Université de Montréal

Combinaison de différents scores génétiques pour mieux évaluer l'impact des CNV sur la cognition

par

Mame Seynabou Diop

Département de Biochimie et de Médecine Moléculaire
Faculté de Médecine

Mémoire présenté en vue de l'obtention du grade de Maîtrise
en Bio-informatique
option Recherche

Août 2021

Centre de recherche du CHU Sainte-Justine,Département de pédiatrie,
Laboratoire Sébastien Jacquemont

Université de Montréal

Faculté de Médecine

Ce mémoire intitulé

Combinaison de différents scores génétiques pour mieux évaluer l'impact des CNV sur la cognition

présenté par

Mame Seynabou Diop

a été évalué par un jury composé des personnes suivantes :

Mark Samuels
(président-rapporteur)

Sylvie Hamel
(directeur de recherche)

Sébastien Jacquemont
(codirecteur)

Marie-Pierre Dube
(membre du jury)

Résumé

Les pratiques de diagnostic génétique standard identifient les variations génétiques délétères (dangereuses pour la santé) rares. Parmi ces variations, on peut citer les variations du nombre de copies (CNV) qui sont présentes chez 10 à 15% des enfants référés aux cliniques de neuro-développement et de pédopsychiatrie [36]. En effet, les CNV sont associées à un risque de troubles neuro-développementaux caractérisés par divers degrés de déficience cognitive, notamment la schizophrénie, les troubles du spectre autistique et la déficience intellectuelle [61]. Dans le laboratoire du Dr Jacquemont, Huguet et al. [36,37] ont effectué la quantification des effets des CNV sur la cognition en utilisant les scores génétiques. En effet, un score génétique est une valeur numérique, continue ou catégorielle pouvant être attribuée à un gène donné dans une expérience quelconque et qui permet d'évaluer son niveau d'intérêt ou de quantifier son importance dans l'expérience en soi [124]. Cependant, cette quantification est faite en utilisant les scores de façon individuelle alors que les scores ont des fonctions biologiques différentes; ce que Huguet et al. [36, 37] ne prennent pas en compte dans leur quantification.

Notre étude vise à quantifier les effets des CNV sur la cognition en combinant plusieurs scores génétiques afin de prendre en compte toutes les dimensions biologiques. Pour ce faire, nous mettons en place un score composite. Ce dernier consiste à regrouper des mesures individuelles en une mesure globale visant à déterminer différents aspects d'un modèle conceptuel.

Nos résultats montrent qu'un score composite est associé à une plus importante réduction du quotient intellectuel (QI) comparé aux études qui utilisent les scores de façon individuelle. Cet effet différentiel montre l'importance d'étudier la perte de QI en prenant en compte toutes les dimensions biologiques (combinant les scores) plutôt qu'une seule dimension biologique.

Mots clés: Cognition, CNV, scores, génétique.

Abstract

Standard genetic diagnostic practices identify rare (unhealthy) deleterious genetic variations such as copy number variants (CNV) in 10 to 15% of children referred to neurodevelopmental and child psychiatry clinics [36]. Indeed, CNV are associated with a risk of neurodevelopmental disorders characterized by varying degrees of cognitive impairment, including schizophrenia, autism spectrum disorders and intellectual disability [61]. In Dr Jacquemont laboratory, Huguet and al. [36,37] quantified the effects of CNV on cognition using genetic scores. Indeed, a genetic score is a numerical, continuous or categorical value that can be attributed to a given gene in any experiment and which makes it possible to evaluate its level of interest or to quantify its importance in the experiment itself [124]. However, this quantification is done by using the scores individually while the scores have different functions; what Huguet and al. [36, 37] do not take into account in their quantification.

Our study aims to quantify the effects of CNV on cognition by combining several genetic scores in order to take into account all the biological dimensions. To do this, we set up a composite score. The latter consists of grouping individual measures into an aggregate measure in order to determine different aspects of a conceptual model.

Our results show that a composite score is associated with a greater reduction in intelligence quotient (IQ) compared to studies that use scores individually. This differential effect shows the importance of studying the loss of IQ by considering all the biological dimensions (combining the scores) rather than a single biological dimension.

MKeywords: Cognition, CNV, scores, genetic.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et des abréviations

AIC : Akaike information criterion
BAF : Fréquence de l'allèle B
CNV : Variation du nombre de copies
NVIQ : Non Verbal Intelligence Quotient
pLI : Probabilité d'être intolérant à la perte de fonction du gène
QI : Quotient intellectuel
TSA : Troubles du spectre autistique
LRR : Ratio du log R
FISH : Fluorescent In Situ Hybridisation
TDAH : Trouble du déficit de l'attention avec ou sans hyperactivité

Dédicaces

À mon héros MOUHAMMAD Rassouloulah PSL,
À ma famille, mon mari,
À mes campeurs avec TDAH du camp Emergo.

Remerciements

J'aimerais exprimer une très profonde gratitude à ma directrice de recherche Sylvie Hamel qui a rendu possible mon voeux de poursuivre la maîtrise avec le cheminement mémoire. Je la remercie aussi pour sa présence à toutes les rencontres, son support, sa patience, ses conseils, tout le temps passé sur les corrections et surtout pour sa bienveillance sans égale.
Je remercie aussi mon co-directeur Sébastien Jacquemont de m'avoir offert l'opportunité de travailler dans un domaine qui m'est très cher.
J'aimerais adresser des remerciements particuliers à Mor Absa Loum pour son implication continu et sans faille. Un simple merci ne saurait suffire mais merci pour ton temps, j'ai beaucoup appris de toi notamment les statistiques. Ton encadrement depuis le début de ma maîtrise m'a permis de développer des compétences qui me suivront tout au long de mon cursus. Merci pour ta présence, tes orientations, tes corrections, tes conseils et surtout ta patience à toute épreuve.
Un énorme merci à Guillaume Huguet. C'est toujours un énorme plaisir de t'écouter car tu rends les choses tellement plus simple. Merci pour tes corrections, tes lessons en génétique et surtout ton appuie ces derniers mois. Je remercie également Jean Louis Martineau et Elise Douard pour vos corrections, vos orientations et votre disponibilité. De vous trois j'ai beaucoup appris.
Mes remerciements à Nadine Younis, Kuldeep Kumar, Catherine Proulx (merci pour ton partage), Zohra Saci, Sayez Kazem, Cécile Poulain et tous les membres du laboratoire Jacquemont.
Toute ma reconnaissance à mon père qui m'a offert le moyen de concrétiser ce projet d'études, merci pour son soutien. Mes remerciements à ma mère, mon frère et mes soeurs pour leur soutien, leurs prières. Un énorme merci à mon époux pour sa patience, sa compréhension, son aide pour appréhender les statistiques et surtout de m'écouter parler de mon projet sans toujours comprendre.
Merci à Jean François Ndiaye qui a rendu ce nouveau système plus facile à appréhender et pour son soutien.
Merci beaucoup à Élaine Meunier pour son aide et à Gierka Kathie.

Chapitre 1

Introduction

Les premières descriptions des troubles neuro-développementaux ont émergé en 1887 avec les travaux du Dr. John Langdon Down qui a décrit pour la première fois le syndrome de Down ou trisomie 21 [40]. Les troubles neuro-développementaux sont des conditions aux facettes multiples qui regroupent beaucoup de troubles, entre autres, les troubles de l'attention, l'autisme, la schizophrénie, les troubles de la communication (verbale et non verbale) et la déficience intellectuelle. Les troubles neuro-développementaux ont une prévalence de 13,87% dans la population générale [80]. La déficience intellectuelle est caractérisée par des capacités cognitives altérées, communément définies par un quotient intellectuel (QI) inférieur à 70 et des déficits sévères de la capacité d'adaptation à l'environnement et au milieu social. Avec une prévalence de 2 à 3 % dans le monde, elle représente l'un des plus grands défis médicaux et sociaux de notre société [60]. Ces troubles peuvent avoir des causes environnementales et/ou génétiques. En effet, une variation dans la séquence génomique peut conduire à la modification des fonctions biologiques.

Avec l'avènement des méthodes de génotypage avec micro-puce, plusieurs associations ont pu être établies entre la variation du nombre de copies (CNV) d'une séquence génomique et ces troubles neuro-développementaux [82]. Les pratiques de diagnostic génétique standard identifient les variations génétiques délétères rares telles que les variantes du nombre de copies (CNV) chez 10 à 15% des enfants référés aux cliniques de neuro-développement et de pédopsychiatrie [36]. En effet, les CNV sont associées à un risque de troubles neuro-développementaux caractérisés par divers degrés de déficience cognitive, notamment la schizophrénie, les troubles du spectre autistique et la déficience intellectuelle [61]. Cependant, bien qu'un lien entre une poignée de CNV et un trouble ait souvent été établi, l'impact quantitatif de la majorité des CNV identifiées chez les patients reste inconnu. Il n'existe actuellement aucune étude ou stratégie convaincante pour comprendre les effets globaux du paysage complexe des variations génomiques rares sur la cognition et le comportement [36]. Pour la plupart des variations rares rapportées aux patients, nous avons peu ou pas de

données caractérisant et quantifiant leurs effets sur la cognition. Étant donné que la taille des effets de variants rares sur des caractères continus n'est pas documentée, il est difficile pour les cliniciens d'estimer la véritable contribution d'une CNV aux symptômes neuro-développementaux d'un patient [36]. Cela est dû à deux problèmes : 1) La plupart des CNV pathogènes sont non récurrentes. Parce qu'elles ne sont observées qu'une ou quelques fois chez les patients, il est impossible d'atteindre la puissance statistique requise pour les études d'association individuelles. 2) La plupart des études se sont concentrées sur l'association de variations à des diagnostics catégoriques complexes tels que l'autisme ou les déficiences intellectuelles. Cependant, on pense que ces diagnostics résultent de l'altération de plusieurs traits cognitifs et comportementaux.

Des travaux antérieurs dans le laboratoire du Dr Jacquemont[36,37] ont conduit à des méthodes nous permettant de quantifier les effets des CNV sur la cognition en fonction de ses caractéristiques intrinsèques que sont les scores génétiques. Ce mémoire est basé sur ces travaux. En effet, nous allons utiliser les scores génétiques pour quantifier les effets des CNV sur la cognition. Pour ce faire, nous allons, en premier lieu, effectuer une revue de la littérature existante afin de mieux comprendre le domaine d'étude. En deuxième lieu, nous allons parler de la problématique, de l'hypothèse et des objectifs de notre étude. En troisième lieu, nous allons présenter les données utilisées (matériels), en quatrième lieu nous allons parler de la gestion des données manquantes des scores génétiques. En cinquième lieu nous allons parler du score composite et des résultats, en sixième lieu nous avons la discussion. Enfin en septième et dernier lieu nous avons la conclusion et les perspectives.

Chapitre 2

Revue de littérature

2.1. Aspects génomiques

Dans le cadre du projet du génome humain [64], le génome a été séquencé en entier grâce à une révolution technologique dans le domaine du séquençage. Cette nouvelle ère a donné naissance au domaine de la génomique qui est la science qui étudie les génomes en se basant sur leur séquence. Elle va permettre aux chercheurs d'en savoir plus sur les facteurs génétiques qui influencent les troubles neuro-développementaux comme la schizophrénie, l'autisme, la déficience intellectuelle et constitue ainsi un outil d'aide à la prise de décision pour le diagnostique. Dans les sous-sections de ce chapitre, nous allons traiter du génome en général et des variants génétiques, plus spécifiquement les variations du nombre de copies (CNV) des gènes. Nous allons également présenter l'état des connaissance sur l'association entre CNV et cognition.

2.1.1. Le génome

L'ensemble du matériel génétique d'un organisme est appelé génome. Chaque être vivant a un génome. Chez l'humain, nous avons le génome mitochondrial et le génome nucléaire. Nous allons nous focaliser sur ce dernier.
Le génome permet la transmission de l'information génétique d'une génération à une autre. Cette information génétique code toutes les fonctions essentielles à la vie et à la reproduction. Le support de l'information génétique est constitué d'acide désoxyribonucléique (ADN). Watson et Crick ont découvert que l'ADN est une macromolécule constituée de deux chaînes (ou brins) enroulées l'une autour de l'autre, en une double hélice de 2.37 nanomètres de diamètre [65]. Sur chaque brin, nous avons une succession de monomères appelés nucléotides. Ces derniers sont constitués de trois éléments: une base azotée, un sucre (le désoxyribose) et un ou plusieurs groupes de phosphates.
Le génome nucléaire de l'humain est constitué d'environ 3,2 milliards de paires de bases (pb)

nucléotidiques et de 20 000 à 25 000 gènes [122, 123]. Un gène est une séquence d'ADN ou encore une séquence de bases azotées qui contient une information particulière.
Ces 3,2 milliards de paires de bases (pb) sont compactées dans 23 paires de chromosomes dont une paire de chromosomes sexuelles et 22 paires d'autosomes. Au niveau des autosomes, sauf dans de rares cas, nous avons deux copies de chaque gène. Chaque gène a une position spécifique dans le chromosome. Les gènes peuvent coder pour des ARN (Acide Ribonucléique) ou pour des protéines. Nous nous intéressons aux gènes codant pour ces dernières.
Les protéines sont des macromolécules construites à partir d'une série de molécules appelées acides aminés qui remplissent diverses fonctions nécessaires au bon fonctionnement de l'organisme. Parmi ces fonctions, nous avons le fait qu'elles peuvent agir en tant que récepteurs, molécules de transport, enzymes, protéines régulatrices pour l'expression des gènes.
L'expression régulée des gènes codés dans le génome humain implique un ensemble d'interactions entre les différents niveaux de contrôle, y compris le dosage génique approprié. Cet ensemble d'interactions est appelé un patron d'expression. Pour certains gènes, des fluctuations dans le niveau de produit génique fonctionnel, dû soit à la variation héritée de la structure d'un gène ou à des changements induits par des facteurs non génétiques, peuvent être importantes ou non. Pour d'autres gènes, les variations dans le niveau de l'expression peuvent avoir des conséquences cliniques désastreuses [39].

2.1.2. Variations structurelles

Dans le génome, les variations génétiques peuvent être des anomalies de nombre (de chromosomes) dites aneuploïdies ou des anomalies de structures. On parle d'aneuploïdie quand une cellule se retrouve avec un nombre anormal de chromosome. Chez l'homme on retrouve ce cas de figure quand la cellule a 45 ou 47 chromosomes au lieu de 46. Ces deux cas de figure se retrouvent respectivement dans **le syndrome de Turner** où on a une absence totale ou partielle d'un des deux chromosomes X chez la femme et dans **la trisomie 21** où on se retrouve avec trois chromosomes 21 au lieu de deux. Cependant la plupart de ces malformations chromosomiques ne sont pas viables comme dans le cas de **la trisomie 13** ou **la trisomie 18**.
Une anomalie de structure définit une grande classe d'altérations génomiques. Ces altérations peuvent être quantitatives (variation du nombre de copies), positionnelles (translocations) ou de type d'orientation (inversions) [23]. La variation structurelle du génome peut influencer directement ou indirectement le dosage des gènes par différents mécanismes [83], et donc peu influencer les traits phénotypiques du porteur voire engendrer un trouble spécifique. Parmi les variations structurelles, figurent les variations avec les délétions (suppressions) et celles avec les duplications.
Une délétion est une perte d'une partie du matériel génétique. Elle peut aller d'un à plusieurs

millions de nucléotides pouvant toucher un à plusieurs gènes. Quant à la duplication, elle correspond à un gain d'une partie du matériel génétique. Tout comme la délétion, elle peut aussi aller d'un à plusieurs millions de nucléotides pouvant toucher un à plusieurs gènes. Ces délétions et duplications font partie des anomalies de structures qui ne sont pas équilibrées (soit un morceau de chromosome en plus ou en moins) au contraire des inversions qui peuvent ne pas engendrer d'altération de quantité de matière. Elles peuvent être *de novo* (une altération de la séquence d'ADN d'un individu qui n'est pas présente chez les parents) ou héritées des parents.
Lorsque des délétions ou duplications sont observées au niveau du matériel génétique, on parle de variation du nombre de copies (CNV) d'un gène.

