Cheng, Chao Liang, Wen Lian Hsu, and Hsien Da Huang. MiRTarBase update 2018 : A resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research*, 46(D1) :D296–D302, 2018.

[37] M. Uhlen, L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K.

Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Ponten. Tissue-based map of the human proteome. *Science*, 347(6220) :1260419–1260419, 2015.

[38] Andreas Ruepp, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Michael Stransky, Brigitte Waegele, Thorsten Schmidt, Octave Noubibou Doudieu, Volker Stümpflen, and H. Werner Mewes. CORUM : The comprehensive resource of mammalian protein complexes. *Nucleic Acids Research*, 36(SUPPL. 1) :646–650, 2008.

[39] Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O.B. Jacobsen, Daniel Danis, Jean Philippe Gourdine, Michael Gargano, Nomi L. Harris, Nicolas Matentzoglu, Julie A. McMurry, David Osumi-Sutherland, Valentina Cipriani, James P. Balhoff, Tom Conlin, Hannah Blau, Gareth Baynam, Richard Palmer, Dylan Gratian, Hugh Dawkins, Michael Segal, Anna C. Jansen, Ahmed Muaz, Willie H. Chang, Jenna Bergerson, Stanley J.F. Laulederkind, Zafer Yüksel, Sergi Beltran, Alexandra F. Freeman, Panagiotis I. Sergouniotis, Daniel Durkin, Andrea L. Storm, Marc Hanauer, Michael Brudno, Susan M. Bello, Murat Sincan, Kayli Rageth, Matthew T. Wheeler, Renske Oegema, Halima Lourghi, Maria G. Della Rocca, Rachel Thompson, Francisco Castellanos, James Priest, Charlotte Cunningham-Rundles, Ayushi Hegde, Ruth C. Lovering, Catherine Hajek, Annie Olry, Luigi Notarangelo, Morgan Similuk, Xingmin A. Zhang, David Gómez-Andrés, Hanns Lochmüller, Hélène Dollfus, Sergio Rosenzweig, Shruti Marwaha, Ana Rath, Kathleen Sullivan, Cynthia Smith, Joshua D. Milner, Dorothée Leroux, Cornelius F. Boerkoel, Amy Klion, Melody C. Carter, Tudor Groza, Damian Smedley, Melissa A. Haendel, Chris Mungall, and Peter N. Robinson. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, 47(D1) :D1018–D1027, 2019.

[40] Denise N. Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, Nuno Nunes, Jonathan Mélius, Elisa Cirillo, Susan L. Coort, Daniela DIgles, Friederike Ehrhart, Pieter Giesbertz, Marianthi Kalafati, Marvin Martens, Ryan Miller, Kozo Nishida, Linda Rieswijk, Andra Waagmeester, Lars M.T. Eijssen, Chris T. Evelo, Alexander R. Pico, and Egon L. Willighagen. WikiPathways : A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(D1) :D661–D667, 2018.

[41] Csardi Gabor and Nepusz Tamas. The igraph software package for complex network research. *InterJournal Complex Systems*, 1695, 2006.

[42] Francisco J. Azuaje. Selecting biologically informative genes in co-expression networks with a centrality score. *Biology Direct*, 9(1) :12, 2014.

[43] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5) :604–632, 1999.

[44] Leonard Kaufmann and Peter Rousseeuw. Clustering by Means of Medoids. *Data Analysis based on the L1-Norm and Related Methods*, pages 405–416, 1987.

[45] Erich Schubert and Peter J Rousseeuw. Faster k-Medoids Clustering : Improving the PAM, CLARA, and CLARANS Algorithms. In Giuseppe Amato, Claudio Gennaro, Vincent Oria, and Radovanović Miloš, editors, *Similarity Search and Applications*, pages 171–187, Cham, 2019. Springer International Publishing.

[46] Scott C. Ritchie, Stephen Watts, Liam G. Fearnley, Kathryn E. Holt, Gad Abraham, and Michael Inouye. A Scalable Permutation Approach Reveals Replication and Preservation Patterns of Network Modules in Large Datasets. *Cell Systems*, 3(1) :71–82, 2016.

[47] Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero : calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

[48] Peter Langfelder, Rui Luo, Michael C. Oldham, and Steve Horvath. Is my network module preserved and reproducible ? *PLoS Computational Biology*, 7(1), 2011.

[49] Bing Li, Yingying Zhang, Yanan Yu, Pengqian Wang, Yongcheng Wang, Zhong Wang, and Yongyan Wang. Quantitative assessment of gene expression network module-validation methods. *Scientific Reports*, 5(1) :15258, December 2015.

[50] Tim O. Nieuwenhuis, Stephanie Y. Yang, Rohan X. Verma, Vamsee Pillalamarri, Dan E. Arking, Avi Z. Rosenberg, Matthew N. McCall, and Marc K. Halushka. Consistent RNA sequencing contamination in GTEx and other data sets. *Nature Communications*, 11(1), 2020.

[51] Judith Somekh, Shai S. Shen-Orr, and Isaac S. Kohane. Batch correction evaluation framework using a-priori gene-gene associations : Applied to the GTEx dataset. *BMC Bioinformatics*, 20(1) :1–10, 2019.

[52] Jennifer Murray, Robert H. Whitson, and Keiichi Itakura. Reduced prostaglandin I2 signaling in Arid5b2/2 primary skeletal muscle cells attenuates myogenesis. *FASEB Journal*, 32(4) :1868–1879, 2018.

[53] Yuri Okazaki, Jennifer Murray, Ali Ehsani, Jessica Clark, Robert H. Whitson, Lisa Hirose, Noriyuki Yanaka, and Keiichi Itakura. Increased glucose metabolism in Arid5b -/-skeletal muscle is associated with the down-regulation of TBC1 domain family member 1 (TBC1D1). *Biological Research*, 53(1) :1–14, 2020.

[54] Susan Gray, Mark W. Feinberg, Sarah Hull, Chay T. Kuo, Masafumi Watanabe, Sucharita Sen, Ana Depina, Richard Haspel, and Mukesh K. Jain. The Krüppel-like factor KLF15 regulates the insulin-sensitive glucose transporter GLUT4. *Journal of Biological Chemistry*, 277(37) :34322–34328, 2002.

[55] Liyan Fan, Paishiun N. Hsieh, David R. Sweet, and Mukesh K. Jain. Krüppel-like factor 15 : Regulator of BCAA metabolism and circadian protein rhythmicity. *Pharmacological Research*, 130 :123–126, 2018.

[56] Robi Tacutu, Daniel Thornton, Emily Johnson, Arie Budovsky, DIogo Barardo, Thomas Craig, Eugene DIana, Gilad Lehmann, Dmitri Toren, Jingwei Wang, Vadim E. Fraifeld, and Joaõ P. De Magalhães. Human Ageing Genomic Resources : New and updated databases. *Nucleic Acids Research*, 46(D1) :D1083–D1090, 2018.

[57] Thomas Craig, Chris Smelick, Robi Tacutu, Daniel Wuttke, Shona H. Wood, Henry Stanley, Georges Janssens, Ekaterina Savitskaya, Alexey Moskalev, Robert Arking, and João Pedro De Magalhães. The Digital Ageing Atlas : Integrating the diversity of age-related changes into a unified resource. *Nucleic Acids Research*, 43(D1) :D873–D878, 2015.

[58] Jonas Zierer, Cristina Menni, Gabi Kastenmüller, and Tim D. Spector. Integration of 'omics' data in aging research : From biomarkers to systems biology. *Aging Cell*, 14(6) :933–944, 2015.

[59] Andreas Kuehne, Janosch Hildebrand, Joern Soehle, Horst Wenck, Lara Terstegen, Stefan Gallinat, Anja Knott, Marc Winnefeld, and Nicola Zamboni. An integrative metabolomics and transcriptomics study to identify metabolic alterations in aged skin of humans in vivo. *BMC Genomics*, 18(1) :169, 2017.

[60] Carlos López-Otín, Maria A. Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The hallmarks of aging. *Cell*, 153(6), 2013.

[61] João Pedro de Magalhães, João Curado, and George M. Church. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, 25(7) :875–881, 2009.

[62] Jun Dong and Steve Horvath. Understanding network concepts in modules. *BMC Systems Biology*, 1 :1–20, 2007.

[63] Roberto Anglani, Teresa M. Creanza, Vania C. Liuzzi, Ada Piepoli, Anna Panza, Angelo Andriulli, and Nicola Ancona. Loss of connectivity in cancer co-expression networks. *PLoS ONE*, 9(1), 2014.

[64] Felix Bormann, Manuel Rodríguez-Paredes, Sabine Hagemann, Himanshu Manchanda, Boris Kristof, Julian Gutekunst, Günter Raddatz, Rainer Haas, Lara Terstegen, Horst Wenck, Lars Kaderali, Marc Winnefeld, and Frank Lyko. Reduced DNA methylation patterning and transcriptional connectivity define human skin aging. *Aging Cell*, 15(3) :563–571, 2016.

[65] Lucinda K. Southworth, Art B. Owen, and Stuart K. Kim. Aging Mice Show a Decreasing Correlation of Gene Expression within Genetic Modules. *PLoS Genetics*, 5(12) :e1000776, December 2009.

[66] Mariam El Assar, Javier Angulo, and Leocadio Rodríguez-Mañas. Oxidative stress and vascular inflammation in aging. *Free Radical Biology and Medicine*, 65 :380–401, 2013.

[67] Suchitra D. Gopinath and Thomas A. Rando. Stem Cell Review Series : Aging of the skeletal muscle stem cell niche. *Aging Cell*, 7(4) :590–598, 2008.

[68] Christopher J. Mann, Eusebio Perdiguero, Yacine Kharraz, Susana Aguilar, Patrizia Pessina, Antonio L. Serrano, and Pura Muñoz-Cánoves. Aberrant repair and fibrosis development in skeletal muscle. *Skeletal Muscle*, 1(1) :1–20, 2011.

[69] Nikola Gligorijević, Martina Zámorová Križáková, Ana Penezić, Jaroslav Katrlík, and Olgica Nedić. Structural and functional changes of fibrinogen due to aging. *International Journal of Biological Macromolecules*, 108 :1028–1034, 2018.

[70] Deisy Morselli Gysi, Andre Voigt, Tiago De Miranda Fragoso, Eivind Almaas, and Katja Nowick. wTO : An R package for computing weighted topological overlap and a consensus network with integrated visualization tool. *BMC Bioinformatics*, 19(1) :1–16, 2018.

[71] Pedro S.T. Russo, Gustavo R. Ferreira, Lucas E. Cardozo, Matheus C. Bürger, Raul Arias-Carrasco, Sandra R. Maruyama, Thiago D.C. Hirata, Diógenes S. Lima, Fernando M. Passos, Kiyoshi F. Fukutani, Melissa Lever, João S. Silva, Vinicius Maracaja-Coutinho, and Helder I. Nakaya. CEMiTool : A Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinformatics*, 19(1) :1–13, 2018.

[72] Bruno M. Tesson, Rainer Breitling, and Ritsert C. Jansen. DiffCoEx : A simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics*, 11, 2010.

[73] Deisy Morselli Gysi, Tiago de Miranda Fragoso, Fatemeh Zebardast, Wesley Bertoli, Volker Busskamp, Eivind Almaas, and Katja Nowick. Whole transcriptomic network analysis using Co-expression Differential Network Analysis (CoDiNA). *PLoS ONE*, 15(10 October) :1–28, 2020.

[74] Michael Watson. CoXpress : Differential co-expression in gene expression data. *BMC Bioinformatics*, 7 :1–12, 2006.

[75] Zhenjun Hu. Using VisANT to analyze networks. (SUPPL.45) :1–39, 2014.

[76] Shannon Paul, Markiel Andrew, Ozier Owen, Baliga Nitin S., Wang Jonathan T., Ramage Daniel, Amin Nada, Schwikowski Benno, and Ideker Trey. Cytoscape : A Software Environment for Integrated Models. *Genome Research*, 13(11) :2498–2504, 2003.

Chapitre 2 - Analyse trans-tissus par réseau de co-expression de gènes pour la détection de fonctions physiologiques communes et spécifiques au vieillissement

Gwenaëlle G. Lemoine[1], Marie Pier Scott-Boyer[2], Arnaud Droit[1,2]

1. Département de médecine moléculaire, Faculté de médecine, Université Laval, 2325 rue de l'Université, Québec G1V 0A6, Canada.
2. Centre de recherche du CHU de Québec-Université Laval, 2705 boulevard Laurier Québec, Québec G1V 4G2, Canada.

3.1 Résumé

Tout tissu du corps humain vieillit avec l'âge mais tous ne le font pas au même rythme. S'il existe des mécanismes communs au vieillissement tel que relevé par López-Otín, d'autres sont spécifiques à certains sous ensembles de tissus. Pour déterminer la plausibilité de détection de telles différences, les données de patients jeunes et âgés sur plusieurs tissus ont été analysées avec GWENA. En effectuant une analyse de co-expression différentielle sur chaque couple tissus/tranche d'âge, on a pu isoler ces différents mécanismes communs et spécifiques du vieillissement. En recoupant les gènes contenus dans un exemple de chacun, on a pu retrouver la topologie typique de désorganisation liée au vieillissement. L'étude des fonctions physiologique des gènes impliqués dans cette altération a enfin permis d'identifier de nouveaux gènes susceptibles de contribuer au vieillissement commun ou spécifique.

3.2 Introduction

La variation de la co-expression des gènes associée au vieillissement est une propriété qu'on a pu observer dans le muscle auparavant [1]. Grâce à elle, des gènes ont pu être isolés et leur fonction physiologique [2] confirmée comme liée au vieillissement du muscle [1]. L'étude de la topologie des réseaux de co-expression de gènes semble ainsi une méthode efficace pour la détection de gènes biomarqueurs du vieillissement dans un tissu.

Cependant, le vieillissement est un phénomène qui ne se limite pas au muscle dans un organisme et touche bien d'autres tissus en parallèle. Dans chacun d'eux, le transcriptome est un témoin des modifications liées à l'avancée dans l'âge [3]. Via leur analyse par expression différentielle et d'autres méthodes d'analyse de gènes discrètes [4], nombre de gènes (biomarqueurs) associés à l'âge ont pu être détectés dans les différents tissus du corps humain. Si pour certains de ces gènes on a déjà étudié leur relation avec des fonctions physiologiques, ce n'est pas le cas de la majorité. Suite à ces études fines de certains gènes, il est également rare bien qu'intéressant de connaître comment leur relation évolue plus globalement, au-delà des fonctions physiologiques considérées. Pour pouvoir investiguer chaque relation et impact de chaque gène, la méthode courante implique le plus souvent de réaliser des expérimentations d'inactivation ou sur-activation d'un ou plusieurs gènes (respectivement *knockout* et *knockin* en anglais). Le coût important de ces manipulations, tant en temps qu'en argent limite alors leur utilisation pour caractériser en masse les relations entre deux ou quelques gènes. L'information apportée par de tels liens permettrait pourtant d'aider à déterminer le type de modification observée [5]. Ainsi on pourrait estimer si l'activation ou répression du gène observé est le fruit d'une réponse à l'activation d'un autre gène voir d'un mécanisme [6] complet, ou plutôt l'origine d'un mécanisme avec ou sans autres gènes impliqués.

Ces questions ont d'autant plus d'intérêt que des études précédentes tendent à démontrer l'existence de bases communes au vieillissement tant en termes de fonctions physiologiques que de mécanismes [7]. On relève ainsi des signatures d'expression de gènes communes à plusieurs tissus dans leur état "âgé". Parmi les fonctions physiologiques détectées comme sur-exprimées par l'expression différentielle, on retrouve notamment des composantes de la réponse immunitaire ou inflammatoire et de la dégradation lysosomale. À l'opposé, dans les fonctions détectées comme sous-exprimées, on retrouve des fonctions associées à l'encodage de protéines mitochondriales ainsi que des gènes responsables de la production de différents types de collagène. En complément, bien que ne concernant pas tous les tissus du corps, d'autres signatures sont communes à des sous groupes de tissus partageant des propriétés tant moléculaires que cellules. On retrouve par exemple une accumulation de marques de réparation de l'ADN chez les tissus se renouvelant rapidement [8], un épuisement des capacités de régénération chez les tissus disposant d'une niche de cellules souches [9], des dommages à l'ADN plus présents chez les tissus soumis à des stress mécaniques [10].

L'analyse de co-expression de gènes a démontré sa capacité à aider à la détermination de nouveaux liens, de nouvelles voies de signalisation dans un tissu [11]. Fort de ces réalisations, on s'est interrogé quant à la capacité de la co-expression différentielle à détecter des acteurs du vieillissement non pas simplement entre deux tranches d'âges, mais ceux-ci transversalement à de multiples tissus humains. Les différences observées pourraient permettre, contrairement à l'analyse mono-tissu, d'explorer spécifiquement les phénomènes du vieillissement communs ou bien uniques aux différents tissus. Par cette démarche, on va ainsi montrer dans chacun de ces cas la détection de nouveaux gènes candidats et à prouver l'intérêt de la co-expression différentielle pour y parvenir. Une première phase de traitement des données issues de différents tissus a tout d'abord permis de s'assurer de leur comparabilité entre tissus. Après une détection des modules de co-expression, on a isolé les modules variant avec l'âge à l'aide d'une analyse de co-expression différentielle pour chaque tissu. Tous ont été recoupés entre eux pour sélectionner les gènes variants en commun. Deux branches ont alors été explorées avec un exemple détaillé dans chacune : les intersections de tissus présentant des phénomènes communs du vieillissement, et les intersections de tissus présentant des phénomènes de vieillissement spécifiques. Nos résultats montrent que dans chacune de ces branches, on est parvenu à une compréhension plus fine de la fonction de gènes jusqu'à là peu étudiés dans le vieillissement et même plus globalement pour certains.

3.3 Matériel et méthodes

3.3.1 Contextualisation des données

Les données de GTEx sont l'un des rares jeux de données à contenir autant de tissus sains différents et en grand nombre. Elles sont le regroupement d'échantillons prélevés sur 54 tissus (+1 tissu qui est en fait un assemblage de lignées cellulaires dérivées de patients atteints de leucémie myéloïde aiguë[12]) et 980 donneurs dans sa dernière version, la v8. Cette variété de tissus vise à être le plus représentatif possible des différents tissus chez l'humain au vu du coût de prélèvement et d'analyse de chaque échantillon. Les biopsies sont effectuées sur des donneurs décédés avec leur accord préalable, l'accord d'un proche ou l'accord du représentant légal. Elles sont réalisées sur 4 centres de collecte, puis analysées sur place ou transférées selon le tissu avec réfrigération durant le transport. Ces échantillons sont évalués sur plusieurs critères pour jauger leur admissibilité et échantillon non conforme est exclu de la cohorte. On retrouve :

— des critères cliniques : absence de contamination au VIH, absence de chimiothérapie dans les 2 ans, absence de transfusion sanguine dans les 48 h, etc.

— des critères anatomopathologiques : absence de tissu cancéreux, absence de pathologie tissulaire, etc.

— des critères analytiques : quantité de tissu prélevé suffisante, quantité d'ARN extrait final supérieur à 500 ng d'ARN total, nombre d'intégrité d'ARN ou RIN supérieur à 5.7) [13].

Sur la majorité des échantillons ont été effectué des séquençages de génome (*Whole Genome Sequencing*, WGS), des séquençages d'exome complet (*Whole Exome Sequencing*, WES), des séquençages de transcriptome (aussi appelé séquençage d'ARN, RNA-Seq), ainsi que des images de coupes histologiques colorées. Des données reformatées sont également mises à disposition telle que les locus de caractères quantitatifs (*quantitative trait loci*, QTL) et l'expression de gène qui est ce que l'on va utiliser ici. Ces données sont disponibles sur le site du consortium GTEx (https://gtexportal.org) accompagnées d'information sur le phénotype des échantillons. En raison du fort potentiel d'identification des donneurs, le phénotype donné publiquement est partiel, et le phénotype complet est disponible sur demande auprès de dbGaP après soumission d'un dossier de projet à renouveler chaque année (Annexe B).

Par ailleurs, tous les donneurs n'ont pas pu être prélevés pour l'ensemble des 54 tissus et ce sont en moyenne 23,4 tissus qui ont été prélevés sur la version 8 de l'étude GTEx. Certains tissus ont été priorisés lors des biopsies : tissu adipeux (sous-cutané), artère tibiale, cœur (ventricule gauche), poumon, muscle (squelettique), nerf tibial, peau (exposée au soleil), thyroïde, sang complet. Parmi les différents tissus biopsiés, il est à noter que certains sont issus d'un même organe et on a ainsi 31 organes biopsiés pour 54 tissus biopsiés en tout.

3.3.2 Sélection des tissus

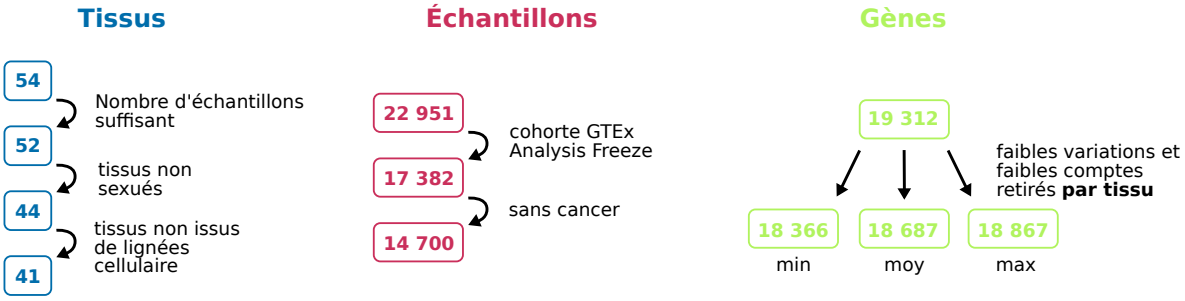

FIGURE 3.1 – Ensemble des filtres appliqués sur les différentes données et impact sur les données utilisées pour la construction ultérieure des réseaux de co-expression.

Le vieillissement est un phénomène dont les altérations moléculaires sont linéaires avec le temps, dont les dommages cellulaires sont super-linéaires [14], et où la mortalité associée augmente de façon exponentielle passée 20 ans [15]. Afin de faciliter la détection de ces altérations grâce à l'analyse de co-expression différentielle, on s'est donc dans un premier temps concentré sur une sélection de tranches d'âges très contrastées. Les échantillons sélectionnés sont issus de donneurs entre 20 et 30 ans pour la tranche qu'on nommera "**jeune**", et entre 60 et 70 ans pour la tranche qu'on nommera "**âgée**". Les données de GTEx ne comportent toutefois pas un nombre d'échantillons similaire pour chacune de ces tranches du fait de la mortalité plus importante chez

les personnes âgées que les personnes jeunes (Figure 3.2). Ce déséquilibre est principalement dû aux causes de décès pour chaque tranche d'âge avec des décès par traumatisme chez les jeunes plutôt que par maladie chronique ou maladie liée à l'âge chez les âgés. Un premier filtre de notre pré-traitement restreint donc la sélection des tissus à ceux comportant au minimum 50 échantillons dans chacune des tranches d'âge, ce qui n'est pas le cas de la Vessie (20) et le Rein - Médulla (3). Ce nombre d'échantillons permet d'assurer un bon compromis entre des réseaux de co-expression de gènes robustes, donc non sensibles à des valeurs aberrantes, et la perte de plus de tissus à étudier dans cette analyse multidimensionnelle qu'on souhaite effectuer [16].

Tous les tissus restant n'étaient pas nécessairement adaptés à l'étude globale du vieillissement chez l'humain. Ainsi on a retiré tous les tissus liés à un seul sexe : Trompes de Fallope, Col de l'utérus, Utérus, Vagin, Sein, Ovaire, Prostate, Testicule. À ce retrait s'ajoute celui des échantillons de lignées cellulaires (dérivées ou non de tissus eux conservés) car non représentatifs du vieillissement biologique : Cellules - Lymphocytes transformés par EBV, Cellules - Fibroblastes cultivés, Cellules - Lignée cellulaire de leucémie (CML).

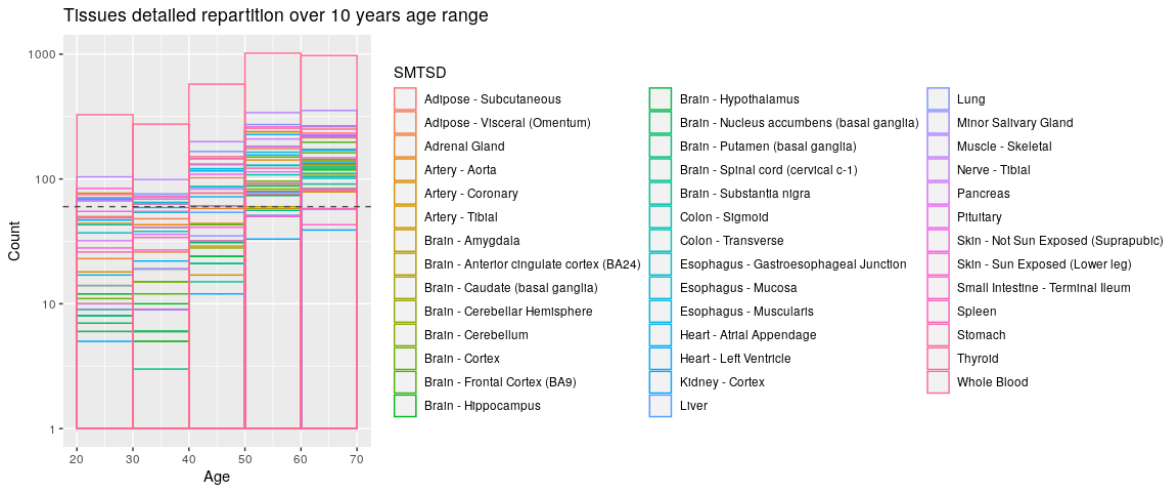

FIGURE 3.2 – Nombre d'échantillons disponibles par tissu et par tranche d'âge de 10 ans dans les données de GTEx.

3.3.3 Filtre des échantillons

En plus de ces sélections de tissus, certains échantillons ont directement nécessité une filtration afin de prévenir de potentiels biais qui pourraient altérer l'analyse de co-expression ou l'interprétation des modules en étant issus. Ainsi, seuls les échantillons répondant au critère d'inclusion dans la cohorte "GTEx Analysis Freeze" ont été retenus en premier lieu (Figure 3.1. Cette cohorte atteste que les échantillons n'étaient pas :

— issus de donneurs ayant des liens de parenté ou de donneurs avec des critères d'exclusion, par exemple des donneurs avec des duplications ou délétions chromosomiques

— affectés d'un syndrome tel que défini par la base de données OMIM [17], hors maladies liées au vieillissement

— ayant effectué une chirurgie de ré-assignation sexuelle

Par ailleurs, l'expression des tissus tumoraux a montré, dans des études préalables, des modèles d'expression génétique différents de ceux non tumoraux [18]. Par conséquent, les échantillons de ces donneurs ont également été supprimées ici ainsi que les échantillons dont le statut cancéreux était inconnu. La répartition finale des échantillons par tissu et par tranche d'âge est visible en Figure 3.3.

3.3.4 Filtre sur les gènes

On a considéré comme base dans cette étude uniquement les gènes codants pour une protéine (d'après GENCODE 26) sur les autosomes. Pour continuer à éviter l'influence du sexe, on a également limité les gènes des gonosomes à ceux du chromosome X. Afin de limiter l'ajout de biais d'origine technique, on a également exclu les gènes dont le compte de lecture (en anglais *count*) était inférieur à 6 dans la totalité des échantillons [19]. Les gènes n'ayant aucune variation se sont vu retirés du jeu de données car leur apport à la co-expression aurait été nul et leur conservation aurait entraîné l'utilisation de ressources inutile. Ces filtres ayant été appliqués par tissu, le nombre de gènes restant varie d'un tissu à un autre avec en moyenne 18687 gènes contre initialement 19 312 (Figure 3.1).

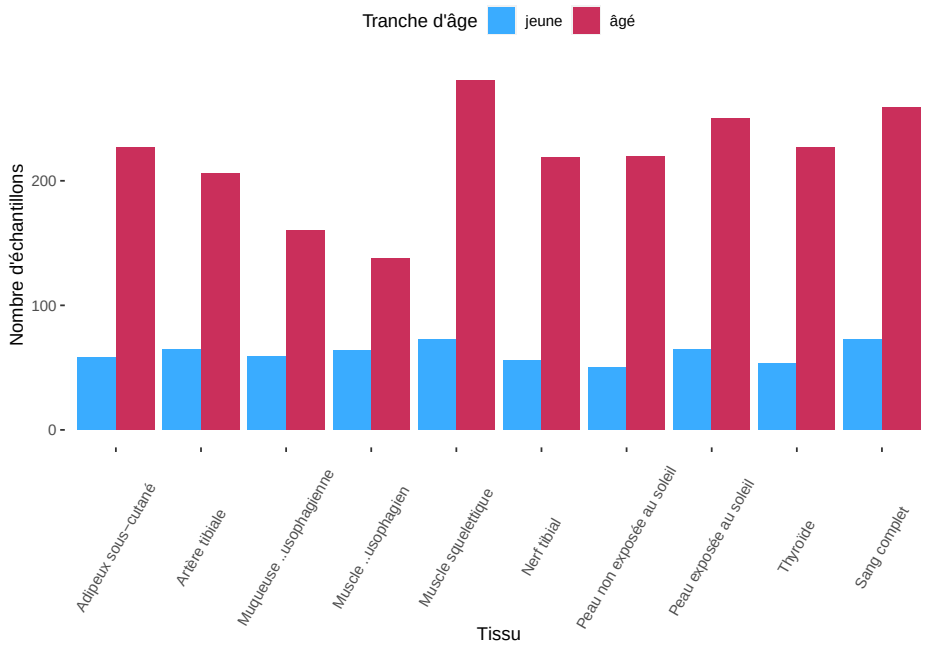

FIGURE 3.3 – Nombre d'échantillons disponibles par tissu dans les tranches d'âge jeune (20-30) et âgé (60-70) après filtration selon la cohorte et le statut cancéreux.

3.3.5 Correction des facteurs confondants

À l'instar de l'analyse d'expression différentielle, l'analyse de co-expression différentielle nécessite des données biaisées au minimum pour une construction de réseau de qualité. Ces biais (ou facteurs confondants) peuvent être tant techniques que biologiques et vont entraîner une augmentation de la variation de façon artificielle. Le risque est alors de détecter une différence artificielle et de l'assumer comme expérimentalement pertinente alors qu'elle est en fait due aux biais. Dans le cas de la co-expression plus spécifiquement, ces biais peuvent entraîner des corrélations erronées mais qui semblent crédibles entre certains gènes. Elles altèrent alors la construction du réseau de co-expression et par la suite la détection des modules qui se base sur un découpage en modules via les corrélations. Ainsi, il est essentiel de corriger ces facteurs confondants au préalable de l'utilisation du package R GWENA [1] sur les données comme on va le faire par la suite.

La complexité de la correction des facteurs confondants réside dans leur suppression sans pour autant altérer la distribution des données ou supprimer le signal expérimental d'intérêt, ici les variations dans le transcriptome dues à l'âge. Il est également important avec l'utilisation de GWENA de veiller à ce que la correction n'altère pas la topologie d'invariance d'échelle (*scale-free topology*) qu'on retrouve dans les réseaux de co-expression construits sur des données de transcriptomique [20]. Cette propriété assure que la distribution des degrés des nœuds suivent en fait une loi de puissance. Briser cette topologie perturberait la détection des modules dans le réseau car la méthode employée pour cela présuppose la présence d'une telle topologie pour fonctionner.

La base de données GTEx n'échappe pas aux biais [21, 22] et doit donc passer par cette étape sensible de correction des facteurs confondants dans le cadre de notre analyse de co-expression. Malgré le progrès de techniques ciblées sur des facteurs confondants connus comme l'effet de lot, l'effet de centre de prélèvement, le sexe, le poids, etc., celles-ci ne sont pas capables de corriger pour des facteurs peu explicites ou diffus comme la classe sociale, l'alimentation, la façon de manipuler des techniciens, etc. La correction par PC vise à répondre à ce type de problématique et a montré de bons résultats selon une évaluation par validation de voies de signalisation au sain de modules détectés [21]. Elle a également montré de meilleurs résultats que d'autres méthodes telles que la régression multiple, le taux exonique, ou encore le numéro d'intégrité d'ARN (*RNA integrity number*, ou RIN). Fait important pour la co-expression en particulier, cette méthode conserve la propriété d'invariance d'échelle des données requise pour la construction des modules de co-expression dans notre méthode.

Cependant, cette correction va également corriger l'âge qui est notre variable d'intérêt. Afin de conserver cette information tout en ayant un effet de correction par PC sur les autres facteurs confondants, on a donc ajusté la méthode estimant le nombre n de PC par lequel corriger les jeux de données. Au lieu d'utiliser une procédure de permutation [23], l'estimation de n s'est faite

Tissu	Nombre de PC corrigées	Nombre d'échantillons	
		Jeune	Âgé
Adipeux sous-cutané	1	58	227
Artère tibiale	3	65	206
Muqueuse œsophagienne	1	59	160
Muscle œsophagien	3	64	138
Muscle squelettique	5	73	281
Nerf tibial	4	56	219
Peau non exposée au soleil	4	50	220
Peau exposée au soleil	3	65	250
Thyroïde	3	53	227
Sang complet	1	73	259

TABLE 3.1 – Résumé du nombre de composantes utilisées pour effectuer la correction de l'expression par tissu, ainsi que le nombre d'échantillons inclus dans chacun pour les deux tranches d'âge.

en deux étapes :

— Un test de corrélation de chaque gène avec l'âge associé à chaque échantillon (donc patient) en fonction de différent n PC corrigées donne une liste de gènes significativement associés à l'âge.

— Cette liste de gènes par n PC corrigées est ensuite croisée avec deux bases de données de gènes connus comme étant associés au vieillissement : GenAge [24] et Digital Aging Atlas [25]. Y sont alors compté le nombre de gènes significatifs recoupés.