2.1.3. Les CNV

Les variations du nombre de copies (CNV) sont une forme courante de variation structurelle. Elles sont définies comme des séquences génomiques dont la taille est supérieure ou égale à 1000 pb et qui diffèrent en nombre de copies de celui d'un génome de référence [56]. Les CNV d'un gène correspondent soit à une délétion ou à une duplication. Ceci peut aller d'une anomalie structurale (évènements sub-microscopiques) à des anomalies du nombre (trisomie ou monosomie). Elles jouent un rôle important dans l'évolution bien qu'elles peuvent aussi être à l'origine de certaines maladies.
Les cytogénéticiens cliniques doivent différencier les variations du nombre de copies (CNV) des gènes qui sont susceptibles d'être pathogènes et les CNV qui sont moins susceptibles de contribuer à la présentation clinique d'un individu affecté. Ils classent les CNV en trois catégories : celles qui sont susceptibles d'être bénignes, celles qui sont susceptibles d'être pathogènes et celles dont la signification clinique est inconnue. Les CNV qui chevauchent des régions critiques de syndromes de micro-délétion ou de micro-duplication connus (ou qui chevauchent d'autres régions génomiques définies comme cliniquement significatives) sont susceptibles d'être de nature pathogène [13].
Les CNV résultant des délétions ont une plus grande probabilité d'être pathogènes que celles provenant des duplications. Les duplications sont donc mieux tolérées. Ceci est cohérent étant donné qu'une délétion est synonyme d'une perte de matériel génomique alors que pour la duplication, le matériel sera toujours présent [36,37,38]. Selon des études effectuées au laboratoire du Dr Jacquemont [36,37,38], les délétions sont associées au trouble du spectre de l'autisme (TSA) et à l'intelligence générale. Cependant, l'impact qu'elles ont sur cette dernière est supérieur.
Les CNV dépassant 500 kb peuvent avoir d'énormes conséquences telles que les troubles du développement et le cancer [23]. Dans l'étiologie d'une maladie ou d'un trouble, la taille des

CNV est très importante. Si la taille des CNV est énorme, cela touche des régions génomiques et ces dernières modifient les niveaux d'ARN et de protéines des gènes importants pour maintenir le développement normal. Cependant, une grande taille de CNV peut ne pas avoir d'impact si elle ne modifie pas le cadre de lecture des gènes. En outre, le nombre de gènes que contient la CNV est plus fiable pour mesurer l'impact. En effet, une CNV peut être grande sans contenir beaucoup de gènes alors qu'une autre peut être petite avec beaucoup de gènes. Donc logiquement cette dernière est susceptible d'être plus pathogène. La probabilité qu'a une CNV d'être pathogène peut aussi être déduite du fait qu'elle soit de novo, rare ou commune.

2.1.4. La taille des CNV

Dans l'étiologie d'une maladie ou d'un trouble, la taille des CNV est très importante. Si la taille des CNV est énorme, cela touche des régions génomiques et ces dernières modifient les niveaux d'ARN et de protéines des gènes importants pour maintenir le développement normal. Cependant, une grande taille de CNV peut ne pas avoir d'impact si elle ne modifie pas le cadre de lecture des gènes.
En outre, le nombre de gènes que contient la CNV est plus fiable pour mesurer l'impact. En effet, une CNV peut être grande sans contenir beaucoup de gènes alors qu'une autre peut être petite avec beaucoup de gènes. Donc logiquement cette dernière est susceptible d'être plus pathogène.
La probabilité qu'a une CNV d'être pathogène peut aussi être déduite du fait qu'elle soit *de novo*, rare ou commune.

2.1.5. CNV rares et communes

Il est essentiel de faire une distinction entre les CNV communes et les CNV rares. Nous avons les CNV communes qui sont partagées par une portion supérieure à 1% de la population. Elles sont dues à des évènements ancestraux et peu d'entre elles ont un impact sur les troubles courants car en général, plus une CNV est fréquente moins elle est dangereuse.
Les CNV rares sont celles dont la fréquence inférieure à 1% [84]. Plus la taille des délétions rares et peu fréquentes est grande, plus l'intelligence psychométrique d'un individu est faible [85]. Il existe de plus en plus d'études qui montrent le rôle des CNV rares dans le développement des troubles neuro-psychiatriques. Il est prouvé que ces CNV ont également un effet sur la variation de la cognition dans ce qui est considéré comme la gamme phénotypique « normale » [84].
Cependant, avant de se prononcer sur cette classification (rare ou commune), sur la taille et la contribution des CNV dans les maladies, il faut d'abord les détecter.

2.1.6. Détection des CNV

La capacité de détecter et de caractériser des variantes structurelles, dans la plage de taille de 1 kb à 3 Mb, de manière robuste à travers le génome n'a pas toujours été possible [83]. Quand la variation du nombre de copies atteint le stade d'aneuploïdie, le chromosome en plus ou en moins peut être détecté grâce au **caryotype**. Ce dernier utilise une culture cellulaire et permet de détecter les CNV avec une résolution supérieure à 5 Mb. Les CNV peuvent aussi être détectées à l'aide d'un **Fish** (l'hybridation in situ en fluorescence) qui permet de trouver le chromosome qui manque ou qui est en trop. Il a une meilleure résolution que le caryotype.
Une autre technique qui offre une meilleure résolution est l'analyse par micro-puce. En effet, elle permet non seulement de définir la taille de l'anomalie mais aussi de montrer les plus petits déséquilibres. Parmi ces analyses par micro-puce nous avons le CGH et le SNP array. **Le CGH** (Comparative genome hybridization), permet de détecter les anomalies chromosomiques retrouvées par le caryotype (sauf les anomalies équilibrées) mais aussi d'autres anomalies plus cryptiques. C'est une méthode fiable et facile d'utilisation car elle est automatisable. Elle permet de tester par hybridation compétitive un nombre très élevé de régions du génome. Elle permet de détecter bon nombre de variations mais ne trouve pas la triploïdie. Quant au **SNP array** (Single nucleotide polymorphism genotyping array) , il permet la détection des anomalies cytogénétiques déséquilibrées dans l'ensemble du génome et à une résolution élevée (25-50 Kb). Cependant, sa résolution dépend de la densité en SNP de la zone d'intérêt. Contrairement au CGH array, il permet de détecter la triploïdie. Les deux technologies peuvent être combinées pour de meilleurs résultats. Plusieurs algorithmes utilisent les données provenant de ces deux technologies. Parmi eux il y'a les chaînes de Markov cachées Hidden Markov Model) qui sont des séquences d'événements aléatoires, faisant partie d'un ensemble possible d'états qui satisfait la propriété de Markov [116]. Cette dernière stipule que le prochain état dans lequel on sera ne dépend que de l'état où l'on est présentement, et non des états passés [116]. Pour la détection des CNV, les états cachés sont le nombre de copies d'ADN à chaque sonde, et les variables observables sont le ratio du log R (LRR) et la fréquence de l'allèle B (BAF) [116]. À partir de ces dernières, on peut inférer pour chaque sonde si on se trouve, par exemple dans l'état normal (2 copies), dans une délétion (1 copie) ou dans une duplication (3 copies) [116]. Il y a aussi les outils QuantiSNP et PennCNV qui utilisent les chaines de Markov cachées. Grâce à ces outils, pour chaque position du génome de l'individu génotypé, on peut retracer les séquences de positions altérées par des CNV [117].
Au cours des 10 dernières années, il y a eu une explosion dans la génération des données génomiques due principalement à l'utilisation du séquençage de nouvelle génération (**NextGen** ou **NGS**) [20]. Ces avancées technologiques comme Illumina, PacBio, Oxford nanopore, etc.,

ont eu un impact significatif sur la détection des CNV. En effet, beaucoup de méthodes de détection des CNV avec une résolution meilleure que les précédentes furent développées sur cette base.
La détection et la caractérisation (taille, rare ou commune) des CNV sont importantes avant de faire des études d'associations sur elles. Nous allons voir dans la section suivante quelques travaux effectués sur l'association entre CNV et cognition.

2.1.7. État des connaissances sur l'association entre CNV et cognition

Les deux dernières décennies ont connu une explosion des données génomiques due au séquençage du génome entier et aussi aux technologies de séquençages qui deviennent de plus en plus pointues et se sont démocratisées au fil des années. Le séquençage du génome entier a permis aux chercheurs d'élargir leurs investigations sur différents domaines dont ceux s'intéressant à l'association entre CNV et cognition.
Des études ont tenté d'associer la cognition et les CNV en étudiant des troubles neuro-psychiatriques tels que l'autisme, l'Alzheimer, la schizophrénie, etc.
Parmi elles, l'étude de Stefansson et al. [29] utilise divers tests de la fonction cognitive effectués sur les individus atteints de schizophrénie et/ou d'autisme et sur les individus qui n'ont pas reçu de diagnostic de schizophrénie et/ou d'autisme mais qui sont porteurs de CNV prédisposant à ces deux troubles neuro-psychiatriques. Les résultats de l'étude montrent comment les anomalies cognitives et les changements dans la structure du cerveau chez les patients atteints de schizophrénie et/ou d'autisme se retrouvent également chez les porteurs témoins de CNV qui présentent un risque élevé de la maladie. Ceci suggère alors que ces anomalies cognitives ne sont pas liées à la maladie en tant que telle.
Andrew K. MacLeod et al. [30] examinent les effets des CNV rares sur la cognition. Pour ce faire, ils ont dérivé trois variables pour chaque individu : le nombre total de CNV qui a passé le contrôle qualité, la longueur totale de ces variants et le nombre de gènes touchés par ces CNV. Un certain nombre de modèles de régression ont été ajustés, en utilisant les scores des facteurs d'intelligence, des corrections par l'âge et le sexe, par rapport aux trois variables ci-dessus. Ils ajustent par la cohorte. Cette étude n'a trouvé aucun effet des CNV sur la cognition générale.
L'étude de Männik et al. [31] rapporte que la présence à la fois de CNV syndromiques récurrentes et CNV rares de taille intermédiaire non récurrentes, qui sont cumulativement fréquentes dans la population générale (10,5%), est associée à une déficience intellectuelle et négativement associée au niveau d'instruction. C'est-à-dire que chez les porteurs de délétion dont la taille est comprise entre 250 kb et 500 kb, la fréquence de la déficience intellectuelle augmente à 4,3 % contre 1,7 % dans la population générale estonienne.

Basé sur une revue systématique et d'une méta-analyse de la littérature existante, Johan H. Thygesen et ses collègues [32] vont tenter d'associer les effets des CNV à la cognition générale. Cette approche ne montre aucune association entre eux. En parallèle, ils mènent une étude sur le rapport des CNV associées à la schizophrénie et la cognition qui montre que les mesures de la mémoire de travail et la mémoire à long terme sont altérées.
Dans cet article, ils utilisent des tests cognitifs et l'indice de privation de Townsend (Townsend a montré que les privations multiples sont fortement corrélées à la pauvreté monétaire). Ils effectuent des analyses de régression linéaire (**GLM**) avec le score cognitif comme variable dépendante. Les chercheurs rapportent que les CNV impliquées dans les troubles neuro-développementaux, y compris la schizophrénie, sont associés à des déficits cognitifs, même chez les individus non diagnostiqués [33].
Certaines études sont parvenues à établir le lien entre la cognition et les CNV, d'autres n'ont pas trouvé d'évidences [30]. Nous voyons que ces études visent à associer les CNV qui sont impliquées à des troubles neuro-développementaux à la cognition mais malheureusement, ils ne quantifient pas la taille des effets des CNV sur la cognition. Donc des modèles ont été développé sur la base des scores génétiques.

2.1.7.1. Les scores génétiques

Pour connaître l'impact d'une CNV, il faut déterminer son effet sur la fonction d'un gène surtout si ce gène est sensible au dosage. On parle de sensibilité au dosage lorsqu'une perte ou un gain de la moitié du nombre de copies d'un gène est délétère. Pour illustrer ce phénomène, nous allons parler de l'haplo-insuffisance.
En effet, pour certains gènes, la délétion d'une copie fonctionnelle d'un génome diploïde (deux copies) modifie le phénotype de l'organisme en un état anormal ou pathologique. Ces gènes sont dits haplo-insuffisants car une seule copie est insuffisante pour produire le phénotype normal [27]. Cette étude a pu trouver dans OMIM et PubMed 299 gènes haplo-insuffisants. Selon l'étude en question, la plupart de ces gènes sont associés à des maladies (ou susceptibles à des maladies) telles que le cancer, le retard mental, les troubles neurologiques, le retard de croissance, etc.
Étant donné qu'une CNV peut affecter un ou plusieurs gènes, son impact est alors lié à son effet sur la fonction des gènes qu'elle a affectés d'autant plus si ces gènes sont sensibles au dosage. Les CNV qui touchent des gènes sensibles au dosage peuvent être dangereuses et causes de maladies tandis que celles qui affectent des gènes insensibles au dosage ne sont pas dangereuses. Plus récemment, de grands jeux de données portant sur le séquençage de grands nombres de sujets ont permis le développement de mesures de sensibilité au dosage des gènes [27]. Ces mesures sont appelées scores génétiques. Les informations provenant de ces scores peuvent être utilisées pour dépister des maladies afin d'identifier les individus à plus haut risque afin de cibler les interventions thérapeutiques ou de prévention [25]. Ces

scores sont associés à chaque gène et chaque CNV renferme plusieurs gènes. Un score de gène est une valeur numérique, continue ou catégorielle pouvant être attribuée à un gène donné dans une expérience quelconque et qui permet d'évaluer son niveau d'intérêt ou de quantifier son importance dans l'expérience en soi [124].
Selon la base de données GnomAD, c'est une fonction du rapport entre le nombre observé et le nombre attendu de variants de perte de fonction dans un gène donné. Ces scores permettent de quantifier l'intolérance d'un gène à la perte de fonction. Les tableaux suivants illustrent les scores génétiques dont nous disposons.

	Scores de contraintes génétiques en fonction des types de variants (scores d'haplo-insuffisance)
Description	Lek et al ont créé le score de probabilité d'intolérance à la perte de fonction (pLI) d'un gène. Ce score a été calculé pour tous les gènes codants en utilisant les séquences d'exomes de 60 706 individus issus du Consortium d'agrégation de l'exome (ExAC). Un algorithme utilisant les nombres de variants tronquant la protéine observés et attendus au sein de chaque gène permet d'attribuer un score pLI aux gènes. Ce score varie de 0 à 1, où 0 représente les gènes les plus tolérants à la perte de fonction et les gènes dont le pLI >=0,9 sont considérés comme intolérants à la perte de fonction (pLI >= 0,9). Karczewski et al ont utilisé environ 141 000 exomes de la base de données d'agrégation du génome (GnomAD) et ont introduit le score LOEUF (loss-of-function observed/expected upper bound fraction) qui est une estimation de la limite supérieure d'un intervalle de confiance du ratio des variants de perte de fonction observé/attendu. Le score LOEUF varie de 0.03 à 2, où les scores <= 0.35 indiquent que les gènes sont intolérants à la perte de fonction.
Pourquoi nous utilisons ces scores?	Ces scores de contraintes génétiques permettent de quantifier l'intolérance à la perte de fonction des gènes codants. Ils aident à évaluer l'impact d'un variant de perte de fonction sur chacun des gènes codants identifiés. En effet, la quantification de l'intolérance génique à la perte de fonction fournit une information supplémentaire aux études sur les maladies. Ces scores peuvent nous aider à quantifier l'impact des effets des variations du nombre de copies des gènes sur la cognition. Huguet et al 2018, Huguet et al 2020 et Douard et al 2020 [36,37,38] ont mené des études dans ce sens.
	Scores d'évolution
Description	Dumas et al ont décrit un criblage exhaustif de tous les gènes codant chez homo sapiens pour la conservation et la divergence par rapport à l'ancêtre commun des primates. Ils ont utilisé les données de séquençage. Ce sont des scores d'évolution des gènes codant pour les protéines à travers divers tissus et divers fonctions biologiques [52].
Pourquoi nous utilisons ces scores?	Ces scores nous permettent d'étudier la dynamique évolutive du gène. [52].

	Scores d'expression génique au niveau du cortex et du cerveau entier
Description	L'Atlas du cerveau humain d'Allen fournit une vue anatomiquement complète de l'expression des gènes dans le cerveau. L'ensemble des données du transcriptome se compose de 58 692 mesures de l'expression génique dans 3702 échantillons de cerveau obtenus à partir de 6 individus adultes. Ces données ont permis à Paus et al de créer des scores mesurant le gradient d'expression de XXX gènes à travers 68 régions corticales (que l'on appellera PC1_Pauss/PC2_Pauss /PC3_Pauss) [49]. Burt et al ont analysé les données transcriptionnelles et neuro-anatomiques T1 et T2 d'humains et de singes pour étudier l'organisation hiérarchique des microcircuits corticaux. Ils ont combiné ces mesures données de neuro-imagerie structurelle et d'expression génétique pour créer le score de corrélation de l'expression des gènes avec le ratio (TMC) qui fournit une approximation non invasive de la hiérarchie de l'expression des gènes dans le cortex [50].
Pourquoi nous utilisons ces scores?	Ces mesures (TMC, PC1_Pauss/PC2_Pauss /PC3_Pauss) nous offrent des informations sur la fonction des gènes dans le développement et la maintenance des différents réseaux cérébraux. Elles nous seront utiles pour quantifier l'impact des effets des variations du nombre de copies sur la cognition tout en prenant en compte la localisation corticale des gènes altérés.
	Scores d'expression temporelle des gènes au niveau du cerveau
Description	Hyo Jung Kang et al ont voulu étudier la dynamique spatio-temporelle du transcriptome du cerveau humain. Ils explorent les transcriptomes de 16 régions comprenant le cortex cérébelleux, le noyau médiodorsal du thalamus, le striatum, l'amygdale, l'hippocampe et 11 zones du néocortex. L'ensemble de données a été généré à partir de 1340 échantillons de tissus prélevés sur 57 cerveaux au cours du développement (pré- et post-natal) et adultes en post-mortem de donneurs représentant des hommes et des femmes de plusieurs ethnies. Les résultats indiquent que la majorité des gènes exprimées dans le cerveau sont temporellement et, dans une moindre mesure, spatialement régulés et que cette régulation se produit principalement pendant le développement prénatal [51].
Pourquoi nous utilisons ces scores?	Elles nous seront utiles pour quantifier l'impact des effets des variations du nombre de copies sur la cognition et prenant en compte la structure temporelle des gènes. Contrairement aux mesures citées précédemment qui ne sont que chez l'adulte, ces mesures fournissent l'information en prénatal et aussi en bas âge.

	Scores de l'expression spatiale des gènes au niveau du cerveau
Description	L'architecture fonctionnelle du cerveau est organisée à travers plusieurs niveaux de résolutions spatiales, des réseaux distribués aux zones localisées qui les composent. Sebastian Urchs et al pensent qu'une parcellisation du cerveau qui définit les nœuds fonctionnels à plusieurs résolutions est nécessaire pour étudier le connectome fonctionnel à ces échelles. Ils présentent un modèle de segmentation intrinsèque multi-résolution (MIST) qui est une parcellisation multi-résolution au niveau du groupe de la matière grise corticale, sous-corticale et cérébelleuse [53]. Ils utilisent les données de Cambridge provenant de 198 sujets (123 femmes) âgés de 18 à 30 ans. Ces données sont disponibles à partir du projet 1000 connectomes fonctionnels. Ces données ont permis à Sebastian Urchs et al de créer les scores PC1_WholeBrain_MIST64, PC2_WholeBrain_MIST64, PC3_WholeBrain_MIST64.
Pourquoi nous utilisons ces scores?	Le développement et le fonctionnement du cerveau dépendent de la régulation précise de l'expression des gènes. Ces mesures spatiales nous offrent des informations sur plusieurs niveaux nous permettant de mieux quantifier l'impact des effets des variations du nombre de copies des gènes sur la cognition

Fig. 2.1. Les scores génétiques. Ces tableaux illustrent les différents scores génétiques et pourquoi nous nous interessons à eux.

2.1.7.2. Association entre CNV et cognition basée sur les scores génétiques

Nous allons à présent explorer les travaux basés sur les scores génétiques effectués au laboratoire du Dr Jacquemont. Ces travaux visent à mesurer la taille des effets des CNV sur la cognition en utilisant les scores d'haplo-insuffisance (voir 2.1).

Huguet et al.[34] ont mené une étude pour prédire l'effet des CNV sur la performance du quotient intellectuel (PIQ) et sur le quotient intellectuel verbal (VIQ). Ils annotent les gènes pour toutes les CNV en se basant sur les scores d'haplo-insuffisance, les scores d'expression temporelle et tissulaire et sur la fonction des gènes [34]. Cela a permis de détecter des délétions rares et des duplications supérieures à 250 kb chez 10% des individus. Ils trouvent que les délétions supérieures à 250Kb sont associées à une réduction du quotient intellectuel (QI) de 6 points avec une p-value de 2.10^{-3}. Ceci suggère que la taille et le contenu génétique des délétions rares sont associées à une réduction du QI [34]. L'effet des duplications rares sur le QI n'était pas significatif. Pour estimer l'effet de toutes les délétions sur le QI, une procédure de modèle linéaire par étapes a convergé vers un modèle comprenant des scores d'haplo-insuffisance [34].

Une autre étude menée dans le laboratoire du Dr Jacquemont vise à mesurer et estimer la taille des effets des CNV sur le QI [36]. Elle est effectuée à l'aide de deux cohortes représentatives de la population générale. Grâce aux outils de génotypage et de contrôle qualité, des CNV de 50kb ou plus sont identifiées. Pour mener à bien cette étude, des annotations fonctionnelles des gènes ainsi que des régressions linéaires sont effectuées afin d'expliquer le QI à l'aide des scores fonctionnels de gènes. Ces annotations concernent les gènes inclus dans les CNV et permettent d'identifier la variable qui donne une meilleure prédiction sur la variation du QI. Les résultats montrent que les scores d'haplo-insuffisance, en particulier le score de probabilité d'être intolérant à la perte de fonction du gène (pLI), expliquent le mieux la taille des effets des délétions sur le QI. En effet, un point de pLI réduit correspond à une réduction du QI de 2.74 points. Ces résultats ont été également observés dans les travaux de Douard et al. [38] sur les populations autistiques et non sélectionnées (un point de pLI réduit correspond à une réduction du QI de 2.6). Ils ont montré que des modèles linéaires, utilisant la somme de pLI de tous les gènes inclus dans une délétion, peuvent prédire la taille de leur effet sur le quotient intellectuel (QI). Ils obtiennent une concordance de 75% avec la littérature. Ceci suggère que l'association des délétions avec le QI peut être modélisée à l'aide de scores d'haplo-insuffisance basés sur une hypothèse linéaire et additive [36]. Cependant cet article montre uniquement la taille des effets des délétions. Celle des duplications n'étant pas significative.

Dans [37], Huguet et al. adaptent le modèle de [36] et utilisent les mesures d'intolérance à la probabilité de perte de fonction (pLoF) pour estimer l'effet des CNV sur l'intelligence générale. Parmi 10 variables, les 2 principaux scores d'haplo-insuffisance que sont le pLI et LOEUF (loss-of-function observed/expected upper bound fraction) expliquent le mieux la

variance de l'intelligence générale [37]. En effet, une sélection de variable par les modèles linéaires basée sur le critère d'information d'Akaike (AIC) a permis de sélectionner ces deux scores comme étant les meilleurs [37]; ce critère permet de sélectionner le meilleur modèle. Les résultats montrent, avec une concordance de 78%, que le ratio de la taille des effets des délétions et des duplications sur la cognition est de 3:1. En effet, les mêmes résultats que [36] sont retrouvés pour les délétions. Pour les duplications, ils trouvent qu'un point de pLI correspond à une diminution de l'intelligence générale de 0.75 points. Contrairement à l'article précédant [36] qui a utilisé le pLI, qui est une variable binaire pour expliquer la variance de l'intelligence générale, l'étude [37] utilise le LOEUF, qui est une variable continue. En définitive, ces différentes études sur les CNV autosomiques de Huguet et al. [35,36,37] et de Douard et al. [38] montrent qu'en utilisant des modèles additifs sur les scores génétiques, nous pouvons estimer la taille des effets des CNV sur l'intelligence générale. Cependant, ces modèles développés par Huguet et al. [35,36,37] et Douard et al. [38] ne prennent en compte que les scores de contrainte génétique (scores d'haplo-insuffisance). De plus dans leur étude, les scores d'haplo-insuffisance sont utilisés de façon individuelle. Il est vraisemblable que des informations sur la fonction des gènes dans le développement et la maintenance des différents réseaux cérébraux est essentielle pour mieux comprendre et prédire l'effet des gènes et des CVN codants sur la cognition. De nouveaux scores d'expression ont été étudié récemment par Burt, Paus, Hyo Jung Kang et Sebastian Urchs (voir 2.1). Ils ont établi des niveaux d'expression des gènes par région. Ils reflètent l'expression des gènes à travers le développement du cortex humain et la distribution spatio-temporelle des gènes dans le cerveau. Nous pensons que ces scores sont très importants car la cognition émerge du cerveau. De ce fait, nous cherchons à rassembler l'ensemble des informations connues sur l'expression spatio-temporelle des gènes dans le cerveau afin d'en extraire un score composite pour expliquer les effets des variations du nombre de copies de ces gènes sur la cognition.

2.1.8. La génétique cognitive et troubles associés

La cognition est la capacité qu'a l'être humain de traiter l'information. Elle permet au cerveau de procéder à l'enregistrement, à la consolidation, à l'accumulation et à la récupération de l'information afin d'interagir avec l'environnement. La cognition fait appel à de nombreux processus mentaux, tels que l'attention, la perception, la mémoire, le langage et le raisonnement [4].

Des études ont montré [12,28] que les troubles cognitifs peuvent avoir pour racine des causes génétiques telles que la délétion, la monosomie (perte d'un chromosome, par exemple avoir 45 chromosomes au lieu de 46 comme le syndrome de Turner) ou la triplication d'un chromosome comme la trisomie 21. Beaucoup de gènes s'expriment dans le cerveau, raison pour laquelle une multitude de troubles neuro-développementaux découle de la non-expression ou

de la sur expression d'un gène.
Les troubles génétiques du développement intellectuel les plus étudiés sont ceux liés au chromosome X (l'$Xfragile$) et à la trisomie 21 (syndrome de Down) [1]. En effet, les CNV associées à ces troubles ont été détectées très tôt car le caryotype et le Fish (les premières méthodes de détection) peuvent détecter une CNV qui touche tout un chromosome. Le syndrome de l'$Xfragile$ est la cause la plus fréquente du retard mental héréditaire chez les enfants. Ce syndrome résulte du fait qu'un seul gène est brisé ou endommagé sur l'un des chromosomes X [88]. Les difficultés cognitives et comportementales sont courantes chez tous les individus atteints du syndrome de l'$Xfragile$, mais sont plus graves chez les garçons (car ils reçoivent un seul X)[88]. Quant à la trisomie 21 ou syndrome de Down, elle est la première cause génétique de retard mental touchant plus de 5 millions de personnes à travers le monde et 500 000 en Europe. La triplication du chromosome 21 décrite en 1959 comme étant responsable du syndrome de Down modifie le développement du système nerveux central et la plasticité neuronale, conduisant à des altérations de la cognition et du comportement.
Au fil des ans, nous assistons à l'avènement des méthodes de détection avec micro-puce qui offrent une meilleure résolution et permettent ainsi de détecter des CNV de plus petite taille. Ces trouvailles permettent d'associer ces CNV aux troubles neuro-développementaux tels que la schizophrénie, l'autisme,etc. L'autisme est un trouble neuro-développemental hautement héritable avec une étiologie hétérogène regroupant les facteurs génétiques, épigénétiques et environnementaux. Des études génétiques ont révélé l'implication de centaines de variants génétiques dont les CNV dans l'autisme [57]. Bien que les autistes présentent entre eux de larges différences vu la largesse du spectre, trois catégories de déficits cognitifs sont relevées : la cognition sociale, la cohérence centrale et les fonctions exécutives. Quant à la schizophrénie, c'est aussi un trouble neuro-développemental hautement héritable dont l'étiologie génétique est complexe. Les personnes qui souffrent de schizophrénie ont des troubles cognitifs dont les troubles de l'attention, des troubles de la mémoire et des fonctions exécutives. Il y a beaucoup d'études sur l'association des CNV à la schizophrénie, cependant la mieux établie est l'association avec la délétion 22q11.2 [58].
À présent, nous savons que les CNV sont associées aux troubles neuro-développementaux. Cependant, la contribution relative des CNV à ces troubles est encore loin d'être comprise; la quantification des effets des CNV sur la cognition reste donc un défi.

2.2. Aspects statistiques

Les statistiques sont un volet important de la bio-informatique car elles nous donnent des outils qui peuvent nous aider à attribuer un niveau de confiance ou un degré d'incertitude à nos estimations/résultats. Dans cette section, nous allons voir le score composite ainsi que tous les aspects nécessaires pour sa construction.

2.2.1. Le score Composite

Le score composite consiste à regrouper des mesures individuelles en une mesure globale visant à déterminer différents aspects d'un modèle conceptuel. Il permet de réduire le nombre de variables à étudier. Dans cette étude, les mesures individuelles sont les scores génétiques et la mesure globale est la combinaison de ces scores génétiques.
Il y a au moins deux raisons pour lesquelles un score composite peut être utile. Premièrement, résumer toutes les mesures en une seule mesure facilite les comparaisons avec d'autres variables sans avoir à relever les défis soulevés en testant plusieurs hypothèses si chacune des mesures était considérée de façon individuelle. Deuxièmement, en incluant plusieurs mesures dans une seule mesure, l'impact de l'erreur de mesure est minimisé [45]. Dans la majorité des cas, un score composite est calculé en prenant la somme pondérée des mesures ou la moyenne des mesures incluses dans chaque domaine [89].
Il existe deux méthodes pour construire un score composite: les constructions réflexives et les constructions formatives. Elles sont décrites dans les sections ci-dessous.

2.2.1.1. Constructions réflexives

Si l'existence de la construction du score composite est indépendante des mesures spécifiques qui ont été utilisées pour le mesurer, on parle de construction réflexive. De ce fait, supprimer ou ajouter des mesures ne change en rien le sens de la conception. Les mesures empiriques et le construit réflexif (score composite construit à partir de la construction réflexive) ont une relation de cause à effet (on peut dire qu'une mesure est élevée si et seulement si le construit est élevé). Ceci implique qu'améliorer la qualité revient à améliorer le concept plutôt que les mesures. Ces dernières sont le plus souvent très corrélées [69].
L'analyse par composantes principales (ACP) aide à trouver des ensembles des mesures corrélées. C'est un outil extrêmement puissant de compression et de synthèse de l'information, très utile lorsque l'on est en présence d'une somme importante de données quantitatives à traiter et interpréter [55]. C'est une méthode statistique qui nous permet de réduire le nombre de variables sans pour autant perdre une grande partie de l'information.
L'analyse par composantes principales est une décomposition d'un ensemble de variables corrélées en un ensemble de variables non corrélées, appelées composantes principales (CP), qui sont organisées par ordre décroissant de variance. Elle peut également être utilisée pour réduire la dimensionnalité en coupant les CP qui sont moins importantes (moins de variance) que les données d'origine [90]. L'ACP est fréquemment utilisée en sciences biologiques pour l'analyse globale des ensembles de données omiques. Elle fournit des informations non supervisées sur les directions dominantes de la plus grande variabilité dans les données et peut donc être utilisée pour étudier les similitudes entre les échantillons individuels ou la formation de regroupements [91].

Le but de la construction réflexive est de voir si ces différentes mesures expriment la même caractéristique.

2.2.1.2. Constructions formatives

À côté des constructions réflexives, nous avons le construit formatif. Contrairement au premier, la construction formative dépend des mesures empiriques. Ces dernières sont utilisées pour mesurer le concept d'intérêt. Si les mesures d'intérêt changent, le construit va changer. Il n'est pas nécessaire que les mesures soient fortement corrélées. Le but d'un construit formatif est d'identifier une gamme de mesures qui captent les différentes dimensions de la performance à laquelle on s'intéresse, plutôt que différentes mesures qui reflètent la même caractéristique ou trait, comme dans le cas d'un construit réflexif [69].
Dans ce cas, nous ne pouvons pas utiliser l'ACP pour construire notre score composite. Ce dernier est plutôt estimé en prenant une moyenne pondérée des mesures empiriques [69].
Partant de ces deux méthodologies, il est clair qu'il faut d'abord savoir ce que l'on veut faire. Est-ce que nous cherchons à expliquer une variable cible à partir de nos données ou voulons-nous tout simplement voir si nos données expriment la même caractéristique ? Dans l'un ou l'autre des deux cas, ce qui est sûr est qu'il faut d'abord comprendre comment nos données corrèlent.
La Figure 2.2 La Figure suivante illustre les deux options pour construire un score composite. Nous avons les mesures (scores génétiques) qui sont utilisées pour produire un score composite. Deux méthodes sont possibles pour le créer: la construction réflexive et la construction formative. Si les mesures sont corrélées, on utilise la construction réflexive. Cette dernière utilise l'analyse en composante principale pour créer un score composite. Si les mesures ne sont pas corrélées, on utilise la construction formative. Elle utilise la méthode de pondération pour créer un score composite. Étant donné que les mesures sont sur des échelles différentes, il est nécessaire de les normaliser peu importe la méthode utilisée. Cependant, un score composite nécessite un jeu de données complet. Un score composite sur des données manquantes risque de biaiser les résultats. De ce fait, nous allons explorer l'imputation des données manquantes.