Finalement, le n de PC à corriger retenu est celui où le nombre de gènes significatifs est le plus haut dans chaque base de données. En cas de divergence de ce n entre les bases de données, on a sélectionné la valeur la plus basse de n afin de ne pas risquer de corriger l'âge (Table 3.1).

3.3.6 Construction du réseau, détection des modules et co-expression différentielle

Une fois les données pré-traitée, nous avons généré un réseau de co-expression de gènes indépendamment sur chacun des 10 tissus et tranche d'âge à l'aide du package GWENA [1]. Les paramètres spécifiés étaient ceux par défaut de la fonction `build_net` à l'exception de la corrélation qui était basée sur Spearman et du seuil d'ajustement de la loi de puissance fixé à 0,8. Dans chaque réseau de tissu, on a ensuite détecté le nombre de modules via la fonction `detect_modules` selon un seuil de coupe de l'arbre hiérarchique identique dans chaque tissu et une taille de module minimale fixée à 20 gènes. Chaque ensemble de modules a ensuite été testé à l'aide de la fonction `compare_conditions` par tissu entre les deux tranches d'âges pour identifier les modules préservés, modérément préservés (MP) et non préservés (NP) avec le vieillissement.

3.3.7 Investigation des phénomènes communs ou spécifiques du vieillissement

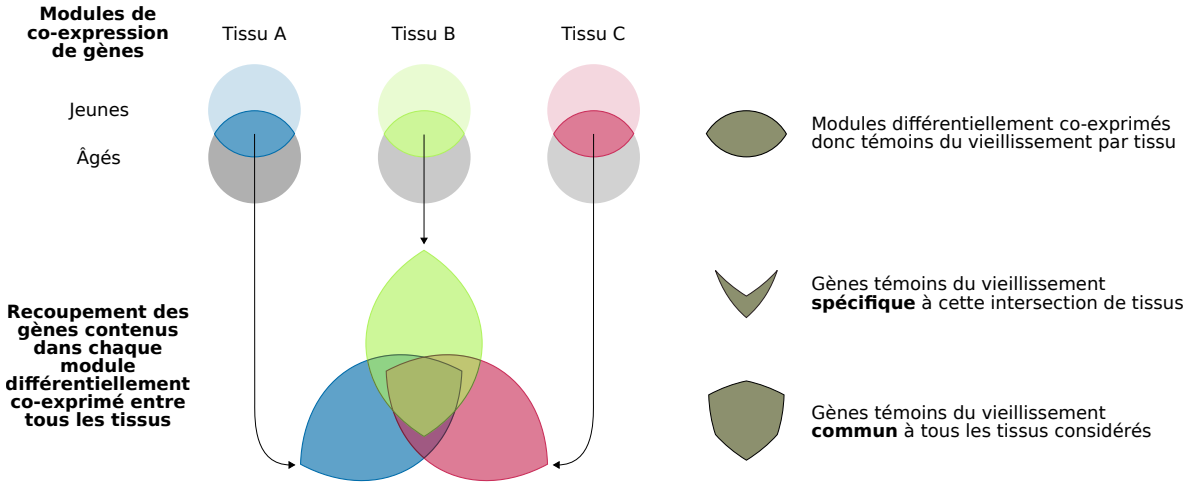

FIGURE 3.4 – Schéma simplifié, avec trois tissus seulement, de la méthode de construction des intersections spécifiques et communes du vieillissement.

Dans chaque tissu on a isolé les modules de co-expression modérément préservés (MP), non préservés (NP) et non conclusifs (NC) entre les deux tranches d'âge. On a ensuite effectué une intersection de leur contenu en gènes entre tous les tissus afin d'isoler les gènes participant à des phénomènes communs à plusieurs tissus et ceux participant à des phénomènes plus spécifiques d'un sous ensemble de tissus partageant des propriétés biologiques (Figure 3.4). Tous ont ensuite été enrichis via la fonction `bio_enrich` de GWENA et leurs résultats revus en regard de la littérature sur le vieillissement. Pour les plus pertinents d'entre eux, un résumé des termes Gene Ontology selon leur similarité sémantique a été réalisé à l'aide du package R rrvgo [26] qui reprend le principe de l'outil REVIGO [27].

3.4 Résultats

3.4.1 Répartition des gènes en fonction de la tranche d'âge et du tissu

La répartition des gènes entre les modules est hétérogène, tant entre les tranches d'âge qu'entre les tissus (Figure 3.5), avec une moyenne à 26 modules tout confondu (25.2 pour les âgés et 26.8 pour les jeunes). Le module 0 présent sur la figure regroupe les gènes sans association à un module. Celui-ci est quasi systématiquement plus grand dans la tranche âgée que jeune, à l'exception des tissus de la thyroïde et du sang complet. Les tissus où cet écart est le plus creusé sont ceux disposant du plus grand nombre de modules dans la tranche d'âge jeune. Ces résultats sont cohérents avec la tendance à la perte de co-expression constatée dans les tranches d'âge âgées [28] : la perturbation des voies de signalisation au cours du vieillissement entraîne une diminution de la co-expression entre gènes de ces voies.

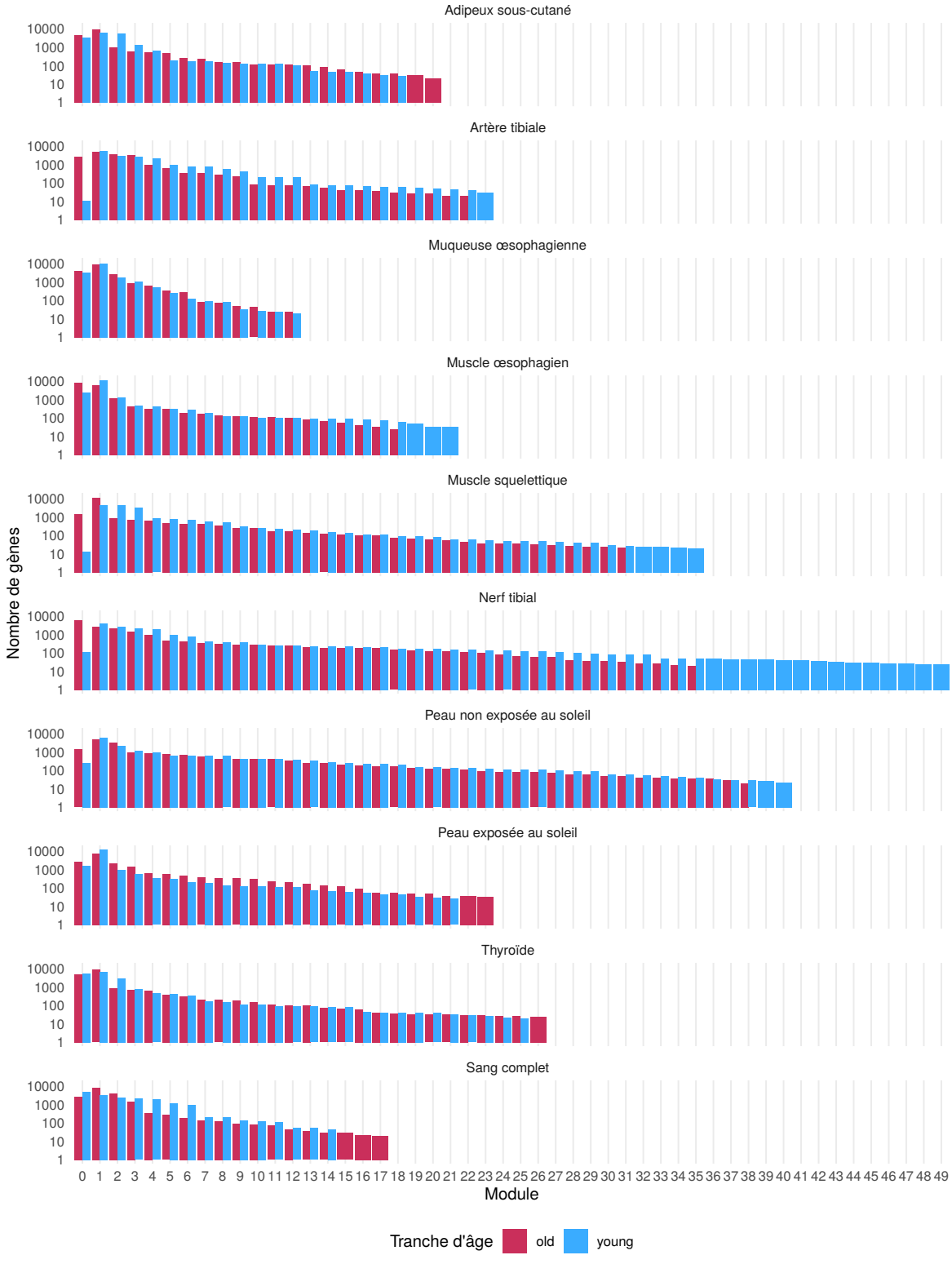

FIGURE 3.5 – Répartition des gènes (en échelle log10) pour chaque tissu entre les deux tranches d'âge jeune (bleu) et âgée (rouge)

3.4.2 Modules associés au vieillissement et recoupement inter-tissus

Le but étant de rechercher des origines communes au vieillissement entre tous les tissus, on a ensuite effectué une étape de co-expression différentielle intra tissu. La tranche d'âge jeune a été prise pour référence, c'est-à-dire que le test indiquera si chaque module détecté dans la tranche d'âge jeune est préservé, MP, (NP, ou (NC) . Les résultats ont été résumés en Table 3.2.

On y constate que certains tissus tendent à avoir proportionnellement moins de modules préservés lors du vieillissement. Avec moins de 50 % de préservation on retrouve notamment le muscle œsophagien et squelettique, ainsi que de la peau exposée au soleil et la thyroïde. À cela s'ajoute que la peau exposée au soleil et le muscle squelettique sont les deux tissus ayant le plus de modules NP proportionnellement à leur nombre de modules.

Tissu	**Préservé**		**Modérement préservé**		**Non préservé**		**Non concluant**		**Total**
	#	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**
Adipeux sous-cutané	12	67	4	22	0	0	2	11	18
Artère tibiale	13	57	8	35	1	4	1	4	23
Muqueuse œsophagienne	3	25	8	67	1	8	0	0	12
Muscle œsophagien	9	43	6	29	0	0	6	29	21
Muscle squelettique	14	40	13	37	5	14	3	9	35
Nerf tibial	35	71	9	18	1	2	4	8	49
Peau non exposée au soleil	36	90	4	10	0	0	0	0	40
Peau exposée au soleil	10	48	8	38	2	10	1	5	21
Thyroïde	12	48	8	32	1	4	4	16	25
Sang complet	9	64	4	29	0	0	1	7	14

TABLE 3.2 – Nombre (#) et ratio (%) de modules par statut de préservation selon chaque tissu.

Afin d'observer des similarités de mécanismes du vieillissement entre ces différents tissus, on a dans un premier temps regardé si des gènes communs existaient entre les modules MP et NP. Le diagramme UpSet [29] visible en Figure 3.6 permet de constater le faible recouvrement entre les gènes contenus dans les modules MP et NP. Ce diagramme révèle également que le maximum de tissus avec des gènes communs est 5. Ce chiffre est à mettre en contraste avec le fait que les modules créés par GWENA pour un même tissu ne sont pas chevauchant. Un gène classé dans un module ne sera donc pas également classé dans un autre et on aurait pu espérer une intersection au maximum de taille 10 étant donné qu'on dispose de 10 tissus.

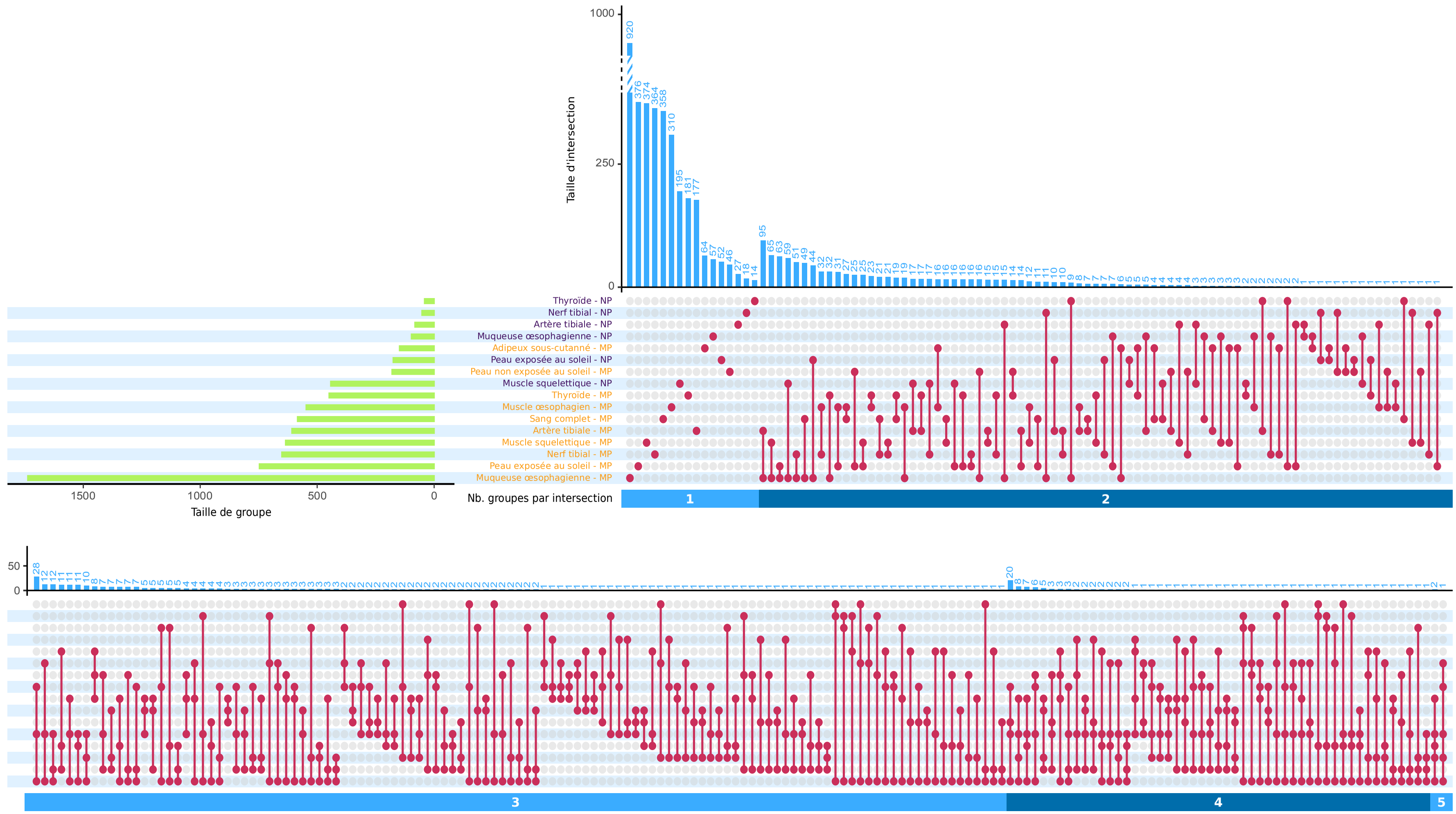

FIGURE 3.6 – Intersections entre tous les jeux de gènes pour chaque couple tissu / statut de préservation. Le diagramme Upset est une représentation qui vise à remplacer une visualisation par diagramme de Venn, ce dernier n'étant pas adaptée à la représentation de plus de 5 catégories. La matrice d'intersection (en orange) indique quel couple tissu / statut est considéré dans l'intersection (rose si pris en compte, gris si non) et quels sont les autres tissus dans cette intersection (ligne rose). L'histogramme des tailles de groupe (gris) à gauche de la matrice indique le nombre total de gène contenu dans le couple tissu / statut en ligne. L'histogramme des tailles d'intersection indique le nombre de gènes contenu dans l'intersection indiquée en matrice d'intersection pour cette colonne.

En plus de ce nombre maximal de tissus présent dans une intersection, certains tissus vont se retrouver plus fréquemment dans ces intersections en raison de leur nombre de modules MP/NP ou de modules contenant plus de gènes (Table 3.2). Ainsi, les modules MP de la muqueuse œsophagienne regroupent 1750 gènes et tendent à se positionner dans plus d'intersection que la moyenne. Une séparation est toutefois visible entre les deux types de statut MP et NP dans les intersections au-delà de 2 tissus et avec un minimum de trois gènes.

3.4.3 Répartition des phénomènes communs liés au vieillissement dans plusieurs tissus

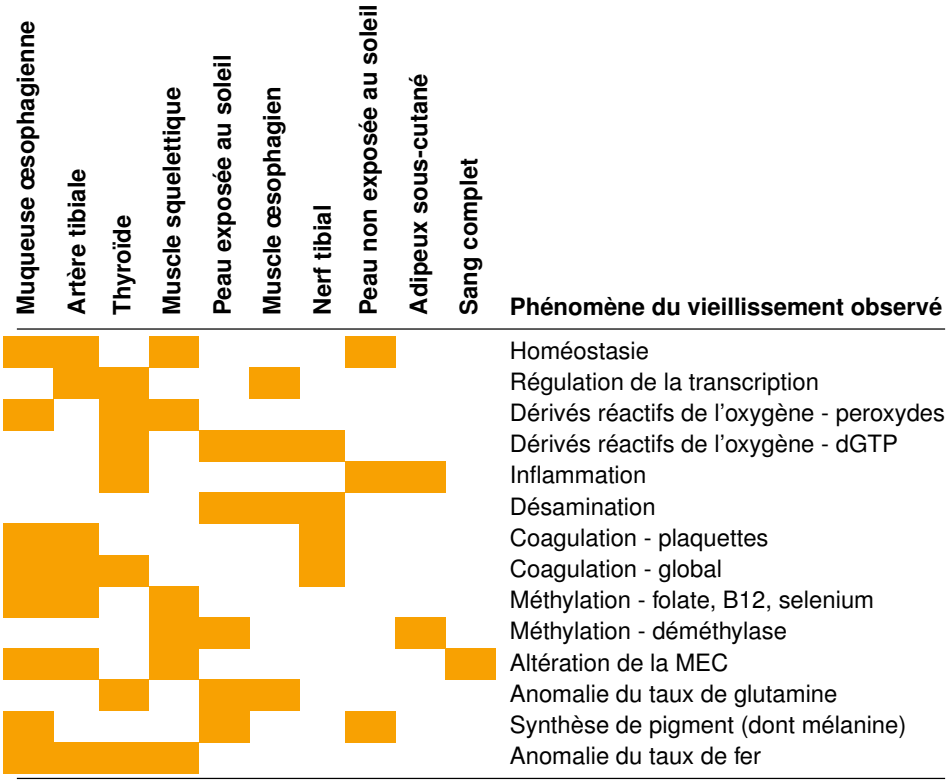

TABLE 3.3 – Phénomènes connus dans le vieillissement et observés dans les intersections de modules MP et NP pour chaque tissu. Les modules MP et NP ont été joints par tissus et les tissus ont été ordonnés par nombre décroissant de phénomènes du vieillissement présent.

Pour mieux comprendre les fonctions physiologiques communes en jeu et pouvant être impliquées dans le vieillissement, les intersections de plus de 3 tissus et ayant au moins 5 gènes ont été enrichies via GWENA. Comme visible en Annexe C dans les tables des enrichissements obtenus, 5 intersections n'ont pas donné d'enrichissement significatif malgré un nombre de gènes équivalent à d'autres intersections avec enrichissement. Les 18 autres intersections ont quant à elles retourné des fonctions physiologiques (Gene Ontology [30], CORUM [31]) et voies d'activation (KEGG [32], REACTOME [33], WikiPathways[34]) connues comme faisant partie du vieillis-

sement global, ou de manifestations spécifiques à plusieurs tissus. Ainsi on retrouve, dans ces intersections de modules peu ou pas préservés dans la tranche âgée, des fonctions associées à l'homéostasie (Figure 3.7.A.), la régulation de la transcription (Figure 3.7.B.), la gestion des dérivés réactifs de l'oxygène (Figure 3.7.C.), l'inflammation (Figure 3.7.D.), etc. (Table 3.3). La muqueuse œsophagienne bien qu'ayant un nombre de gènes plus élevé que les autres tissus ne se retrouve pas sur-représentée dans ces phénomènes observés. Cela indique donc que la majorité de ses gènes se trouvent dans les intersections de faible taille, tant en gènes qu'en nombre de tissus recoupés. Le sang complet, le tissu adipeux sous-cutané et la peau non exposée au soleil sont eux les tissus les moins présents dans ces intersections avec des phénomènes liés au vieillissement. Le sang complet est notablement absent des intersections avec des phénomènes liés à la coagulation. Les tissus ayant un fort renouvellement (voir en Table 1.3) que sont les épithéliums (peau exposée au soleil, artère tibiale, muqueuse œsophagienne) sont porteurs d'un grand nombre de phénomènes, ainsi que la thyroïde et le muscle squelettique.

FIGURE 3.7 – Exemple de résumé des enrichissements sur GO pour 4 des intersections par carte proportionnelle. Chaque ensemble de GO terme partageant une ontologie parente commune (selon un score de similarité) est groupé sous une même couleur. Chaque carte présente ici un phénomène connu dans le vieillissement : A. Homéostasie (GO *Biological Process*), B. Régulation de la transcription (GO *Biological Process*), C. Dérivés réactifs de l'oxygène (GO *Molecular Function*), D. Inflammation (GO *Biological Process*)

Des anomalies phénotypiques ont été relevées (via enrichissement sur HPO [35]) et présentent

un lien parfois direct avec les altérations physiologiques précédemment relevées, et parfois plus éloigné car issues d'une chaine d'événements et réactions physiologiques. À ces anomalies du taux de fer détectés dans les fonctions physiologiques coïncident ainsi des phénotypes d'anémie (déficience en taux de globules rouges ou en concentration d'hémoglobine) ou de défaut de coagulation avec hémorragies diverses (épistaxis, saignement gingival, menstruations hors cycle, hémorragie cérébrale). De même, les tissus présentant une altération de la MEC ont été associés avec des taux élevés de protéine C-réactive et d'anomalies de la fonction exocrine du pancréas ainsi que de malabsorption des lipides et ce qui en découle (stéatorrhée, pancréatite, thrombose veineuse). Ces résultats confirment le lien entre vieillissement et dérèglement progressif de l'homéostasie entrainant en cascade des pathologies de trouble de la coagulation [36, 37] et de la dégradation des lipides circulants [38, 39]. Cependant d'autres résultats semblent moins intuitifs comme dans le cas de pathologies liées au sexe (azoospermie, maladie d'hérédité gonosomale) relevées parallèlement à la déméthylation dans le tissu adipeux, le muscle squelettique et la peau non exposée au soleil. Ces informations, en l'absence d'erreur technique, semblent continuer de montrer la complexité des relations entre pathologies et vieillissement.

Enfin, on a exploré le vieillissement et son phénomène de perturbation de la régulation de la transcription en évaluant l'enrichissement de nos intersections en éléments de régulation via les bases de données MiRTarBase [40] et TRANSFAC [41]. L'intersection présentant des phénomènes de coagulation globale est ainsi fortement enrichi en micro ARN (miARN ou en anglais miRNA) qui ont un rôle de régulateurs post-transcriptionnels. Les miRNA détectés ici sont notamment associés aux 3 gènes de la génération de fibrinogène FGA, FGB, FGG détectés précédemment dans un phénomène de compensation du vieillissement [1]. Conjointement, deux facteurs de transcription, HNF1A et son isoforme HNF1B, sont associés à cette intersection liée à la coagulation globale. Si la perturbation des gènes de la fibrinogénèse est connue dans le vieillissement, l'impact d'une altération parallèle de ces TF HNF1A et HNF1B est peu étudié. Ils sont pourtant connus comme impliqués dans des malformations des canaux biliaires et des artères [42]. L'intersection associée à l'inflammation quant à elle ne présente qu'un seul miARN agissant ici sur RSAD2, IFI44, OASL, et EPSTI1 présents dans l'intersection. RSAD2 et OASL sont également connus pour être stimulé par des interférons (IFN) de type I (α et β) qui ont entre autres IRF1 (pour *IFNeron related factor 1*) pour FT [43]. Ce facteur IRF1 fait par ailleurs parti de la liste des FT détectés dans cette intersection. On y trouve un large panel d'autres membres de la famille d'IRF (de 1 à 9 à l'exception d'IRF6) dont l'implication dans l'inflammation s'exerce tant dans la réaction innée qu'acquise [44]. Mais tous les FT détectés via enrichissement n'étaient pas des IRF. Bien que n'étant pas directement liés à la régulation des IFN, STAT2 est présent dans la voie de signalisation des IFN α et β détecté plus tôt (R-HSA-909733). Le dernier FT identifié, FOXP1, n'est quant à lui pas contenu dans cette voie, mais agit sur l'engagement des cellules souches mésenchymateuses et la sénescence au cours du vieillissement [45] et sur de la régulation d'autres cytokines, l'interleukine 1 et 12 d'après UniprotKB (Q9H334).

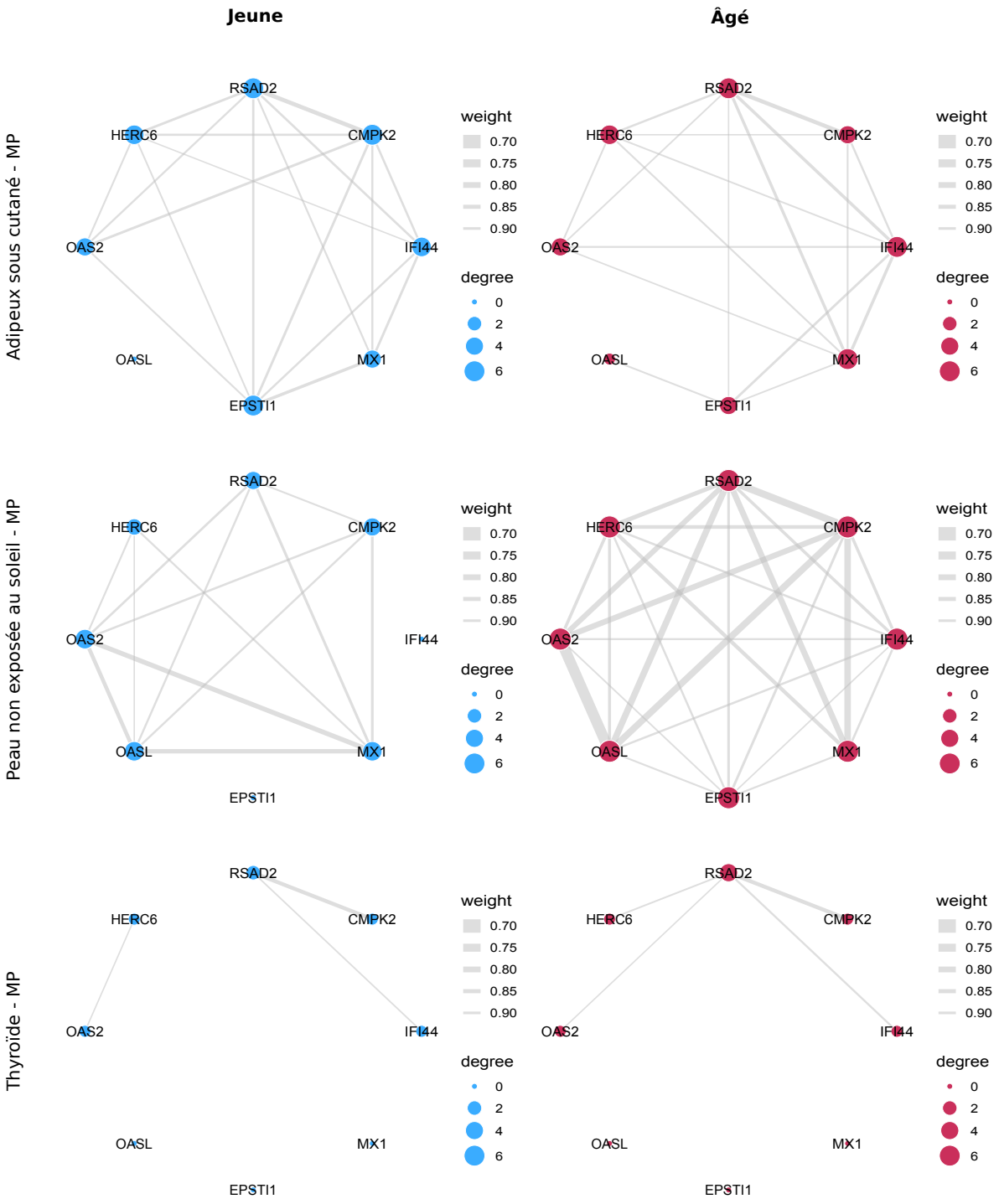

FIGURE 3.8 – Réseaux de co-expression des gènes de l'intersection associée au phénomène d'inflammation lors du vieillissement entre les deux tranches d'âge. Le réseau est filtré à 0.95 de dissimilarité (sur une échelle de 0 à 1) pour des questions de lisibilité et 3 tissus sont présentés : tissu adipeux, peau non exposée au soleil et thyroïde.

3.4.4 Les variation de co-expression dans l'intersection liée à l'inflammation

L'inflammation systémique chronique de faible intensité est une des marques connues du vieillissement [5, 36]. La complexité des mécanismes d'immunité et d'inflammation rend ce phénomène particulièrement difficile à étudier et les publications actuelles tendent donc à cibler peu d'acteurs en même temps (gène, protéine, miARN, etc.) et à se concentrer sur des contextes très précis [46]. Bien qu'apportant un niveau de preuve inférieur à toute validation expérimentale, les réseaux de co-expression de gènes permettent de comprendre bien plus d'acteurs. À titre d'exemple, on a souhaité ici détailler le cas de l'intersection de gènes associés à l'inflammation lors du vieillissement (Figure 3.7.D).

Les variations de motif dans la co-expression sont bien souvent la première cause de non-préservation des modules [47]. Des changements de score de centralité, des modifications de gène pivot (*hub gene*), des suppressions totales de signal sont bien souvent à l'origine de cette différence de topologie détectée entre les modules à l'aide de la co-expression différentielle [47, 20]. Cet effet se manifeste de plusieurs façons dans le cas de l'intersection sur l'inflammation comme visible en figure 3.8. Bien que la variation de la co-expression ne soit pas identique entre tous les tissus qui forment cette intersection, on observe des acteurs communs. Le gène RSAD2, précédemment mis en avant comme gène simulé par le facteur de transcription IRF1 et un miARN, devient ainsi un gène pivot dans la condition âgée selon le classement de score de centralité. Cependant, les sources de progression dans le classement selon la centralité divergent selon le tissu. Dans le tissu adipeux on constate que c'est dû à un renfort de la co-expression entre RSAD2 et IF44 associé à une perte partielle de co-expression globale par CMPK2, gène pivot dans la condition jeune. Dans la peau non exposée au soleil, on constate un renfort global de la co-expression avec toutefois une centralisation vers RSAD2. Une augmentation localisée entre OAS2 et OASL est également à noter. Enfin, dans la thyroïde, c'est une déconnexion entre HERC6 et OAS2 au profit de RSAD2 qui le rend gène pivot dans la condition âgée. Si cette réorganisation de la co-expression vers RSAD2 aurait pu être issue d'une contribution d'IRF1 ou tout autre IRF étant donné les informations précédentes, des analyses complémentaires ont précisé qu'aucun des IRF détectés comme FT auparavant n'était présent dans le module où se trouve RSAD2 (jeune comme âgé). L'origine de la variation de RSAD2 et OAS2/OASL pourrait donc être d'une autre nature qu'une modification de la transcription d'un ou plusieurs FT.

3.4.5 Cas particulier : le vieillissement spécifique à la peau

Si le vieillissement partage des mécanismes et altérations communes à travers plusieurs tissus, il existe parallèlement des altérations spécifiques à certains tissus. Bien que ne permettant pas l'enrichissement global de la compréhension du vieillissement, ces altérations locales nécessitent de s'y intéresser afin de prévenir et prendre en charge les pathologies qui s'y rattachent. En

étudiant les intersections ne présentant pas d'enrichissements liés aux mécanismes communs du vieillissement et en tenant compte des tissus inclus, on peut ainsi étudier l'impact du vieillissement spécifique du tissu.

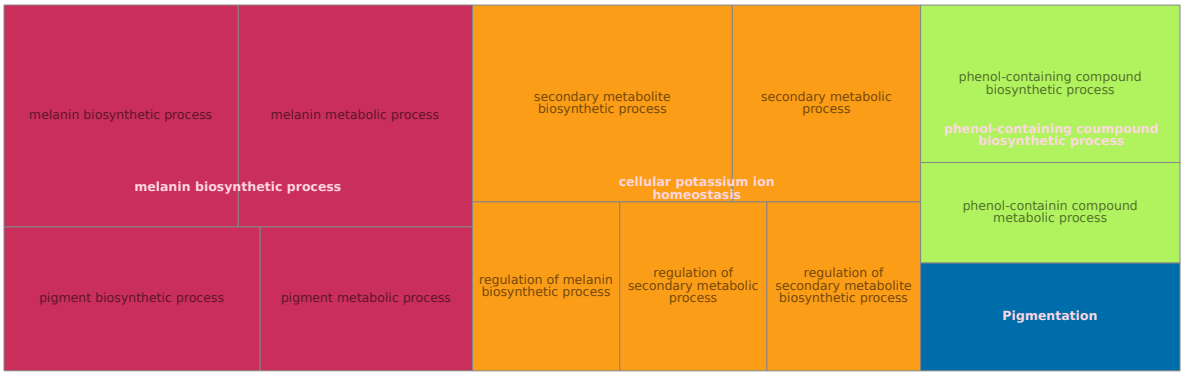

FIGURE 3.9 – Résumé des enrichissements sur GO par carte proportionnelle de l'intersection peau exposée au soleil, peau non exposée au soleil et muqueuse œsophagienne. Chaque ensemble de GO terme partageant une ontologie parente commune (selon un score de similarité) est groupé sous une même couleur.

Une intersection faite de la peau exposée au soleil, la peau non exposée au soleil et la muqueuse œsophagienne fait état dans notre cas d'un tel type de vieillissement. Elle est en effet composée principalement d'enrichissements sur des fonctions physiologiques directement liées à la synthèse et régulation de mélanine comme exposé en Figure 3.9. C'est le gène TYR, impliqué dans cette synthèse qu'on trouve dans le réseau de co-expression de notre intersection qui entraîne notamment cet enrichissement (Figure 3.10). Il est détecté comme gène pivot (score de centralité) dans chacun des tissus et chacune des tranches d'âge (Figure 3.10), attestant ainsi de son rôle majeur dans la régulation de la mélanogénèse. On constate qu'il est lié dans la condition jeune par une relation forte (dissimilarité < 0.8 en moyenne) avec un autre gène, SLC24A5. Ce gène est connu pour coder pour un échangeur de cations (sodium, potassium, calcium) de façon générale et a été plusieurs fois soupçonné d'une contribution à la régulation de la mélanogénèse [48, 49]. C'est notamment à lui que se rattachent les enrichissements de type *secondary metabolite biosynthetic process*. À ces deux gènes s'ajoute une relation de chacun avec CLEC12B dans le cas des deux tissus de peau et pas de la muqueuse œsophagienne. Ce gène appartient à la famille des récepteurs de lectine de type C qui jouent un rôle essentiel dans l'immunité et l'homéostasie à laquelle contribue la mélanine localement. La fonction de CLEC12B elle-même reste cependant encore mal caractérisée malgré la présence d'un motif d'inhibition basé sur la tyrosine d'un immunorécepteur dans son domaine intracellulaire. Ceci lui permettrait donc de recruter SHP-1/SHP-2 [50, 51], deux phosphatases suspectées d'avoir un rôle dans la progression des mélanomes [52].

La Figure 3.10 permet de constater que la relation TYR/SLC24A5 tend à augmenter avec l'âge dans la peau exposée ou non au soleil. Si l'augmentation est très marquée la peau non exposée au soleil, elle est bien moindre dans la peau non exposée. Ces résultats coïncident avec l'augmentation de la production de mélanine constatée chez la personne âgée pour compenser la

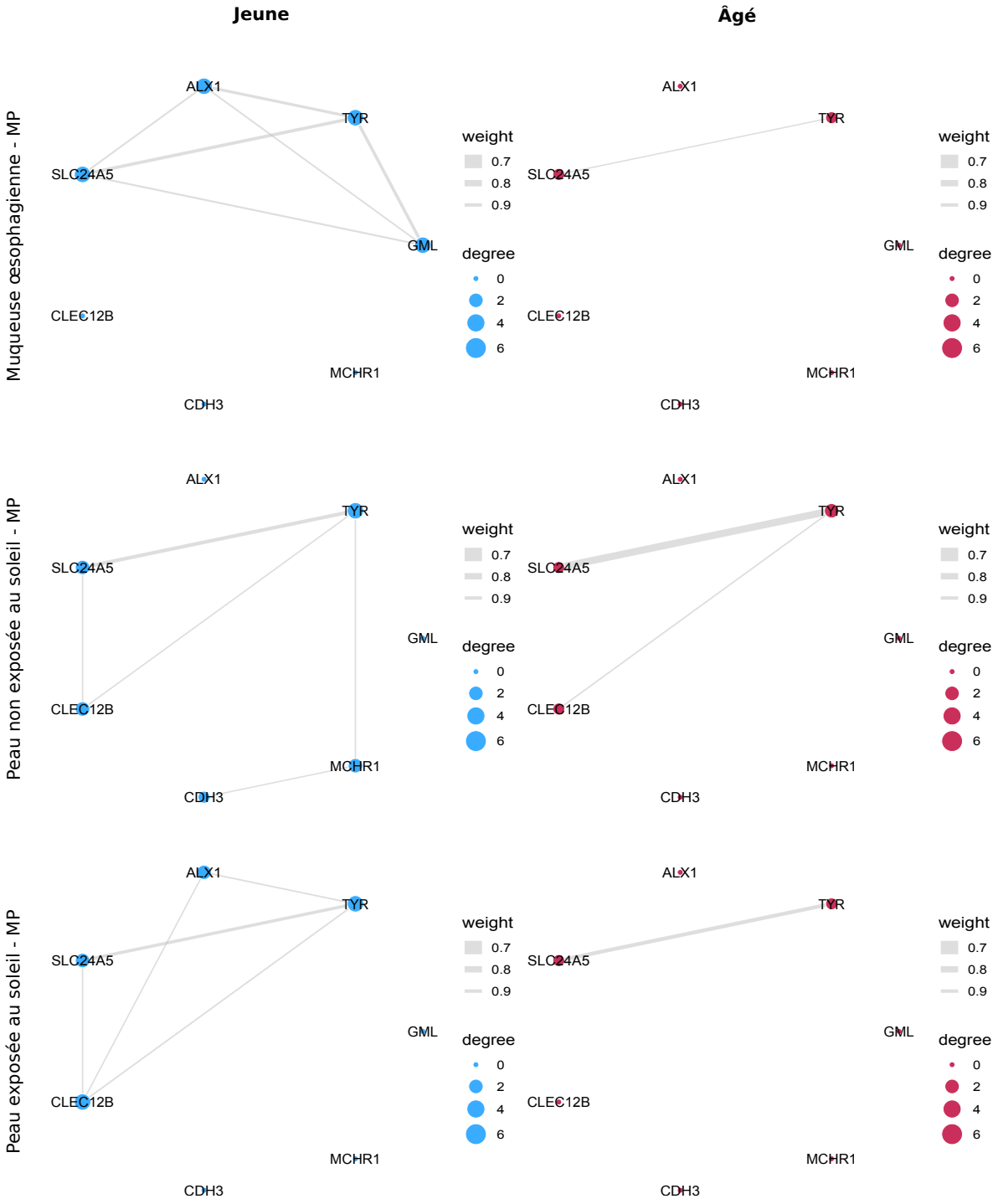

FIGURE 3.10 – Réseaux de co-expression des gènes de l'intersection associée à la surproduction de mélanine lors du vieillissement entre les deux tranches d'âge. Le réseau est filtré à 0.99 de dissimilarité (sur une échelle de 0 à 1) pour des questions de lisibilité et 3 tissus sont présentés : muqueuse œsophagienne, peau non exposée au soleil et peau exposée au soleil.

diminution globale de la densité de mélanocytes [53]. À l'inverse, la co-expression de CLEC12B avec SLC24A5 et TYR tend à diminuer avec l'âge, ainsi que l'intégralité de la co-expression de l'intersection.

La muqueuse œsophagienne elle semble interagir plus étroitement avec deux autres gènes, GML et ALX1 avec toutefois une interaction notable de la peau exposée au soleil avec ALX1 également. ALX1 est un gène jouant un rôle dans la migration cellulaire dans les structures craniofaciale lors du stade embryon, tandis que la fonction de GML est peu connu en dehors d'une potentielle contribution dans la voie d'activation de l'apoptose via p53 en cas de dommage à l'ADN. Comme visible en Figure 3.10, la relation entre ces gènes ainsi qu'avec le gène TYR tend à diminuer dans la tranche âgée, qu'il s'agisse de la muqueuse œsophagienne comme de la peau exposée au soleil.

3.5 Discussion

3.5.1 La susceptibilité des tissus aux variations liées au vieillissement

Le vieillissement peut être vu comme une imbrication de multiples phénomènes qui découlent d'une dérégulation du fonctionnement physiologique ou d'une réponse à cette dérégulation. En prenant en compte le système dans sa quasi-intégralité, les réseaux de co-expression permettent d'acquérir de nouvelles connaissances au-delà d'acteurs transcriptomiques individuels connus. Les variations au cours du temps des relations de co-expression entre gènes sont alors les témoins d'autant de voies d'activations qui s'altèrent. Leur suivi et mise en contraste au travers de la co-expression différentielle sont donc à leur tour un moyen supplémentaire de comprendre le déroulé des altérations [54, 28]. En comparant ces dernières entre de multiples tissus, il est possible d'en déduire des mécanismes communs au vieillissement, et à l'inverse ceux spécifiques. Dans cette étude, on s'est donc attachés à étudier ce vieillissement d'un point de vue multidimensionnel en comparant les réseaux issus de multiples tissus, et ce, entre deux tranches d'âge extrêmes.

Si le vieillissement tend à perturber de nombreuses fonctions physiologiques, il semble cependant que peu de gènes soient impactés ou impactant dans ce processus. Ainsi, comme cela a pu être observé auparavant [55], une minorité de gènes varie significativement entre les deux conditions (modules NP) et une proportion légèrement plus grande varie modérément (modules MP) (Figure 3.6). Ces gènes suffisent pourtant à représenter une large majorité des phénomènes connus du vieillissement : perte d'homéostasie, instabilité génomique, inflammation, dérivés réactifs de l'oxygène, etc. (Table 3.3 et Figure 3.7). Si leur répartition est inégale entre les tissus, ceux en portant le plus coïncident avec les tissus ayant un fort taux de renouvellement [8, 56, 57] (Table 3.3) à l'exception dans notre cas de la thyroïde dont le renouvellement cellulaire minimal

est de 8 ans [58]. Malgré une évidence des modifications que subit le système endocrinien avec le vieillissement, l'impact de celui-ci sur la thyroïde reste peu documenté hors pathologies graves [59]. En effet, les altérations endocrines thyroïdiennes (hypo ou hyperthyroïdies) sont souvent associées à un vieillissement "naturel" de la personne [60]. Il est donc présomptueux ici de s'essayer à trouver une cause d'autant de marques du vieillissement sans plus d'information. Pourtant, les indices d'un rôle de la thyroïde sur la longévité s'accumulent ces dernières années [61, 62]. Renforcer les connaissances sur les altérations de la fonction thyroïdienne chez la personne âgée semble donc une voie intéressante pour l'amélioration de la durée de vie chez l'homme.

3.5.2 La variation de la réponse inflammatoire issue du vieillissement

L'inflammation chronique de faible intensité est une caractéristique majeure du vieillissement qu'on nomme en anglais "*inflammaging*" [63, 64] et dont la complexité limite encore les connaissances à son sujet [46]. Si l'inflammation est un mécanisme bénéfique dans la réponse aigüe et temporaire à un pathogène, elle se retrouve délétère dans le cas du vieillissement malgré une faible intensité car elle est persistante. Ce phénomène assumé comme étant un précurseur de plusieurs autres phénomènes du vieillissement [5] s'est retrouvé ici plutôt exprimé dans le tissu adipeux, la peau non exposée au soleil, ainsi que la thyroïde. Les enrichissements fonctionnels sont venu préciser cette composante de l'inflammation comme une réponse à une réaction virale. En cause, de nombreux régulateurs d'IFN de type I significativement enrichis malgré leur absence de l'intersection. Ceci s'explique à contrario par la présence de nombreux gènes de réponse aux IFN, stimulés en temps normal par la présence d'ARN ou ADN viral.

Les IFN de type I (IFN-I) sont des cytokines regroupant tous les IFN à l'exception des IFN-γ et dont les IFN majoritaires sont les IFN-α et IFN-β. Ces IFN-I ont pour rôle, dans la réponse inflammatoire normale, d'entraver à la fois la réplication du virus et les cellules hôtes infectées. Pour cela, un large panel de gènes de réponse aux IFN, dont ceux détectés ici, est induit avec différents objectifs : interférer avec le trafic intracellulaire des vésicules, limiter la stabilité et la traduction des ARNm viraux, et entraîner une apoptose de la cellule au besoin via la voie d'activation de la protéine p53[44]. Cependant l'intervention de ce mécanisme antiviral dans le cadre du vieillissement est encore mal compris. Une hypothèse prometteuse est que des éléments transposables, plus particulièrement les LINE-1 (*long interspersed nuclear elements*) se retrouvent non réprimés avec l'âge en raison d'une dégradation des acteurs de leur répression (ex : des petits ARN ou *small RNA* (smRNA) en anglais) [65, 66, 44]. Ces LINE-1 codent alors pour diverses transcriptases inverses et autres protéines de rétro-transposition qui vont déclencher la réponse anti-virale associée à une sénescence [67]. S'il est admis que cette réponse varie de la "normale" antivirale, il reste toutefois difficile de comprendre comment les gènes antiviraux se comportent dans ce cadre non-viral [44] et la raison pour laquelle des marqueurs de la sénescence sont produits.

Dans le cadre de notre étude, on a constaté que le gène antiviral RSAD2 induit par les IFN officie

comme gène pivot de co-expression. Dans le cas de la peau non exposée au soleil, il collabore également étroitement avec deux autres gènes antiviraux de la famille OAS : OAS2 et OASL. Les gènes OAS ont déjà été auparavant associés à une réponse de type IFN-I induite par la sénescence et plus particulièrement l'accumulation d'ADN dans la cellule [66] qui est une composante délétère du vieillissement. Concernant RSAD2, aucun lien direct chez l'humain n'a été mis en avant à ce jour, seulement chez la souris [68]. La protéine produite par RSAD2 est connue pour limiter la réplication de l'ADN viral ou de l'ARN double brin viral bien que le mécanisme par lequel elle opère n'est pas entièrement clair [69]. En plus d'être un potentiel marqueur de cette accumulation d'ADN liée aux éléments transposables, RSAD2 pourrait donc être un potentiel gène candidat pour le développement de médicaments anti-âge.

3.5.3 L'altération de la régulation des mélanocytes et de la mélanogénèse avec l'âge

Au-delà des manifestations communes du vieillissement, chaque tissu est affecté de façon plus spécifique dans sa physiologie. L'altération avec l'âge de la mélanogénèse est un de ces changements qu'on retrouve en toute logique uniquement dans les tissus disposant de mélanocytes. La peau est l'organe en disposant du plus grand nombre et se retrouve ainsi d'autant plus affecté par cette altération avec l'âge. Dans notre intersection liée à la mélanogénèse, l'hyperpigmentation connue dans le vieillissement [70] s'est retranscrite par l'augmentation de la co-expression entre TYR et SLC24A5, respectivement catalyseur et régulateurs de la production d'eumélanine [71, 49], dans la peau exposée ou non au soleil. L'augmentation nettement plus grande dans la peau non exposée au soleil était toutefois inattendue étant donné que l'hyperpigmentation a plutôt été relevée dans les peaux exposées au soleil en raison de la stimulation de la mélanogénèse par les UV [53].

En plus d'une co-expression entre TYR et SLC24A5, une forte co-expression de CLEC12B a été montrée dans les deux tissus de peau. Une investigation de ce gène a révélé que peu d'information est connue sur ses fonctions et son implication dans la mélanogénèse. Des résultats préliminaires conséquents ont toutefois montré sa capacité à recruter les phosphorylases SHP-1/SHP-2 via son domaine ITIM [72], enzymes notamment connues pour leur implication dans la pigmentation dans le syndrome LEOPARD [73]. Ce nouveau gène impliqué dans la pigmentation de la peau coïncide ainsi avec les estimations d'un plus grand nombre de gènes contribuant au spectre de pigmentation que ceux actuellement connus et majeurs comme MC1R [74]. Contrairement à la co-expression entre TYR et SLC24A5, CLEC12B tend à perdre sa synergie avec l'âge. L'inactivation de CLEC12B ayant démontré une diminution de la pigmentation dans un modèle de peau humaine reconstruite [72], on peut alors se demander si ce gène ne serait pas impliqué dans le phénomène de dépigmentation global de la peau lors du vieillissement (phénomène parallèle à l'hyper-pigmentation locale). Plus précisément, CLEC12B serait impliqué dans les voies

d'inhibition de la prolifération des mélanocytaire par le biais de p53/p21/p27 [75] et on pourrait envisager alors que la diminution de la densité de mélanocytes avec l'âge serait en partie dû à la diminution de co-expression de CLEC12B avec TYR et SLC24A5. Une validation expérimentale par une limitation partielle de l'expression de CLEC12B pourrait être à même de confirmer cette hypothèse.

La muqueuse œsophagienne et la peau étant toutes deux des tissus dotés d'un épithélium squameux stratifié dérivé de la crête neurale [76], il est cohérent de les retrouver au sein d'une même intersection. Pierson et al. ont d'ailleurs montré la proximité de leur expression dans leur étude du jeu de données GTEx [77]. S'il est commun que la peau soit étudiée pour son contenu en mélanocytes, il est beaucoup moins courant que le contenu en mélanocytes de la muqueuse œsophagienne soit étudié. Le rôle de ce type cellulaire dans ce tissu reste ainsi encore mal compris en dehors des parallèles avec les mécanismes déjà connus dans la peau comme la réduction des dérivés réactifs de l'oxygène ou la fixation de certaines molécules organiques et ions métalliques [78].

Dans cette intersection côté muqueuse œsophagienne donc, la co-expression de TYR et SLC24A5 avec CLEC12B fait place à une co-expression croisée de TYR et SLC24A45A avec ALX1 et GML. GML est un homologue des protéines de membrane ancrées de type glycosyl-phosphatidylinositol (GPI), mais dont la fonction encore inconnue est suspectée d'être liée à l'apoptose et la régulation du cycle cellulaire [79]. Il a été constaté que son expression tend à ralentir la progression des cellules cancéreuses de l'œsophage et à augmenter la sensibilité des cellules cancéreuses à certaines chimiothérapies [80]. De son côté, ALX1 est un facteur de transcription impliqué dans la régulation de gènes liés au développement embryonnaire et à la migration cellulaire [81]. Il a été détecté comme sous exprimé chez des embryons de souris irradié entraînant un retard de pigmentation de l'épithélium rétinien. Pris ensemble, ces gènes pourraient donc être soupçonnés d'être à l'origine de la régulation de la prolifération des mélanocytes dans la muqueuse œsophagienne, bien qu'aucune direction de régulation ne puisse être extrapolée de nos réseaux de co-expression en tant que tel. En revanche, on peut avancer que si une telle régulation existe, elle est perturbée avec le vieillissement d'après la diminution notable de la co-expression entre TYR et SLC24A5 (donc contraire à l'évolution de la co-expression dans la peau) coïncidant avec la forte diminution de leur co-expression avec ALX1 et GML ainsi qu'entre eux deux. Ces altérations et les fonctions pour lesquelles codent ces gènes évoquent logiquement le développement de mélanomes qui sont parmi les pathologies associées au vieillissement et où les fonctions de régulation de la prolifération des mélanocytes par ALX1 et GML pourraient entrer en jeu. Toutefois, le mélanome œsophagien reste une pathologie rare, même si la fréquence des mélanomes œsophagiens d'origine métastatique est plus élevée que celle des mélanomes œsophagiens d'origine primitive. Il reste donc compliqué d'évaluer l'impact de la diminution de co-expression de ALX1 et GML ainsi que SLC24A5 et TYR dans le vieillissement de la muqueuse œsophagienne. Si des études sont menées ultérieurement sur ces gènes ou sur le contenu en mélanocytes de la mu-

queuse œsophagienne, elles pourront potentiellement donner de nouvelles perspectives à ces résultats.

3.6 Conclusion

Grâce à leur capacité d'étude à plus large échelle, les réseaux de co-expression sont un outil à privilégier pour l'exploration multi-tissus d'un phénomène aussi global et complexe que le vieillissement. Notre mise en évidence de groupes gènes impliqués dans les deux phénomènes du vieillissement sur lesquels on s'est concentrés est encourageant quant à l'utilisation de la co-expression différentielle et l'étude des motifs impliqués. Des tissus portant les mêmes gènes d'altérés peuvent ainsi être étudiées via la co-expression et permettre d'établir un lien direct entre les gènes et les marques du vieillissement. Des phénomènes communs comme spécifiques de différents tissus ont pu être mis en évidence. Des limitations restent toutefois présentes sur la nécessité pour se faire d'un grand nombre d'échantillons dans les tranches d'âge jeunes, ou en tout cas pas sur tout type de tissu. De même, une validation de ces résultats via expérimentalement reste un passage obligatoire et coûteux. Notre profilage par co-expression différentielle aura toutefois permis de réduire leur coût financier en ciblant un sous ensemble de gènes à étudier par rapport à un criblage à haut débit sans pré-sélection.

3.7 Références

[1] Gwenaëlle G. Lemoine, Marie-Pier Scott-Boyer, Bathilde Ambroise, Olivier Périn, and Arnaud Droit. GWENA : gene co-expression networks analysis and extended modules characterization in a single Bioconductor package. *BMC Bioinf.*, 22(1) :1–20, Dec 2021.

[2] Diane Marie Keeling, Patricia Garza, Charisse Michelle Nartey, and Anne-Ruxandra Carvunis. Philosophy of Biology : The meanings of 'function' in biology and the problematic case of de novo gene emergence. *eLife*, Nov 2019.

[3] João Pedro de Magalhães, Caleb E. Finch, and Georges Janssens. Next-generation sequencing in aging research : Emerging applications, problems, pitfalls and possible solutions. *Ageing Research Reviews*, 9(3) :315–323, 2010.

[4] Albert László Barabási and Zoltán N. Oltvai. Network biology : Understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2) :101–113, 2004.

[5] Carlos López-Otín, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The hallmarks of aging. *Cell*, 153(6), 2013.

[6] William Bechtel. Understanding Biological Mechanisms : Using Illustrations from Circadian Rhythm Research. *History, Philosophy and Theory of the Life Sciences*, 1 :487–510, 2013.

[7] João Pedro de Magalhães, João Curado, and George M. Church. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, 25(7) :875–881, 2009.

[8] Mary Armanios and Elizabeth H. Blackburn. The telomere syndromes. *Nature reviews. Genetics*, 13(10) :693–704, 2012.

[9] Mariusz Z. Ratajczak, Janina Ratajczak, Malwina Suszynska, Donald M. Miller, Magda Kucia, and Dong Myung Shin. A Novel View of the Adult Stem Cell Compartment from the Perspective of a Quiescent Population of Very Small Embryonic-Like Stem Cells. *Circulation Research*, 120(1) :166–178, 2017.

[10] Nard Kubben and Tom Misteli. Shared molecular and cellular mechanisms of premature ageing and ageing-associated diseases. *Nature Reviews Molecular Cell Biology*, 18(10) :595–609, 2017.

[11] Timothy R Hughes, Matthew J Marton, Allan R Jones, Christopher J Roberts, Roland Stoughton, Christopher D Armour, Holly A Bennett, Ernest Coffey, Hongyue Dai, Yudong D He, Matthew J Kidd, Amy M King, Michael R Meyer, David Slade, Pek Y Lum, Sergey B Stepaniants, Daniel D Shoemaker, Daniel Gachotte, Kalpana Chakraburtty, Julian Simon, Martin Bard, and Stephen H Friend. Functional Discovery via a Compendium of Expression Profiles. *Cell*, 102(1) :109–126, jul 2000.

[12] Gregory P. Way, Michael Zietz, Vincent Rubinetti, Daniel S. Himmelstein, and Casey S. Greene. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. *Genome Biology*, 21(1), May 2020.

[13] Latarsha J. Carithers, Kristin Ardlie, Mary Barcus, Philip A. Branton, Angela Britton, Stephen A. Buia, Carolyn C. Compton, David S. Deluca, Joanne Peter-Demchok, Ellen T. Gelfand, Ping Guan, Greg E. Korzeniewski, Nicole C. Lockhart, Chana A. Rabiner, Abhi K. Rao, Karna L. Robinson, Nancy V. Roche, Sherilyn J. Sawyer, Ayellet V. Segrè, Charles E. Shive, Anna M. Smith, Leslie H. Sobin, Anita H. Undale, Kimberly M. Valentino, Jim Vaught, Taylor R. Young, Helen M. Moore, Laura Barker, Margaret Basile, Alexis Battle, Joy Boyer, Debra Bradbury, Jason P. Bridge, Amanda Brown, Robin Burges, Christopher Choi, Deborah Colantuoni, Nancy Cox, Emmanouil T. Dermitzakis, Leslie K. Derr, Michael J. Dinsmore, Kenyon Erickson, Johnelle Fleming, Timothée Flutre, Barbara A. Foster, Eric R. Gamazon, Gad Getz, Bryan M. Gillard, Roderic Guigó, Kenneth W. Hambright, Pushpa Hariharan, Rick Hasz, Hae K. Im, Scott Jewell, Ellen Karasik, Manolis Kellis, Pouya Kheradpour, Susan Koester, Daphne Koller, Anuar Konkashbaev, Tuuli Lappalainen, Roger Little, Jun Liu, Edmund Lo, John T. Lonsdale, Chunrong Lu, Daniel G. MacArthur, Harold Magazine, Julian B. Maller, Yvonne Marcus, Deborah C. Mash, Mark I. McCarthy, Jeffrey McLean, Bernadette Mestichelli, Mark Miklos, Jean Monlong, Magboeba Mosavel, Michael T. Moser, Sara Mostafavi, Dan L. Nicolae, Jonathan Pritchard, Liqun Qi, Kimberly Ramsey, Manuel A. Rivas,

Barnaby E. Robles, Daniel C. Rohrer, Mike Salvatore, Michael Sammeth, John Seleski, Saboor Shad, Laura A. Siminoff, Matthew Stephens, Jeff Struewing, Timothy Sullivan, Susan Sullivan, John Syron, David Tabor, Mehran Taherian, Jorge Tejada, Gary F. Temple, Jeffrey A. Thomas, Alexander W. Thomson, Denee Tidwell, Heather M. Traino, Zhidong Tu, Dana R. Valley, Simona Volpi, Gary D. Walters, Lucas D. Ward, Xiaoquan Wen, Wendy Winckler, Shenpei Wu, Jun Zhu, Assya Abdallah, Anjene Addington, James M. Anderson, Patrick K. Bender, Mark Cosentino, Norma Diaz-Mayoral, Theresa Engel, Fernando Garci, Allen Green, Tiffanie Hammond, Katherine Jaffe, Judy Keen, Mary Kennedy, Peter Kigonya, Brent Lander, Sreenath Nampally, Cathy Ny, James Robb, Vikram Santhanum, Nataliya Sharopova, Shilpi Singh, Conrado Soria, Anne Sturcke, Surendra Sukari, Elizabeth J. Thomson, Magda Tomaszewski, Casandra Trowbridge, Ferdinand Udoye, David Vanscoy, Negin Vatanian, Elizabeth L. Wilder, and Penelope Williams. A Novel Approach to High-Quality Postmortem Tissue Procurement : The GTEx Project. *Biopreservation and Biobanking*, 13(5) :311–317, 2015.

[14] Michael E. Todhunter, Rosalyn W. Sayaman, Masaru Miyano, and Mark A. LaBarge. Tissue aging : the integration of collective and variant responses of cells to entropic forces over time. *Current Opinion in Cell Biology*, 54 :121–129, 2018.

[15] Caleb E. Finch and Eileen M. Crimmins. Constant molecular aging rates vs. the exponential acceleration of mortality. *Proceedings of the National Academy of Sciences*, 113(5) :1121–1123, jan 2016.

[16] Franziska Liesecke, Johan Owen De Craene, Sébastien Besseau, Vincent Courdavault, Marc Clastre, Valentin Vergès, Nicolas Papon, Nathalie Giglioli-Guivarc'h, Gaëlle Glévarec, Olivier Pichon, and Thomas Dugé de Bernonville. Improved gene co-expression network quality through expression dataset down-sampling and network aggregation. *Scientific Reports*, 9(1) :1–16, 2019.

[17] Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl_1) :D514–D517, 01 2005.

[18] Zefang Tang, Chenwei Li, Boxi Kang, Ge Gao, Cheng Li, and Zemin Zhang. GEPIA : a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Research*, 45(W1) :W98–W102, April 2017.

[19] David M. Rocke and Blythe Durbin. A model for measurement error for gene expression arrays. *Journal of Computational Biology*, 8(6) :557–569, November 2001.

[20] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[21] Princy Parsana, Claire Ruberman, Andrew E. Jaffe, Michael C. Schatz, Alexis Battle, and Jeffrey T. Leek. Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biology*, 20(1) :94, 2019.

[22] Tim O. Nieuwenhuis, Stephanie Y. Yang, Rohan X. Verma, Vamsee Pillalamarri, Dan E. Arking, Avi Z. Rosenberg, Matthew N. McCall, and Marc K. Halushka. Consistent RNA sequencing contamination in GTEx and other data sets. *Nature Communications*, 11(1), 2020.

[23] Andreas Buja and Nermin Eyuboglu. Remarks on Parallel Analysis. *Multivariate Behavioral Research*, 27(4) :509–540, Oct 1992.

[24] João Pedro De Magalhães and Olivier Toussaint. GenAge : A genomic and proteomic network map of human ageing. *FEBS Letters*, 571(1-3) :243–247, 2004.

[25] Thomas Craig, Chris Smelick, Robi Tacutu, Daniel Wuttke, Shona H. Wood, Henry Stanley, Georges Janssens, Ekaterina Savitskaya, Alexey Moskalev, Robert Arking, and João Pedro De Magalhães. The Digital Ageing Atlas : Integrating the diversity of age-related changes into a unified resource. *Nucleic Acids Research*, 43(D1) :D873–D878, 2015.

[26] rrvgo, Jun 2021. [Online ; accessed 22. Jun. 2021].

[27] Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS One*, 6(7) :e21800, Jul 2011.

[28] Lucinda K. Southworth, Art B. Owen, and Stuart K. Kim. Aging Mice Show a Decreasing Correlation of Gene Expression within Genetic Modules. *PLoS Genetics*, 5(12) :e1000776, dec 2009.

[29] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. UpSet : Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12) :1983–1992, 2014.

[30] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology : tool for the unification of biology. *Nature Genetics*, 25(1) :25–29, may 2000.

[31] Andreas Ruepp, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Michael Stransky, Brigitte Waegele, Thorsten Schmidt, Octave Noubibou Doudieu, Volker Stümpflen, and H. Werner Mewes. CORUM : The comprehensive resource of mammalian protein complexes. *Nucleic Acids Research*, 36(SUPPL. 1) :646–650, 2008.

[32] Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, 47(D1) :D590–D595, 2019.

[33] Antonio Fabregat, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, Steven Jupe, Florian Korninger, Sheldon McKay, Lisa Matthews, Bruce May, Marija Milacic, Karen Rothfels, Veronica Shamovsky, Marissa Webber, Joel Weiser, Mark Williams, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 44(D1) :D481–D487, 2016.

[34] Denise N. Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, Nuno Nunes, Jonathan Mélius, Elisa Cirillo, Susan L. Coort, Daniela DIgles, Friederike Ehrhart, Pieter Giesbertz, Marianthi Kalafati, Marvin Martens, Ryan Miller, Kozo Nishida, Linda Rieswijk, Andra Waagmeester, Lars M.T. Eijssen, Chris T. Evelo, Alexander R. Pico, and Egon L. Willighagen. WikiPathways : A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(D1) :D661–D667, 2018.

[35] Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O B Jacobsen, Daniel Danis, Jean Philippe Gourdine, Michael Gargano, Nomi L Harris, Nicolas Matentzoglu, Julie A McMurry, David Osumi-Sutherland, Valentina Cipriani, James P Balhoff, Tom Conlin, Hannah Blau, Gareth Baynam, Richard Palmer, Dylan Gratian, Hugh Dawkins, Michael Segal, Anna C Jansen, Ahmed Muaz, Willie H Chang, Jenna Bergerson, Stanley J F Laulederkind, Zafer Yüksel, Sergi Beltran, Alexandra F Freeman, Panagiotis I Sergouniotis, Daniel Durkin, Andrea L Storm, Marc Hanauer, Michael Brudno, Susan M Bello, Murat Sincan, Kayli Rageth, Matthew T Wheeler, Renske Oegema, Halima Lourghi, Maria G Della Rocca, Rachel Thompson, Francisco Castellanos, James Priest, Charlotte Cunningham-Rundles, Ayushi Hegde, Ruth C Lovering, Catherine Hajek, Annie Olry, Luigi Notarangelo, Morgan Similuk, Xingmin A Zhang, David Gómez-Andrés, Hanns Lochmüller, Hélène Dollfus, Sergio Rosenzweig, Shruti Marwaha, Ana Rath, Kathleen Sullivan, Cynthia Smith, Joshua D Milner, Dorothée Leroux, Cornelius F Boerkoel, Amy Klion, Melody C Carter, Tudor Groza, Damian Smedley, Melissa A Haendel, Chris Mungall, and Peter N Robinson. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, 47(D1) :D1018—-D1027, 2019.

[36] Massimo Franchini. Hemostasis and aging. *Critical Reviews in Oncology/Hematology*, 60(2) :144–151, November 2006.

[37] Kazuomi Kario and Takefumi Matsuo. Lipid-related hemostatic abnormalities in the elderly : imbalance between coagulation and fibrinolysis. *Atherosclerosis*, 103(2) :131–138, November 1993.

[38] Kazushi Yamamoto, Yasuna Kitano, Shuang E, Yu Hatakeyama, Yu Sakamoto, Taro Honma, and Tsuyoshi Tsuduki. Decreased lipid absorption due to reduced pancreatic lipase activity in aging male mice. *Biogerontology*, 15(5) :463–473, July 2014.

[39] G.M. Hirschfield and M.B. Pepys. C-reactive protein and cardiovascular disease : new insights from an old molecule. *QJM : An International Journal of Medicine*, 96(11) :793–807, November 2003.

[40] Chih Hung Chou, Sirjana Shrestha, Chi Dung Yang, Nai Wen Chang, Yu Ling Lin, Kuang Wen Liao, Wei Chi Huang, Ting Hsuan Sun, Siang Jyun Tu, Wei Hsiang Lee, Men Yee Chiew, Chun San Tai, Ting Yen Wei, Tzi Ren Tsai, Hsin Tzu Huang, Chung Yu Wang, Hsin Yi Wu, Shu Yi Ho, Pin Rong Chen, Cheng Hsun Chuang, Pei Jung Hsieh, Yi Shin Wu, Wen Liang Chen, Meng Ju Li, Yu Chun Wu, Xin Yi Huang, Fung Ling Ng, Waradee Buddhakosai, Pei Chun Huang, Kuan Chun Lan, Chia Yen Huang, Shun Long Weng, Yeong Nan Cheng, Chao Liang, Wen Lian Hsu, and Hsien Da Huang. MiRTarBase update 2018 : A resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research*, 46(D1) :D296–D302, 2018.

[41] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel : transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(Database issue) :108–110, 2006.