2.2.2. Données manquantes

Lorsque pour une observation donnée une ou plusieurs valeurs manquent, on parle de données manquantes (DM). Celles-ci affectent la plupart des bases de données y compris les bases de données génomiques. Dans notre étude, la base de données contenant les scores des gènes (scores génétiques) contient plusieurs données manquantes. En effet, pour un gène donné, un ou plusieurs scores peuvent être absents.
Étant donné que la plupart des modèles statistiques ne fonctionnent que sur des observations complètes des variables d'exposition et de résultat, il est nécessaire de traiter les données

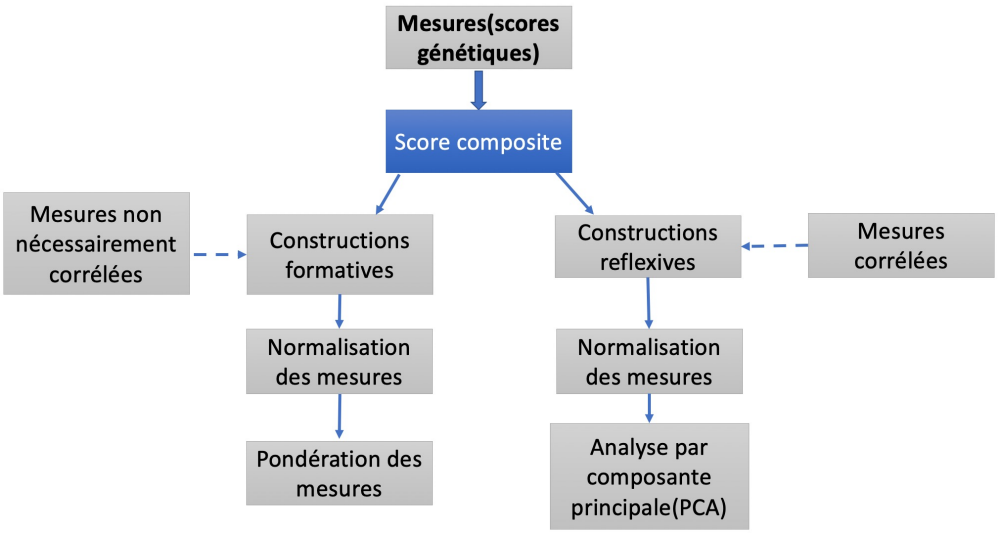


Fig. 2.2. Processus simplifié du score composite, adapté de [69].
Cette figure illustre les deux options pour construire un score composite.

manquantes, soit en supprimant les observations manquantes, soit en remplaçant les valeurs manquantes par une valeur donnée basée sur les autres informations disponibles. Ce processus est appelé imputation. Les deux méthodes peuvent biaiser les conclusions qui peuvent être tirées des données [42]. En effet, en présence d'un gros volume de données manquantes, le fait de supprimer les données manquantes peut biaiser les résultats car les analyses sont effectuées sur des observations incomplètes; dans le contexte où les données manquantes sont imputées, le biais est causé par le fait que la valeur qui remplace la donnée manquante n'est pas réelle mais prédite. Cependant, il est important de comprendre la nature des données manquantes afin de définir la méthode adéquate qui sera utilisée pour résoudre le problème [8].

Il faut alors étudier ces données en déterminant leur typologie, en essayant de les visualiser afin de les comprendre et ensuite voir quelles méthodes utiliser afin de trouver une solution quant aux valeurs manquantes.

2.2.2.1. Types de données manquantes

Afin de déterminer le type de données manquantes de notre étude, il faut se poser la question à savoir ce qui a conduit à leur absence. Nous avons trois types de données manquantes :

MCAR (Missing completely at random):

Nous sommes dans ce cas de figure si la probabilité que la valeur soit manquante est indépendante des valeurs prises par les autres variables de l'observation. Ils ne dépendent donc pas des paramètres et variables du système. En d'autres termes, les données manquantes constituent un sous-ensemble aléatoire de l'ensemble des données [8].

MAR (Missing at random) :

On parle de MAR si la probabilité d'absence est liée à une ou plusieurs autres variables du

jeu de données. Cette probabilité ne dépend pas de la valeur hypothétique qu'aurait pu prendre la valeur absente dans le cas où elle aurait été présente.

MNAR (Missing not at random):

Les données sont dites manquantes non au hasard MNAR si la probabilité que la valeur d'une donnée soit manquante dépend que des valeurs prises par cette donnée. Il n'existe aucun test pour vérifier si le processus qui génère les données manquantes est MNAR [8].

A présent que nous avons la démarche pour la typologie de nos données manquantes (DM), nous allons voir comment les gérer. Cependant, il est important de prendre en compte la structure de la base de données à imputer. Sommes-nous en présence d'un gros volume de données? Les variables sont-elles colinéaires?

La gestion des DM comme nous l'avons vu plus haut peut se faire soit en supprimant les données manquantes ou en les imputant. Pour la méthode de suppression des données, nous allons adopter celle appelée suppression par liste qui consiste à supprimer toute observation ayant une donnée manquante. Nous la verrons plus en détail dans les méthodes. En outre, nous avons aussi une méthode qui consiste à remplacer la donnée manquante par une valeur neutre (dans ce cas valeur neutre est zéro). C'est cette dernière qui est utilisée dans le aboratoire du Dr Jacquemont. À côté des méthodes de suppressions et de remplacement par zéro, nous avons les méthodes d'imputations.

2.2.2.2. Imputation de données manquantes

L'imputation des données manquantes consiste à remplacer les valeurs manquantes par d'autres valeurs estimées sur la base des autres informations disponibles. Dans notre cas, une donnée manquante va correspondre à un score qui n'est pas présent. Les données manquantes pour un gène peuvent être imputées en se basant sur les données présentes pour ce gène.

2.2.2.2.1. Méthodes paramétriques et non paramétriques

Pour les méthodes paramétriques, il est nécessaire d'évaluer la distribution des variables à partir des données disponibles. Cela va servir à imputer les données manquantes. Cependant, pour faire cette distribution, il faut prendre en compte un certain nombre de paramètres fixes. Les interprétations faites à partir d'une distribution n'étant pas toujours exactes, l'imputation avec les méthodes paramétriques peut introduire du biais. Une colinéarité peut empêcher une méthode paramétrique de bien imputer des données manquantes. À l'inverse, les méthodes non paramétriques ne sont pas régies par des lois de probabilités paramétriques et ne font donc pas de supposition sur la distribution des données [3]. Elles sont beaucoup moins contraignantes.

En nous basant sur la littérature notamment sur les travaux de Càtia M. Salgado et al.[42] et de Jabir. M [8], nous aborderons deux types de méthodes d'imputations: Les méthodes d'imputation simple et les méthodes élaborées d'analyse en présence des données manquantes [8].

2.2.2.2.2. Méthodes d'imputation

Méthodes à imputation simple

La méthode à imputation simple consiste à imputer une donnée manquante une seule fois ; une seule valeur lui est donc associée. La valeur imputée est considérée comme étant réelle, observée ; c'est-à-dire que la méthode ne prend pas en compte l'erreur qui se produit durant l'imputation. Les méthodes à imputation simple standard sont: L'imputation par la moyenne, l'imputation par la médiane, l'imputation par régression, etc.

Étant en présence d'un gros volume de données manquantes, ces méthodes ne sont pas adéquates pour les besoins de notre étude.

Méthodes élaborées d'analyse en présence des données manquantes

Dans la littérature [3,8,42,43,44], les méthodes populaires et qui ont fait l'objet de comparaison entre elles sont la méthode d'imputation multiple, la méthode du plus proche voisin et l'approche par forêts aléatoires.

La méthode à imputation multiple permet de trouver une valeur avec le moins de biais possible pour une donnée manquante. Plusieurs imputations sont effectuées dans cette méthode. La donnée manquante est imputée plusieurs fois et à la fin, les résultats sont combinés pour donner lieu à la donnée imputée finale (voir 2.3)[3,8]. Elle prend en compte l'erreur contrairement à l'imputation simple et réduit ainsi le biais à chaque imputation. La méthode d'imputation multiple est efficace pour les types de données MAR ou MNAR. Dans le cas d'un gros volume de données manquantes, cette méthode donne des estimations très satisfaisantes.

La méthode MICE (multiple imputation by chained equations) de Buuren et Oudshoorn [47] est une méthode d'imputation multiple paramétrique. C'est une méthode itérative, c'est-à-dire que MICE impute les données manquantes une nouvelle fois à chaque itération en se basant sur les résultats obtenus à l'itération précédente jusqu'à ce que le critère d'arrêt soit atteint [8].

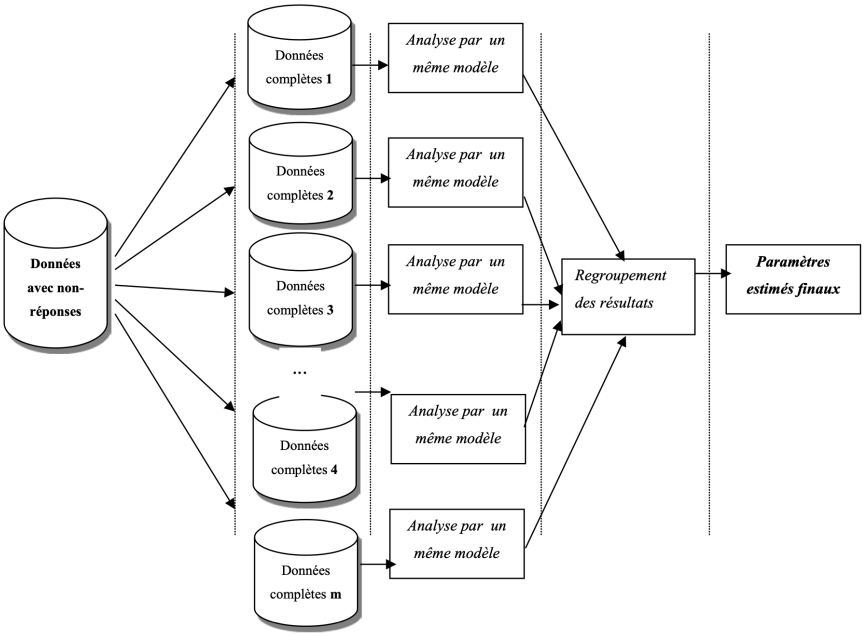


Fig. 2.3. Processus de l'imputation multiple. Issue de [8].
Cette figure illustre le processus de l'imputation multiple. Nous avons une base de données contenant des données manquantes. Pour une valeur manquante de la base de données initiale, plusieurs valeurs sont estimées. Une analyse utilisant le même modèle est effectuée sur ces valeurs et les résultats sont combinés pour donner la valeur imputée.

La méthode des K plus proches voisins KNN de Troyanskaya et al., 2001 [71] consiste à choisir une valeur de remplacement parmi les unités similaires qui coexistent dans la même base de données. Elle est non paramétrique. Ces valeurs de remplacement représentent la moyenne des K valeurs provenant des K observations présentes les plus similaires. K est un nombre entier. La similitude de deux observations est déterminée, après normalisation du jeu de données, à l'aide d'une fonction de distance [42]. Le principal avantage de l'algorithme KNN est qu'avec suffisamment de données, il est rapide et peut prédire avec une précision raisonnable.

Finalement nous avons l'approche basée sur les forêts aléatoires. Ces dernières sont une combinaison de prédicteurs d'arbres tels que chaque arbre dépend des valeurs d'un vecteur aléatoire échantillonné indépendamment et avec la même distribution pour tous les arbres de la forêt [41]. Elle est non paramétrique comme KNN. Un modèle de forêt aléatoire est construit pour chaque variable. Les prédictions de ce modèle vont permettre de remplacer les valeurs manquantes. L' approche appelée MissForest, introduite par Stekhoven et Bühlmann [72], utilise les arbres de décision (arbre de classification et de régression) qui sont des modèles de prédiction. Cette approche par forêt aléatoire fournit une estimation de l'erreur d'imputation.

Ces trois méthodes (MICE, MissForest et KNN) peuvent être modélisées à l'aide de l'outil informatique R. Des études de comparaison entre MICE et KNN ont trouvé que MICE est

la meilleure [42,43]. Cependant, des études de comparaison entre les trois approches MICE, KNN et MissForest [8,3] ont trouvé que MissForest est la meilleure approche sur le plan des erreurs d'imputation.
Ces méthodes d'imputations doivent être choisies minutieusement en prenant en compte le type de données et leur colinéarité, la quantité de données manquantes et le fait que certaines sont paramétriques et d'autres non. Dans le chapitre suivant, nous allons traiter de cette problématique puis soulever une hypothèse et définir nos objectifs.

Chapitre 3

Problématique, hypothèse et objectifs

Le laboratoire du Dr Jacquemont a développé des modèles pour estimer la taille des effets des CNV codants sur la cognition [36,37].
Les modèles publiés par ces premières études permettent d'estimer la taille des effets d'une CNV avec une précision proche de 78% [37]. Cependant ces modèles reposent uniquement sur un score de contrainte génétique (le pLI ou le LOEUF). Par ailleurs, de nombreuses CNV sont mal estimées par ces modèles.
La performance des modèles initiaux est acceptable. Cependant, ces modèles ne sont basés que sur un score de contrainte génétique (score d'haplo-insuffisance). Il est vraisemblable que des informations sur la fonction des gènes dans le développement et la maintenance des différents réseaux cérébraux est essentielle pour mieux comprendre et prédire l'effet des gènes et des CVN codants sur la cognition.

Hypothèse :
La combinaison des informations sur la distribution spatiale et temporelle des gènes dans le cerveau est une information critique qui permettrait de mieux comprendre et prédire l'effet des CNV codants sur la cognition.

Objectif global :
Développer de nouvelles annotations des gènes codants pour mieux comprendre et prédire l'impact (la taille de l'effet) des CNV sur la cognition.
Objectifs spécifiques :
1. Caractérisation et gestion des données manquantes pour 54 scores (scores d'expression spatial, temporel des gènes exprimés dans le cerveau, scores de contraintes génétiques qui quantifient la génétique fitness, scores d'évolution).
2. Sélection des scores et du jeu de données complet pour le score composite.
3. Analyse de regroupement afin d'identifier un ou des scores composites qui captent le

maximum d'information sur les scores sélectionnés dans l'objectif 2.
4. Estimation de la taille des effets des CNV sur la cognition par des scores composites.

Méthode :
Nous utiliserons des méthodes de regroupement et de corrélation canonique pour combiner tous les scores disponibles qui peuvent également mesurer la taille des effets des CNV sur la cognition. Pour ce faire, nous allons d'abord sélectionner un jeu de données complet à l'aide de la réplication de l'étude précédente [37] ensuite nous allons mettre en place le score composite.
Nous présentons plus en détail ces méthodes dans le chapitre suivant.

Chapitre 4

Données utilisées (matériels)

4.1. Populations étudiées

Le jeu de données utilisé dans notre étude comporte 41109 individus provenant de différentes cohortes de la population générale et de cohortes de cas d'autistes. Notez que pour chacune des cohortes nous n'avons gardé que les individus retenus suite aux contôles de qualité de chaque étude. Le détail des cohortes est le suivant:
- Imagen [109], cohorte européenne de la population générale, 1790 adolescents.
- Saguenay Youth Study (SYS) [110], cohorte de canadien-français de la population générale, 1893 individus dont 1032 enfants et 951 parents.
- Génération Scotland (G-Scot)[112], cohorte de la population générale, 14160 individus.
- CartaGene (CaG)[111], cohorte d'individus recrutés au Québec non apparentés de la population génarale [37], 5764 individus.
- Lothian Birth Cohort (LBC ou cohorte de naissance Lothian) [113], cohorte de la population générale, 554 individus.
- Simon Simplex Collection (SSC) [114], cohorte de cas d'autisme dont un seul enfant est atteint d'autisme (le proband) dans la famille alors que les parents et frères et soeurs sont non atteints, 10163 individus.
- MSSNG (consortium international)[115], cohorte de cas d'autisme dont moins un à cinq enfants atteints d'autisme dans la famille, 6785 individus.
Ces données nous sont fournis par l'équipe de Huguet et al. [36,37] et Douard et al. [38].