[42] Robert H. Costa, Vladimir V. Kalinichenko, Ai-Xuan L. Holterman, and Xinhe Wang. Transcription factors in liver development, differentiation, and regeneration. *Hepatology*, 38(6) :1331–1347, December 2003.

[43] John W Schoggins and Charles M Rice. Interferon-stimulated genes and their antiviral effector functions. *Current Opinion in Virology*, 1(6) :519–525, December 2011.

[44] Steven M. Frisch and Ian P. MacFawn. Type i interferons and related pathways in cell senescence. *Aging Cell*, 19(10), September 2020.

[45] Arantza Infante and Clara I. Rodríguez. Osteogenesis and aging : lessons from mesenchymal stem cells. *Stem Cell Research & Therapy*, 9(1), September 2018.

[46] Claudio Franceschi, Paolo Garagnani, Giovanni Vitale, Miriam Capri, and Stefano Salvioli. Inflammaging and 'garb-aging'. *Trends in Endocrinology & Metabolism*, 28(3) :199–212, March 2017.

[47] Scott C. Ritchie, Stephen Watts, Liam G. Fearnley, Kathryn E. Holt, Gad Abraham, and Michael Inouye. A Scalable Permutation Approach Reveals Replication and Preservation Patterns of Network Modules in Large Datasets. *Cell Systems*, 3(1) :71–82, 2016.

[48] Zhao Zhang, Juanjuan Gong, Elena V. Sviderskaya, Aihua Wei, and Wei Li. Mitochondrial NCKX5 regulates melanosomal biogenesis and pigment production. *Journal of Cell Science*, January 2019.

[49] Rebecca S. Ginger, Sarah E. Askew, Richard M. Ogborne, Stephen Wilson, Dudley Ferdinando, Tony Dadd, Adrian M. Smith, Shubana Kazi, Robert T. Szerencsei, Robert J. Winkfein, Paul P.M. Schnetkamp, and Martin R. Green. SLC24a5 encodes a trans-golgi network protein with potassium-dependent sodium-calcium exchange activity that regulates human epidermal melanogenesis. *Journal of Biological Chemistry*, 283(9) :5486–5495, February 2008.

[50] Sabrina C. Hoffmann, Carola Schellack, Sonja Textor, Stephanie Konold, Debora Schmitz, Adelheid Cerwenka, Stefan Pflanz, and Carsten Watzl. Identification of CLEC12B, an Inhibitory Receptor on Myeloid Cells $*$. *J. Biol. Chem.*, 282(31) :22370–22375, Aug 2007.

[51] Kazuya Tone, Mark H.T. Stappers, Janet A. Willment, and Gordon D. Brown. C-type lectin receptors of the dectin-1 cluster : Physiological roles and involvement in disease. *European Journal of Immunology*, 49(12) :2127–2133, November 2019.

[52] Tao Zhang, Wenjie Guo, Yang Yang, Wen Liu, Lele Guo, Yanhong Gu, Yongqian Shu, Lu Wang, Xuefeng Wu, Zichun Hua, Yuehai Ke, Yang Sun, Yan Shen, and Qiang Xu. Loss of SHP-2 activity in CD4+ t cells promotes melanoma progression and metastasis. *Scientific Reports*, 3(1), October 2013.

[53] Barbara A. Gilchrest, Frederick B. Blog, and George Szabo. Effects of aging and chronic sun exposure on melanocytes in human skin. *Journal of Investigative Dermatology*, 73(2) :141–143, August 1979.

[54] Roded Sharan and Trey Ideker. Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24(4) :427–433, April 2006.

[55] Roberto A. Avelar, Javier Gómez Ortega, Robi Tacutu, Eleanor Tyler, Dominic Bennett, Paolo Binetti, Arie Budovsky, Kasit Chatsirisupachai, Emily Johnson, Alex Murray, Samuel Shields, Daniela Tejada-Martinez, Daniel Thornton, Vadim E. Fraifeld, Cleo L. Bishop, and João Pedro de Magalhães. A multidimensional systems biology analysis of cellular senescence in ageing and disease. *bioRxiv*, pages 1–22, 2019.

[56] Nick Barker, Sina Bartfeld, and Hans Clevers. Tissue-resident adult stem cell populations of rapidly self-renewing organs. *Cell Stem Cell*, 7(6) :656–670, December 2010.

[57] C. P. Leblond and Bruce E. Walker. Renewal of cell populations. *Physiological Reviews*, 36(2) :255–276, April 1956.

[58] J. Coclet, F. Foureau, P. Ketelbant, P. Galand, and J. E. Dumont. CELL POPULATION KINETICS IN DOG AND HUMAN ADULT THYROID. *Clin. Endocrinol.*, 31(6) :655–666, Dec 1989.

[59] A. Faggiano, M. Del Prete, F. Marciello, V. Marotta, V. Ramundo, and A. Colao. Thyroid diseases in elderly. *Minerva Endocrinol.*, 36(3) :211–231, Sep 2011.

[60] David S. Cooper. Thyroid disease in the oldest old : the exception to the rule. *JAMA*, 292(21) :2651–2654, Dec 2004.

[61] Sabrina Garasto, Alberto Montesanto, Andrea Corsonello, Fabrizia Lattanzio, Sergio Fusco, Giuseppe Passarino, Valeria Prestipino Giarritta, and Francesco Corica. Thyroid hormones in extreme longevity. *Mech. Ageing Dev.*, 165 :98–106, Jul 2017.

[62] Beatrice Arosio, Daniela Monti, Daniela Mari, Giuseppe Passarino, Rita Ostan, Evelyn Ferri, Francesco De Rango, Claudio Franceschi, Matteo Cesari, and Giovanni Vitale. Thyroid hormones and frailty in persons experiencing extreme longevity. *Exp. Gerontol.*, 138 :111000, Sep 2020.

[63] Claudio Franceschi and Judith Campisi. Chronic Inflammation (Inflammaging) and Its Potential Contribution to Age-Associated Diseases. *J. Gerontol. A Biol. Sci. Med. Sci.*, 69(Suppl_1) :S4–S9, Jun 2014.

[64] Paola Lucia Minciullo, Antonino Catalano, Giuseppe Mandraffino, Marco Casciaro, Andrea Crucitti, Giuseppe Maltese, Nunziata Morabito, Antonino Lasco, Sebastiano Gangemi, and Giorgio Basile. Inflammaging and anti-inflammaging : The role of cytokines in extreme longevity. *Archivum Immunologiae et Therapiae Experimentalis*, 64(2) :111–126, December 2015.

[65] Jill A. Kreiling, Brian C. Jones, Jason G. Wood, Marco De Cecco, Steven W. Criscione, Nicola Neretti, Stephen L. Helfand, and John M. Sedivy. Contribution of Retrotransposable Elements to Aging. In *Human Retrotransposons in Health and Disease*, pages 297–321. Springer, Cham, Switzerland, Jan 2017.

[66] Marco De Cecco, Takahiro Ito, Anna P. Petrashen, Amy E. Elias, Nicholas J. Skvir, Steven W. Criscione, Alberto Caligiana, Greta Brocculi, Emily M. Adney, Jef D. Boeke, Oanh Le, Christian Beauséjour, Jayakrishna Ambati, Kameshwari Ambati, Matthew Simon, Andrei Seluanov, Vera Gorbunova, P. Eline Slagboom, Stephen L. Helfand, Nicola Neretti, and John M. Sedivy. L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature*, 566 :73–78, Feb 2019.

[67] # Qiujing Yu, # Yuliya V. Katlinskaya, Christopher J. Carbone, Bin Zhao, Kanstantsin V. Katlinski, Hui Zheng, Manti Guha, Ning Li, Qijun Chen, Ting Yang, Christopher J. Lengner, Roger A. Greenberg, F. Brad Johnson, and Serge Y. Fuchs. DNA-damage-induced type I interferon promotes senescence and inhibits stem cell function. *Cell Rep.*, 11(5) :785–797, May 2015.

[68] Hongming Ma, Wei Qian, Monika Bambouskova, Patrick L. Collins, Sofia I. Porter, Andrea K. Byrum, Rong Zhang, Maxim Artyomov, Eugene M. Oltz, Nima Mosammaparast, Jonathan J. Miner, and Michael S. Diamond. Barrier-to-Autointegration Factor 1 Protects against a Basal cGAS-STING Response. *mBio*, 11(2), Mar 2020.

[69] Katherine A. Fitzgerald. The Interferon Inducible Gene : Viperin. *J. Interferon Cytokine Res.*, 31(1) :131–135, Jan 2011.

[70] Tomohiro Hakozaki, Cheri L. Swanson, and Donald L. Bissett. Hyperpigmentation in Aging Skin. In *Textbook of Aging Skin*, pages 1017–1026. Springer, Berlin, Germany, Sep 2016.

[71] Andrew R. Cullinane, Thierry Vilboux, Kevin O'Brien, James A. Curry, Dawn M. Maynard, Hannah Carlson-Donohoe, Carla Ciccone, NISC Comparative Sequencing Program, Thomas C. Markello, Meral Gunay-Aygun, Marjan Huizing, and William A. Gahl. Homozygosity mapping and whole-exome sequencing to detect SLC45A2 and G6PC3 mutations in a single patient with oculocutaneous albinism and neutropenia. *J. Invest. Dermatol.*, 131(10) :2017–2025, Oct 2011.

[72] Laura Sormani. *Identification d'un nouveau gène dans la pigmentation cutanée - CLEC12B*. PhD thesis, Université Côte d'Azur (ComUE), 2019.

[73] S Motegi, Y Yokoyama, S Ogino, K Yamada, A Uchiyama, B Perera, Y Takeuchi, H Ohnishi, and O Ishikawa. Pathogenesis of multiple lentigines in LEOPARD syndrome with PTPN11 gene mutation. *Acta Dermato Venereologica*, 95(8) :978–984, 2015.

[74] E. J. Parra, R. A. Kittles, and M. D. Shriver. Implications of correlations between skin color and genetic ancestry for biomedical research. *Nat. Genet.*, 36(11), Nov 2004.

[75] Henri Montaudié. *CLEC12B un gène de la famille des lectines impliqué dans le processus de melanomagénèse en agissant comme un gène suppresseur de tumeurs : CLEC12B un gène suppresseur de tumeurs impliqué dans le mélanome*. PhD thesis, Université Côte d'Azur (ComUE), 2019. 2019AZUR6013.

[76] Samuel de la Pava, Goryun Nigogosyan, John W. Pickren, and Aurelio Cabrera. Melanosis of the esophagus. *Cancer*, 16(1) :48–50, January 1963.

[77] Emma Pierson, Daphne Koller, Alexis Battle, and Sara Mostafavi. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLoS Computational Biology*, 11(5) :1–19, 2015.

[78] William H. Tolleson. Human melanocyte biology, toxicology, and pathology. *J. Environ. Sci. Health, Part C Environ. Carcinog. Ecotoxicol. Rev.*, 23(2) :105–161, 2005.

[79] T. Furuhata, T. Tokino, T. Urano, and Y. Nakamura. Isolation of a novel GPI-anchored gene specifically regulated by p53 ; correlation between its expression and anti-cancer drug sensitivity. *Oncogene*, 13(9) :1965–1970, Nov 1996.

[80] V. Catalano, Baldelli A. M., P. Giordani, and S. Cascinu. Molecular markers predictive of response to chemotherapy in gastrointestinal tumors. *Crit. Rev. Oncol. Hematol.*, 38(2) :93–104, May 2001.

[81] Chris T. Dee, Christoph R. Szymoniuk, Peter E. D. Mills, and Tokiharu Takahashi. Defective neural crest migration revealed by a Zebrafish model of Alx1-related frontonasal dysplasia. *Hum. Mol. Genet.*, 22(2) :239–251, Jan 2013.

Discussion

5.1 Apport des travaux et retour critique

5.1.1 GWENA

Les travaux de ce doctorat ont débutés dans un contexte où l'analyse par réseaux de co-expression, bien qu'existante depuis plus de 15 ans, n'était pas aussi démocratisé qu'elle l'aurait pu au vu des bénéfices qu'elle apporte (Figure 5.1). Pour contribuer à sa démocratisation et plus particulièrement à celle de la co-expression différentielle, le progiciel R GWENA a été développé avec comme objectif d'être facilement utilisable et de contenir de nombreux avertissements sur un mésusage. Son dépôt sur le répertoire Bioconductor complétait cette volonté car celui-ci constitue une référence en bio-informatique et favorise une meilleure visibilité. Il assure également une qualité plus haute qu'un package déposé sur l'autre répertoire populaire, le CRAN, grâce à un contrôle qualité du code bien plus strict et une obligation de maintenance et réponse aux demandes des utilisateurs.

Cependant d'autres outils existaient déjà pour des usages similaires bien que moins complets. Une étude des besoins d'analyse via une revue de nombreux article utilisant ces différents outils a amené à plusieurs choix et ajouts dans GWENA. L'ajout majeur réside dans l'implémentation dans le pipeline d'une analyse de co-expresssion différentielle utilisable avec une unique fonction R qui suit le déroulé des fonctions précédentes, qu'il s'agisse d'analyse d'enrichissement, de topologie, ou simplement de détection de modules. Un second ajout est celui d'une visualisation automatisée du réseau de gènes d'un module avec la possibilité de le colorer selon des groupes. Par exemple, une personne pourra si elle le souhaite colorer le réseau en fonction de termes d'enrichissements qu'elle aura sélectionnés, de gènes différentiellement exprimés qu'elle aura préalablement calculés, ou encore de gènes pivots qu'elle aura détectés avec GWENA. Face à la modularité hiérarchique des réseaux, une fonction de sous-partitionnement des modules a également été ajoutée pour permettre des enrichissements ciblés sur des sous-modules et ainsi mieux comprendre les fonctions physiologiques associées aux gènes qu'ils contiennent. Enfin, le choix d'une architecture modulaire pour le pipeline, c'est-à-dire en groupes de fonctions exécu-

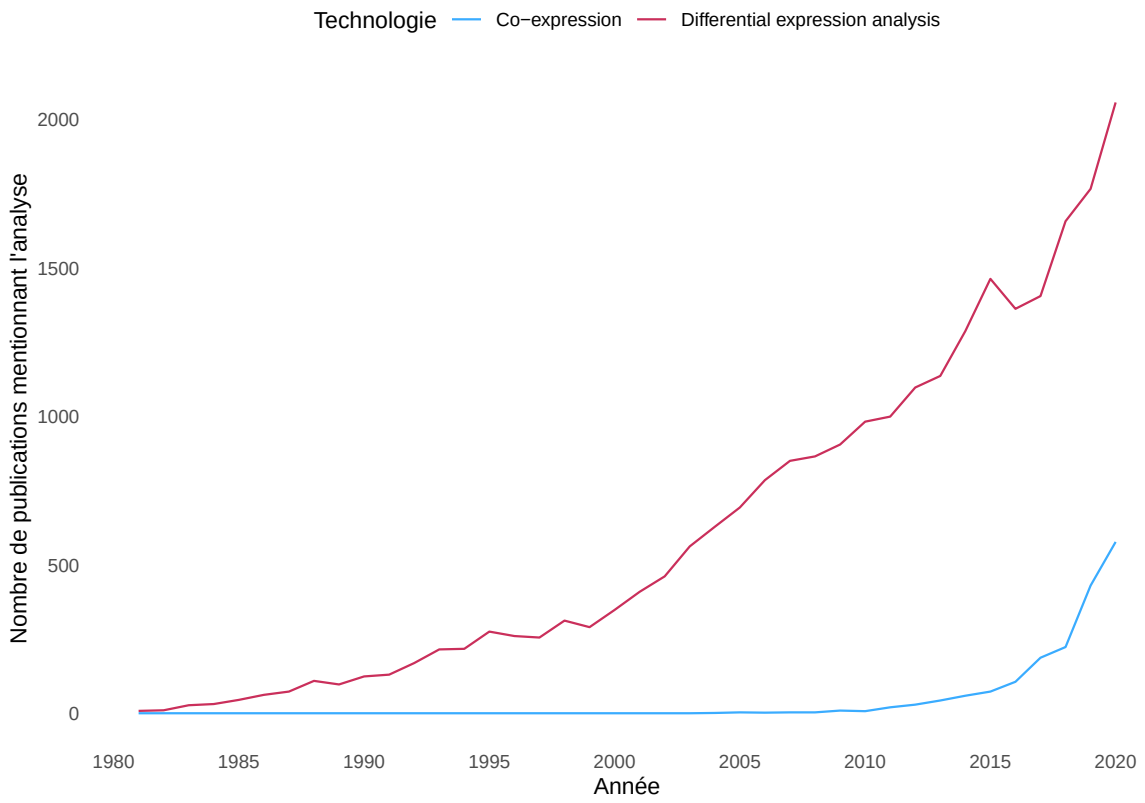

FIGURE 5.1 – Évolution de l'utilisation de l'analyse d'expression différentielle et l'analyse par réseaux de co-expression de gènes en se basant sur le nombre de publications les mentionnant. Les résultats proviennent du site PubMed (https://pubmed.ncbi.nlm.nih.gov) avec les requêtes suivantes : **Expression différentielle** = "differential expression" analysis, **Co-expression** = "gene co-expression network"

tables indépendamment des autres et réutilisable dans d'autres pipelines, rajoute la pérennité qui manquait à tous les outils développés jusque-là. GWENA répond donc à l'objectif qu'on s'était fixé en premier point de ces travaux de doctorat.

GWENA est toutefois perfectible, tant dans le code que dans les méthodes qui pourraient lui être ajoutées. Il gagnerait par exemple à embarquer d'autres méthodes pour réaliser les différentes étapes, par exemple proposer une méthode par score de similarité mais par modélisation selon un modèle de mélange gaussien, ou bien proposer une méthode de découpage de module par décomposition. Les réseaux de co-expression ne permettent pas de connaître le sens du lien présent entre deux nœuds, aussi appelée causalité. Des méthodes d'inférence de réseaux de régulation pourraient donc être ajoutée à GWENA pour compenser cela. Cependant l'utilisation de telles méthodes requière de travailler sur des données comportant peu de boucles de régulation et nécessitent des données extrêmement peu bruitées [193]. Certaines régulations annoncées peuvent également être le fruit non pas d'un seul gène mais de plusieurs [104]. Un ajout d'une méthode d'inférence de réseau de régulation nécessiterait donc une utilisation plus experte de GWENA ou bien demanderait de rajouter de nombreux tests de validation mais qui ne rempla-

ceront jamais une expertise. La création d'un progiciel en R demande également de maintenir cet outil au fur et à mesure des versions du langage à venir, ainsi que d'adapter les fonctions de GWENA à d'éventuelles modifications internes aux autres progiciels qu'il embarque. L'auteur se doit donc de poursuivre une veille technologie malgré qu'il quitte le laboratoire à la fin de son doctorat, ou bien un autre membre du laboratoire se doit d'apprendre le fonctionnement du progiciel pour effectuer les futures modifications. Ce progiciel est de plus prévu pour fonctionner sur des machines disposant de puissance processeur et mémoire vive bien plus large que celles présentes sur un ordinateur de bureau dès lors que le jeu de donnée dépasse une certaine taille [1]. Ceci est toutefois améliorable à l'avenir en reprenant une technique de calcul par bloc introduite par le progiciel WGCNA [133] et en l'adaptant à l'architecture modulaire

5.1.2 L'analyse par réseaux de co-expression

Si l'analyse par réseaux de co-expression telle qu'on l'a présentée et utilisée dans cette thèse présente une valeur ajoutée non négligeable, il faut toutefois l'utiliser avec précaution et connaître ses limites. Ce type d'analyses, bien que relativement robuste aux artéfacts est ainsi fortement influencée par d'autres effets dans les données si elles sont mal normalisées ou filtrées. Ainsi, un jeu de donnée avec un effet de lot se retrouvera très rapidement avec des modules bruités sous l'effet de corrélations fallacieuses, ou bien engendrera l'ajustement d'une loi de puissance à un paramètre très élevé. Cet ajustement de la loi de puissance est d'ailleurs assez délicat car un paramètre suffisamment élevé permettra quasi systématiquement d'ajuster la loi sur des données. P. Langfelder précisera par la suite dans une foire aux questions dédiée à WGCNA [2] quelles étaient les différentes valeurs du paramètre qu'on pouvait attendre.

La loi de puissance a par ailleurs été l'objet de plusieurs critiques elle-même pour son utilisation dans le calcul de la matrice d'adjacence [111]. En effet d'autres lois aux propriétés similaires comme la loi log-normale et la loi exponentielle tendent à être ajustable sur des données d'expression de gènes. Dans les travaux non présentés de ce doctorat car ne donnant pas lieu à des résultats, cette non-spécificité a notamment entravé le développement d'une méthode de détection automatisée d'une perte ou variation significative de connectivité qui se basait sur la loi de puissance entre deux conditions. Un test de différence de lois de puissances ajustées préconise au préalable de tester la spécificité de la loi de puissance par rapport à d'autres hypothèses de loi comme la loi log-normale ou exponentielle. Dans la majorité des tests, la loi log-normale était ajustable sur les données au même titre que la loi de puissance. Même dans le cas où la loi de puissance était prouvée comme la seule ajustable parmi les lois testées des incohérences ont été relevées. Des modules préservés par co-expression différentielle étaient ainsi détectés comme

1. Des tests sur un ordinateur portable avec Intel® Core™ i5-6200U CPU @ 2.30GHz × 4, et 8Go de mémoire DDR4, ce qui représente un pc moyen dans un laboratoire ont montré une taille maximale de jeux de données d'environ 50 échantillons et 3000 gènes
2. https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/faq.html

significativement différents au vu des lois ajustées, et inversement des modules non préservés n'étaient pas détectés comme différent.

Enfin, pour assurer la fiabilité des réseaux produits, il est nécessaire d'utiliser un nombre d'échantillon conséquent, idéalement plus de 50 par condition, ce qui représente un coût non négligeable. De nombreux jeux de données de puces à ADN et RNA-seq sont disponibles sur GEO et ArrayExpress sans avoir été exploités par co-expression, mais trop peu disposent d'assez d'échantillons lorsqu'il s'agit de l'humain.

5.1.3 L'étude du vieillissement

Le vieillissement est un phénomène dont la complexité n'a d'égal que la multiplicité de ses manifestations. Tantôt origine du dérèglement fonctionnel, tantôt conséquence, les marques majeures du vieillissement concernent de nombreux mécanismes cellulaire et moléculaires. Pour distinguer de façon certaine chacune des altérations dues au vieillissement chez les individus de chaque tranche d'âge, un nombre important d'échantillons est nécessaire. C'est pourquoi ces travaux se sont appuyés sur les données de l'étude GTEx qui est à ce jour le plus gros regroupement de séquençage en terme d'individus/tissus non ciblé sur une pathologie. Toutefois, ces données ne sont pas indemnes de biais et un contrôle qualité attentif doit être effectué. Elles sont également réparties de façon hétérogène selon les différents phénotypes fournis. Ainsi, l'âge qui a été notre phénotype d'intérêt s'est retrouvé avec bien plus d'échantillons chez les personnes âgées et bien moins chez les personnes jeunes en raison de la mortalité plus faible de ces derniers (les décès étant majoritairement issus de traumatismes). Si cette inégalité de répartition est souhaitable humainement parlant, cela rajoute des biais potentiels sur le plan de l'analyse biostatistique. Ainsi on a été particulièrement attentif à l'équivalence de réseaux construits sur la totalité des échantillons âgés ou sur un sous ensemble de taille identique à celle des échantillons jeunes. Cette démarche n'étant et ne pouvant pas être intégrée à l'outil ou à la méthode de co-expression, un oeil expert sera toujours requis.

Outre la découverte de gènes candidats au vieillissement du muscle squelettique pouvant aider à la compréhension de la sarcopénie, nos études ont permis d'explorer un phénomène moins bien connu : la perte de coordination entre gènes avec l'âge. Si ce phénomène est déjà bien étudié dans les cancers [192], il l'est beaucoup moins dans le vieillissement alors que les deux conditions sont connues pour partager de nombreux points communs comme les déficits de réparation de l'ADN, les altérations de la sénescence, ou encore la perte de protéostasie [194]. Soutworth fut une des premières à se pencher sur la question en 2009 [153] en remarquant une perte de densité de la connectivité chez la souris. Quelques études se sont ensuite poursuivies sur d'autres organismes modèles [195] ou sur des tissus humains ciblés [196, 197] mais aucune n'avait jusque-là essayé de s'intéresser à ce phénomène de façon générale chez l'humain. D'autres approches cohérentes avec la perte de connectivité ont toutefois été réalisées

dans ce temps. Ainsi des recherches sur la localisation dans le réseau des gènes associés au vieillissement a montré que les gènes pivots étaient une catégorie surreprésentée [198]. En tenant compte de la tendance des gènes pivots à être des éléments de régulation [142], la perte de connectivité périphérique dans le réseau fait sens. Parallèlement, des travaux se penchant sur la variabilité transcriptomique croissante entre cellules lors du vieillissement indiquent qu'elle pourrait faire suite à une stimulation de type immunitaire chez la souris [199]. Chacune de ces publications vient alors compléter peu à peu la vision globale du phénomène de déconnexion. Toutefois, aucune d'elles n'explique le phénomène inverse de reconnexion observé dans nos travaux (Chapitre 1) et qui semble être lié à des mécanismes de compensation. Ces derniers restent d'ailleurs peu étudiés d'un point de vue topologique et restent focalisés sur quelques gènes [200].

Suite aux études menées avec notre collaborateur sur un autre tissu pour l'étude du vieillissement, on s'est intéressé aux mécanismes de celui-ci qui pouvaient être partagés entre différents tissus ou qui à l'inverse leur étaient spécifiques. L'analyse de co-expression différentielle incluse dans GWENA sur les couples tissus et tranche d'âge a permise simultanément l'isolement de phénomènes communs et spécifiques du vieillissement comme en témoignent les enrichissements obtenus. À titre d'exemple, un mécanisme commun et un mécanisme spécifiques ont été explorés et ont permis la priorisation de gènes par analyse topologique. Une revue de la littérature à leur sujet est ensuite venu étayer ces déductions sans pour autant les affirmer avec certitude. Cependant, la revue détaillée de ces gènes priorisés dans le vieillissement ou tout autre condition est l'affaire d'une expertise par des experts biologistes, ce qui dépasse donc la portée ces travaux de doctorat bien qu'ils contribuent à leur compréhension. L'estimation de la cohérence d'un gène priorisé par rapport au contexte étudié nécessite en effet d'être au fait de bien plus d'informations qu'une revue de littérature réalisée par un bio-informaticien ne saurait égaler. Si l'étude des mécanismes communs retrouvés par notre analyse pourrait être réalisée par un expert en vieillissement, l'analyse de chaque mécanisme spécifique de sous groupes de tissus requerrait elle plutôt un ensemble de collaborations avec des experts de chaque tissu ou mécanisme étudié. Par ailleurs, les gènes priorisables avec GWENA reste putatifs et nécessiteront toujours une validation expérimentale. L'horizon des recherches aura toutefois été restreint vers les plus prometteurs, entrainant des coûts de recherche réduits.

5.2 Perspectives de recherches

5.2.1 L'intégration d'autres omiques

Avec l'accroissement constant des jeux de données de transcriptomique, la possibilité de construction de réseaux de co-expression augmente ainsi que leur fiabilité. Des tissus qui n'auraient pu être explorés comme dans le Chapitre 2 pourraient êtres explorés et permettre de compléter cette caractérisation des phénomènes spécifiques et communs du vieillissement. Les données

de transcriptomiques ne sont d'ailleurs pas les seules à voir leur nombre augmenter et de plus en plus d'approches multi-omique se développent. Les analyses par réseau ne font pas exception et des applications de la co-expression sur des données de protéomique ou de métabolomique sont très vite apparues [201]. Ces démarches sont toutefois à réaliser avec précaution car les technologies actuelles de quantification protéiques et métabolomiques sont sujettes aux valeurs manquantes et à une faible couverture de l'ensemble des protéines ou métabolites, ce qui va bruiter le réseau construit [202]. Elles offrent toutefois des possibilités d'étude étendues et sont attendue comme un moyen de préciser la causalité de nombreux phénomène du vieillissement grâce à l'analyse de la coordination de la dégradation des omiques [203]. Différentes méthodes de combinaisons de données existent alors telles que la fusion de réseaux de similarité [204] qui permet de détecter des modules non pas de gènes mais patients aux caractéristiques semblables.

Les réseaux de co-expression sur d'autres omiques ne sont par ailleurs pas les seuls types de réseau pouvant être utilisés. Ainsi sans réaliser de réseaux de co-expression de protéines, il est possible d'ajouter de la connaissance protéique en incorporant l'information de réseaux protéine-protéine [145]. Similairement, les voies de régulations de métabolites connues ou des profils d'expression de ceux-ci *in situ* peuvent être intégrés [205]. Cette dernière démarche est d'ailleurs particulièrement intéressante dans l'étude de l'inflammation chronique de faible intensité (*inflammaging* en anglais) présente dans le vieillissement. Cette marque principale du vieillissement représente actuellement une des meilleures sources d'explication de la coordination et la propagation du vieillissement. En essayant d'observer la relation existante entre les variations de co-expression transcriptomique dans le réseau et les métabolites sur ou sous représentés dans les échantillons, il serait potentiellement possible de comprendre quelles dégradations initiales du transcriptome entrainent telle ou telle accumulation de métabolites ou transcrits va déclencher une réponse immunitaire. Parallèlement les mécanismes défaillant dans l'élimination de ces déchets du vieillissement (*garb-aging* en anglais) [206] pourraient faire l'objet d'une enquête grâce au transcriptome pour pareillement comprendre les origines de leur dégradation avec l'âge.

5.2.2 Compléter l'information des réseaux de co-expression

Outre les méthodes impliquant d'autres omiques, l'analyse de co-expression peut venir être complétée par des informations de la transcriptomique elle-même. Des quantifications de l'expression sur plusieurs points dans le temps peuvent ainsi permettre de suivre l'évolution de perturbations du réseau au cours du temps [207]. Concernant le vieillissement, on pourrait ainsi envisager de construire un réseau par différentes tranches d'âge croissantes pour estimer quelle période est la plus propice à l'inflexion de la dégradation liée au vieillissement. Il serait alors possible de déterminer laquelle ou lesquelles des marques principales du vieillissement sont le plus souvent le point de départ.

Le RNA-seq, contrairement au puces à ADN, permet une quantification de la totalité du transcriptome. Grâce à cela, il est alors possible, une fois les modules d'intérêt identifiés, de revenir au détail des transcrits qui forment l'expression d'un gène pour enquêter sur un potentiel épissage alternatif qui serait à l'origine de la variation de co-expression par rapport à la condition de contrôle [208, 209].

Enfin, les analyses des différents tissus réalisées dans le vieillissement dans ces travaux sont réalisées sur des données dites en vrac car contenant de multiples types cellulaires. Avec la fiabilité croissante des quantifications d'expression par RNA-seq mono cellule (*single cell RNA-seq* ou scRNA-seq en anglais), il est à présent possible d'affiner les analyses pour démêler la contribution de chaque type cellulaire à une condition [104]. Si elles sont particulièrement intéressante pour étudier des types celluaires soupçonnés d'être la source de certains phénomènes du vieillissement [210, 211, 212], ces analyses restent toutefois couteuses, particulièrement si une analyse par réseaux de co-expression est envisagées sur elle.

5.2.3 Autres approches du vieillissement

Un point commun des études sur le vieillissement à de nombreuses études actuelles est la focalisation sur un certain type de population majoritaire, à savoir les personnes européennes ou descendantes de celles-ci. Si l'environnement joue un rôle non négligeable, la composante génétique du vieillissement est importante. Face à la longévité constatée de certaines populations, des hypothèses sur des variations génétiques spécifiques influant le vieillissement ont été émises [213]. On pourrait alors envisager une étude de préservation de la co-expression entre ces différentes populations pour essayer de détecter de nouveaux biomarqueurs de la longévité ou d'autres mécanismes du vieillissement de compensation du vieillissement. Malheureusement les sources de données avec suffisamment d'échantillons manquent actuellement pour ce faire.

Le cancer possède une grande proximité avec le vieillissement de par ses phénomènes de dérégulation de la machinerie cellulaire [214]. Si les phénomènes de perte de connectivité sont connus dans chacun, le phénomène de reconnexion qu'on a pu identifier n'a encore été étudié dans le cancer. On pourrait donc envisager d'isoler les gènes d'un même tissu qui ont été identifiés comme se reconnectant dans le vieillissement, et les observer dans un réseau de co-expression issu du cancer pour comprendre quels mécanismes diffèrent. Sachant que le cancer et le vieillissement partagent tous deux le phénomène d'inflammation chronique de faible intensité [215], une exploration comparative de celui-ci dans les deux conditions pourrait amener de nouveaux éléments de compréhension [192]. L'influence de la protéine C réactive (CRP en anglais) est ainsi une piste privilégiée car elle contribue à la pathogenèse du cancer et du vieillissement. Il est toutefois encore nécessaire de la démêler de son rôle de marqueur inflammatoire lors d'infections.

Conclusion

Après avoir présenté une vue d'ensemble des technologies de transcriptomique et leur intérêt dans la recherche en médecine moléculaire, cette thèse s'est attardée sur une présentation détaillée de l'analyse par réseaux de co-expression de gènes. Les méthodes de préparation des données, de construction du réseau, de détection des modules et leur exploitation avec ou sans connaissance a priori ont été examinées par rapport aux connaissances dans la littérature actuelle. Chaque étape demandant un niveau de maîtrise poussé en biostatistique ou théorie des graphes, il était important d'expliciter chacun des choix fait pour mieux comprendre la démarche poursuivie durant ce doctorat. L'adéquation de l'analyse par réseaux de co-expression de gènes pour l'étude du vieillissement a ensuite été mise en avant après une présentation concise du vieillissement, de ses manifestations cellulaire et moléculaire, ainsi que les enjeux qu'il représente.

Face à la difficulté d'emploi des outils existant sans expertise, la pénibilité de leur combinaison et de la faible pérennité d'une analyse avec les outils actuels pour réaliser une analyse par réseaux de co-expression de gènes, on avait formulé une première hypothèse qui s'interrogeait sur la faisabilité d'un outil pipeline sous la forme d'un progiciel R déposé sur Bioconductor pour répondre à ces problèmes. Grâce au développement du progiciel GWENA présenté en Chapitre 1 on a démontré qu'il était possible d'avoir un outil facilitant l'analyse par réseaux de co-expression de gènes de bout en bout et qui, grâce à une architecture modulaire, était voué à s'adapter aux dernières avancées méthodologiques sur ce type d'analyse. Avec une étude de cas dédiée au vieillissement du muscle, il a également été possible de montrer la valeur ajoutée de la co-expression différentielle dans la priorisation de gènes d'une condition donnée. L'analyse de topologie a elle permit de venir préciser un phénomène déjà observé chez la souris mais pas encore chez l'homme à ce jour : une déconnexion modulaire de gènes dans le réseaux avec le vieillissement. Mieux encore, il a été mis en évidence que cette déconnexion modulaire s'accompagnait d'une reconnexion centralisée autour des gènes pivots dans le réseau.

Fort de cette capacité de GWENA à innover dans la recherche sur le vieillissement, la seconde hypothèse supposait qu'il serait à même de trouver de nouveaux gènes candidats par co-expression différentielle non pas simplement entre deux tranches d'âge, mais en plus à travers plusieurs tis-

sus. Le Chapitre 2 présente donc l'utilisation des différents outils intégrés à GWENA pour réaliser une analyse parallélisée de couples tissu et tranche d'âge. L'identification parmi les modules modérément ou non préservés de nombreux phénomènes du vieillissement tels qu'énoncés dans les marques principales du vieillissement établies par López-Otín *et al.* a permit de valider la démarche de co-expression différentielle et d'aller plus en avant dans l'exploitation de ces modules. Par un recoupement des gènes contenus dans chacun des modules, des gènes impliqués dans des phénomènes du vieillissement spécifiques à quelques tissus ou à l'inverse commun à la majorité ont été détectés. Une analyse topologique faite dans un exemple de gènes spécifiques et un exemple de gènes communs a finalement retourné de nouveaux gènes pertinents dans la compréhension du vieillissement au vu des annotations dont ils bénéficiaient déjà.

Chacun de ces chapitres aura donc, en plus de montrer l'intérêt de GWENA, permis de détecter de nouveaux gènes candidats au vieillissement humain. Ceux-ci devront à l'avenir faire l'objet de validation expérimentale.

Bibliographie

[1] A. L. Barabási and Z. N. Oltvai, “Network biology : Understanding the cell’s functional organization,” *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[2] R.-t. Hu, Q. Yu, S.-d. Zhou, Y.-x. Yin, R.-g. Hu, H.-p. Lu, and B.-l. Hu, “Co-expression Network Analysis Reveals Novel Genes Underlying Alzheimer’s Disease Pathogenesis,” *Front. Aging Neurosci.*, vol. 12, Nov 2020.

[3] D. García-Cortés, G. de Anda-Jáuregui, C. Fresno, E. Hernández-Lemus, and J. Espinal-Enríquez, “Gene Co-expression Is Distance-Dependent in Breast Cancer,” *Front. Oncol.*, vol. 10, Jul 2020.

[4] D. J. Morris, “Cell formation by myxozoan species is not explained by dogma,” *Proc. R. Soc. B.*, vol. 277, pp. 2565–2570, Aug 2010.

[5] E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, S. Perez-Amodio, P. Strippoli, and S. Canaider, “An estimation of the number of cells in the human body,” *Annals of Human Biology*, vol. 40, pp. 463–471, July 2013.

[6] Y. Panina, P. Karagiannis, A. Kurtz, G. N. Stacey, and W. Fujibuchi, “Human Cell Atlas and cell-type authentication for regenerative medicine,” *Exp. Mol. Med.*, vol. 52, pp. 1443–1451, Sep 2020.

[7] T. Yuki, A. Haratake, H. Koishikawa, K. Morita, Y. Miyachi, and S. Inoue, “Tight junction proteins in keratinocytes : localization and contribution to barrier function,” *Exp. Dermatol.*, vol. 16, pp. 324–330, Apr 2007.

[8] F. Yuan, S. Chien, and S. Weinbaum, “A New View of Convective-Diffusive Transport Processes in the Arterial Intima,” *J. Biomech. Eng.*, vol. 113, pp. 314–329, Aug 1991.

[9] A. Trounson, R. Gosden, and U. Eichenlaub-Ritter, *Biology and pathology of the oocyte : role in fertility, medicine and nuclear reprograming*. Cambridge University Press, 2013.

[10] I. Hekselman and E. Yeger-Lotem, “Mechanisms of tissue and cell-type specificity in heritable traits and diseases,” *Nat. Rev. Genet.*, vol. 21, pp. 137–150, Mar 2020.

[11] F. R. Bettley, “Textbook of Dermatology,” *BMJ*, vol. 1, pp. 178–178, jan 1965.

[12] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, “Gene Ontology : tool for the unification of biology,” *Nature Genetics*, vol. 25, pp. 25–29, May 2000.

[13] M. Christopher K., v. H. Kensal E., A. Dean R., and A.-C. Spencer J., *Biochemistry*. Pearson, 4th ed., 2012.

[14] A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, and P. Flicek, “GENCODE reference annotation for the human and mouse genomes,” *Nucleic Acids Res.*, vol. 47, pp. D766–D773, Jan 2019.

[15] V. M. Weake and J. L. Workman, “Inducible gene expression : diverse regulatory mechanisms - Nature Reviews Genetics,” *Nat. Rev. Genet.*, vol. 11, pp. 426–437, Jun 2010.

[16] M. Levo and E. Segal, “In pursuit of design principles of regulatory sequences,” *Nat. Rev. Genet.*, vol. 15, pp. 453–468, Jul 2014.

[17] A. Chen and A. N. Koehler, “Transcription Factor Inhibition : Lessons Learned and Emerging Targets,” *Trends Mol. Med.*, vol. 26, pp. 508–518, May 2020.

[18] S. Kadauke and G. A. Blobel, “Chromatin loops in gene regulation,” *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, vol. 1789, pp. 17–25, Jan 2009.

[19] T. R. Mercer, M. E. Dinger, C. P. Bracken, G. Kolle, J. M. Szubert, D. J. Korbie, M. E. Askarian-Amiri, B. B. Gardiner, G. J. Goodall, S. M. Grimmond, and J. S. Mattick, “Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome,” *Genome Res.*, vol. 20, pp. 1639–1650, Nov 2010.

[20] M. Gutierrez-Arcelus, T. Lappalainen, S. B. Montgomery, A. Buil, H. Ongen, A. Yurovsky, J. Bryois, T. Giger, L. Romano, A. Planchon, E. Falconnet, D. Bielser, M. Gagnebin, I. Padioleau, C. Borel, A. Letourneau, P. Makrythanasis, M. Guipponi, C. Gehrig, S. E. Antonarakis, and E. T. Dermitzakis, “Passive and active DNA methylation and the interplay with genetic variation in gene regulation,” *eLife*, Jun 2013.

[21] J. Yu and J. E. Russell, “Structural and functional analysis of an mRNP complex that mediates the high stability of human β-globin mRNA,” *Molecular and Cellular Biology*, vol. 21, pp. 5879–5888, Sept. 2001.

[22] V. S. Patil, R. Zhou, and T. M. Rana, "Gene regulation by non-coding RNAs," *Crit. Rev. Biochem. Mol. Biol.*, vol. 49, pp. 16–32, Jan 2014.

[23] S. Krishna, S. Raghavan, R. DasGupta, and D. Palakodeti, "tRNA-derived fragments (tRFs) : establishing their turf in post-transcriptional gene regulation," *Cell. Mol. Life Sci.*, vol. 78, pp. 2607–2619, Mar 2021.

[24] R. K. Khajuria, M. Munschauer, J. C. Ulirsch, C. Fiorini, L. S. Ludwig, S. K. McFarland, N. J. Abdulhay, H. Specht, H. Keshishian, D. R. Mani, M. Jovanovic, S. R. Ellis, C. P. Fulco, J. M. Engreitz, S. Schütz, J. Lian, K. W. Gripp, O. K. Weinberg, G. S. Pinkus, L. Gehrke, A. Regev, E. S. Lander, H. T. Gazda, W. Y. Lee, V. G. Panse, S. A. Carr, and V. G. Sankaran, "Ribosome Levels Selectively Regulate Translation and Lineage Commitment in Human Hematopoiesis," *Cell*, vol. 173, pp. 90–103.e19, Mar 2018.

[25] H. A. Meijer, Y. W. Kong, W. T. Lu, A. Wilczynska, R. V. Spriggs, S. W. Robinson, J. D. Godfrey, A. E. Willis, and M. Bushell, "Translational Repression and eIF4A2 Activity Are Critical for MicroRNA-Mediated Gene Regulation," *Science*, vol. 340, pp. 82–85, Apr 2013.

[26] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend, "Functional Discovery via a Compendium of Expression Profiles," *Cell*, vol. 102, pp. 109–126, jul 2000.

[27] N. Cloonan, A. R. R. Forrest, G. Kolle, B. B. A. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, A. J. Robertson, A. C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M. Manning, K. J. McKernan, and S. M. Grimmond, "Stem cell transcriptome profiling via massive-scale mRNA sequencing," *Nat. Methods*, vol. 5, pp. 613–619, Jul 2008.

[28] R. N. Munji, A. L. Soung, G. A. Weiner, F. Sohet, B. D. Semple, A. Trivedi, K. Gimlin, M. Kotoda, M. Korai, S. Aydin, A. Batugal, A. C. Cabangcala, P. G. Schupp, M. C. Oldham, T. Hashimoto, L. J. Noble-Haeusslein, and R. Daneman, "Profiling the mouse brain endothelial transcriptome in health and disease models reveals a core blood–brain barrier dysfunction module," *Nat. Neurosci.*, vol. 22, pp. 1892–1902, Nov 2019.

[29] A. K. Sarkar, P.-Y. Tung, J. D. Blischak, J. E. Burnett, Y. I. Li, M. Stephens, and Y. Gilad, "Discovery and characterization of variance QTLs in human induced pluripotent stem cells," *PLos Genet.*, vol. 15, p. e1008045, Apr 2019.

[30] J. Segalés, A. B. M. M. K. Islam, R. Kumar, Q.-C. Liu, P. Sousa-Victor, F. J. Dilworth, E. Ballestar, E. Perdiguero, and P. Muñoz-Cánoves, "Chromatin-wide and transcriptome profiling integration uncovers p38α MAPK as a global regulator of skeletal muscle differentiation," *Skeletal Muscle*, vol. 6, pp. 1–15, Dec 2016.

[31] P. Godoy, W. Schmidt-Heck, B. Hellwig, P. Nell, D. Feuerborn, J. Rahnenführer, K. Kattler, J. Walter, N. Blüthgen, and J. G. Hengstler, "Assessment of stem cell differentiation based on genome-wide expression profiles," *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 373, Jul 2018.

[32] K. Jividen, K. Z. Kedzierska, C.-S. Yang, K. Szlachta, A. Ratan, and B. M. Paschal, "Genomic analysis of DNA repair genes and androgen signaling in prostate cancer," *BMC Cancer*, vol. 18, pp. 1–20, Dec 2018.

[33] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO : archive for functional genomics data sets—update," *Nucleic Acids Res.*, vol. 41, pp. D991–D995, Jan 2013.

[34] A. Athar, A. Füllgrabe, N. George, H. Iqbal, L. Huerta, A. Ali, C. Snow, F. N. A., R. Petryszak, I. Papatheodorou, U. Sarkans, and A. Brazma, "ArrayExpress update - from bulk to single-cell expression data.," *Nucleic Acids Res.*, vol. 47, pp. D711–D715, Jan 2019.

[35] M. Lukk, M. Kapushesky, J. Nikkilä, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen, and A. Brazma, "A global map of human gene expression," *Nat. Biotechnol.*, vol. 28, pp. 322–324, Apr 2010.

[36] I. Papatheodorou, P. Moreno, J. Manning, A. M.-P. Fuentes, N. George, S. Fexova, N. A. Fonseca, A. Füllgrabe, M. Green, N. Huang, L. Huerta, H. Iqbal, M. Jianu, S. Mohammed, L. Zhao, A. F. Jarnuczak, S. Jupp, J. Marioni, K. Meyer, R. Petryszak, C. A. Prada Medina, C. Talavera-López, S. Teichmann, J. A. Vizcaino, and A. Brazma, "Expression Atlas update : from tissues to single cells," *Nucleic Acids Res.*, vol. 48, pp. D77–D83, Jan 2020.

[37] O. Morozova, M. Hirst, and M. A. Marra, "Applications of New Sequencing Technologies for Transcriptome Analysis," *Annu. Rev. Genomics Hum. Genet.*, vol. 10, pp. 135–151, Aug 2009.

[38] M. Pennycooke, N. Chaudary, I. Shuralyova, Y. Zhang, and I. R. Coe, "Differential Expression of Human Nucleoside Transporters in Normal and Tumor Tissue," *Biochem. Biophys. Res. Commun.*, vol. 280, pp. 951–959, Jan 2001.

[39] S.-H. Ahn, J.-H. Ahn, D.-R. Ryu, J. Lee, M.-S. Cho, and Y.-H. Choi, "Effect of Necrosis on the miRNA-mRNA Regulatory Network in CRT-MG Human Astroglioma Cells," *Cancer Res. Treat.*, vol. 50, p. 382, Apr 2018.

[40] A. M. Martin, A. L. Lumsden, R. L. Young, C. F. Jessup, N. J. Spencer, and D. J. Keating, "The nutrient-sensing repertoires of mouse enterochromaffin cells differ between duodenum and colon," *Neurogastroenterol. Motil.*, vol. 29, p. e13046, Jun 2017.

[41] C. Ventura, I. E. Leon, A. Asuaje, P. Martín, N. Enrique, M. Núñez, C. Cocca, and V. Milesi, "Differential expression of the long and truncated Hv1 isoforms in breast-cancer cells," *J. Cell. Physiol.*, vol. 235, pp. 8757–8767, Nov 2020.

[42] D. C. Collins, R. Sundar, J. S. J. Lim, and T. A. Yap, "Towards Precision Medicine in the Clinic : From Biomarker Discovery to Novel Therapeutics," *Trends Pharmacol. Sci.*, vol. 38, pp. 25–40, Jan 2017.

[43] H. M. Temin and S. Mizutani, "Viral RNA-dependent DNA Polymerase : RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus," *Nature*, vol. 226, pp. 1211–1213, Jun 1970.

[44] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler, "Serial Analysis of Gene Expression," *Science*, vol. 270, pp. 484–487, Oct 1995.

[45] F. Sanger and A. R. Coulson, "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase," *J. Mol. Biol.*, vol. 94, pp. 441–448, May 1975.

[46] M. A. Marra, L. Hillier, and R. H. Waterston, "Expressed sequence tags — ESTablishing bridges between genomes," *Trends Genet.*, vol. 14, pp. 4–7, Jan 1998.

[47] P. G. N. Jeppesen, J. A. Steitz, R. F. Gesteland, and P. F. Spahr, "Gene Order in the Bacteriophage R17 RNA : 5′– ;A Protein–Coat Protein–Synthetase–3′," *Nature*, vol. 226, pp. 230–237, Apr 1970.

[48] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. M. Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert, and M. Ysebaert, "Complete nucleotide sequence of bacteriophage MS2 RNA : primary and secondary structure of the replicase gene," *Nature*, vol. 260, pp. 500–507, Apr 1976.

[49] M. Becker-André and K. Hahlbrock, "Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcipt titration assay (PATTY)," *Nucleic Acids Res.*, vol. 17, pp. 9437–9446, Nov 1989.

[50] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee, "Transcriptomics technologies," *PLoS Comput. Biol.*, vol. 13, p. e1005457, May 2017.

[51] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," *Science*, vol. 270, pp. 467–470, Oct 1995.

[52] T. Lenoir and E. Giannella, "The emergence and diffusion of DNA microarray technology," *J. Biomed. Discovery Collab.*, vol. 1, p. 11, 2006.

[53] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq : a revolutionary tool for transcriptomics," *Nat. Rev. Genet.*, vol. 10, pp. 57–63, Jan 2009.

[54] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing," *Science*, vol. 320, pp. 1344–1349, Jun 2008.

[55] F. Cheung, B. J. Haas, S. M. D. Goldberg, G. D. May, Y. Xiao, and C. D. Town, "Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology," *BMC Genomics*, vol. 7, pp. 1–10, Dec 2006.

[56] H. Liu, "Microarray probes and probe sets," *Frontiers in Bioscience*, vol. E2, no. 1, pp. 325–338, 2010.

[57] A. Thompson, S. Lucchini, and J. C. D. Hinton, "It's easy to build your own microarrayer!," *Trends Microbiol.*, vol. 9, pp. 154–156, Apr 2001.

[58] R. Bumgarner, "Overview of DNA Microarrays : Types, Applications, and Their Future," *Current Protocols in Molecular Biology*, vol. 101, pp. 22.1.1–22.1.11, Jan 2013.

[59] H. Koltai and C. Weingarten-Baror, "Specificity of DNA microarray hybridization : characterization, effectors and approaches for data correction," *Nucleic Acids Res.*, vol. 36, pp. 2395–2405, Apr 2008.

[60] A. Nesterov-Mueller, F. Maerkle, L. Hahn, T. Foertsch, S. Schillo, V. Bykovskaya, M. Sedlmayr, L. K. Weber, B. Ridder, M. Soehindrijo, B. Muenster, J. Striffler, F. R. Bischoff, F. Breitling, and F. F. Loeffler, "Particle-Based Microarrays of Oligonucleotides and Oligopeptides," *Microarrays*, vol. 3, pp. 245–262, Oct 2014.

[61] R. Lukac, K. N. Plataniotis, B. Smolka, and A. N. Venetsanopoulos, "cDNA microarray image processing using fuzzy vector filtering framework," *Fuzzy Sets Syst.*, vol. 152, pp. 17–35, May 2005.

[62] A. Petrov and S. Shams, "Microarray Image Processing and Quality Control," *The. Journal of VLSI Signal Processing-Systems. for Signal, Image, and Video Technology*, vol. 38, pp. 211–226, Nov 2004.

[63] G. K. Smyth and T. Speed, "Normalization of cDNA microarray data," *Methods*, vol. 31, pp. 265–273, Dec 2003.

[64] W. S. Cleveland and S. J. Devlin, "Locally Weighted Regression : An Approach to Regression Analysis by Local Fitting," *J. Am. Stat. Assoc.*, vol. 83, pp. 596–610, Sep 1988.

[65] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res.*, vol. 43, p. e47, Apr 2015.

[66] J. Quackenbush, "Microarray data normalization and transformation - Nature Genetics," *Nat. Genet.*, vol. 32, pp. 496–501, Dec 2002.

[67] T. Wilkes, H. Laux, and C. A. Foy, “Microarray Data Quality—Review of Current Developments,” *OMICS : A Journal of Integrative Biology*, vol. 11, pp. 1–13, Apr 2007.

[68] C. Argyropoulos, A. A. Chatziioannou, G. Nikiforidis, A. Moustakas, G. Kollias, and V. Aidinis, “Operational criteria for selecting a cDNA microarray data normalization algorithm,” *Oncol. Rep.*, vol. 15, pp. 983–996, Apr 2006.

[69] V. Bolón-Canedo, A. Alonso-Betanzos, I. López-de Ullibarri, and R. Cao, “Challenges and Future Trends for Microarray Analysis,” in *Microarray Bioinformatics*, pp. 283–293, New York, NY, USA : Humana, New York, NY, 2019.

[70] T. E. P. Consortium, “An integrated encyclopedia of DNA elements in the human genome,” *Nature*, vol. 489, pp. 57–74, Sept. 2012.

[71] A. J. Bergsma, E. van der Wal, M. Broeders, A. T. van der Ploeg, and W. W. M. Pim Pijnappel, “Alternative Splicing in Genetic Diseases : Improved Diagnosis and Novel Treatment Options,” in *International Review of Cell and Molecular Biology*, vol. 335, pp. 85–141, Cambridge, MA, USA : Academic Press, Jan 2018.

[72] L. Yi, L. Liu, P. Melsted, and L. Pachter, “A direct comparison of genome alignment and transcriptome pseudoalignment,” *bioRxiv*, p. 444620, Oct 2018.

[73] A. Srivastava, L. Malik, H. Sarkar, M. Zakeri, F. Almodaresi, C. Soneson, M. I. Love, C. Kingsford, and R. Patro, “Alignment and mapping methodology influence transcript abundance estimation,” *Genome Biol.*, vol. 21, pp. 1–29, Dec 2020.

[74] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, “STAR : ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, pp. 15–21, Jan 2013.

[75] H. Li, “Minimap2 : pairwise alignment for nucleotide sequences,” *Bioinformatics*, vol. 34, pp. 3094–3100, Sep 2018.

[76] B. Li and C. N. Dewey, “RSEM : accurate transcript quantification from RNA-Seq data with or without a reference genome,” *BMC Bioinf.*, vol. 12, pp. 1–16, Dec 2011.

[77] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, “Salmon provides fast and bias-aware quantification of transcript expression - Nature Methods,” *Nat. Methods*, vol. 14, pp. 417–419, Apr 2017.

[78] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, “Near-optimal probabilistic RNA-seq quantification - Nature Biotechnology,” *Nat. Biotechnol.*, vol. 34, pp. 525–527, May 2016.

[79] G. P. Wagner, K. Kin, and V. J. Lynch, “Measurement of mRNA abundance using RNA-seq data : RPKM measure is inconsistent among samples,” *Theory Biosci.*, vol. 131, pp. 281–285, Dec 2012.

[80] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Genome Biol.*, vol. 11, pp. 1–12, Oct 2010.

[81] M. D. Robinson and A. Oshlack, “A scaling normalization method for differential expression analysis of RNA-seq data,” *Genome Biol.*, vol. 11, pp. 1–9, Mar 2010.

[82] M.-A. Dillies, on behalf of The French StatOmique Consortium, A. Rau, on behalf of The French StatOmique Consortium, J. Aubert, on behalf of The French StatOmique Consortium, C. Hennequet-Antier, on behalf of The French StatOmique Consortium, M. Jeanmougin, on behalf of The French StatOmique Consortium, N. Servant, on behalf of The French StatOmique Consortium, C. Keime, on behalf of The French StatOmique Consortium, G. Marot, on behalf of The French StatOmique Consortium, D. Castel, on behalf of The French StatOmique Consortium, J. Estelle, on behalf of The French StatOmique Consortium, G. Guernec, on behalf of The French StatOmique Consortium, B. Jagla, on behalf of The French StatOmique Consortium, L. Jouneau, on behalf of The French StatOmique Consortium, D. Laloë, on behalf of The French StatOmique Consortium, C. Le Gall, on behalf of The French StatOmique Consortium, B. Schaëffer, on behalf of The French StatOmique Consortium, S. Le Crom, on behalf of The French StatOmique Consortium, M. Guedj, on behalf of The French StatOmique Consortium, F. Jaffrézic, and on behalf of The French StatOmique Consortium, “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis,” *Briefings Bioinf.*, vol. 14, pp. 671–683, Nov 2013.

[83] L. C. Gandolfo and T. P. Speed, “RLE plots : Visualizing unwanted variation in high dimensional data,” *PLoS One*, vol. 13, p. e0191629, Feb 2018.

[84] A. S. Booeshaghi and L. Pachter, “Normalization of single-cell RNA-seq counts by $\log(x + 1)$ or $\log(1 + x)$,” *Bioinformatics*, Mar 2021.

[85] E. Maza, “In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design,” *Front. Genet.*, vol. 0, 2016.

[86] C. Filloux, M. Cédric, P. Romain, F. Lionel, K. Christophe, R. Dominique, M. Abderrahman, and P. Daniel, “An integrative method to normalize RNA-Seq data,” *BMC Bioinf.*, vol. 15, pp. 1–11, Dec 2014.

[87] Z. Zhou, P. Cong, Y. Tian, and Y. Zhu, “Using RNA-seq data to select reference genes for normalizing gene expression in apple roots,” *PLoS One*, vol. 12, p. e0185288, Sep 2017.

[88] J. B. de Kok, R. W. Roelofs, B. A. Giesendorf, J. L. Pennings, E. T. Waas, T. Feuth, D. W. Swinkels, and P. N. Span, “Normalization of gene expression measurements in tumor tissues : comparison of 13 endogenous control genes - Laboratory Investigation,” *Lab. Invest.*, vol. 85, pp. 154–159, Jan 2005.

[89] C. Soneson and M. Delorenzi, “A comparison of methods for differential expression analysis of RNA-seq data,” *BMC Bioinf.*, vol. 14, pp. 1–18, Dec 2013.

[90] D. Spies, P. F. Renz, T. A. Beyer, and C. Ciaudo, “Comparative analysis of differential gene expression tools for RNA sequencing time course data,” *Briefings Bioinf.*, vol. 20, pp. 288–298, Jan 2019.

[91] J. Costa-Silva, D. Domingues, and F. M. Lopes, “RNA-Seq differential expression analysis : An extended review and a software tool,” *PLoS One*, vol. 12, p. e0190152, Dec 2017.

[92] G. Östlund and E. L. Sonnhammer, “Avoiding pitfalls in gene (co)expression meta-analysis,” *Genomics*, vol. 103, pp. 21–30, Jan. 2014.

[93] A. de la Fuente, “From ‘differential expression’ to ‘differential networking’ – identification of dysfunctional regulatory networks in diseases,” *Trends Genet.*, vol. 26, pp. 326–333, Jul 2010.

[94] P. Parsana, C. Ruberman, A. E. Jaffe, M. C. Schatz, A. Battle, and J. T. Leek, “Addressing confounding artifacts in reconstruction of gene co-expression networks,” *Genome Biology*, vol. 20, no. 1, p. 94, 2019.

[95] B. Zhang and S. Horvath, “A general framework for weighted gene co-expression network analysis,” *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, 2005.

[96] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, “Network medicine : a network-based approach to human disease - Nature Reviews Genetics,” *Nat. Rev. Genet.*, vol. 12, pp. 56–68, Jan 2011.

[97] J. A. Barnes and F. Harary, “Graph theory in network analysis,” *Social Networks*, vol. 5, pp. 235–244, Jun 1983.

[98] S. Van Dam, *Development and exploitation of GeneFriends : An online database for gene and transcript co-expression analysis*. PhD thesis, University of Liverpool, Liverpool, England, UK, Feb 2017.

[99] A. Vandenbon, “Guidance for RNA-seq co-expression estimates : the importance of data normalization, batch effects, and correlation measures,” *bioRxiv*, p. 2021.03.11.435043, Mar 2021.

[100] A. Reverter, W. Barris, S. McWilliam, K. A. Byrne, Y. H. Wang, S. H. Tan, N. Hudson, and B. P. Dalrymple, “Validation of alternative methods of data normalization in gene co-expression studies,” *Bioinformatics*, vol. 21, pp. 1112–1120, Apr 2005.

[101] W. K. Lim, K. Wang, C. Lefebvre, and A. Califano, “Comparative analysis of microarray normalization procedures : effects on reverse engineering gene networks,” *Bioinformatics*, vol. 23, pp. i282–i288, Jul 2007.

[102] F. M. Giorgi, C. Del Fabbro, and F. Licausi, “Comparative study of RNA-seq- and Microarray-derived coexpression networks in Arabidopsis thaliana,” *Bioinformatics*, vol. 29, pp. 717–724, Mar 2013.

[103] S. Ballouz, W. Verleyen, and J. Gillis, “Guidance for RNA-seq co-expression network construction and analysis : Safety in numbers,” *Bioinformatics*, vol. 31, no. 13, pp. 2123–2130, 2015.

[104] H. A. Chowdhury, D. K. Bhattacharyya, and J. K. Kalita, “(Differential) Co-Expression Analysis of Gene Expression : A Survey of Best Practices,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. c, pp. 1–1, 2019.

[105] F. Liesecke, J. O. De Craene, S. Besseau, V. Courdavault, M. Clastre, V. Vergès, N. Papon, N. Giglioli-Guivarc’h, G. Glévarec, O. Pichon, and T. Dugé de Bernonville, “Improved gene co-expression network quality through expression dataset down-sampling and network aggregation,” *Scientific Reports*, vol. 9, no. 1, pp. 1–16, 2019.

[106] S. L. Carter, C. M. Brechbuhler, M. Griffin, and A. T. Bond, “Gene co-expression network topology provides a framework for molecular characterization of cellular state,” *Bioinformatics*, vol. 20, pp. 2242–2250, sep 2004.

[107] E. A. R. Serin, H. Nijveen, H. W. M. Hilhorst, and W. Ligterink, “Learning from Co-expression Networks : Possibilities and Challenges,” *Frontiers in Plant Science*, vol. 7, no. April, pp. 1–18, 2016.

[108] A. Kuehne, J. Hildebrand, J. Soehle, H. Wenck, L. Terstegen, S. Gallinat, A. Knott, M. Winnefeld, and N. Zamboni, “An integrative metabolomics and transcriptomics study to identify metabolic alterations in aged skin of humans in vivo,” *BMC Genomics*, vol. 18, no. 1, p. 169, 2017.

[109] L. Song, P. Langfelder, and S. Horvath, “Comparison of co-expression measures : Mutual information, correlation, and model based indices,” *BMC Bioinformatics*, vol. 13, no. 1, 2012.

[110] S. Kullback, *Information theory and statistics*. Courier Corporation, 1978.

[111] A. D. Broido and A. Clauset, “Scale-free networks are rare - Nature Communications,” *Nat. Commun.*, vol. 10, pp. 1–10, Mar 2019.

[112] A.-L. Barabási and R. Albert, “Emergence of Scaling in Random Networks,” *Science*, vol. 286, pp. 509–512, Oct 1999.

[113] R. Cohen and S. Havlin, “Scale-Free Networks Are Ultrasmall,” *Phys. Rev. Lett.*, vol. 90, p. 058701, Feb 2003.

[114] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães, “Gene co-expression analysis for functional classification and gene-disease predictions,” *Briefings in Bioinformatics*, vol. 19, no. 4, pp. 575–592, 2018.

[115] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási, “Hierarchical organization of modularity in metabolic networks,” *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[116] V. Batagelj and A. Mrvar, “Pajek-program for large network analysis,” *Connect*, vol. 21, pp. 47–57, 01 1998.

[117] D. M. Lorenz, A. Jeng, and M. W. Deem, “The emergence of modularity in biological systems,” *Physics of Life Reviews*, vol. 25, pp. 289–313, feb 2011.

[118] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, “Cluster analysis and display of genome-wide expression patterns,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 95, pp. 14863–14868, Dec 1998.

[119] A. M. Yip and S. Horvath, “Gene network interconnectedness and the generalized topological overlap measure,” *BMC Bioinf.*, vol. 8, pp. 1–14, Dec 2007.

[120] D. M. Gysi, A. Voigt, T. D. M. Fragoso, E. Almaas, and K. Nowick, “wTO : An R package for computing weighted topological overlap and a consensus network with integrated visualization tool,” *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–16, 2018.

[121] W. Saelens, R. Cannoodt, and Y. Saeys, “A comprehensive evaluation of module detection methods for gene expression data,” *Nature Communications*, vol. 9, no. 1, 2018.

[122] M. Filteau, S. A. Pavey, J. St-Cyr, and L. Bernatchez, “Gene coexpression networks reveal key drivers of phenotypic divergence in lake whitefish,” *Molecular Biology and Evolution*, vol. 30, no. 6, pp. 1384–1396, 2013.

[123] S. Sundarrajan and M. Arumugam, “Weighted gene co-expression based biomarker discovery for psoriasis detection,” *Gene*, vol. 593, no. 1, pp. 225–234, 2016.

[124] L. J. Kogelman, S. Cirera, D. V. Zhernakova, M. Fredholm, L. Franke, and H. N. Kadarmideen, “Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA Sequencing in a porcine model,” *BMC Medical Genomics*, vol. 7, no. 1, 2014.

[125] J. Ruan and W. Zhang, “Identification and Evaluation of Functional Modules in Gene Co-expression Networks,” in *Systems Biology and Computational Proteomics*, pp. 57–76, Berlin, Germany : Springer, Dec 2006.

[126] Z. Shi, C. K. Derow, and B. Zhang, “Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression,” *BMC Syst. Biol.*, vol. 4, pp. 1–14, Dec 2010.

[127] A. Rau and C. Maugis-Rabusseau, “Transformation and model choice for RNA-seq co-expression analysis,” *Briefings Bioinf.*, vol. 19, pp. 425–436, May 2018.

[128] J. Tang, D. Kong, Q. Cui, K. Wang, D. Zhang, Y. Gong, and G. Wu, “Prognostic genes of breast cancer identified by gene co-expression network analysis,” *Frontiers in Oncology*, vol. 8, no. SEP, pp. 1–13, 2018.

[129] L. Mao, J. L. Van Hemert, S. Dash, and J. A. Dickerson, “Arabidopsis gene co-expression network and its functional modules,” *BMC Bioinformatics*, vol. 10, pp. 1–24, 2009.

[130] M. Rotival and E. Petretto, “Leveraging gene co-expression networks to pinpoint the regulation of complex traits and disease, with a focus on cardiovascular traits,” *Briefings in Functional Genomics*, vol. 13, no. 1, pp. 66–78, 2014.

[131] F. Murtagh and P. Contreras, “Algorithms for hierarchical clustering : an overview,” *WIREs Data Min. Knowl. Discovery*, vol. 2, pp. 86–97, Jan 2012.

[132] P. Langfelder, B. Zhang, and S. Horvath, “Defining clusters from a hierarchical cluster tree : The Dynamic Tree Cut package for R,” *Bioinformatics*, vol. 24, no. 5, pp. 719–720, 2008.

[133] P. Langfelder and S. Horvath, “WGCNA : An R package for weighted correlation network analysis,” *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[134] M. L. Green and P. D. Karp, “The outcomes of pathway database computations depend on pathway ontology,” *Nucleic Acids Res.*, vol. 34, pp. 3687–3697, Aug 2006.

[135] P. Khatri, M. Sirota, and A. J. Butte, “Ten Years of Pathway Analysis : Current Approaches and Outstanding Challenges,” *PLoS Computational Biology*, vol. 8, p. e1002375, feb 2012.