4.2. Phénotypes: QI et facteur g

Le QI (quotient intellectuel) est une mesure de capacité cognitive générale, normalisée selon l'âge, qui fournit une estimation de la façon dont une personne se classe par rapport à ses pairs du même âge. Quant au facteur g (g-factor en anglais), c'est une mesure indirecte de l'intelligence générale, obtenue en extrayant la première composante principale de

l'analyse par composantes principales (ACP) de différentes mesures cognitives standardisées [37]. Le QI dans cette étude fait référence au QI non verbal (NVIQ) et si ce dernier n'est pas disponible dans nos données, nous utilisons le facteur g s'il est disponible. Le QI et le facteur g sont normalisés en utilisant le score Z . En effet, les scores Z sont un type de scores standards dont la moyenne et l'écart-type de la distribution sont conventionnels et connus de ceux qui les utilisent, ce qui facilite leur interprétation [48]. Le score Z du QI a une moyenne de 100 et un écart type de 15 tandis que le score Z du facteur g d'une cohorte est calculé selon la moyenne et l'écart type de chaque cohorte. Les études précédentes ont montré que le score Z du QI et score Z du facteur g sont fortement corrélés [36,37], nous allons donc utiliser l'écart type du QI (15) pour calculer la perte de QI estimée.
Afin de déterminer le QI, différents tests ont été effectués sur les individus de ces cohortes. Pour les cohortes de la population générale, nous avons le test Wechsler Intelligence Scale Children (WISC-IV[94] , WISC-III[95]) et le test Moray house[96][97]. Pour les cohortes des malades, nous avons les tests suivants: Mullen scales of early learning[98], Leiter international performance scale [99, 100], Raven progressive matrices[101], Stanford-Binet intelligence scale[102], WISC-IV[103], WISC-V[104], Wechsler Abbreviated Scale of Intelligence (WASI-I; WASI-II)[105–106], Wechsler Preschool and Primary Scale of Intelligence (WPPSI-IV) [107], Leiter-reviewed et le Differential Ability Scales (DAS-II [108]).
Parmi les 41109 individus retenus par le contrôle qualité (Tableau 4.1), 24074 ont un score Z (score Z du QI ou score Z du facteur g) dénoté $ZScore\_IQ$. C'est sur ces derniers que nous allons étudier l'impact des CNV sur la cognition.

4.3. Détection des CNV

Toutes les données que nous utilisons sont issues des travaux de Huguet et al. et Douard et al. [36, 37, 38]. La détection et le filtrage des CNV y sont détaillés. Dans notre étude, nous ne détectons pas les CNV.
Pour les cohortes SSC et de la population générale (Imagen, LBC, CaG, G-Scot, SYS), le génotypage a été effectué, dans différents centres de génotypage, en utilisant l'approche SNP array. Les critères de contrôle de la qualité sont : un taux d'appel de 95%; un écart type du LRR <0,35; un écart type de la BAF <0,08 et la valeur absolue du facteur d'ondulation <0,05.
Sur les données génotypées, ils utilisent les outils PennCNV[118] et QuantiSNP[119] pour identifier les CNV. Ces outils, basés sur un modèle de Markov caché (Hidden Markov model), permettent de détecter avec une haute résolution les CNV en utilisant des données provenant du génotypage par SNP. Les CNV détectées par ces algorithmes sont ensuite fusionnées à l'aide de l'algorithme CNVISION (www.CNVision.org)[40].
Pour la cohorte MSSNG, les données de séquençage proviennent d'Illumina. Elles ont été

analysées en utilisant l'outil GATK (Broad institute Genome Analysis ToolKit)[120]. Pour la détection des CNV, ils ont utilisé la méthode d'alignement des lectures "read-depth" en se basant sur l'approche proposée par Trost et al. [121].
Des filtres supplémentaires ont été appliqués pour éliminer les faux positifs (Huguet et al. [36,37]). Il est à noter que du fait de la technologie utilisée dans le projet, plus les critères de filtres appliqués par Huguet et al [36,38], les résultats obtenus au cours de ce projet ne concernent que l'estimation des CNV de plus de 50Kb. Il faut noter également que seul les CNV autosomiques sont étudiées. En effet, le chromosome X est associé à plusieurs troubles neuro-développementaux et ceci pourrait biaiser les résultats de la quantification des effets des CNV sur la cognition. En outre, le chromosome X ayant un nombre de copies dépendant du sexe, il ne peut être étudié la même façon que les autosomes.

Cohorte	Technologie génotypage	N (contrôle qualité)	N femme	Age (SD) en mois	Phénotype cognitif	score Z(SD) (QI ou facteur g)
Cohortes de la population générale (N = 24251)						
Imagen	610Kq-660Wq	1790	898	173,4 (4,4)	QI	0,44 (0,98)
SYS (enfants)	610Kq,HOE-12V	1032	503	180,7 (22,7)	QI	0,29 (0,87)
SYS (parents)	610Kq, HOE-12V	951	323	591,6 (66,9)	facteur g	$p = 8{,}82e^{-3}$ (1)
LBC	610Kq	554	247	-	QI	0,05 (0,96)
CaG	GSA,Omni2,5, AXIOM	5764	1094	651,8 (91,2)	facteur g	-0,02 (1,03)
GS	610Kq	14160	8101	560,7 (179,9)	facteur g	$p = -2{,}53e^{-3}$ (0,99)
cohortes de malade (N = 16948)						
SSC	1MV1,1MV3, Omni2,5	10163	341	108,4(42,9)	QI	-0,55 (1,59)
MSSNG	WGS	6785	275	110,2 (52,7)	QI	-0,43 (1,58)

Tableau 4.1. Description des cohortes. SD= écart type. Saguenay Youth Study (SYS) ; Lothian Birth Cohort (LBC) ; Generation Scotland (GS); CartaGene(CaG);HOE-12V: HumanOmni-Express-12V. GSA: Global Screening Array; Omni2.5: HumanOmni2.5; WGS: Whole Genome Sequencing.
Pour chaque cohorte, sont présentés: la technologie de génotypage utilisée, le nombre d'individus après les filtres de contrôle qualité, le nombre de femmes, l'âge en mois, le phénotype cognitif de chaque cohorte et enfin le score Z du QI ou du facteur g. Dans toutes les cohortes sauf MSSNG, les technologies utilisées sont des méthodes d'hybridation de SNP sur les puces à ADN. Pour MSSNG, la méthode de séquençage du génome entier (séquençage pan-génomique) a été utilisée.

Chapitre 5

Gestion des données manquantes des scores génétiques

Notre approche consiste à mettre en place un score composite en utilisant les scores génétiques. Pour ce faire, nous avons besoin d'un jeu de données complet.
Notre jeu de données est un fichier de scores génétiques qui contient tous les gènes codants ou non codants qui ont au moins un score génétique. Ce fichier contient un gros volume de données manquantes (DM). Ces dernières peuvent induire un biais dans des analyses statistiques si une solution pour les gérer n'est pas mise en place. Nous allons utiliser l'environnement de développement R pour nos analyses statistiques afin de voir la distribution des DM dans notre jeu de données et comment nos scores corrèlent entre eux.

5.1. Distribution des DM et classification hiérarchique

Afin d'avoir un aperçu sur nos données et de voir la distribution des données manquantes, nous allons utiliser le package R **naniar** avec la fonction vis_miss. Cette distribution est étudiée sur tout le génome et pas uniquement sur les gènes contenus dans les CNV des cohortes.
Sur la Figure 5.1, en noir nous avons le pourcentage de données génétiques manquantes et en gris le pourcentage de données génétiques présentes. Nous constatons que presque la moitié de nos données sont manquantes. Il est donc impératif de trouver une solution pour éviter de biaiser nos résultats. À présent que nous avons une idée de la quantité de données manquantes que nous avons, nous allons déterminer le type de données dont nous disposons. Nous avons des données manquantes au niveau des scores de génétiques. Leur disponibilité dépend alors des données génomiques ou des mesures non trouvées par les chercheurs et non des scores en tant que tels. De ce fait, nous pouvons dire que nous sommes en présence du type de données MNAR(Missing not at random). La présence ou l'absence d'un score ne dépend pas d'un autre score.

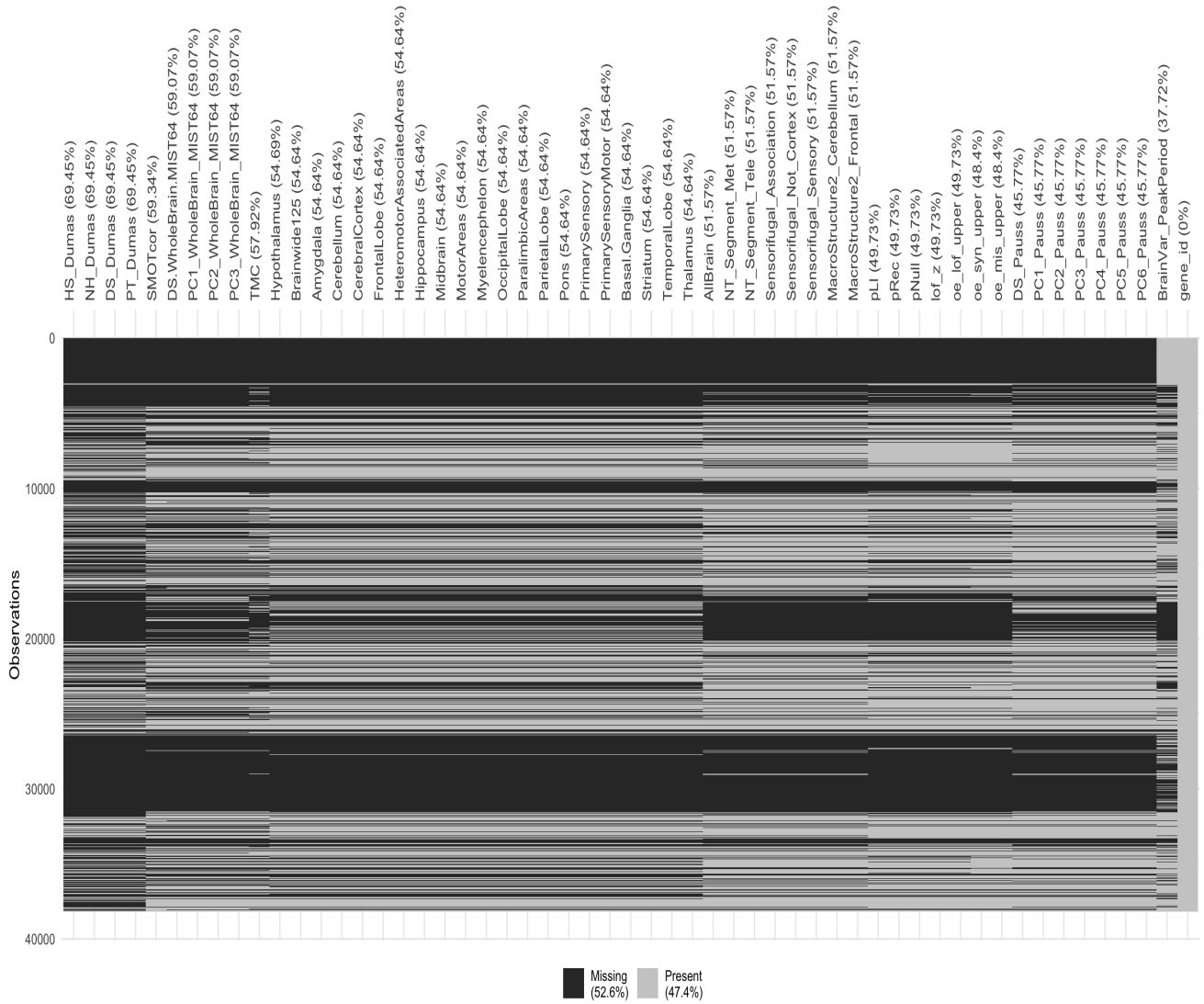


Fig. 5.1. Distribution des données manquantes dans notre jeu de données. En ordonnée nous avons tous les gènes codants et non codants du génome qui ont au moins un score génétique et en abscisse les scores de gènes. Nous avons 38184 observations (gènes codants et non codants) et 54 variables (scores génétiques). En noir nous avons les données manquantes et en gris les données présentes.

Nous allons aborder trois approches de gestion des données manquantes applicables au type de données MNAR.

5.2. Méthode de remplacement par une valeur neutre (zéro)

Cette méthode consiste à remplacer toutes les données manquantes au niveau des scores génétiques par une valeur neutre (zéro). C'est cette méthode qui est utilisée dans toutes les études effectuées dans le laboratoire du Dr Jacquemont.

Dans la suite, le jeu de données provenant de cette méthode, qui est le jeu de données initial avec données manquantes remplacées par zéro, est appelée "jeu de données suivant la méthode de remplacement par zéro". Il va nous permettre de répliquer l'étude [37].

5.3. Suppression des données manquantes

Une approche connue dans la littérature est de supprimer toutes les lignes contenant une donnée manquante. Cette approche peut induire des biais car elle nous fait perdre des données qui étaient disponibles. En effet, si une cellule contient une donnée manquante, on supprime toute la ligne même si les données de toutes les autres cellules de cette ligne sont disponibles. Avec cette approche, nous perdons beaucoup de gènes. En effet, nous passons de 38184 à 8112 observations avec toujours 54 variables. Nous appelons le jeu de données résultant de cette méthode "jeu de données supprimées".

5.4. Imputation de données manquantes: Étude comparative

Dans la section 2.2.2.2.2, nous avons vu que les études [42,43] ont comparé MICE et KNN et ont sélectionné MICE comme la meilleure méthode. Les études [3,8] quant à elles, ont effectué des analyses comparatives entre les algorithmes MICE, MissForest et KNN. Elles sélectionnent MissForest comme la meilleure méthode. Comme nous l'avons vu dans la revue de littérature, MICE est une méthode paramétrique tandis que MissForest et KNN sont des méthodes non paramétriques. Nous avons fait le choix de faire des tests sur une méthode paramétrique et une méthode non paramétrique. Pour la méthode KNN les valeurs qui remplacent les DM représentent la moyenne des K valeurs provenant des K observations présentes les plus similaires (K étant un nombre entier). Le choix d'un K optimal pouvant s'avérer difficile, nous avons donc décidé de garder MissForest plutôt que KNN. De plus dans la littérature [3,8,42,43], KNN n'est pas la meilleure méthode parmi les trois mais MissForest. Notre étude comparative se fera alors avec une méthode paramétrique (MICE) et une méthode non paramétrique (MissForest).

5.4.1. MICE et MISSFOREST

Deux approches sont utilisées: imputation multiple (MICE) et imputation basée sur les forêts aléatoires (MissForest). Pour mettre en place ces approches, nous allons utiliser les packages R de ces méthodes et qui portent les mêmes noms (MICE et MissForest). Nous allons voir le processus d'exécution des méthodes MICE et MissForest avant de passer à la méthodologie utilisée pour imputer nos données.
Selon [3], MICE s'exécute suivant 5 étapes:

1- Une première imputation temporaire est effectuée pour imputer les variables manquantes du système. En général c'est une imputation par la moyenne (imputation simple).
2- Les données précédemment imputées sont réinitialisées à « manquantes » pour une seule des variables du système.
3- Les données manquantes de cette variable sont ensuite imputées à nouveau grâce aux autres variables (présentes) du système qui servent de prédicteurs à son modèle d'imputation spécifique. Ces données nouvellement imputées seront utilisées lorsque cette variable servira de prédicteur aux autres variables du système.
4- Les étapes 2 et 3 sont répétées pour toutes les variables du système jusqu'à ce que chaque donnée manquante soit imputée. À cet instant, une itération est terminée.
5- Les étapes 2 jusqu'à 4 sont répétées le nombre de fois précisé par l'utilisateur et les données imputées sont mises à jour à chaque itération. Les prédicteurs sont donc de plus en plus précis et les paramètres responsables de l'imputation sont de plus en plus stables.

Comme nous l'avons vu dans la revue de littérature, MissForest, quant à elle, est basée sur une approche utilisant les forêts aléatoires et s'exécute suivant 5 étapes:
1- Pour imputer les données manquantes du système, une première imputation temporaire est effectuée. Le plus souvent, une imputation par la moyenne est utilisée.
2- Les valeurs imputées en 1 sont marquées avec le mot "à prédire" et les autres avec le mot "variable d'entrainement".
3- Les données "à prédire" sont ensuite imputées à nouveau grâce à un modèle forêt aléatoire qui utilise les "variables d'entrainement".
4- Pour toutes les variables, les étapes 2 et 3 sont répétées jusqu'à ce que chaque donnée "à prédire" soit imputée. Une itération est terminée à cet instant.
5- Les étapes 2 jusqu'à 4 sont répétées jusqu'à ce que la différence entre la matrice dernièrement imputée et la matrice imputée à l'itération précédente cesse de diminuer.
À présent que nous en savons plus sur le processus d'exécution de ces deux approches, nous allons passer à la méthode utilisée pour les comparer.