[136] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Bioinformatics enrichment tools : paths toward the comprehensive functional analysis of large gene lists,” *Nucleic Acids Res.*, vol. 37, pp. 1–13, Jan 2009.

[137] M. Ackermann and K. Strimmer, “A general modular framework for gene set enrichment analysis,” *BMC Bioinf.*, vol. 10, pp. 1–20, Dec 2009.

[138] J. Rahnenführer, F. S. Domingues, J. Maydt, and T. Lengauer, “Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data,” *Stat. Appl. Genet. Mol. Biol.*, vol. 3, Jun 2004.

[139] G. G. Lemoine, M.-P. Scott-Boyer, B. Ambroise, O. Périn, and A. Droit, “GWENA : gene co-expression networks analysis and extended modules characterization in a single Bioconductor package,” *BMC Bioinf.*, vol. 22, pp. 1–20, Dec 2021.

[140] C. J. Wolfe, I. S. Kohane, and A. J. Butte, “Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks,” *BMC Bioinformatics*, vol. 6, pp. 1–10, 2005.

[141] J. A. Timmons, K. J. Szkop, and I. J. Gallagher, “Multiple sources of bias confound functional enrichment analysis of global -omics data,” *Genome Biol.*, vol. 16, pp. 1–3, Dec 2015.

[142] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, “Lethality and centrality in protein networks - Nature,” *Nature*, vol. 411, pp. 41–42, May 2001.

[143] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, “The human disease network,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 104, pp. 8685–8690, May 2007.

[144] P. Langfelder, P. S. Mischel, and S. Horvath, "When Is Hub Gene Selection Better than Standard Meta-Analysis ?," *PLoS ONE*, vol. 8, no. 4, 2013.

[145] P. S. Russo, G. R. Ferreira, L. E. Cardozo, M. C. Bürger, R. Arias-Carrasco, S. R. Maruyama, T. D. Hirata, D. S. Lima, F. M. Passos, K. F. Fukutani, M. Lever, J. S. Silva, V. Maracaja-Coutinho, and H. I. Nakaya, "CEMiTool : A Bioconductor package for performing comprehensive modular co-expression analyses," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–13, 2018.

[146] A. Saha, Y. Kim, A. D. Gewirtz, B. Jo, C. Gao, I. C. McDowell, B. E. Engelhardt, and A. Battle, "Co-expression networks reveal the tissue-specific regulation of transcription and splicing," *Genome Research*, vol. 27, pp. 1843–1858, nov 2017.

[147] S. Das, P. K. Meher, A. Rai, L. M. Bhar, and B. N. Mandal, "Statistical Approaches for Gene Selection, Hub Gene Identification and Module Interaction in Gene Co-Expression Network Analysis : An Application to Aluminum Stress in Soybean (Glycine max L.)," *PLoS One*, vol. 12, no. 1, 2017.

[148] S. Horvath and J. Dong, "Geometric interpretation of gene coexpression network analysis," *PLoS Computational Biology*, vol. 4, p. e1000117, aug 2008.

[149] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, "Dynamic modularity in protein interaction networks predicts breast cancer outcome - Nature Biotechnology," *Nat. Biotechnol.*, vol. 27, pp. 199–204, Feb 2009.

[150] P. Langfelder, R. Luo, M. C. Oldham, and S. Horvath, "Is my network module preserved and reproducible ?," *PLoS Computational Biology*, vol. 7, no. 1, 2011.

[151] Y. Lai, B. Wu, L. Chen, and H. Zhao, "A statistical method for identifying differential gene–gene co-expression patterns," *Bioinformatics*, vol. 20, pp. 3146–3155, Nov 2004.

[152] E. Gov and K. Y. Arga, "Differential co-expression analysis reveals a novel prognostic gene module in ovarian cancer," *Scientific Reports*, vol. 7, no. 1, pp. 1–10, 2017.

[153] L. K. Southworth, A. B. Owen, and S. K. Kim, "Aging Mice Show a Decreasing Correlation of Gene Expression within Genetic Modules," *PLoS Genetics*, vol. 5, p. e1000776, Dec. 2009.

[154] S. C. Ritchie, S. Watts, L. G. Fearnley, K. E. Holt, G. Abraham, and M. Inouye, "A Scalable Permutation Approach Reveals Replication and Preservation Patterns of Network Modules in Large Datasets," *Cell Systems*, vol. 3, no. 1, pp. 71–82, 2016.

[155] G. Berrut, *Patient âgé : particularités de la consultation*. Édition du Collège de la Médecine Générale, 2013.

[156] S. S. Khan, B. D. Singer, and D. E. Vaughan, "Molecular and physiological manifestations and measurement of aging in humans," *Aging Cell*, vol. 16, pp. 624–633, Aug 2017.

[157] M. E. Todhunter, R. W. Sayaman, M. Miyano, and M. A. LaBarge, “Tissue aging : the integration of collective and variant responses of cells to entropic forces over time,” *Current Opinion in Cell Biology*, vol. 54, pp. 121–129, 2018.

[158] M. A. van Buchem, “Introduction : Successful, Usual, and Pathological Aging,” *Top. Magn. Reson. Imaging*, vol. 15, p. 341, Dec 2004.

[159] J. Belmin, *Gériatrie*. Issy-les-Moulineaux : Elsevier-Masson, 2014.

[160] S. Horvath, “DNA methylation age of human tissues and cell types,” *Genome Biol.*, vol. 14, pp. 1–20, Oct 2013.

[161] J. P. de Magalhães, “Is mammalian aging genetically controlled ?,” *Biogerontology*, vol. 4, pp. 119–120, Mar 2003.

[162] M. J. Jones, S. J. Goodman, and M. S. Kobor, “DNA methylation and healthy human aging,” *Aging Cell*, vol. 14, pp. 924–932, Dec 2015.

[163] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, “The hallmarks of aging,” *Cell*, vol. 153, no. 6, 2013.

[164] A. A. Moskalev, M. V. Shaposhnikov, E. N. Plyusnina, A. Zhavoronkov, A. Budovsky, H. Yanai, and V. E. Fraifeld, “The role of DNA damage and repair in aging through the prism of Koch-like criteria,” *Ageing Res. Rev.*, vol. 12, pp. 661–684, Mar 2013.

[165] C. J. Lord and A. Ashworth, “The DNA damage response and cancer therapy - Nature,” *Nature*, vol. 481, pp. 287–294, Jan 2012.

[166] L. Hayflick and P. S. Moorhead, “The serial cultivation of human diploid cell strains,” *Exp. Cell Res.*, vol. 25, pp. 585–621, Dec 1961.

[167] P. Ilmonen, A. Kotrschal, and D. J. Penn, “Telomere Attrition Due to Infection,” *PLoS One*, vol. 3, p. e2143, May 2008.

[168] M. F. Fraga and M. Esteller, “Epigenetics and aging : the targets and the marks,” *Trends Genet.*, vol. 23, pp. 413–418, Aug 2007.

[169] C. Jin, J. Li, C. D. Green, X. Yu, X. Tang, D. Han, B. Xian, D. Wang, X. Huang, X. Cao, Z. Yan, L. Hou, J. Liu, N. Shukeir, P. Khaitovich, C. D. Chen, H. Zhang, T. Jenuwein, and J.-D. J. Han, “Histone Demethylase UTX-1 Regulates C. elegans Life Span by Targeting the Insulin/IGF-1 Signaling Pathway,” *Cell Metab.*, vol. 14, pp. 161–172, Aug 2011.

[170] H. Koga, S. Kaushik, and A. M. Cuervo, “Protein homeostasis and aging : The importance of exquisite quality control,” *Ageing Res. Rev.*, vol. 10, pp. 205–215, Apr 2011.

[171] R. H. Houtkooper, R. W. Williams, and J. Auwerx, “Metabolic Networks of Longevity,” *Cell*, vol. 142, pp. 9–14, Jul 2010.

[172] M. Laplante and D. M. Sabatini, “mTOR Signaling in Growth Control and Disease,” *Cell*, vol. 149, pp. 274–293, Apr 2012.

[173] D. R. Green, L. Galluzzi, and G. Kroemer, “Mitochondria and the Autophagy–Inflammation–Cell Death Axis in Organismal Aging,” *Science*, vol. 333, pp. 1109–1112, Aug 2011.

[174] D. V. Ziegler, C. D. Wiley, and M. C. Velarde, “Mitochondrial effectors of cellular senescence : beyond the free radical theory of aging,” *Aging Cell*, vol. 14, pp. 1–7, Feb 2015.

[175] T. Kuilman, C. Michaloglou, W. J. Mooi, and D. S. Peeper, “The essence of senescence,” *Genes Dev.*, vol. 24, pp. 2463–2479, Nov 2010.

[176] C. Franceschi and J. Campisi, “Chronic Inflammation (Inflammaging) and Its Potential Contribution to Age-Associated Diseases,” *J. Gerontol. A Biol. Sci. Med. Sci.*, vol. 69, pp. S4–S9, Jun 2014.

[177] N. E. Sharpless and R. A. DePinho, “How stem cells age and why this makes us grow old - Nature Reviews Molecular Cell Biology,” *Nat. Rev. Mol. Cell Biol.*, vol. 8, pp. 703–713, Sep 2007.

[178] D. J. Baker, B. G. Childs, M. Durik, M. E. Wijers, C. J. Sieben, J. Zhong, R. A. Saltness, K. B. Jeganathan, G. C. Verzosa, A. Pezeshki, K. Khazaie, J. D. Miller, and J. M. van Deursen, “Naturally occurring p16Ink4a-positive cells shorten healthy lifespan - Nature,” *Nature*, vol. 530, pp. 184–189, Feb 2016.

[179] M. Lavasani, A. R. Robinson, A. Lu, M. Song, J. M. Feduska, B. Ahani, J. S. Tilstra, C. H. Feldman, P. D. Robbins, L. J. Niedernhofer, and J. Huard, “Muscle-derived stem/progenitor cell dysfunction limits healthspan and lifespan in a murine progeria model - Nature Communications,” *Nat. Commun.*, vol. 3, pp. 1–12, Jan 2012.

[180] L. Ferrucci, M. Gonzalez-Freire, E. Fabbri, E. Simonsick, T. Tanaka, Z. Moore, S. Salimi, F. Sierra, and R. de Cabo, “Measuring biological aging in humans : A quest,” *Aging Cell*, vol. 19, p. e13080, Feb 2020.

[181] L. Hayflick, “The future of ageing - Nature,” *Nature*, vol. 408, pp. 267–269, Nov 2000.

[182] N. Kubben and T. Misteli, “Shared molecular and cellular mechanisms of premature ageing and ageing-associated diseases,” *Nature Reviews Molecular Cell Biology*, vol. 18, no. 10, pp. 595–609, 2017.

[183] D. R. Phillips and R. M. Gyasi, “Global Aging in a Comparative Context,” *Gerontologist*, vol. 61, pp. 476–477, Apr 2021.

[184] A. Blasimme, “Physical frailty, sarcopenia, and the enablement of autonomy : philosophical issues in geriatric medicine,” *Aging Clin. Exp. Res.*, vol. 29, pp. 59–63, Feb 2017.

[185] M. Saint and P. C. Rath, “Transcription and Aging,” in *Models, Molecules and Mechanisms in Biogerontology*, pp. 43–66, Singapore : Springer, Jun 2020.

[186] D. Melzer, L. C. Pilling, and L. Ferrucci, “The genetics of human ageing,” *Nature Reviews Genetics*, vol. 21, no. 2, pp. 88–101, 2020.

[187] J. P. De Magalhães and O. Toussaint, “GenAge : A genomic and proteomic network map of human ageing,” *FEBS Letters*, vol. 571, no. 1-3, pp. 243–247, 2004.

[188] T. Craig, C. Smelick, R. Tacutu, D. Wuttke, S. H. Wood, H. Stanley, G. Janssens, E. Savitskaya, A. Moskalev, R. Arking, and J. P. De Magalhães, “The Digital Ageing Atlas : Integrating the diversity of age-related changes into a unified resource,” *Nucleic Acids Research*, vol. 43, no. D1, pp. D873–D878, 2015.

[189] J. P. de Magalhães, J. Curado, and G. M. Church, “Meta-analysis of age-related gene expression profiles identifies common signatures of aging,” *Bioinformatics*, vol. 25, no. 7, pp. 875–881, 2009.

[190] M. Armanios and E. H. Blackburn, “The telomere syndromes.,” *Nature reviews. Genetics*, vol. 13, no. 10, pp. 693–704, 2012.

[191] J. Zierer, C. Menni, G. Kastenmüller, and T. D. Spector, “Integration of ’omics’ data in aging research : From biomarkers to systems biology,” *Aging Cell*, vol. 14, no. 6, pp. 933–944, 2015.

[192] R. Anglani, T. M. Creanza, V. C. Liuzzi, A. Piepoli, A. Panza, A. Andriulli, and N. Ancona, “Loss of connectivity in cancer co-expression networks,” *PLoS ONE*, vol. 9, no. 1, 2014.

[193] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, “ARACNE : An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context,” *BMC Bioinformatics*, vol. 7, no. SUPPL.1, pp. 1–15, 2006.

[194] D. Hanahan and R. A. Weinberg, “The Hallmarks of Cancer,” *Cell*, vol. 100, pp. 57–70, Jan 2000.

[195] H. R. Woo, H. J. Koo, J. Kim, H. Jeong, J. O. Yang, I. H. Lee, J. H. Jun, S. H. Choi, S. J. Park, B. Kang, Y. W. Kim, B.-K. Phee, J. H. Kim, C. Seo, C. Park, S. C. Kim, S. Park, B. Lee, S. Lee, D. Hwang, H. G. Nam, and P. O. Lim, “Programming of Plant Leaf Senescence with Temporal and Inter-Organellar Coordination of Transcriptome in Arabidopsis,” *Plant Physiol.*, vol. 171, pp. 452–467, May 2016.

[196] F. Bormann, M. Rodríguez-Paredes, S. Hagemann, H. Manchanda, B. Kristof, J. Gutekunst, G. Raddatz, R. Haas, L. Terstegen, H. Wenck, L. Kaderali, M. Winnefeld, and F. Lyko, “Reduced DNA methylation patterning and transcriptional connectivity define human skin aging,” *Aging Cell*, vol. 15, no. 3, pp. 563–571, 2016.

[197] D. Derous, S. E. Mitchell, C. L. Green, Y. Wang, J. D. J. Han, L. Chen, D. E. L. Promislow, D. Lusseau, J. R. Speakman, and A. Douglas, "The effects of graded levels of calorie restriction : VII. Topological rearrangement of hypothalamic aging networks," *Aging (Albany NY)*, vol. 8, p. 917, May 2016.

[198] Q. Zhang, R. Nogales-Cadenas, J.-R. Lin, W. Zhang, Y. Cai, J. Vijg, and Z. D. Zhang, "Systems-level analysis of human aging genes shed new light on mechanisms of aging," *Hum. Mol. Genet.*, vol. 25, pp. 2934–2947, Jul 2016.

[199] C. P. Martinez-Jimenez, N. Eling, H.-C. Chen, C. A. Vallejos, A. A. Kolodziejczyk, F. Connor, L. Stojic, T. F. Rayner, M. J. T. Stubbington, S. A. Teichmann, M. de la Roche, J. C. Marioni, and D. T. Odom, "Aging increases cell-to-cell transcriptional variability upon immune stimulation," *Science*, vol. 355, pp. 1433–1436, Mar 2017.

[200] E. Scheller, L. V. Schumacher, J. Peter, J. Lahr, J. Wehrle, C. P. Kaller, C. Gaser, and S. Klöppel, "Brain Aging and APOE ϵ4 Interact to Reveal Potential Neuronal Compensation in Healthy Older Adults," *Front. Aging Neurosci.*, vol. 0, 2018.

[201] J. S. Hawe, F. J. Theis, and M. Heinig, "Inferring Interaction Networks From Multi-Omics Data," *Front. Genet.*, vol. 0, 2019.

[202] G. Pei, L. Chen, and W. Zhang, "WGCNA Application to Proteomic and Metabolomic Data Analysis," in *Methods in Enzymology*, vol. 585, pp. 135–158, Cambridge, MA, USA : Academic Press, Jan 2017.

[203] I. Solovev, M. Shaposhnikov, and A. Moskalev, "Multi-omics approaches to human biological age estimation," *Mech. Ageing Dev.*, vol. 185, p. 111192, Jan 2020.

[204] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale - Nature Methods," *Nat. Methods*, vol. 11, pp. 333–337, Mar 2014.

[205] H. Yuan, X. Zeng, Q. Yang, Q. Xu, Y. Wang, D. Jabu, Z. Sang, and N. Tashi, "Gene coexpression network analysis combined with metabonomics reveals the resistance responses to powdery mildew in Tibetan hulless barley - Scientific Reports," *Sci. Rep.*, vol. 8, pp. 1–13, Oct 2018.

[206] C. Franceschi, P. Garagnani, G. Vitale, M. Capri, and S. Salvioli, "Inflammaging and 'garbaging'," *Trends in Endocrinology & Metabolism*, vol. 28, pp. 199–212, Mar. 2017.

[207] Y.-C. Liu, C.-P. Cheng, and V. S. Tseng, "Mining differential top-k co-expression patterns from time course comparative gene expression datasets," *BMC Bioinf.*, vol. 14, pp. 1–13, Dec 2013.

[208] A. Saha, Y. Kim, A. D. H. Gewirtz, B. Jo, C. Gao, I. C. McDowell, The GTEx Consortium, B. E. Engelhardt, A. Battle, F. Aguet, K. G. Ardlie, B. B. Cummings, E. T. Gelfand, G. Getz,

K. Hadley, R. E. Handsaker, K. H. Huang, S. Kashin, K. J. Karczewski, M. Lek, X. Li, D. G. MacArthur, J. L. Nedzel, D. T. Nguyen, M. S. Noble, A. V. Segrè, C. A. Trowbridge, T. Tukiainen, N. S. Abell, B. Balliu, R. Barshir, O. Basha, A. Battle, G. K. Bogu, A. Brown, C. D. Brown, S. E. Castel, L. S. Chen, C. Chiang, D. F. Conrad, N. J. Cox, F. N. Damani, J. R. Davis, O. Delaneau, E. T. Dermitzakis, B. E. Engelhardt, E. Eskin, P. G. Ferreira, L. Frésard, E. R. Gamazon, D. Garrido-Martín, A. D. H. Gewirtz, G. Gliner, M. J. Gloudemans, R. Guigo, I. M. Hall, B. Han, Y. He, F. Hormozdiari, C. Howald, H. K. Im, B. Jo, E. Y. Kang, Y. Kim, S. Kim-Hellmuth, T. Lappalainen, G. Li, X. Li, B. Liu, S. Mangul, M. I. McCarthy, I. C. McDowell, P. Mohammadi, J. Monlong, S. B. Montgomery, M. Muñoz-Aguirre, A. W. Ndungu, D. L. Nicolae, A. B. Nobel, M. Oliva, H. Ongen, J. J. Palowitch, N. Panousis, P. Papasaikas, Y. Park, P. Parsana, A. J. Payne, C. B. Peterson, J. Quan, F. Reverter, C. Sabatti, A. Saha, M. Sammeth, A. J. Scott, A. A. Shabalin, R. Sodaei, M. Stephens, B. E. Stranger, B. J. Strober, J. H. Sul, E. K. Tsang, S. Urbut, M. van de Bunt, G. Wang, X. Wen, F. A. Wright, H. S. Xi, E. Yeger-Lotem, Z. Zappala, J. B. Zaugg, Y.-H. Zhou, J. M. Akey, D. Bates, J. Chan, L. S. Chen, M. Claussnitzer, K. Demanelis, M. Diegel, J. A. Doherty, A. P. Feinberg, M. S. Fernando, J. Halow, K. D. Hansen, E. Haugen, P. F. Hickey, L. Hou, F. Jasmine, R. Jian, L. Jiang, A. Johnson, R. Kaul, M. Kellis, M. G. Kibriya, K. Lee, J. B. Li, Q. Li, X. Li, J. Lin, S. Lin, S. Linder, C. Linke, Y. Liu, M. T. Maurano, B. Molinie, S. B. Montgomery, J. Nelson, F. J. Neri, M. Oliva, Y. Park, B. L. Pierce, N. J. Rinaldi, L. F. Rizzardi, R. Sandstrom, A. Skol, K. S. Smith, M. P. Snyder, J. Stamatoyannopoulos, B. E. Stranger, H. Tang, E. K. Tsang, L. Wang, M. Wang, N. Van Wittenberghe, F. Wu, R. Zhang, C. R. Nierras, P. A. Branton, L. J. Carithers, P. Guan, H. M. Moore, A. Rao, J. B. Vaught, S. E. Gould, N. C. Lockart, C. Martin, J. P. Struewing, S. Volpi, A. M. Addington, S. E. Koester, A. R. Little, L. E. Brigham, R. Hasz, M. Hunter, C. Johns, M. Johnson, G. Kopen, W. F. Leinweber, J. T. Lonsdale, A. McDonald, B. Mestichelli, K. Myer, B. Roe, M. Salvatore, S. Shad, J. A. Thomas, G. Walters, M. Washington, J. Wheeler, J. Bridge, B. A. Foster, B. M. Gillard, E. Karasik, R. Kumar, M. Miklos, M. T. Moser, S. D. Jewell, R. G. Montroy, D. C. Rohrer, D. R. Valley, D. A. Davis, D. C. Mash, A. H. Undale, A. M. Smith, D. E. Tabor, N. V. Roche, J. A. McLean, N. Vatanian, K. L. Robinson, L. Sobin, M. E. Barcus, K. M. Valentino, L. Qi, S. Hunter, P. Hariharan, S. Singh, K. S. Um, T. Matose, M. M. Tomaszewski, L. K. Barker, M. Mosavel, L. A. Siminoff, H. M. Traino, P. Flicek, T. Juettemann, M. Ruffier, D. Sheppard, K. Taylor, S. J. Trevanion, D. R. Zerbino, B. Craft, M. Goldman, M. Haeussler, W. J. Kent, C. M. Lee, B. Paten, K. R. Rosenbloom, J. Vivian, and J. Zhu, "Co-expression networks reveal the tissue-specific regulation of transcription and splicing," *Genome Res.*, vol. 27, pp. 1843–1858, Oct 2017.

[209] H.-Z. Sun, Z. Zhu, M. Zhou, J. Wang, M. E. R. Dugan, and L. L. Guan, "Gene co-expression and alternative splicing analysis of key metabolic tissues to unravel the regulatory signatures of fatty acid composition in cattle," *RNA Biol.*, vol. 18, pp. 854–862, Jun 2021.

[210] B. Uyar, D. Palmer, A. Kowald, H. M. Escobar, I. Barrantes, S. Möller, A. Akalin, and G. Fuellen, "Single-cell analyses of aging, inflammation and senescence," *Ageing Research Re-*

views, vol. 64, p. 101156, Dec. 2020.

[211] S. S. Fonseca Costa, M. Robinson-Rechavi, and J. A. Ripperger, "Single-cell transcriptomics allows novel insights into aging and circadian processes," *Brief. Funct. Genomics*, vol. 19, pp. 343–349, Dec 2020.

[212] M. Menon, S. Mohammadi, J. Davila-Velderrain, B. A. Goods, T. D. Cadwell, Y. Xing, A. Stemmer-Rachamimov, A. K. Shalek, J. C. Love, M. Kellis, and B. P. Hafler, "Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration - Nature Communications," *Nat. Commun.*, vol. 10, pp. 1–9, Oct 2019.

[213] J. Deelen, D. S. Evans, D. E. Arking, N. Tesi, M. Nygaard, X. Liu, M. K. Wojczynski, M. L. Biggs, A. van der Spek, G. Atzmon, E. B. Ware, C. Sarnowski, A. V. Smith, I. Seppälä, H. J. Cordell, J. Dose, N. Amin, A. M. Arnold, K. L. Ayers, N. Barzilai, E. J. Becker, M. Beekman, H. Blanché, K. Christensen, L. Christiansen, J. C. Collerton, S. Cubaynes, S. R. Cummings, K. Davies, B. Debrabant, J.-F. Deleuze, R. Duncan, J. D. Faul, C. Franceschi, P. Galan, V. Gudnason, T. B. Harris, M. Huisman, M. A. Hurme, C. Jagger, I. Jansen, M. Jylhä, M. Kähönen, D. Karasik, S. L. R. Kardia, A. Kingston, T. B. L. Kirkwood, L. J. Launer, T. Lehtimäki, W. Lieb, L.-P. Lyytikäinen, C. Martin-Ruiz, J. Min, A. Nebel, A. B. Newman, C. Nie, E. A. Nohr, E. S. Orwoll, T. T. Perls, M. A. Province, B. M. Psaty, O. T. Raitakari, M. J. T. Reinders, J.-M. Robine, J. I. Rotter, P. Sebastiani, J. Smith, T. I. A. Sørensen, K. D. Taylor, A. G. Uitterlinden, W. van der Flier, S. J. van der Lee, C. M. van Duijn, D. van Heemst, J. W. Vaupel, D. Weir, K. Ye, Y. Zeng, W. Zheng, H. Holstege, D. P. Kiel, K. L. Lunetta, P. E. Slagboom, and J. M. Murabito, "A meta-analysis of genome-wide association studies identifies multiple longevity genes - Nature Communications," *Nat. Commun.*, vol. 10, pp. 1–14, Aug 2019.

[214] J. R. Aunan, W. C. Cho, and K. Søreide, "The biology of aging and cancer : A brief overview of shared and divergent molecular hallmarks," *Aging and Disease*, vol. 8, no. 5, pp. 628–642, 2017.

[215] A. Moreira-Pais, R. Ferreira, P. A. Oliveira, and J. A. Duarte, "Sarcopenia versus cancer cachexia : the muscle wasting continuum in healthy and diseased aging," *Biogerontology*, pp. 1–19, Jul 2021.

[216] R. Tacutu, D. Thornton, E. Johnson, A. Budovsky, D. Barardo, T. Craig, E. DIana, G. Lehmann, D. Toren, J. Wang, V. E. Fraifeld, and J. P. De Magalhães, "Human Ageing Genomic Resources : New and updated databases," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1083–D1090, 2018.

[217] A. Verguet, *Développements méthodologiques et informatiques pour la microscopie électronique en transmission appliqués à des échantillons biologiques*. PhD thesis, Université Paris Saclay (COmUE), Paris, France, Dec 2019.

[218] D. M. Keeling, P. Garza, C. M. Nartey, and A.-R. Carvunis, “Philosophy of Biology : The meanings of ’function’ in biology and the problematic case of de novo gene emergence,” *eLife*, Nov 2019.

[219] W. Bechtel, “Understanding Biological Mechanisms : Using Illustrations from Circadian Rhythm Research,” *History, Philosophy and Theory of the Life Sciences*, vol. 1, pp. 487–510, 2013.

Annexes

Fichier additionnel associé au chapitre 2

Additional file 1

GWENA : gene co-expression networks analysis and extended modules characterization in a single Bioconductor package

Gwenaëlle G. Lemoine, Marie-Pier Scott-Boyer, Bathilde Ambroise, Olivier Périn, Arnaud Droit

A.1 Supplementary Material and Method

A.1.1 Z summary detail and combination with NetRep

As NetRep uses a permutation test with the null hypothesis of the module being not preserved, it can only return if the module is preserved or not significant. To determine if a module is not preserved or moderately preserved, a Z summary statistic is computed using the topological metrics defined by Langfelder et al. [150] and renamed by NetRep [154] such as :

$$Z_{summary} = \frac{Z_{density} + Z_{connectivity}}{2}$$

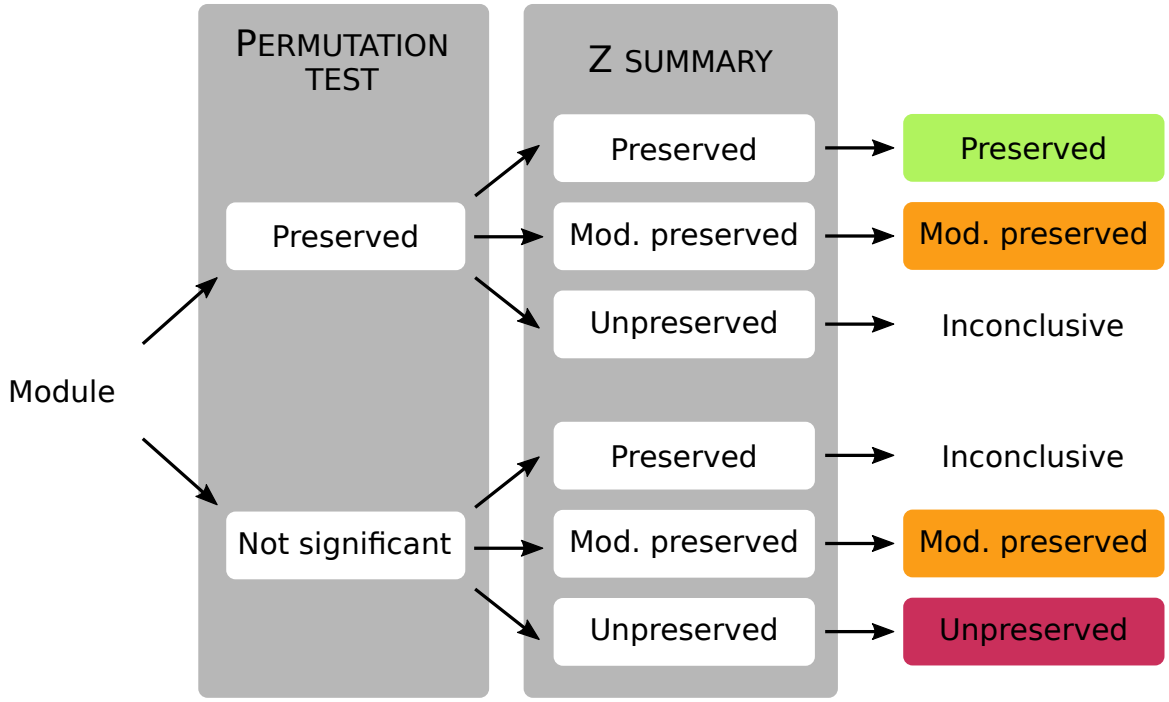

FIGURE A.1 – Combination of the permutation test result ans the $Z_{summary}$ result in GWENA to return a final result on the module comparison.

With the NetRep notation :

$$Z_{density} = median(Z_{cor.density}, Z_{avg.edge.wei}, Z_{mod.coh}, Z_{avg.node.contrib})$$
$$Z_{connectivity} = median(Z_{concor.wei.deg}, Z_{concor.nod.contrib}, Z_{concor.cor})$$

Where :

$$vect.Matrix(A) = (a_{2,1}, a_{3,1}, ..., a_{n,1}, a_{n,n-1})$$
$$r_{ij}^{[ref]} = \text{cor}(x_i^{[ref]}, x_j^{[ref]})$$
$$r_{ij}^{[test]} = \text{cor}(x_i^{[test]}, x_j^{[test]})$$
$$cor.density = \text{mean}(vect.Matrix(\text{sign}(r_{ij}^{[ref](q)} rij^{[test](q)})))$$
$$avg.edge.wei = density^{[test](q)} = \text{mean}(vect.Mat(A^{[test](q)}))$$
$$mod.coh = \text{mean}_{i \in Mq}((kME_i^{[test](q)})^2)$$
$$avg.node.contrib = \text{mean}_{i \in Mq}(\text{sign}(kME_i^{[ref](q)}) kME_i^{[test](q)})$$
$$concor.wei.deg = \text{cor}(kIM)^{[ref](q)}, kIM^{[test](q)}$$
$$concor.nod.contrib = \text{cor}_{i \in Mq}(kME_i^{[ref](q)}, kME_i^{[test](q)})$$
$$concor.cor = \text{cor}(vect.Matrix(r^{[ref](q)}), vect.Matrix(r^{[test](q)}))$$

This score returns :

— **Preserved** if the $Z_{summary}$ is above 10

— **Moderately preserved** if the $Z_{summary}$ is between 2 and 10

— **Unpreserved** if the $Z_{summary}$ is below 2

The results from both NetRep permutation test and the $Z_{summary}$ are then combined in GWENA as shown in Figure A.1 and return a final result on the module comparison.

A.1.2 Details on case study data

Public data were obtained from GTEx v8 version on the GTEx Portal on 09/20/2020. Access to private data was subject to a request to dbGaP on accession number phs000424.v8.p2. Data were obtained on 10/21/2020.

Data	File
Gene expression	GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_reads.gct.gz
Public annotation	GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt
Private annotation	phs000424.v8.pht002742.v8.p2.c1.GTEx_Subject_Phenotypes.GRU.txt
Phenotype	GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt

TABLE A.1 – Correspondence between file names and their contents

A.1.3 GTEx data normalization with PC-correction method

In order to limit batch effect and handle the maximum of other co-founding effects, we chose to use a method based on PC-correction as recommended by Parsana et al. [94] for GTEx data. However age is usually included in this confounding factors, therefore is corrected. Since we're interested in gene changes we adapted the method to remove only the top n PC correlated to age and which removed the least of genes correlating with age. The n number of PC to remove was estimated by calculating the loss of correlation between phenotype and genes expression (Figure A.2) and confirmed by looking for the number of significantly correlated genes with two ageing gene databases (Figure A.3) : GenAge [216] and Digital Aging Atlas [188].

Correlation density in Figure A.2 suggest a similarity between the corrections from 2 to 5 PC removal. Combined with the proportion of overlapping known ageing genes in Figure A.3 we determined the optimal number of PC n to remove to be 4.

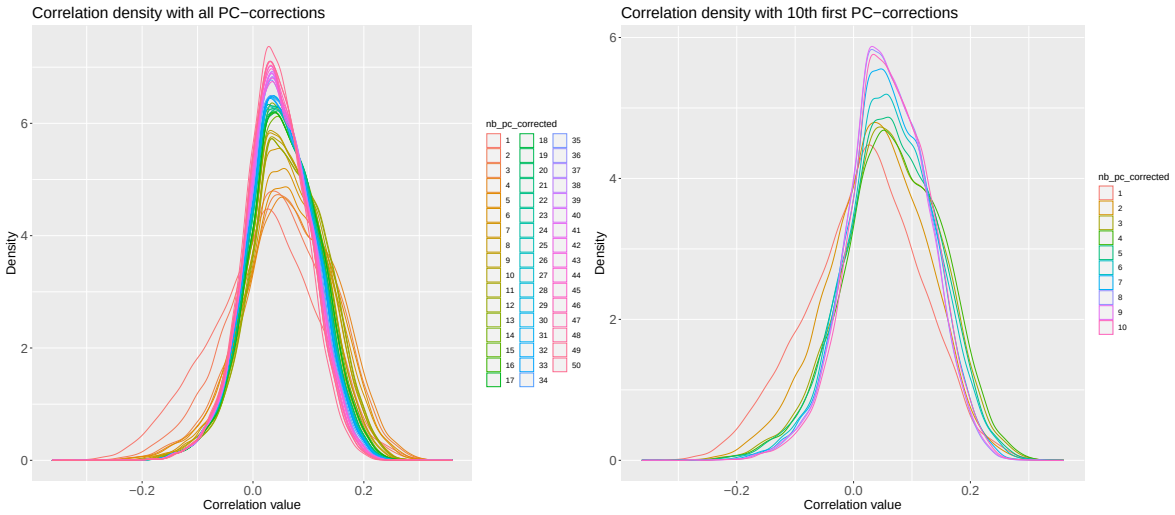

FIGURE A.2 – Ageing genes correlation density with phenotype depending on the number of PC corrected. Left figure contains all PC correction tested. For clarity we filtered on the first 10 PC corrected on the right figure.

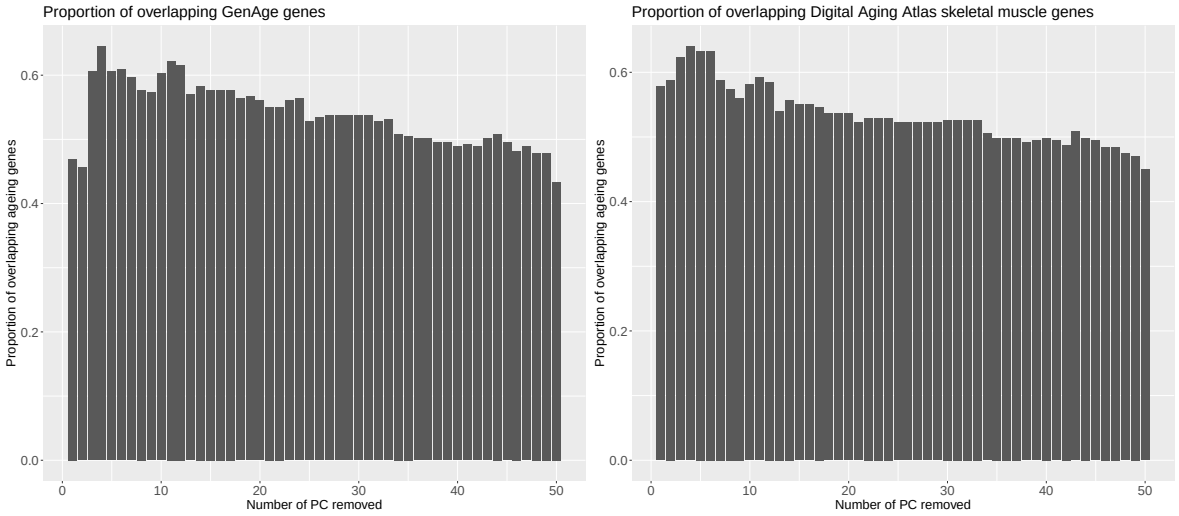

FIGURE A.3 – Number of genes known to be associated with ageing.

A.2 Supplementary Results

A.2.1 Connectivity drop on all modules

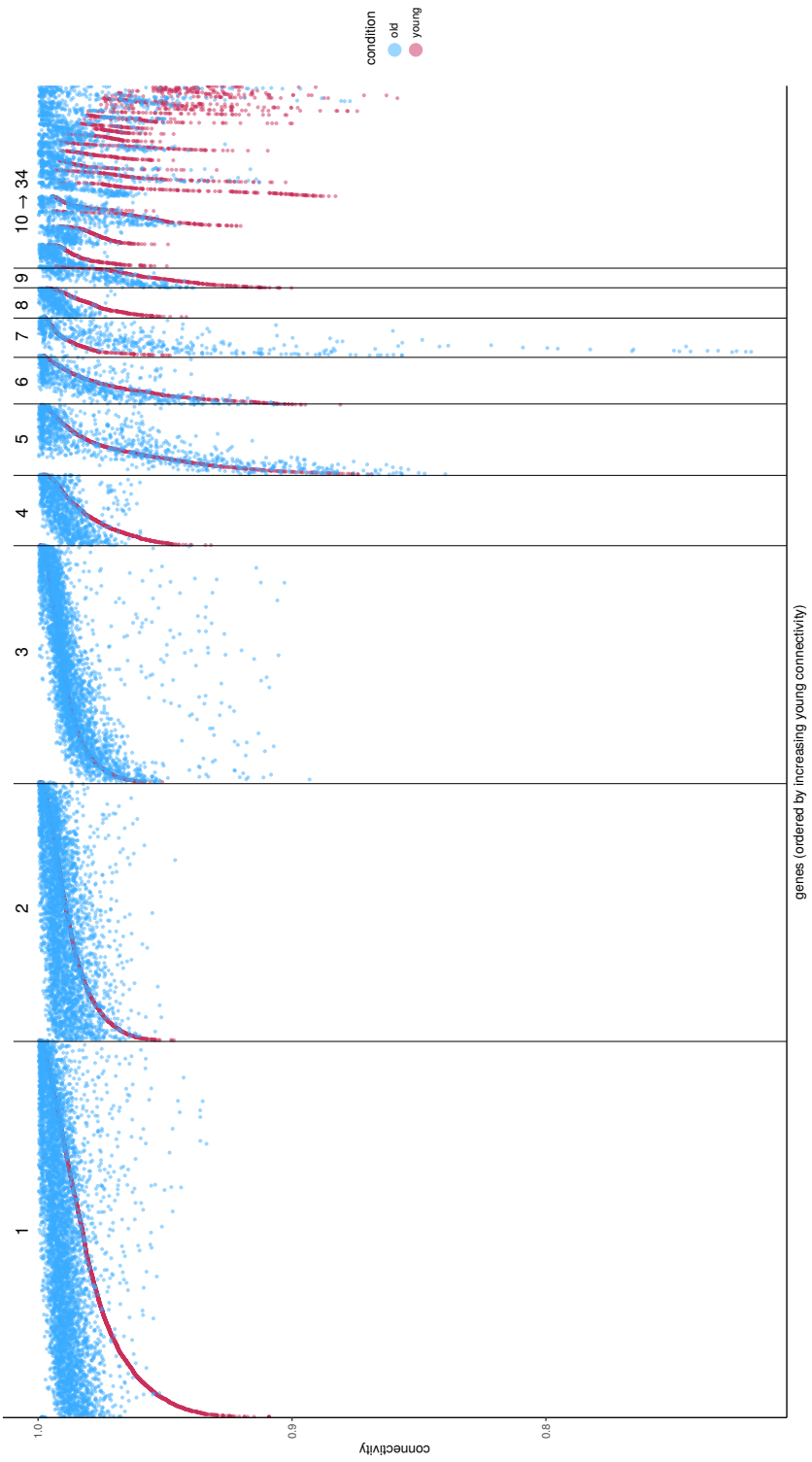

FIGURE A.4 – Distribution of the connectivity for each gene by module between the two age range. Genes connectivity is ordered by increasing connectivity in the young condition (red).

A.2.2 New enrichment terms in sub module 6 from module 7 old age range

TABLE A.2 – Enrichment table from module 7 sub module 6 in old condition. Terms are sorted along their novelty (is the enrichment new compared to the enrichments from sub modules in the young age range) and then the source. Source is the enrichment database used on the gene set (GO :BP = Gene Ontology : Biological Process, GO :CC = Gene Ontology Cellular Compartment, GO :MF = Gene Ontology : Molecular Function, HP : Human Phenotype Ontology, WP = WikiPathway, KEGG = Kyoto Encyclopedia of Genes and Genomes, REAC = Reactome, TF = Transfac).

source	term name	is new	source	term name	is new
CORUM	Fibrinogen complex	no	GO :BP	regulation of vasoconstriction	yes
GO :BP	fibrinolysis	no	GO :BP	heterotypic cell-cell adhesion	yes
GO :BP	negative regulation of blood coagulation	no	GO :BP	platelet aggregation	yes
GO :BP	negative regulation of hemostasis	no	GO :BP	regulation of endothelial cell apoptotic process	yes
GO :BP	negative regulation of coagulation	no	GO :BP	endothelial cell apoptotic process	yes
GO :BP	negative regulation of wound healing	no	GO :BP	protein processing	yes
GO :BP	negative regulation of response to wounding	no	GO :BP	cell-matrix adhesion	yes
GO :BP	regulation of body fluid levels	no	GO :BP	positive regulation of response to wounding	yes
GO :CC	blood microparticle	no	GO :BP	positive regulation of blood circulation	yes
GO :CC	platelet alpha granule lumen	no	GO :BP	vasoconstriction	yes
GO :CC	platelet alpha granule	no	GO :BP	regulation of vesicle-mediated transport	yes
GO :CC	collagen-containing extracellular matrix	no	GO :BP	cell adhesion	yes
GO :CC	fibrinogen complex	no	GO :BP	biological adhesion	yes
GO :CC	extracellular space	no	GO :BP	homotypic cell-cell adhesion	yes
GO :CC	extracellular exosome	no	GO :BP	vascular process in circulatory system	yes
GO :CC	extracellular vesicle	no	GO :BP	extrinsic apoptotic signaling pathway via death domain receptors	yes

GO :CC	extracellular organelle	no	GO :CC	secretory granule lumen	yes
GO :CC	extracellular region	no	GO :CC	cytoplasmic vesicle lumen	yes
GO :CC	secretory granule	no	GO :CC	vesicle lumen	yes
GO :CC	secretory vesicle	no	GO :CC	extracellular matrix	yes
GO :CC	vesicle	no	GO :CC	cell surface	yes
GO :MF	enzyme inhibitor activity	no	GO :CC	endoplasmic reticulum lumen	yes
HP	Splenic rupture	no	GO :CC	cytoplasmic vesicle	yes
WP	COVID-19, thrombosis and anticoagulation	no	GO :CC	intracellular vesicle	yes
GO :BP	platelet degranulation	yes	GO :CC	chylomicron	yes
GO :BP	regulation of blood coagulation	yes	GO :CC	very-low-density lipoprotein particle	yes
GO :BP	regulation of hemostasis	yes	GO :CC	triglyceride-rich plasma lipoprotein particle	yes
GO :BP	regulation of coagulation	yes	GO :CC	external side of plasma membrane	yes
GO :BP	regulation of wound healing	yes	GO :CC	high-density lipoprotein particle	yes
GO :BP	plasminogen activation	yes	GO :CC	plasma lipoprotein particle	yes
GO :BP	regulation of response to wounding	yes	GO :CC	lipoprotein particle	yes
GO :BP	protein activation cascade	yes	GO :CC	protein-lipid complex	yes
GO :BP	blood coagulation, fibrin clot formation	yes	GO :MF	signaling receptor binding	yes
GO :BP	vesicle-mediated transport	yes	GO :MF	chaperone binding	yes
GO :BP	regulated exocytosis	yes	GO :MF	immunoglobulin binding	yes
GO :BP	negative regulation of fibrinolysis	yes	GO :MF	lipoprotein particle receptor binding	yes
GO :BP	exocytosis	yes	GO :MF	extracellular matrix structural constituent	yes
GO :BP	zymogen activation	yes	HP	Menometrorrhagia	yes
GO :BP	blood coagulation	yes	HP	Abnormality of the common coagulation pathway	yes

GO :BP	hemostasis	yes	HP	Spontaneous abortion	yes
GO :BP	coagulation	yes	HP	Abnormality of coagulation	yes
GO :BP	regulation of fibrinolysis	yes	HP	Hypofibrinogenemia	yes
GO :BP	positive regulation of heterotypic cell-cell adhesion	yes	HP	Abnormality of circulating fibrinogen	yes
GO :BP	regulation of cell-substrate adhesion	yes	HP	Joint swelling	yes
GO :BP	negative regulation of response to external stimulus	yes	HP	Abnormality of the coagulation cascade	yes
GO :BP	negative regulation of blood vessel diameter	yes	HP	Abnormal delivery	yes
GO :BP	negative regulation of response to stimulus	yes	HP	Abnormal thrombosis	yes
GO :BP	regulation of heterotypic cell-cell adhesion	yes	KEGG	Complement and coagulation cascades	yes
GO :BP	positive regulation of blood coagulation	yes	KEGG	Platelet activation	yes
GO :BP	positive regulation of hemostasis	yes	KEGG	Cholesterol metabolism	yes
GO :BP	positive regulation of coagulation	yes	MIRNA	hsa-miR-409-3p	yes
GO :BP	wound healing	yes	MIRNA	hsa-miR-144-3p	yes
GO :BP	negative regulation of multicellular organismal process	yes	REAC	Platelet degranulation	yes
GO :BP	positive regulation of cell-substrate adhesion	yes	REAC	Response to elevated platelet cytosolic Ca2+	yes
GO :BP	secretion by cell	yes	REAC	Platelet activation, signaling and aggregation	yes
GO :BP	negative regulation of endothelial cell apoptotic process	yes	REAC	Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)	yes

GO :BP	positive regulation of vasoconstriction	yes	REAC	Post-translational protein phosphorylation	yes
GO :BP	export from cell	yes	REAC	Hemostasis	yes
GO :BP	negative regulation of cellular process	yes	REAC	GRB2 :SOS provides linkage to MAPK signaling for Integrins	yes
GO :BP	regulation of response to stress	yes	REAC	p130Cas linkage to MAPK signaling for integrins	yes
GO :BP	positive regulation of substrate adhesion-dependent cell spreading	yes	REAC	Regulation of TLR by endogenous ligand	yes
GO :BP	regulation of blood vessel diameter	yes	REAC	Common Pathway of Fibrin Clot Formation	yes
GO :BP	regulation of tube diameter	yes	REAC	Integrin signaling	yes
GO :BP	negative regulation of extrinsic apoptotic signaling pathway via death domain receptors	yes	REAC	Signaling by high-kinase activity BRAF mutants	yes
GO :BP	regulation of tube size	yes	REAC	Platelet Aggregation (Plug Formation)	yes
GO :BP	cell-substrate adhesion	yes	REAC	Formation of Fibrin Clot (Clotting Cascade)	yes
GO :BP	post-translational protein modification	yes	REAC	MAP2K and MAPK activation	yes
GO :BP	response to wounding	yes	REAC	Signaling by moderate kinase activity BRAF mutants	yes
GO :BP	secretion	yes	REAC	Signaling downstream of RAS mutants	yes
GO :BP	regulation of response to external stimulus	yes	REAC	Paradoxical activation of RAF signaling by kinase inactive BRAF	yes
GO :BP	platelet activation	yes	REAC	Signaling by RAS mutants	yes
GO :BP	transport	yes	REAC	Signaling by BRAF and RAF fusions	yes

GO :BP	response to stress	yes	REAC	Oncogenic MAPK signaling	yes
GO :BP	establishment of localization	yes	REAC	Dissolution of Fibrin Clot	yes
GO :BP	negative regulation of epithelial cell apoptotic process	yes	REAC	Integrin cell surface interactions	yes
GO :BP	regulation of cell adhesion	yes	TF	Factor : HNF1A ; motif : GGTTAATNATTAMC	yes
GO :BP	regulation of substrate adhesion-dependent cell spreading	yes	TF	Factor : HNF-1alpha ; motif : GGTTAATNWTTAMCN	yes
GO :BP	induction of bacterial agglutination	yes	TF	Factor : Sox-2 ; motif : NNNNNAACAAWGN ; match class : 1	yes
GO :BP	regulation of response to stimulus	yes	WP	Folate Metabolism	yes
GO :BP	proteolysis	yes	WP	Selenium Micronutrient Network	yes
GO :BP	negative regulation of biological process	yes	WP	Human Complement System	yes
GO :BP	regulation of cell-cell adhesion	yes	WP	Blood Clotting Cascade	yes
GO :BP	regulation of extrinsic apoptotic signaling pathway via death domain receptors	yes	WP	Fibrin Complement Receptor 3 Signaling Pathway	yes
GO :BP	positive regulation of wound healing	yes	WP	Vitamin B12 Metabolism	yes

The distribution of the newly and previously found terms in the enrichment analysis across the sub-modules from young and old age range (Figure A.5).

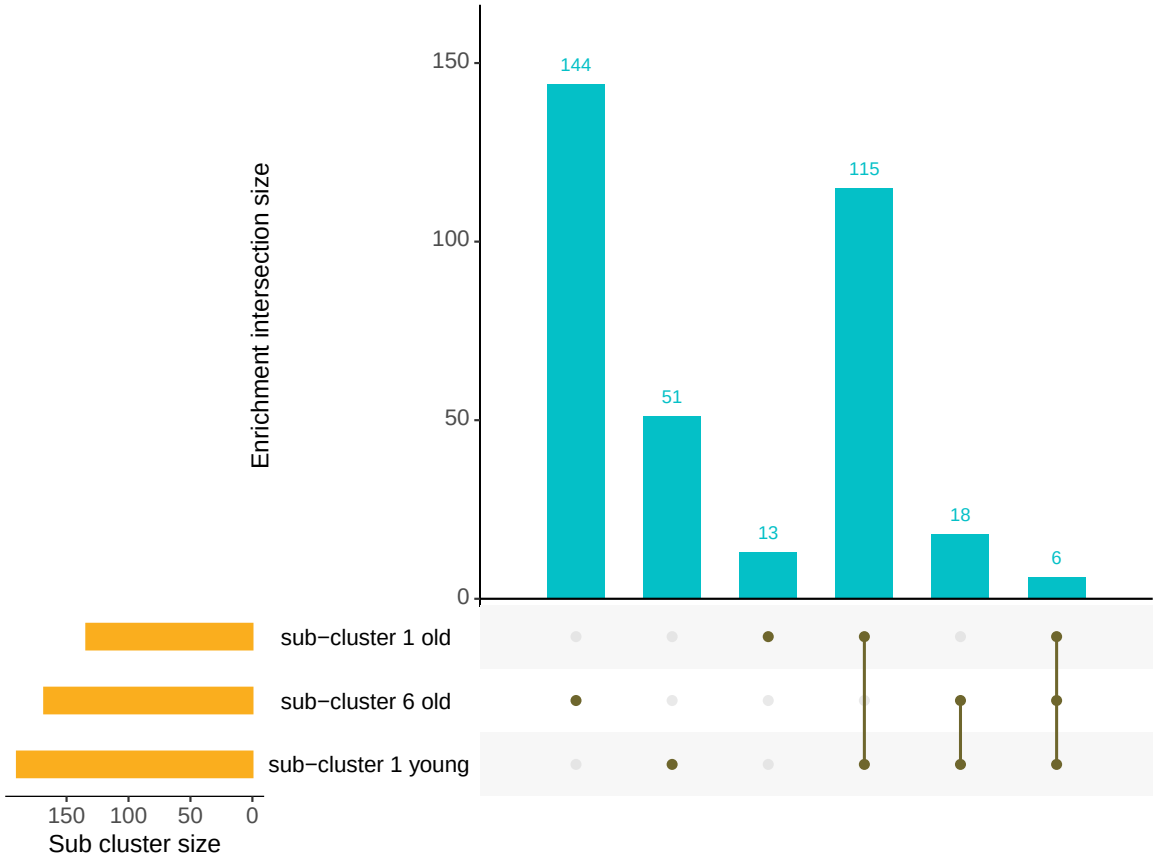

FIGURE A.5 – Overlap between the enrichments found in sub-cluster 1 young, sub-cluster 1 old, and sub-cluster 6 old.(Upset diagram)

Demande d'accès dbGaP aux données protégées de GTEx

B.1 Query title

Muscle gene co-expression in multiple age range for sarcopenia condition exploration

B.2 Research use statement

As a multi factorial phenomena, aging is a complex condition to study. Current analysis of transcriptomic data joined with phenotypic ones have unravel few genes and environmental variables impacting it. However, aging remains not fully understood. These may come from current approaches focusing on single factors while aging is more about interaction between many of them. This is why we would like to study it through the spectrum of the co-expression networks. However, because aging is already complex by itself, we will focus on a single tissue in the beginning : muscle. Reasons are myopenia and dynapenia (known together as sarcopenia) are responsible for loss of autonomy, weak metabolic aggression resistance, and increased mortality. By using our yet to be published pipeline developed in our lab, we aim to build gene co-expression modules and network from transcriptomic data and characterize them with external resources. This pipeline begin with a quality assessment of the data, based on the technique used to get transcriptomic information (either microarray or RNA-Seq). It then compute co-expression levels and detects modules by using the WGCNA package. Additionnal steps are then performed to fully charachterize the modules : biological enrichment, topological study, phenotipic association, differentially expressed and condition-specific gene positionning.The external resources used to it include : enrichment databases (GO, KEGG, etc.), phenotypic information (exact age, condition of death, ethnicity), and age related databases (Digital aging atlas, Aging map, etc.). The use of these complementary information represents no additional risk to participants since only summa-

rized gene information we'll be used in the form of modules. No other raw transcriptomics data will be combined to the GTEx data, but only a comparison of the modules built on each datasets in order to study the reproductibility of our modules, and the comparison with modules built on sarcopenia dedicated datasets. Phenotipyc information is the main reason for our current demand regarding protected datasets because we think precise age and cause of death will impact the aging expression. No collaboration with other institution is planned. Finnally, this methodological work will advance the understanding of the genetic bases of the aging processes occurring in the muscle.

B.3 Non-technical summary

Aging affect every one of us. It is defined as the progressive degradation of biological functions inside the body. In the muscle, this results in a decreasing in muscle density and strength called sarcopenia. They are responsible for mobility difficulties, therefore autonomy, and a deficit in body protection against aggression, sometimes leading to death. With this risks at stake, it is understandable that a better understanding of aging processes in muscle is important for public health. Few single genes linked to it have already been discovered, however we still don't fully understand the mechanisms occurring and how to limit them. Gene expression is a witness of the changes in the body mechanisms. By studying the expression of the genes and the coordination between this levels of expression, therefore patterns of expression, we aim to detect news genes involved in muscle aging. Do to so, we regroup the most coordinated genes in entities called module and try to unravel the biological interactions which link them. Because some of this genes are already involved into other biological interactions, we can have an insight on the regulations that take place in this case by extrapolating information. These, will lead to the discovery of new genes associated with sarcopenia.

Réalisé le 30/09/2019

B.4 Access update : Research Progress

Current analysis of the GTEx data focused on the skeletal muscle samples as we are studying sarcopenia and other aging impact on muscle. The RNAseq data was then split in two opposite age range (young and old) to emphase and capture the differences in aging through our newly developped differential gene co-expression network pipeline GWENA (https://www.bioconductor.org/packages/release/bioc/html/GWENA.html). A first analysis solely on the modules (gene groups)

detected by GWENA in the young condition. A selection of interesting modules for muscle activity was achieved through a combinaison of moduels phenotypic association and gene sets enrichment. Inside one promissing module, hub genes connected in the network to genes involved in muscle activity allowed the identification of genes with strong evidence of contribution to muscle developpemnt and growth. A second analysis used the full potential of the differential co-expression analysis offered by GWENA to find specific co-expression modules to aging. Important topological variation lead us to focus on one module. Further investigation is currently underway to characterise the origin of the specificity and this variations.

Réalisé le 01/12/2020

Liste des enrichissements des intersections de gènes en chapitre 2

Résultats des tests d'enrichissement effectués grâce à GWENA sur les intersections ayant au moins 3 tissus et 5 gènes lors du chapitre 3. Pour rappel :

- **GO** : Gene Ontology
- **GO :MF** : GO Molecular Function
- **GO :BP** : GO Biological Process
- **GO :CC** : GO Cellular Compartment
- **KEGG** : Kyoto Encyclopedia of Genes and Genomes
- **REAC** : Reactome
- **WP** : WikiPathway
- **TF** : TRANSFAC
- **MIRNA** : mirTarBase
- **HPA** : Human Protein Atlas
- **CORUM** : Comprehensive Resource of Mammalian protein complexes
- **HP** : Human Phenotype Ontology

esophagus_muscularis_mod_pres
skin_sun_exposed_mod_pres
thyroid_mod_pres

Source	Nom du terme enrichit

CORUM	Cell-cell junction complex (ARHGAP10-CTNNA1)
GO :BP	dicarboxylic acid catabolic process
HP	Hypoglutaminemia
HP	Abnormal CSF glutamine family amino acid concentration
HP	Abnormal CSF glutamine concentration
HP	Decreased CSF glutamine concentration

esophagus_muscularis_mod_pres
nerve_tibial_mod_pres
skin_sun_exposed_mod_pres

Source	Nom du terme enrichit
CORUM	Seipin-lipin1 complex
GO :BP	adenosine to inosine editing
GO :BP	base conversion or substitution editing
GO :MF	phenylethanolamine N-methyltransferase activity
GO :MF	creatine :sodium symporter activity
GO :MF	XTP binding
GO :MF	ITP binding

esophagus_muscularis_mod_pres
nerve_tibial_mod_pres
skin_sun_exposed_mod_pres
thyroid_mod_pres

Source	Nom du terme enrichit
GO :MF	dGTPase activity
GO :MF	triphosphoric monoester hydrolase activity
GO :MF	guanyl deoxyribonucleotide binding
GO :MF	dGTP binding
HPA	skin 1 ; melanocytes[High]
HPA	testis ; peritubular cells[High]

esophagus_mucosa_mod_pres
skin_not_sun_exposed_mod_pres
skin_sun_exposed_mod_pres

Source	Nom du terme enrichit
GO :BP	melanin biosynthetic process
GO :BP	melanin metabolic process
GO :BP	secondary metabolite biosynthetic process
GO :BP	phenol-containing compound biosynthetic process
GO :BP	secondary metabolic process
GO :BP	pigment biosynthetic process
GO :BP	pigment metabolic process
GO :BP	pigmentation
GO :BP	phenol-containing compound metabolic process
GO :BP	regulation of secondary metabolic process
GO :BP	regulation of melanin biosynthetic process
GO :BP	regulation of secondary metabolite biosynthetic process
GO :MF	melanin-concentrating hormone receptor activity
HPA	retina ; pigment epithelial cells[High]

esophagus_mucosa_mod_pres
muscle_skeletal_mod_pres
thyroid_mod_pres

Source	Nom du terme enrichit
GO :BP	oxygen transport
GO :BP	gas transport
GO :BP	hydrogen peroxide catabolic process
GO :CC	hemoglobin complex
GO :CC	haptoglobin-hemoglobin complex
GO :MF	hemoglobin binding
GO :MF	haptoglobin binding
GO :MF	oxygen carrier activity
GO :MF	oxygen binding
GO :MF	peroxidase activity

GO :MF	oxidoreductase activity, acting on peroxide as acceptor
GO :MF	molecular carrier activity

esophagus_mucosa_mod_pres

muscle_skeletal_mod_pres

skin_sun_exposed_mod_pres

Source	Nom du terme enrichit
GO :BP	response to nicotine
GO :MF	V1B vasopressin receptor binding
REAC	ADORA2B mediated anti-inflammatory cytokines production

esophagus_mucosa_mod_pres

muscle_skeletal_unpres

skin_sun_exposed_mod_pres

Source	Nom du terme enrichit
HP	Gonadal tissue inappropriate for external genitalia or chromosomal sex

artery_tibial_mod_pres

skin_not_sun_exposed_mod_pres

skin_sun_exposed_mod_pres

Source	Nom du terme enrichit
GO :MF	neuropeptide receptor activity

artery_tibial_mod_pres

esophagus_muscularis_mod_pres

thyroid_mod_pres

Source	Nom du terme enrichit
GO :BP	response to corticotropin-releasing hormone
GO :BP	cellular response to corticotropin-releasing hormone stimulus
GO :BP	regulation of type B pancreatic cell proliferation
GO :BP	type B pancreatic cell proliferation
GO :CC	transcription regulator complex
GO :CC	chromatin
GO :MF	DNA-binding transcription activator activity, RNA polymerase II-specific
GO :MF	DNA-binding transcription activator activity
GO :MF	glucocorticoid receptor binding
GO :MF	nuclear receptor activity
GO :MF	ligand-activated transcription factor activity
GO :MF	DNA-binding transcription factor activity, RNA polymerase II-specific
GO :MF	DNA-binding transcription factor activity
GO :MF	steroid hormone receptor binding
TF	Factor : ATF ; motif : CNSTGACGTNNNYC ; match class : 1

artery_tibial_mod_pres
esophagus_mucosa_mod_pres
thyroid_mod_pres

No enrichment found

artery_tibial_mod_pres
esophagus_mucosa_mod_pres
skin_sun_exposed_mod_pres

Source	Nom du terme enrichit
GO :BP	positive regulation of gastrulation
GO :BP	primitive streak formation
GO :BP	regulation of gastrulation
KEGG	Viral protein interaction with cytokine and cytokine receptor

artery_tibial_mod_pres
esophagus_mucosa_mod_pres
skin_sun_exposed_unpres

No enrichment found

artery_tibial_mod_pres
esophagus_mucosa_mod_pres
nerve_tibial_mod_pres

Source	Nom du terme enrichit
GO :CC	blood microparticle
GO :CC	platelet alpha granule lumen
GO :CC	platelet alpha granule
GO :CC	collagen-containing extracellular matrix
GO :CC	extracellular matrix
GO :CC	external encapsulating structure
KEGG	Complement and coagulation cascades
REAC	Platelet degranulation
REAC	Response to elevated platelet cytosolic Ca2+
TF	Factor : C/EBP ; motif : NTTRCNNAANNN
WP	Complement and Coagulation Cascades

artery_tibial_mod_pres
esophagus_mucosa_mod_pres
nerve_tibial_mod_pres
thyroid_mod_pres

Source	Nom du terme enrichit
CORUM	Fibrinogen complex
GO :BP	negative regulation of wound healing
GO :BP	plasminogen activation
GO :BP	negative regulation of response to wounding

GO :BP fibrinolysis
GO :BP blood coagulation, fibrin clot formation
GO :BP protein activation cascade
GO :BP platelet degranulation
GO :BP regulation of wound healing
GO :BP negative regulation of blood coagulation
GO :BP regulation of response to wounding
GO :BP negative regulation of hemostasis
GO :BP negative regulation of coagulation
GO :BP zymogen activation
GO :BP regulation of blood coagulation
GO :BP regulation of hemostasis
GO :BP regulation of coagulation
GO :BP positive regulation of heterotypic cell-cell adhesion
GO :BP regulation of heterotypic cell-cell adhesion
GO :BP negative regulation of response to external stimulus
GO :BP positive regulation of vasoconstriction
GO :BP negative regulation of endothelial cell apoptotic process
GO :BP negative regulation of extrinsic apoptotic signaling pathway via death domain receptors
GO :BP positive regulation of substrate adhesion-dependent cell spreading
GO :BP wound healing
GO :BP negative regulation of epithelial cell apoptotic process
GO :BP negative regulation of multicellular organismal process
GO :BP induction of bacterial agglutination
GO :BP regulation of substrate adhesion-dependent cell spreading
GO :BP regulation of extrinsic apoptotic signaling pathway via death domain receptors
GO :BP protein processing
GO :BP heterotypic cell-cell adhesion
GO :BP regulation of endothelial cell apoptotic process
GO :BP platelet aggregation
GO :BP regulation of vasoconstriction
GO :BP endothelial cell apoptotic process
GO :BP response to wounding
GO :BP positive regulation of cell morphogenesis involved in differentiation
GO :BP extrinsic apoptotic signaling pathway via death domain receptors
GO :BP vasoconstriction
GO :BP protein maturation

GO :BP homotypic cell-cell adhesion
GO :BP positive regulation of exocytosis
GO :BP regulated exocytosis
GO :BP regulation of cell morphogenesis involved in differentiation
GO :BP blood coagulation
GO :BP hemostasis
GO :BP coagulation
GO :BP positive regulation of peptide hormone secretion
GO :BP regulation of epithelial cell apoptotic process
GO :BP negative regulation of extrinsic apoptotic signaling pathway
GO :BP substrate adhesion-dependent cell spreading
GO :BP exocytosis
GO :BP positive regulation of cell-substrate adhesion
GO :BP epithelial cell apoptotic process
GO :BP positive regulation of hormone secretion
GO :BP blood vessel diameter maintenance
GO :BP regulation of tube diameter
GO :BP regulation of tube size
GO :BP positive regulation of protein secretion
GO :BP response to calcium ion
GO :BP regulation of response to external stimulus
GO :BP positive regulation of peptide secretion
GO :BP regulation of extrinsic apoptotic signaling pathway
GO :BP platelet activation
GO :BP toll-like receptor signaling pathway
GO :BP regulation of body fluid levels
GO :CC blood microparticle
GO :CC fibrinogen complex
GO :CC collagen-containing extracellular matrix
GO :CC platelet alpha granule lumen
GO :CC extracellular matrix
GO :CC external encapsulating structure
GO :CC platelet alpha granule
GO :CC secretory granule lumen
GO :CC cytoplasmic vesicle lumen
GO :CC vesicle lumen
GO :CC extracellular exosome
GO :CC extracellular vesicle
GO :CC extracellular organelle