5.4.2. Méthode de comparaison

Nous avons un jeu de données de 38184 observations et 54 variables que nous allons imputer en utilisant les packages **MICE** et **MissForest** de R.
Les algorithmes d'imputation se basent sur les données présentes pour estimer les données manquantes. Sachant que nos scores expriment des mesures différentes, selon qu'ils soient des scores d'haplo-insuffisance, des scores de spécificité de l'expression cérébrale (localisation dans le cerveau) ou des scores temporels (expression prénatale ou postanale), l'imputation est faite sur la base des scores les plus corrélés. De ce fait, nous allons nous intéresser à la

similarité de nos scores et comment ils se regroupent. Toujours avec R, nous mettons en place une classification hiérarchique en utilisant la fonction hclust du package **stats** de R. En effet, la détection de groupes (grappes) d'objets étroitement liés est un problème important en bioinformatique et en exploration de données en général. La classification hiérarchique organise les objets en un dendrogramme dont les branches sont les regroupements souhaités. Le processus de détection de grappes est appelé coupe d'arbres, coupe de branches ou élagage de branches[46].
La figure 5.2 représente la classification hiérarchique effectuée sur nos données. Nous avons au préalable omis les données manquantes. Nous remarquons que nous avons cinq regroupements.

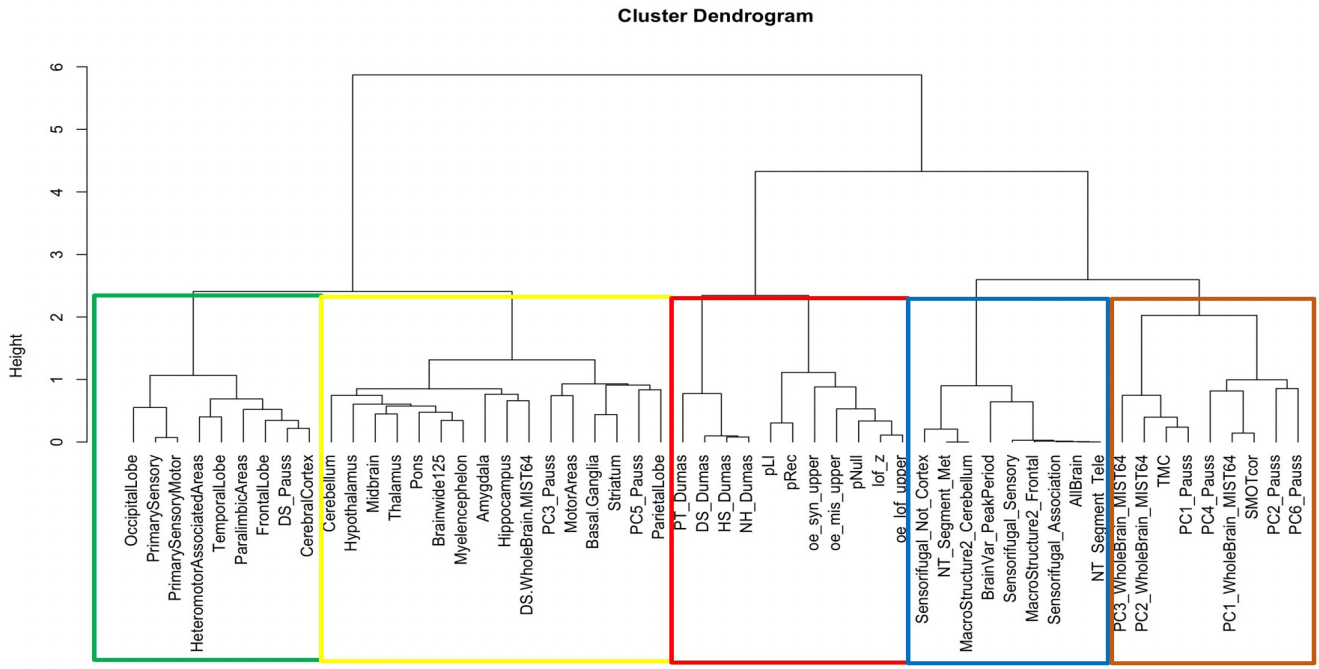


Fig. 5.2. Classification hiérarchique effectuée sur nos scores génétiques. Sur l'axe vertical, nous avons la dissemblance entre les regroupements et sur l'axe horizontal, nous avons les regroupements avec les scores. Nous avons utilisé des couleurs différentes pour mettre en évidence le nombre de regroupements que nous avons. Nous avons le regroupement 1 en vert, le regroupement 2 en jaune, le regroupement 3 en rouge, le regroupement 4 en bleu et le regroupement 5 en marron.

Nous séparons notre jeux de données de scores génétiques par regroupement et nous nous retrouvons avec 5 jeux de données. Nous avons respectivement 9, 16, 11, 9 et 9 scores dans les regroupements 1, 2, 3, 4 et 5. Notre méthodologie pour l'imputation consiste à prendre un jeu de données complet, d'y insérer des DM de façon aléatoire et d'imputer ce jeu de

données manquantes avec les deux méthodes d'imputation. Nous allons par la suite utiliser des métriques pour comparer les valeurs du jeu de données initial qui est complet et celles qui sont imputées. Ce processus est illustré dans la Figure 5.3.

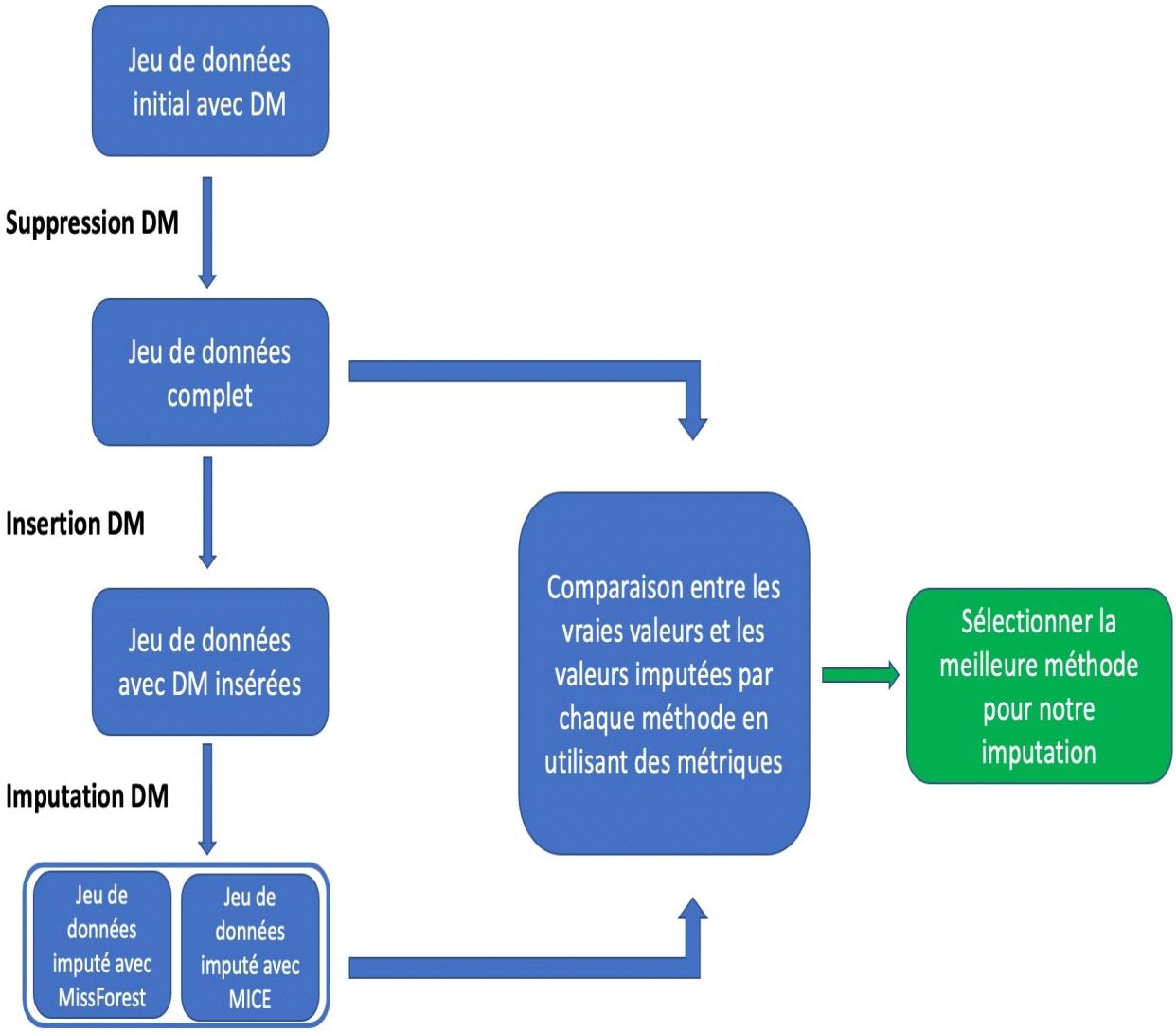


Fig. 5.3. Processus pour le choix de la meilleure méthode d'imputation.

Cependant, nous allons d'abord effectuer des tests avec un petit jeu de données pour bien valider notre approche, puis avec toutes les données de chaque regroupement. Les étapes suivantes sont appliquées sur chaque regroupement mais dans un soucis de clarté, allons les présenter seulement sur le regroupement 1.

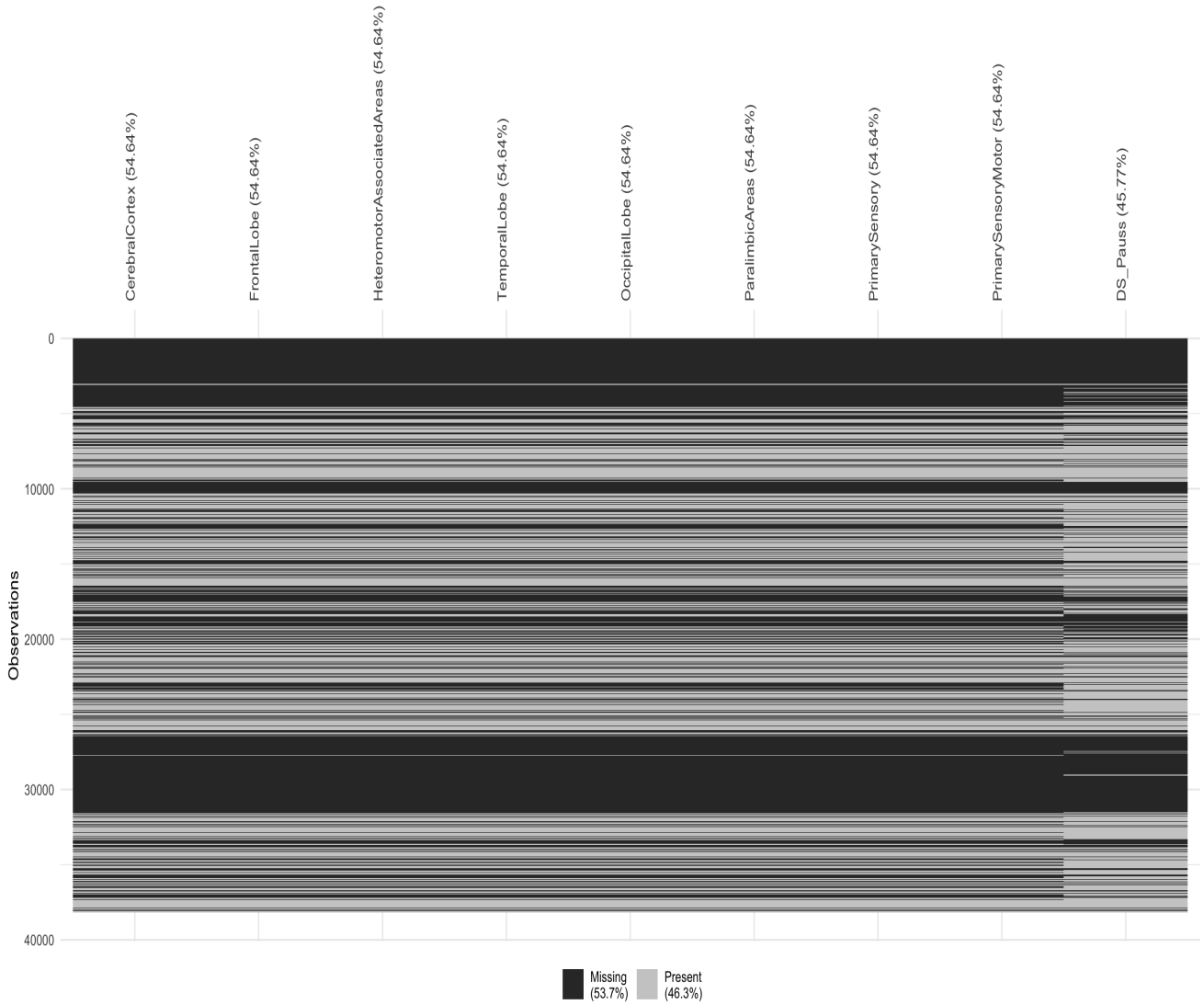


Fig. 5.4. Quantité des DM présente dans le regroupement 1. En ordonnée nous avons les gènes et en abscisse les scores de gènes. En noir nous avons les données manquantes et en gris les données présentes.

Nous prélevons de façon aléatoire 2000 gènes du regroupement 1. Après avoir retenu le pourcentage de données manquantes présent dans ces 2000 gènes, nous enlevons les données manquantes qui s'y trouvent afin d'avoir un jeu de données complet (sans données manquantes). Ce dernier est gardé en mémoire. À partir de ce jeu de données complet dénoté C, nous créons le jeu de données M, dans lequel nous remplaçons aléatoirement des données existantes par des valeurs manquantes. Le pourcentage final de données manquantes dans le jeu de données M est le même que celui du jeu de données initial (Figure 5.6). L'objectif de cette opération est de comparer les valeurs fournies par l'imputation pour les valeurs manquantes de M aux données réelles de C.

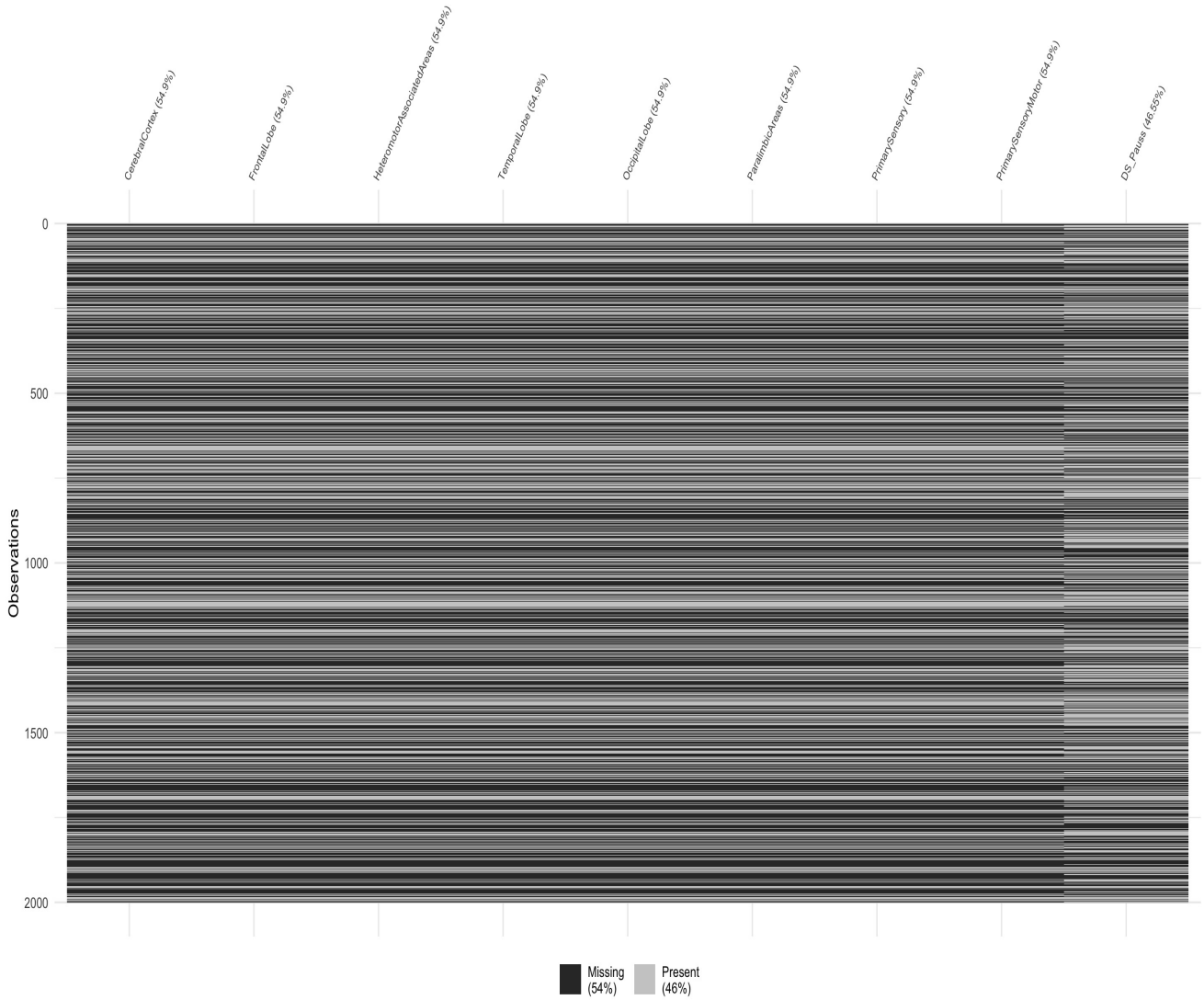


Fig. 5.5. Quantité des DM présente dans le regroupement 1 pour les 2000 gènes. En ordonnée nous avons les gènes et en abscisse les scores de gènes. En noir nous avons les données manquantes et en gris les données présentes.