GO :CC	secretory granule
GO :CC	secretory vesicle
GO :CC	extracellular space
GO :CC	vesicle
GO :CC	extracellular region
GO :CC	cell surface
GO :MF	extracellular matrix structural constituent
HP	Splenic rupture
HP	Menometrorrhagia
HP	Spontaneous abortion
HP	Abnormality of circulating fibrinogen
HP	Hypofibrinogenemia
HP	Joint swelling
HP	Abnormality of the common coagulation pathway
HP	Gingival bleeding
HP	Cerebral hemorrhage
HP	Abnormality of the cerebral vasculature
HP	Abnormal delivery
HP	Venous thrombosis
HP	Epistaxis
HP	Abnormality of the coagulation cascade
KEGG	Complement and coagulation cascades
KEGG	Platelet activation
KEGG	Neutrophil extracellular trap formation
KEGG	Coronavirus disease - COVID-19
MIRNA	hsa-miR-409-3p
MIRNA	hsa-miR-144-3p
MIRNA	hsa-miR-29c-3p
MIRNA	hsa-miR-29b-3p
MIRNA	hsa-miR-29a-3p
REAC	Platelet degranulation
REAC	Response to elevated platelet cytosolic Ca2+
REAC	GRB2 :SOS provides linkage to MAPK signaling for Integrins
REAC	p130Cas linkage to MAPK signaling for integrins
REAC	Regulation of TLR by endogenous ligand
REAC	Common Pathway of Fibrin Clot Formation
REAC	Platelet activation, signaling and aggregation
REAC	Integrin signaling
REAC	Signaling by high-kinase activity BRAF mutants

REAC	Platelet Aggregation (Plug Formation)
REAC	Formation of Fibrin Clot (Clotting Cascade)
REAC	MAP2K and MAPK activation
REAC	Signaling by RAF1 mutants
REAC	Signaling by RAS mutants
REAC	Paradoxical activation of RAF signaling by kinase inactive BRAF
REAC	Signaling by moderate kinase activity BRAF mutants
REAC	Signaling downstream of RAS mutants
REAC	Signaling by BRAF and RAF fusions
REAC	Oncogenic MAPK signaling
REAC	Integrin cell surface interactions
REAC	Hemostasis
REAC	Toll-like Receptor Cascades
TF	Factor : HNF1B ; motif : NRTTAATNATTAACN
TF	Factor : HNF1A ; motif : GGTTAATNATTAMC
WP	COVID-19, thrombosis and anticoagulation
WP	Blood Clotting Cascade
WP	Human Complement System
WP	Fibrin Complement Receptor 3 Signaling Pathway
WP	Folate Metabolism
WP	Selenium Micronutrient Network

artery_tibial_mod_pres
esophagus_mucosa_mod_pres
muscle_skeletal_mod_pres

No enrichment found

artery_tibial_mod_pres
esophagus_mucosa_mod_pres
muscle_skeletal_mod_pres
thyroid_mod_pres

Source	Nom du terme enrichit
GO :MF	IgA receptor activity

HP	Hypochromic microcytic anemia
HP	Abnormality of iron homeostasis
HP	Iron deficiency anemia
HP	Abnormal blood transition element cation concentration
HP	Microcytic anemia
HP	Hypochromic anemia

artery_tibial_mod_pres

esophagus_mucosa_mod_pres

muscle_skeletal_unpres

Source	Nom du terme enrichit
CORUM	TCL1 homotrimer complex
GO :BP	cell wall disruption in other organism
GO :CC	high-density lipoprotein particle
GO :CC	plasma lipoprotein particle
GO :CC	lipoprotein particle
GO :CC	protein-lipid complex
GO :CC	zymogen granule
GO :MF	oligosaccharide binding
GO :MF	peptidoglycan binding
GO :MF	glycosaminoglycan binding
GO :MF	carbohydrate binding
HP	Malabsorption
HP	Cholestasis
HP	Pancreatic pseudocyst
KEGG	Vitamin digestion and absorption
REAC	Scavenging by Class B Receptors
TF	Factor : GATA-1 ; motif : NTGNNNNNNNSAGATAAGR
TF	Factor : GATA-1 ; motif : NTGNNNNNNNNAGATAAGN
TF	Factor : GATA-X ; motif : NGATAAGNMNN
TF	Factor : GATA-2 ; motif : NTGNNNNNNNNAGATAAGN
TF	Factor : SRY ; motif : AACAATNR ; match class : 1
TF	Factor : GATA-1 ; motif : MNAGATAANR
TF	Factor : JunD ; motif : NATGASTCATS
TF	Factor : GATA-3 ; motif : AGATAA
WP	Vitamin B12 Metabolism

WP	Folate Metabolism
WP	Selenium Micronutrient Network

artery_tibial_mod_pres

esophagus_mucosa_mod_pres

muscle_skeletal_unpres

whole_blood_mod_pres

Source	Nom du terme enrichit
GO :BP	digestion
GO :BP	proteolysis
GO :BP	lipid catabolic process
GO :BP	lipid digestion
GO :BP	cobalamin metabolic process
GO :BP	extracellular matrix disassembly
GO :CC	extracellular region
GO :CC	extracellular space
GO :MF	peptidase activity
GO :MF	serine-type endopeptidase activity
GO :MF	hydrolase activity
GO :MF	serine-type peptidase activity
GO :MF	serine hydrolase activity
GO :MF	endopeptidase activity
GO :MF	triglyceride lipase activity
GO :MF	metallocarboxypeptidase activity
GO :MF	lipase activity
GO :MF	carboxypeptidase activity
GO :MF	carboxylic ester hydrolase activity
GO :MF	catalytic activity, acting on a protein
GO :MF	catalytic activity
GO :MF	metalloexopeptidase activity
GO :MF	exopeptidase activity
HP	Pancreatic calcification
HP	Recurrent pancreatitis
HP	Splanchnic vein thrombosis
HP	Abnormal pancreas morphology
HP	Steatorrhea

HP	Abnormality of pancreas physiology
HP	Fat malabsorption
HP	Pancreatic pseudocyst
HP	Abnormality of exocrine pancreas physiology
HP	Exocrine pancreatic insufficiency
HP	Elevated C-reactive protein level
HP	Abnormal C-reactive protein level
HP	Acute phase response
HP	Pancreatitis
HP	Abnormality of the pancreas
HP	Venous thrombosis
HPA	pancreas ; exocrine glandular cells[High]
KEGG	Pancreatic secretion
KEGG	Protein digestion and absorption
KEGG	Fat digestion and absorption
KEGG	Glycerolipid metabolism
REAC	Digestion
REAC	Digestion of dietary lipid
REAC	Digestion and absorption
REAC	Activation of Matrix Metalloproteinases
REAC	Cobalamin (Cbl, vitamin B12) transport and metabolism
REAC	Metabolism of vitamins and cofactors
REAC	Degradation of the extracellular matrix
TF	Factor : GATA-X ; motif : NGATAAGNMNN ; match class : 1
TF	Factor : FIGLA ; motif : NMCACCTGKN ; match class : 1
TF	Factor : FIGLA ; motif : NMCACCTGN ; match class : 1
TF	Factor : FIGLA ; motif : NMCACCTGK ; match class : 1
TF	Factor : GATA-3 ; motif : WGATAASN
TF	Factor : GATA3 ; motif : NGATAANN
TF	Factor : GATA-X ; motif : NGATAAGNMNN
WP	Glucose Homeostasis

artery_tibial_mod_pres

esophagus_mucosa_mod_pres

muscle_skeletal_unpres

skin_not_sun_exposed_mod_pres

Source	Nom du terme enrichit
CORUM	ATP4A-ATP4B complex
GO :BP	establishment or maintenance of transmembrane electrochemical gradient
GO :BP	cellular potassium ion homeostasis
GO :BP	digestion
GO :BP	sodium ion export across plasma membrane
GO :BP	cellular sodium ion homeostasis
GO :BP	potassium ion homeostasis
GO :CC	multivesicular body lumen
GO :CC	late endosome lumen
GO :CC	endosome lumen
GO :CC	multivesicular body
GO :MF	aspartic-type endopeptidase activity
GO :MF	aspartic-type peptidase activity
GO :MF	potassium :proton exchanging ATPase activity
GO :MF	potassium transmembrane transporter activity, phosphorylative mechanism
GO :MF	P-type sodium :potassium-exchanging ATPase activity
GO :MF	proton-exporting ATPase activity, phosphorylative mechanism
GO :MF	P-type transmembrane transporter activity
GO :MF	hydrolase activity
GO :MF	ion transmembrane transporter activity, phosphorylative mechanism
GO :MF	ATPase-coupled cation transmembrane transporter activity
GO :MF	ATPase-coupled ion transmembrane transporter activity
GO :MF	ATPase-coupled transmembrane transporter activity
GO :MF	primary active transmembrane transporter activity
GO :MF	endopeptidase activity
HPA	stomach 2 ; glandular cells[High]
HPA	stomach 1 ; glandular cells[High]
KEGG	Collecting duct acid secretion
KEGG	Gastric acid secretion
REAC	Surfactant metabolism
REAC	Ion transport by P-type ATPases
TF	Factor : ZNF626 ; motif : CCTGCTGAWGSA
TF	Factor : MafA ; motif : NAWWNTGCTGACN
WP	Secretion of Hydrochloric Acid in Parietal Cells

artery_tibial_unpres
esophagus_mucosa_mod_pres
muscle_skeletal_mod_pres

No enrichment found

artery_tibial_unpres
esophagus_mucosa_mod_pres
muscle_skeletal_unpres

No enrichment found

adipose_subcutaneous_mod_pres
skin_not_sun_exposed_mod_pres
thyroid_mod_pres

Source	Nom du terme enrichit
GO :BP	negative regulation of viral genome replication
GO :BP	regulation of viral genome replication
GO :BP	negative regulation of viral process
GO :BP	type I interferon signaling pathway
GO :BP	cellular response to type I interferon
GO :BP	response to type I interferon
GO :BP	response to virus
GO :BP	viral genome replication
GO :BP	regulation of viral life cycle
GO :BP	regulation of viral process
GO :BP	regulation of biological process involved in symbiotic interaction
GO :BP	defense response to symbiont
GO :BP	defense response to virus
GO :BP	response to other organism
GO :BP	response to external biotic stimulus
GO :BP	response to biotic stimulus
GO :BP	regulation of ribonuclease activity

GO :BP	viral life cycle
GO :BP	biological process involved in interspecies interaction between organisms
GO :BP	regulation of nuclease activity
GO :MF	2'-5'-oligoadenylate synthetase activity
GO :MF	adenylyltransferase activity
KEGG	Hepatitis C
KEGG	Influenza A
MIRNA	hsa-miR-146a-5p
REAC	Interferon alpha/beta signaling
REAC	Interferon Signaling
REAC	Antiviral mechanism by IFN-stimulated genes
REAC	OAS antiviral response
REAC	Cytokine Signaling in Immune system
TF	Factor : IRF-9 ; motif : NCGAAACYGAAACYN
TF	Factor : IRF-5 ; motif : NCGAAACCGAAACY
TF	Factor : IRF-9 ; motif : NYGAAACYGAAACYN
TF	Factor : IRF5 ; motif : CCGAAACCGAAACY
TF	Factor : STAT2 ; motif : RRGRAANNGAAACTGAAAN
TF	Factor : ISGF-3 ; motif : CAGTTTCWCTTTYCC
TF	Factor : IRF-5 ; motif : NCGAAACCGAAACY
TF	Factor : IRF-5 ; motif : NYGAAACCGAAACY
TF	Factor : IRF-4 ; motif : NCGAAACCGAAACYA ; match class : 1
TF	Factor : IRF-3 ; motif : NGGAAACNGAAACCGAAACN
TF	Factor : IRF8 ; motif : NCGAAACCGAAACT
TF	Factor : ICSBP ; motif : RAARTGAAACTG ; match class : 1
TF	Factor : IRF-4 ; motif : NCGAAACCGAAACYA
TF	Factor : ICSBP ; motif : RAARTGAAACTG
TF	Factor : IRF-5 ; motif : NCGAAACCGAAACY
TF	Factor : IRF ; motif : NNGAAANTGAAANN
TF	Factor : FOXP1 ; motif : TNTGTTTMY ; match class : 1
TF	Factor : IRF3 ; motif : NNRRAANGGAAACCGAAACYR
TF	Factor : IRF-8 ; motif : NCGAAACYGAAACYN
TF	Factor : STAT2 ; motif : RRGRAANNGAAACTGAAAN ; match class : 1
TF	Factor : IRF-2 ; motif : NGAAASYGAAAS
TF	Factor : IRF-4 ; motif : KRAAMNGAAANYN
TF	Factor : IRF-7 ; motif : TNSGAAWNCGAAANTNNN
TF	Factor : IRF1 ; motif : NNNYASTTTCACTTTCNNTTT
TF	Factor : IRF ; motif : RRAANTGAAASYGNV
TF	Factor : IRF-1 ; motif : STTTCACTTTCNNT

adipose_subcutaneous_mod_pres

muscle_skeletal_mod_pres

skin_sun_exposed_mod_pres

Source	Nom du terme enrichit
GO :MF	histone demethylase activity
GO :MF	protein demethylase activity
GO :MF	demethylase activity
HP	Y-linked inheritance
HP	Gonosomal inheritance
HP	Azoospermia
REAC	HDMs demethylate histones
TF	Factor : A-Myb :HOXA13 ; motif : CTCGTAAAWNNNNRMCGTTR

Projets annexes scientifique hors académique

D.1 Bioinfo-fr.net

Le doctorat est l'occasion d'acquérir et par la suite de partager de nombreuses connaissances. Co-administratrice et auteur sur le blog bioinfo-fr.net, ce blog vise à partager des tutoriels sur des analyses, à répondre à des questions communément posées aux bio-informaticiens, et à faire profiter la communauté francophone d'astuces amenées par tous. Ce blog communautaire est également l'occasion de parler de façon plus légère et décontractée de sujet important tout en restant rigoureux dans les explications. J'ai donc moi-même pu continuer de contribuer à cet effort durant mon doctorat avec les articles suivant qui ont été notamment inspirés par mes travaux :

- Packrat ou comment gérer ses packages R par projet, mars 2017
- Enquête Bioinfo-fr 2018 : un portrait de la bioinformatique, juin 2018
- Les grandes questions du doctorant en devenir/débutant, novembre 2018
- La bio-informatique au service de l'antibiorésistance, février 2019
- Analyses bioinformatiques du coronavirus 2019-nCoV : pourquoi et comment ?, février 2020, co-auteure
- Contrarié par les diagrammes de Venn ? Découvrez les diagrammes UpSet, février 2020
- Pourquoi et comment déposer un package R sur Bioconductor ?, juin 2020.

D.2 Illustration scientifique

Comme visible dans les figures parcourant cette thèse et mes articles de blog, cette thèse a également été l'occasion de développer ma compétente d'illustration scientifique. À titre d'exemple, voici quelques autres travaux originaux réalisés dans divers contextes scientifiques :

CANDIDATE GENES DISCOVERY THROUGH AUGMENTED GENE CO-EXPRESSION ANALYSIS - A CASE STUDY IN SKIN AGING

Gwenaëlle Lemoine*, Marie-Pier Scott-Boyer, Arnaud Droit

* gwenaelle.lemoine@protonmail.com

Genomics Center, Centre Hospitalier Universitaire de Québec – Université Laval Research Center, Québec, QC, Canada

AD Lab
compbio.ulaval.ca

BACKGROUND

The rapid increase in size of transcriptomic datasets requires the development of new methods able to manage them[1]. However, methods such as differential expression analysis only focuses on few genes.

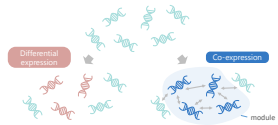

Figure 1 : Comparison of differential expression and gene co-expression. Most strongly interacting genes are grouped into clusters (modules) and assumed to be involved in a same biological function (guilt-by-association paradigm[2]).

Network-based methods such as **gene co-expression** unveil a new dimension in the study of complex molecular processes[3] (Figure 1) and multiple studies have been successfully using them[4–6].

The R package WGCNA[7] is a widely used implementation of this method and has proven its effectiveness. However, it require the users to have certain experience of co-expression subtleties to work with it and don't implement comparison of condition along with help to interpretation.

Here we present **Gene Whole co-Expression Network Analysis (GWENA)**, an R package improving information which can be retrieved from co-expression analysis applied to transcriptomic data. We applied GWENA to **skin aging** data and retrieved potentially useful information about processes involved.

IMPLEMENTATION

WEN — **G**ene **W**hole co-**E**xpression **N**etwork **A**nalysis

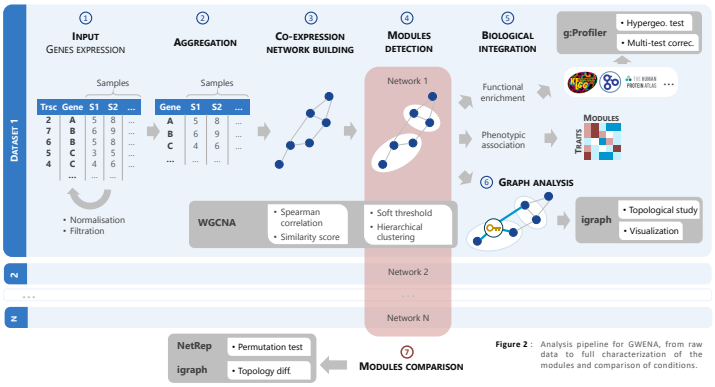

Figure 2 : Analysis pipeline for GWENA, from raw data to full characterization of the modules and comparison of conditions.

Network building

1. Filter & normalization: adapted to either microarray or RNA-seq (based on limma[8] and DESeq2[9]).
2. Aggregation from transcript to gene (based on collapseRow method).
3. • New correlation implemented among WGCNA : spearman correlation.
   • Similarity score computation based on correlation.
4. Module detection : soft-threshold based on power law because of scale free topology.

Modules characterization

5. Knowledge-driven:
   • Functional enrichment with g:Profiler[10] (GO, KEGG, REACTOME, WikiPathways, miRTarBase, Human Protein Atlas, CORUM, TRANSFAC, Human Phenotype Ontology).
   • Phenotypic association by eigengenes.
6. Data-driven:
   • Detection of patterns of interest, such as hub genes.
   • Network visualization and layer system to display additional information (like enrichments) with igraph[11].

Modules comparison

7. • Detection of (un)preserved modules through a permutation test performed through NetRep[12] with one condition as reference.
   • Analysis of changes in modules conformation and biological functions.

CASE STUDY IN SKIN AGING

Aim : Comparing two aging ranges in order to find new potential candidate gene able to explain biological processes involved in human skin aging.

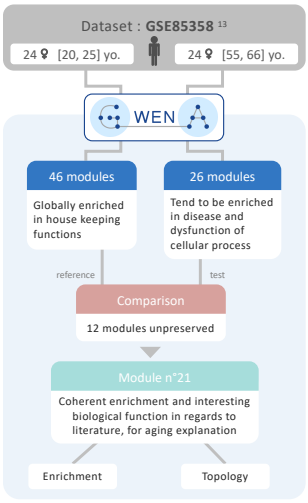

Figure 3 : Application of GWENA on GSE85358 data set and results summary.

OBJECTIVES

1. Providing a complete pipeline for co-expression analysis, from transcriptomic data to extended characterization of genes modules.
2. Bringing novel method allowing module comparison between conditions.
3. Discovering new genes of interest into skin aging.

Further exploration of module n°21

- 144 genes, with only the 69 most strongly interacting, plotted in Figure 4.
- Multiple enrichment linked to immunity processes and cellular stress.
- Proteins coded by genes into this module are mostly localized in membranes of different vacuoles, as trans-membranous proteins. In case of cell membrane, important amount of various cytokine receptor = signaling. Coherent with the immune aspect known to be related to skin aging.
- Topology:
  - 3 genes detected as hubs in young: HLA-DRB1, HLA-DMA, HLA-DM.
  - Density is significantly higher in young module compared to the old. Meaning some interactions, therefore biological function, disappear.

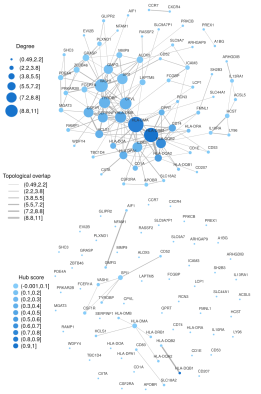

Figure 4 : Module 21's network with intensity of interactions as edge thickness. Upper module is from young population and lower one is the same genes in the old population. Highly connected genes show an important hub score.

CONCLUSION & FUTURE WORKS

- **Co-expression pipeline** WEN
  - Our tool allow **modules comparison** and biological discrimination between conditions.
  - This functionality is integrated in a **complete pipeline facilitating access to co-expression analysis** and interpretation while providing advanced customization for researchers with advanced bioinformatics skills.
  - We aim to develop a graphical interface based on the same package to go further in the user-friendly aspect.
- **Skin aging**
  - Our method appears to be appropriate for highlighting potential candidate genes in skin aging processes exploration.
  - Skin aging is known to be related with immunity and inflammation. Next step will be to determine which pathways or biological function are implied and how they contribute to skin degradation.

REFERENCES

1. van Dam, S., Võsa, U., van der Graaf, A., Franke, L. & de Magalhães, J. P. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics* (2017).
2. Barabási, A. L. & Oltvai, Z. N. Network biology: Understanding the cell's functional organization. *Nature Review Genetics* (2004).
3. López-Kleine, L., Leal, L. & López, C. Biostatistical approaches for the reconstruction of gene co-expression networks based on transcriptomic data. *Briefings in Functional Genomics* (2013).
4. Li, J. *et al.* Application of Weighted Gene Co-expression Network Analysis for Data from Paired Design. *Scientific Report* (2018).
5. MacLennan, N. K. *et al.* Weighted gene co-expression network analysis identifies biomarkers in glycerol kinase deficient mice. *Molecular Genetics and Metabolism* (2009).
6. Kogelman, L. J. A. *et al.* Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA Sequencing in a porcine model. *BMC Medical Genomics*, (2014).
7. Langfelder, P. & Horvath, S. WGCNA : An R package for weighted correlation network analysis. *BMC Bioinformatics* (2008).
8. Ritchie, ME., Phipson, B., Wu, D., Hu, Y, Law, CW, Shi, W, and Smyth, GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* (2015).
9. M. I. Love, W. Huber, S. Anders: Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *BMC Genome Biology* (2014).
10. Reimand J *et al.* g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* (2007).
11. Csardi, Gabor, and Tamas Nepusz. The igraph software package for complex network research. Interjournal, Complex Systems (2006).
12. Ritchie, Scott C, *et al.* A scalable permutation approach reveals replication and preservation patterns of network modules in large datasets. Cell systems (2016).
13. Kuehne, Andreas, *et al.* An integrative metabolomics and transcriptomics study to identify metabolic alterations in aged skin of humans in vivo. *BMC Genomics* (2017)

ACKNOWLEDGMENT

Thank you to L'Oréal – Recherche et Innovation for financing my thesis subject.
Also many thanks to Audrey Lemaçon, Régis Ongaro-Carcy, and Julien Prunier for their help along this project and beyond.

UNIVERSITÉ LAVAL
CHAIRE DE RECHERCHE ET D'INNOVATION L'ORÉAL EN BIOLOGIE NUMÉRIQUE
CHU de Québec

Designed by Gwenaëlle Lemoine

FIGURE D.1 – Affiche présentée à la bi-conférence internationale ISMB/ECCB en 2019

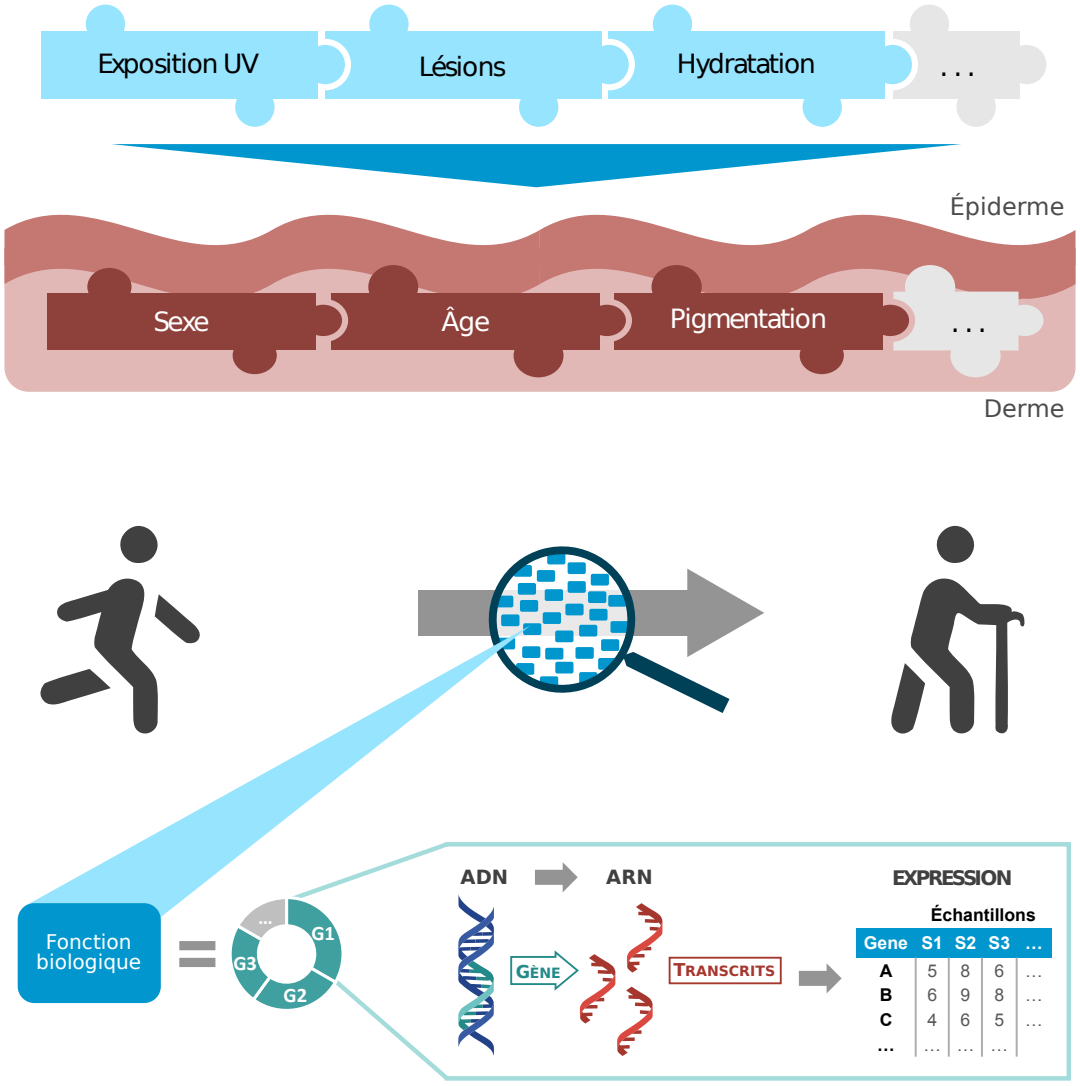

FIGURE D.2 – Illustration de la contribution des facteurs intrinsèques et extrinsèques au vieillissement de la peau et leur perception via l'expression des gènes. Réalisé pour une présentation de séminaire étudiant au CHU de Québec-Université Laval.

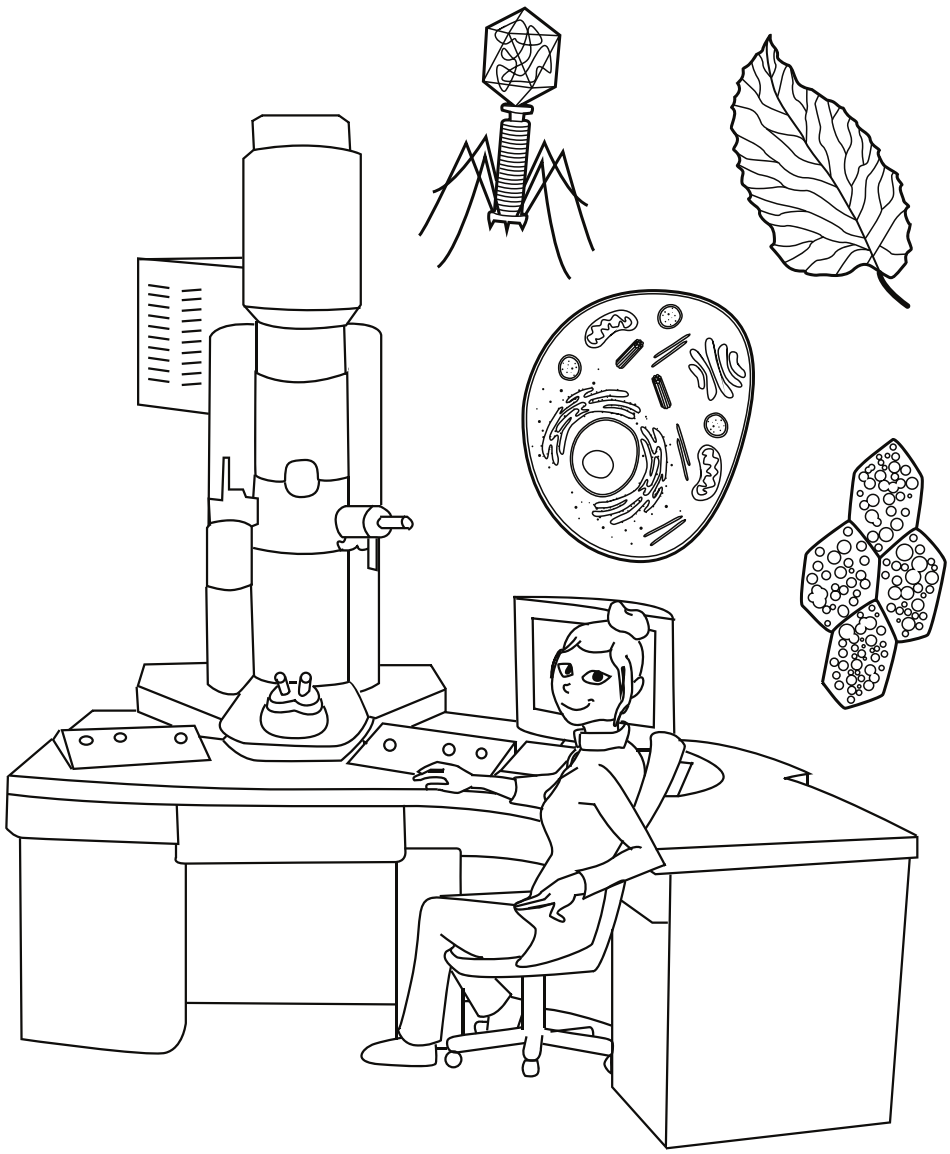

FIGURE D.3 – Ensemble d'illustrations réalisées pour la thèse d'Amandine Verguet en microscopie électronique en transmission appliquée à des échantillons biologiques [217]

Glossaire

condition contexte pratique dont l'existence est nécessaire pour l'observation dans un échantillon d'un phénomène. Ce terme vise dans cette thèse à englober "maladie" et "phénotype" sous un même mot.

connectivité somme des valeurs des liens d'un nœud. Différente définitions existent dépendant du contexte, mais c'est celle-ci qui sera entendue dans cette thèse.

degré nombre de liens (ou arêtes) reliant un nœud (ou sommet), avec les boucles comptées deux fois.

expression de gènes [W] ensemble des processus biochimiques par lesquels l'information stockée dans un gène est lue pour aboutir à la fabrication d'ARN ou transcrit.

fonction biologique terme ambigu pouvant désigner de nombreux concepts en biologie. Aussi dans cette thèse on s'appliquera à préciser au maximum son sens à l'aide du modèle de Pittsburg [218] :

Implications évolutives : l'influence de l'objet sur la dynamique de la population au cours de générations successives, telle qu'elle est rendue possible par ses implications physiologiques et leur interaction avec les pressions environnementales.

Implications physiologiques : l'implication de l'objet dans les processus biologiques, telle qu'elle est rendue possible par un ensemble de capacités, d'interactions et de modèles d'expression, indépendamment des considérations transgénérationnelles.

Interactions : les contacts physiques, directs ou indirects, entre l'objet étudié et les autres composants d'un système, y compris les contacts qui servent de médiateurs aux transformations chimiques.

Capacités : les propriétés physiques intrinsèques de l'objet étudié ; la nécessité du comportement de l'objet compte tenu de son environnement (par exemple, les contraintes structurelles)

Expression : la présence ou la quantité de l'objet étudié (objet ARN ou protéine), ou la présence ou la quantité de ses produits de transcription ou de traduction (objet ADN).

Vague : on n'a pas trouvé de preuves suffisantes pour déduire une ou plusieurs significations de la fonction dans ce modèle, ni pour dériver une nouvelle signification.

.

graphe modèle abstrait mathématique fait de sommets reliés par des arrêtes pour conceptualiser des intéractions.

gène propre première composante d'une analyse par composantes principales sur un module de gènes.

invariance d'échelle [*scale free network*] (mathématiques) réseau dont les degrés suivent une loi de puissance. C'est à dire un réseau où la proportion de nœuds de degré k est proportionnelle à $k^{-\gamma}$ pour k grand, où γ est un paramètre.

longévité durée de vie potentielle d'un organisme.

mécanisme ensemble des acteurs cellulaires et moléculaires intervenant dans la réalisation d'une manifestation d'un phénomène [219].

marque principale représentation des causes et conséquence du vieillissement en neuf marques réalisée par López-Otín [163].

module groupe de gènes co-exprimés au sein d'un réseau.

organe [W] groupe de tissus collaborant à une même fonction physiologique.

organisme [W] ensemble des organes d'un être vivant et, par métonymie, l'être vivant lui-même.

phénotype ensemble des traits observables de l'échelle moléculaire à macroscopique d'un organisme.

plus court chemin suite de liens d'un nœud à un autre de longueur la plus petite possible.

réseau ensemble interconnecté de nœud par des liens ou graphe qui représentent l'organisation d'un phénomène..

tissu [W] ensemble fonctionnel de cellules au type cellulaire différent regroupées en amas, réseau ou faisceau.

transcriptome ensemble des ARN transcrits depuis l'ADN chez un organisme donné.

transcriptomique [W] étude de l'ensemble des ARN messagers produits lors du processus de transcription d'un génome.

transcrit version d'ARN messager issue d'un gène.

type cellulaire spécialisation d'une cellule au cours de sa différenciation pour la réalisation d'une fonction biologique spécifique.

voisin nœuds reliés à un nœud considéré. On peut parler à voisins à une distance x où x représente la distance en terme de liens qu'on prendra en plus en considération.

0. Les définitions précédées d'un [W] sont issues de Wikipédia