Fig. 5.6. Processus de création d'un jeu de données pour la simulation de l'imputation. Nous sélectionnons de façon aléatoire 2000 gènes dans le regroupement 1. Dans ces 2000 gènes, nous avons 54% de DM. En supprimant ces dernières, nous obtenons le jeu de données complet C avec 837 gènes sans DM. Nous réinsérons aléatoirement 54% de DM dans ce jeu de données complet pour obtenir un nouveau jeu de données manquantes M.

Une fois que nous avons nos deux jeux de données, nous effectuons une imputation par MICE puis par MissForest sur le jeu de données manquantes M.
Après avoir consulté la littérature [3, 8, 42, 43] et effectué un certains nombre de tests avec différents paramètres, nous retenons les paramètres qui nous offrent de meilleurs résultats.
Pour MICE, nous utilisons sa méthode par défaut PMM (Predictive Mean Matching). Nous avons le paramètre m qui est le nombre d'imputations multiples et le paramètre maxit qui correspond au nombre maximal d'itérations pour chaque imputation multiple. Nous avons effectué des tests et changé les paramètres avec m=5, 10, 20, 25, 50, 100. Pour chaque m, nous testons avec maxit=5, 10, 20, 25, 50, 100. Par ces tests nous visons à trouver les paramètres optimaux. Nous trouvons que m=5 et maxit =20 donne de meilleurs résultats. Les détails de cette évaluation se trouvent dans l'annexe A.1.1.
Pour MissForest, nous avons toujours le paramètre maxit et nous avons aussi ntree qui est le nombre d'arbres générés aléatoirement à chaque itération. Dans la littérature[72], le ntree le plus utilisé est de 100. Nous avons fait des tests avec maxit=5, 10, 20, 25, 50, 100 et pour chaque valeur, nous utilisons ntree= 100, 150, 200, 300. Les résultats de ces tests sont présentés dans l'annexe A.1.2. Sur ces résultats, nous remarquons qu'à partir de 100, ntree est stable c'est-à-dire que pour une valeur de maxit donnée, si on change le ntree après 100, le résultat ne change plus. La combinaison ayant fournie de meilleurs résultats est maxit=10 et ntree=100.

Maintenant, nous avons notre jeu de données complet C, le jeu de données imputées par MICE et celui des données imputées par MissForest. Nous allons utiliser trois métriques pour évaluer la performace des deux modèles en effectuant une comparaison entre les données disponibles et celles imputées. Comme métriques nous avons le temps d'exécution de l'imputation, l'erreur quadratique moyenne RMSE (Root Mean Square Error) et l'erreur absolue moyenne MAE (Mean Absolute Error). Pour le temps d'exécution, nous fixons une horloge avec la fonction tictoc de R avant l'exécution de chaque fonction d'imputation et nous relevons le temps que prend chaque méthode pour s'exécuter.
RMSE et MAE permettent d'évaluer les performances d'un modèle. Ils mesurent la différence entre les vraies valeurs et les valeurs prédites par le modèle. Celle-ci est appelée erreur. Plus la valeur de ces deux métriques est proche de 0, meilleur est le modèle. MAE est la moyenne des différences absolues entre la valeur réelle et celle prédite par l'imputation. Quant à RMSE, c'est la racine carrée de la moyenne des différences au carré entre la valeur réelle et celle prédite par l'imputation. RMSE est très intéressant si on veut éviter les différences énormes entre les valeurs imputées et les valeurs réelles. En effet, il donne de grandes mesures à l'imputation qui contient d'énormes erreurs (une grande différence entre les vraies valeurs et les valeurs prédites).
Nous présentons les résultats de l'imputation effectuée sur le jeu de données manquantes M.

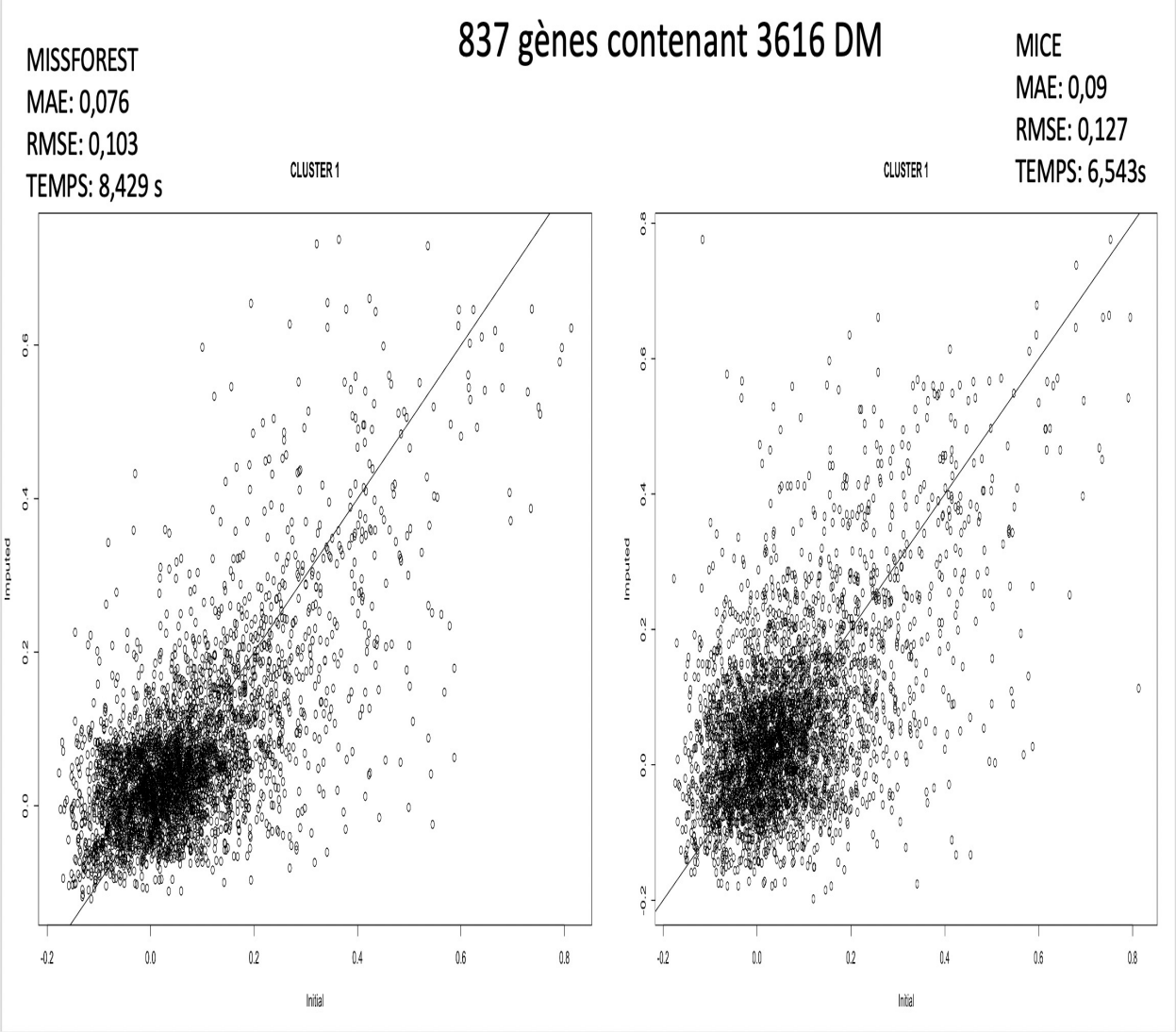


Fig. 5.7. Résultats de l'imputation pour le jeu de données manquantes M. Aperçu du regroupement 1. Ici nous avons tracé la droite d'équation y=aX + b afin de faire une comparaison entre les valeurs imputées et les vraies valeurs issues du jeu de données C. En abscisse nous avons les vraies valeurs (initiales) et en ordonnée nous avons les valeurs imputées. Nous avons pour chaque méthode, le résultats des métriques temps d'exécution, RMSE et MAE. Dans le regroupement 1, nous avons 837 gènes et 9 scores ce qui nous donne 7533 valeurs. Parmi elles, nous avons 3616 données manquantes.

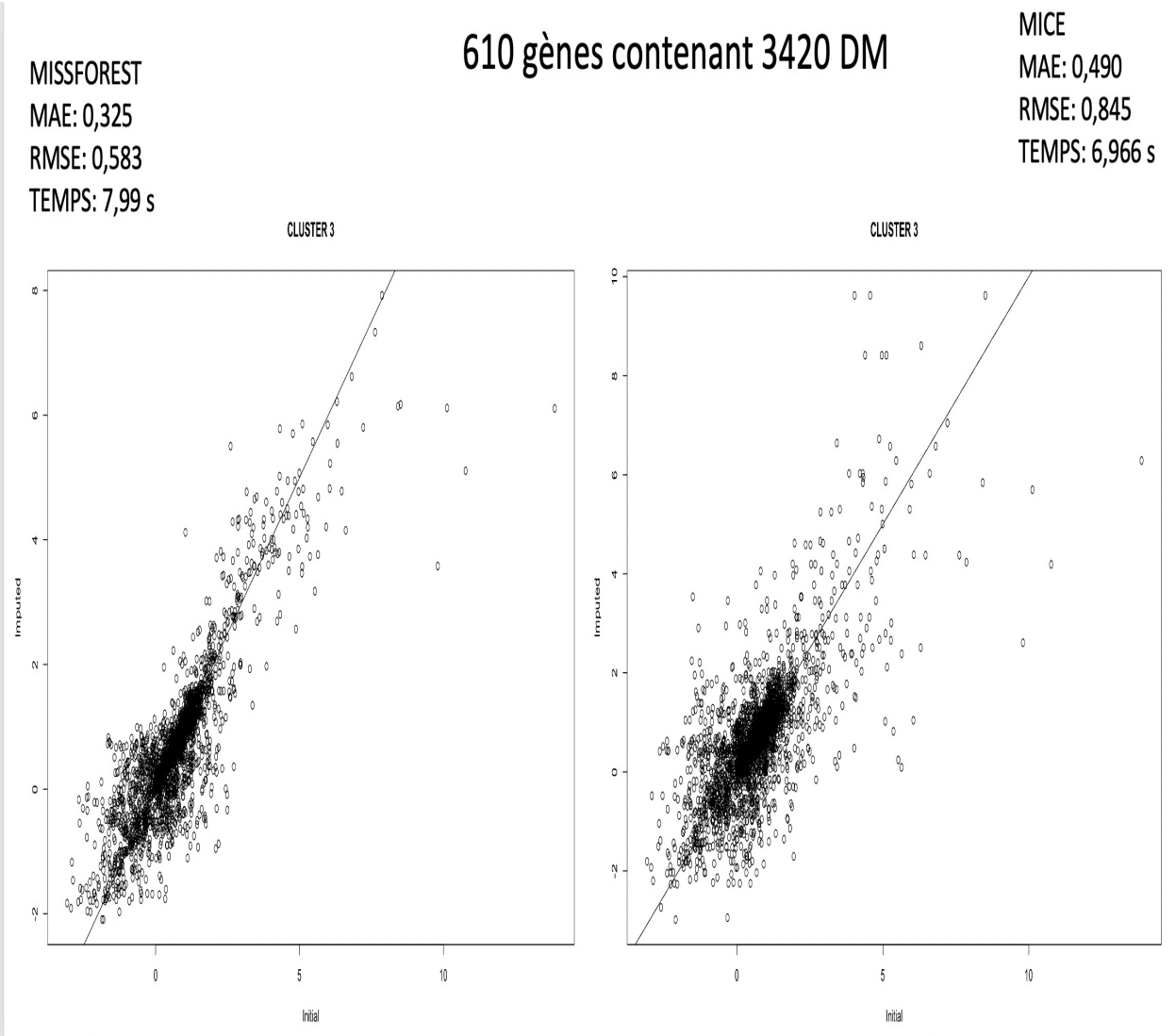


Fig. 5.8. Résultats de l'imputation pour le jeu de données manquantes M. Aperçu du regroupement 3. Ici nous avons tracé la droite d'équation y=aX + b afin de faire une comparaison entre les valeurs imputées et les vraies valeurs issues du jeu de données C. En abscisse nous avons les vraies valeurs (initiales) et en ordonnée nous avons les valeurs imputées. Nous avons pour chaque méthode, le résultats des métriques temps d'exécution, RMSE et MAE. Dans le regroupement 3, nous avons 610 gènes et 11 scores ce qui nous donne 6710 valeurs. Parmi elles, nous avons 3420 données manquantes.

Les Figures 5.7 et 5.8 présentent les résultats de cette imputation avec les vraies (initiales) valeurs en abscisse et les valeurs imputées en ordonné. Plus les points sont proches de la droite, plus l'imputation est meilleure car si on a une vraie valeur qui se confond avec une valeur imputée, ça signifie que la valeur imputée a été bien prédite. Nous remarquons que MissForest impute mieux. Nous avons aussi les métriques RMSE et MAE qui sélectionnent aussi MissForest comme étant la meilleure méthode. Cependant le temps d'exécution de

MICE est meilleur. Les résultats des autres regroupements se trouvent dans les annexes A.3 et A.4.
La Figure 5.9 nous montre les résultats des métriques sur tous les regroupements. Ils peuvent être arrondis par rapport aux résultats fournis au niveau des Figures (5.7 et 5.8 et au niveau des annexes A.3, A.4). Nous remarquons toujours que MissForest est meilleur pour les métriques RMSE et MAE. Nous rappelons que plus la valeur de ces deux métriques est proche de 0, meilleure est l'imputation. À part le regroupement 5, MICE présente un meilleur temps d'exécution mais la différence avec le temps d'exécution de MissForest est minime.

Clusters-Méthodes / Métriques	Cluster 1 (837 gènes/ 3616 DM)		Cluster 2 (731 gènes/ 5835 DM)		Cluster 3 (610 gènes/ 3420 DM)		Cluster 4 (799 gènes/ 3232 DM)		Cluster 5 (754 gènes/ 3192 DM)	
	MICE	MissForest	MICE	MissForest	MICE	MissForest	MICE	MissForest	MICE	MissForest
RMSE	0.128	0.103	0.260	0.198	0.846	0.583	1.605	1.265	2.75	2.06
MAE	0.095	0.076	0.194	0.148	0.490	0.326	0.458	0.381	1.17	0.881
Temps d'exécution (s)	6.54	8.42	16.67	14.65	6.9	7.9	6.18	6.64	6.53	5.21

Fig. 5.9. Résultats de l'imputation pour le jeu de données manquantes M. Pour chaque regroupement nous avons le nombre de gènes dans le jeu de données M et le nombre de données manquantes qu'il contient. Par exemple pour le regroupement 1, nous avons 837 gènes et puisqu'il y'a 9 scores dans le regroupement 1, nous avons 7533 valeurs. Parmi elles, 3616 données manquantes; pour le regroupement 2, nous avons 731 gènes et puisqu'il y'a 16 scores dans le regroupement 2, nous avons 11696 valeurs dont 5835 sont manquantes.
Pour chaque méthode d'imputation, nous avons les résultats des métriques RMSE, MAE et temps d'exécution.

Ce test sur les 2000 gènes sélectionnés de façon aléatoire nous montre que pour un petit jeu de données, à part le temps d'exécution, MissForest est la meilleure méthode pour nos données. Nous allons à présent effectuer la même opération avec toutes les données de chaque regroupement. Comme précédemment, nous enlevons toutes les données manquantes pour obtenir un jeu de données complet dénoté C'; nous insérons aléatoirement la même quantité de données manquantes pour obtenir un jeu de données manquantes dénoté M'; Après imputation avec MICE et MissForest, nous utilisons les métriques RMSE, MAE et le temps d'exécution pour mesurer les performances de chaque modèle. Le but de cette opération est de voir si les résultats sur la performance sont différents d'une méthode à une autre selon le fait qu'on impute un petit ou un gros volume de données.

Clusters-Méthodes / Métriques	Cluster 1 (16098 gènes/ 69160 DM)		Cluster 2 (13930 gènes/ 112410 DM)		Cluster 3 (11590 gènes/ 65720 DM)		Cluster 4 (15371 gènes/ 61488 DM)		Cluster 5 (14405 gènes/ 61192 DM)	
	MICE	MissForest	MICE	MissForest	MICE	MissForest	MICE	MissForest	MICE	MissForest
RMSE	0.127	0.099	0.261	0.203	0.811	0.542	1.676	1.224	2.7	2.022
MAE	0.094	0.073	0.193	0.148	0.477	0.283	0.474	0.361	1.168	0.848
Temps d'exécution (s)	32.34	794.16	65,13	1216.48	32.89	420.1	33.1	988.8	25.6	654.9

Fig. 5.10. Résultats de l'imputation pour le jeu de données manquantes M'. Pour chaque regroupement nous avons le nombre de gènes dans le jeu de données manquantes M' et le nombre de données manquantes qu'il contient. Par exemple pour le regroupement 1, nous avons 16098 gènes et puisqu'il y'a 9 scores dans le regroupement 1, nous avons 144882 valeurs dont 69160 données manquantes.
Pour chaque méthode d'imputation, nous avons les résultats des métriques RMSE, MAE et temps d'exécution.

Dans la Figure 5.10, voyons que MissForest est la meilleure méthode d'imputation pour nos données, selon RMSE et MAE, que ce soit avec un petit jeu de données sélectionné de façon aléatoire ou en prenant toutes les données de chaque regroupement. Certes son temps d'exécution est plus long et varie entre 7 et 20 minutes selon le regroupement mais c'est acceptable pour nos données. De ce fait, nous retenons MissForest et nous procédons à l'imputation de nos données manquantes par regroupement avec MissForest. La fonction **rbind** de R nous permet de regrouper nos 5 regroupements imputés en un seul jeu de données. Ce dernier est appelé dans la suite de ce mémoire "le jeu de données imputées".
Après ces trois approches de gestion de nos données manquantes, nous nous retrouvons avec trois jeux de données provenant d'un seul: Le jeu de données suivant la méthode de remplacement par zéro, contenant des DM, avec 38184 gènes; le jeu de données supprimées avec 8112 gènes; le jeu de données imputées avec 38184 gènes. Nous avons vu que ces jeux de données de scores de gènes renferment 54 variables (scores). Nous allons effectuer la sélection de nos scores génétiques pour mettre en place le score composite.

Chapitre 6

Score composite

6.1. Sélection des scores pertinents pour le score composite

Nous avons vu que les scores représentent des mesures d'expression des gènes dans plusieurs régions cérébrales; d'expressions temporelles; de contrainte génomique; etc. Cependant, il peut arriver que plusieurs scores soient fortement corrélés. Parmi nos 54 scores, nous allons juste retenir les scores principaux car bon nombre d'entre eux représentent la même mesure. Pour ce faire nous allons d'abord présenter nos scores groupés selon leur provenance. Le tableau 6.1 présente ces scores. Pour les scores d'expression corticale de l'article de T. Pauss et al. [49], nous gardons les scores $PC1\_Pauss$, $PC2\_Pauss$, $PC3\_Pauss$). Selon [54], ces scores sont les composantes principales qui expliquent le plus la variance de l'expression spatiale des gènes dans le cortex cérébral. Le score TMC issue de l'article de Burt et al. [50] est le seul donc nous allons le garder. Quant aux scores d'expression des gènes à travers tout le cerveau ($PC1\_WholeBrain\_MIST64$, $PC2\_WholeBrain\_MIST64$, $PC3\_WholeBrain\_MIST64$) de l'étude [53], nous ne gardons que les 3 premières composantes expliquant le plus la variance de l'expression spatiale des gènes au niveau cérébral. Nous gardons également le DS_Pauss [49] et le DS.WholeBrain.MIST64 [53] car ce sont deux scores de stabilité différentielle par tissu provenant de deux études différentes.

Pour le choix des autres scores, nous allons effectuer une corrélation de Pearson [63] afin de déterminer les scores fortement corrélés qui expriment la même mesure et retenir que les scores principaux. En effet, la corrélation de Pearson permet de déterminer la corrélation linéaire entre deux variables. Elle nous fournit un aperçu sur la relation entre deux variables. Nous allons la mettre en place en utilisant la fonction corrplot de R. Nous l'appliquons à notre jeux de données initial (le jeu de données qu'on avait avant l'application des méthodes de gestion des données manquantes) après avoir enlevé les données manquantes (avec la

Scores	Description
pLI, pRec, pNull, *oe_syn_upper*, *oe_mis_upper*, *lof_z*, LOEUF	scores de contrainte en fonction des types de variants (scores d'haplo-insuffisance) [https://gnomad.broadinstitute.org/].
*PC*1_*Pauss*, *PC*2_*Pauss*, *PC*3_*Pauss*, *PC*4_*Pauss*, *PC*5_*Pauss*, *PC*6_*Pauss*	Scores d'expression génique au niveau du cortex de T. Pauss [49]
TMC: T1w/T2wMapCorrelation	Scores d'expression génique au niveau du cortex Burt 2018 [50]
Brainwide125 [54], Amygdala [54], Cerebellum [54], CerebralCortex [54], FrontalLobe [54], HeteromotorAssociatedAreas [54], Hippocampus [54], Hypothalamus [54], Midbrain [54], MotorAreas [54], Myelencephelon [54], OccipitalLobe [54], ParalimbicAreas [54], ParietalLobe [54], Pons [54], PrimarySensory [54], PrimarySensoryMotor [54], Basal.Ganglia [54], Striatum [54], TemporalLobe [54], Thalamus [54],*DS_Pauss* [49], DS.WholeBrain.MIST64 [53]	Scores de stabilité différentielle par tissus
AllBrain, *NT_Segment_Met*, *NT_Segment_Tele*, *Sensorifugal_Association*, *Sensorifugal_Not_Cortex*, *Sensorifugal_Sensory*, *MacroStructure*2_*Cerebellum*, *MacroStructure*2_*Frontal*, *BrainVar_PeakPeriod*	scores d'expression temporelle [51]
HS_Dumas, *NH_Dumas* ,*DS_Dumas*, *PT_Dumas*	Scores d'évolution [52]
*PC*1_*WholeBrain_MIST*64, *PC*2_*WholeBrain_MIST*64, *PC*3_*WholeBrain_MIST*64	Scores de l'expression spatiale des gènes au niveau cérébral [53]

Tableau 6.1. Les différents scores de notre jeu de données classés selon leur provenance et le fait qu'ils expriment la même mesure.

fonction **na.omit** de R) car cette approche nécessite un jeu de données complet.
La corrélation de Pearson produit un coefficient de corrélation qui est compris entre -1 et 1. Les valeurs négatives signifient que les deux scores sont négativement corrélés et les valeurs positives signifient que les deux scores sont positivement corrélés. Si la valeur absolue du coefficient est proche de 1, nous avons donc une forte relation linéaire entre les deux scores; inversement si le coefficient est égal à 0, nous n'avons pas de relation linéaire entre les deux scores. Les scores qui ont une forte relation linéaire (une forte corrélation) expriment la même mesure. Nous allons alors nous baser sur ces résultats, sur la littérature [49, 50, 51, 52, 53, 54] et sur des travaux du laboratoire du Dr Jacquemont non publiés, pour sélectionner nos scores.

La Figure 6.1 nous montre les scores d'expression temporelle qui sont très fortement corrélés. Ceci s'explique par le fait que tous ces scores expriment une mesure du cerveau. AllBrain représente la pente à la courbe de l'expression temporelle dans le cerveau entier (est-ce que l'expression augmente au cours du temps, diminue, ou reste stable?) alors que les autres scores mesurent chacun l'expression temporelle des gènes dans une zone spécifique du cerveau. De ce fait, nous retenons le score AllBrain.

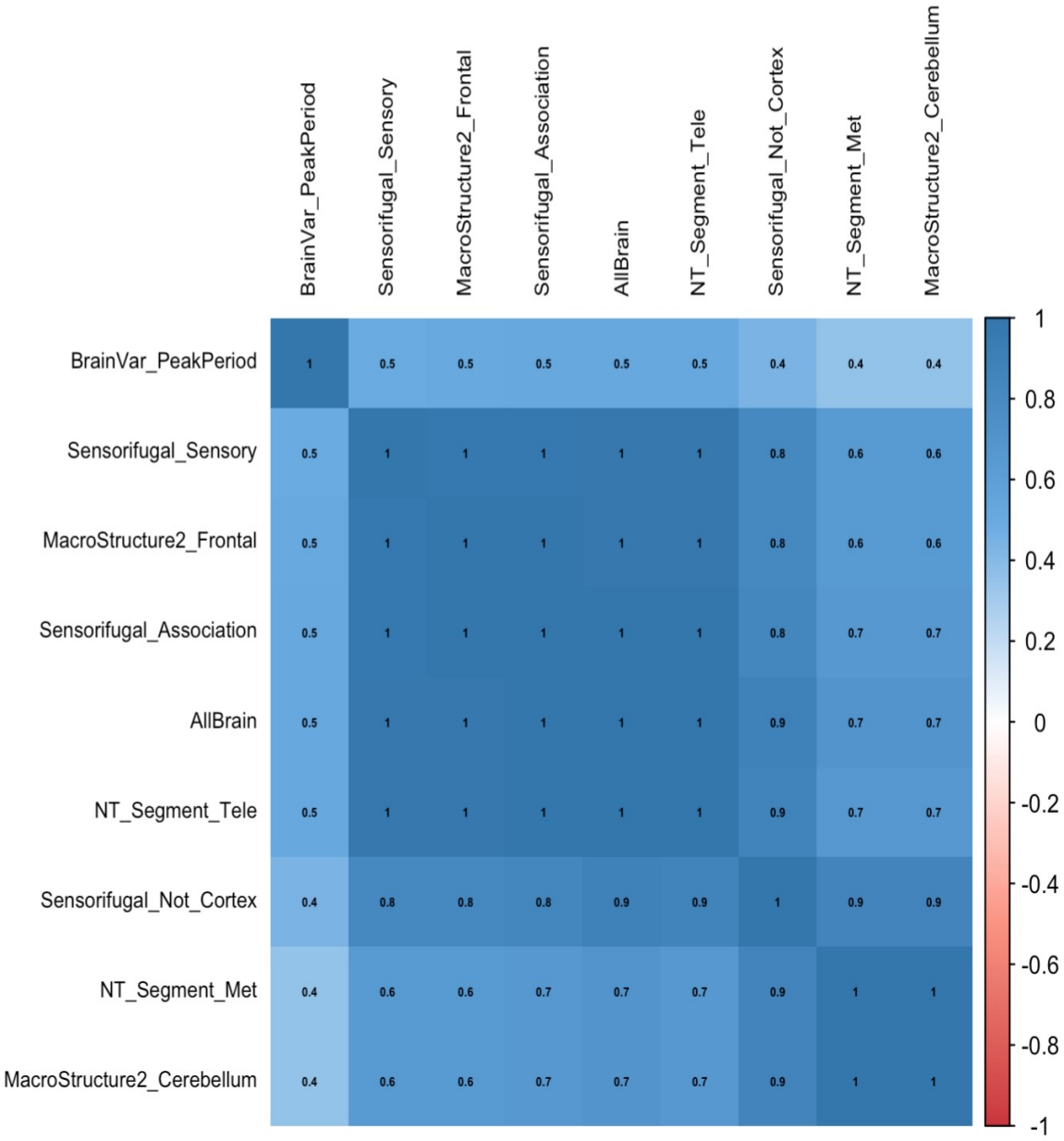


Fig. 6.1. Matrices de corrélation de Pearson pour la sélection de scores d'expression temporelle. En rouge nous avons les corrélations négatives et en bleu celles positives. Plus la valeur est proche de 1, plus ces scores sont corrélés.

Sur la Figure 6.2, nous avons les scores de stabilité différentielle. Nous avons le score brainwide125 qui est une mesure effectuée sur tout le cerveau alors que le reste des scores de cette figure expriment des mesures sur des sous structures cérébrales. Nous retenons alors le brainwide125.

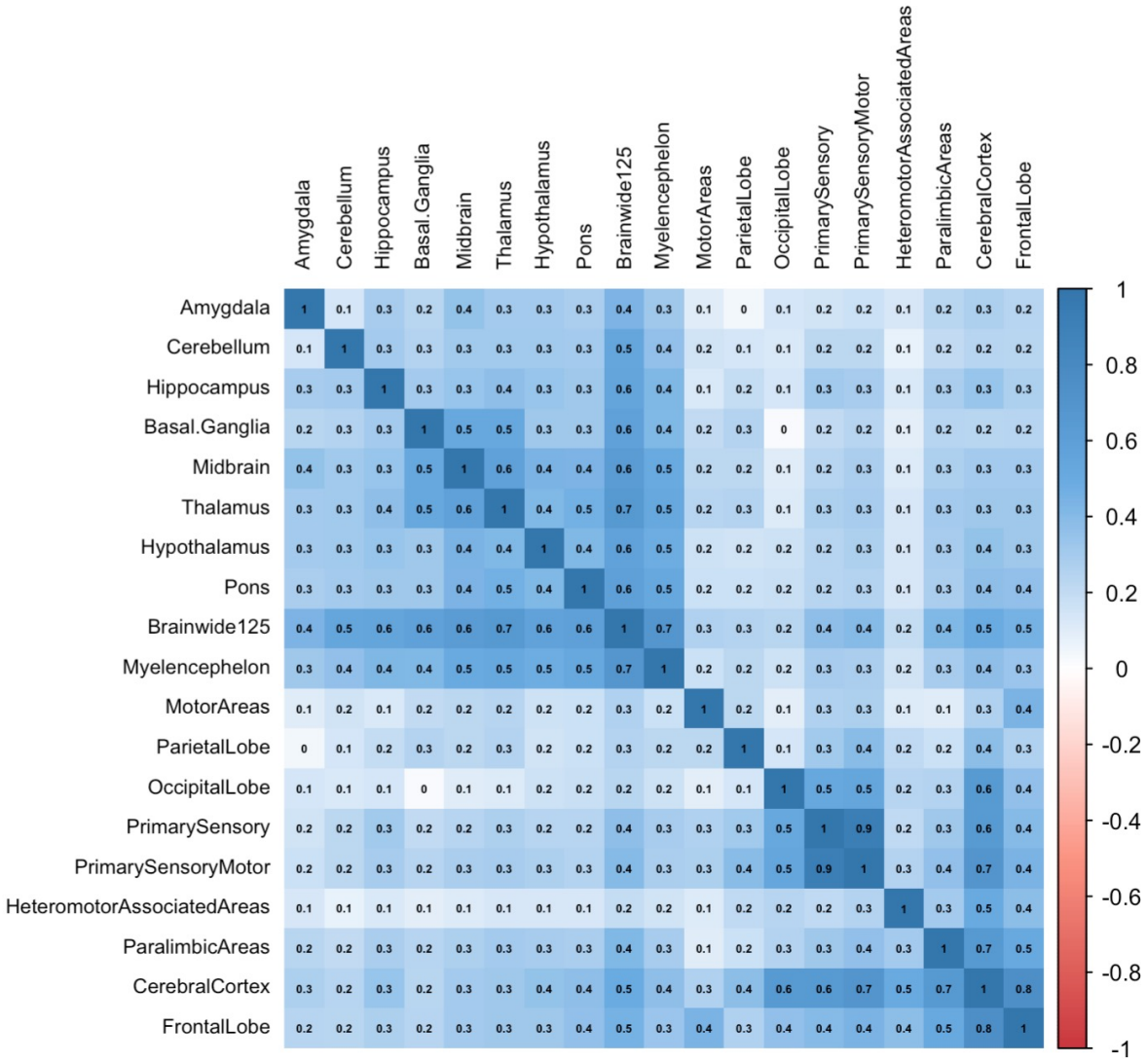


Fig. 6.2. Matrices de corrélation de Pearson pour la sélection de scores de stabilité différentielle. En rouge nous avons les corrélations négatives et en bleu celles positives. Plus la valeur est proche de 1, plus ces scores sont corrélés.

La figure 6.3 nous montre les scores de contrainte génétique en fonction des types de variants missense (*oe_mis_upper*), synonyme (*oe_syn_upper*), ou perte de fonction (*lof_z*, LOEUF, pLI, pRec, pNull). Nous retenons les scores des variants missense et synonyme. Pour les scores des variants de perte de fonction, le score pLI et le score LOEUF sont fortement corrélés aux autres. Cependant, ces deux scores expriment la même mesure; la seule différence est que le score pLI est plus binaire et moins évolutif que le score LOEUF. Nous gardons alors le score LOEUF.

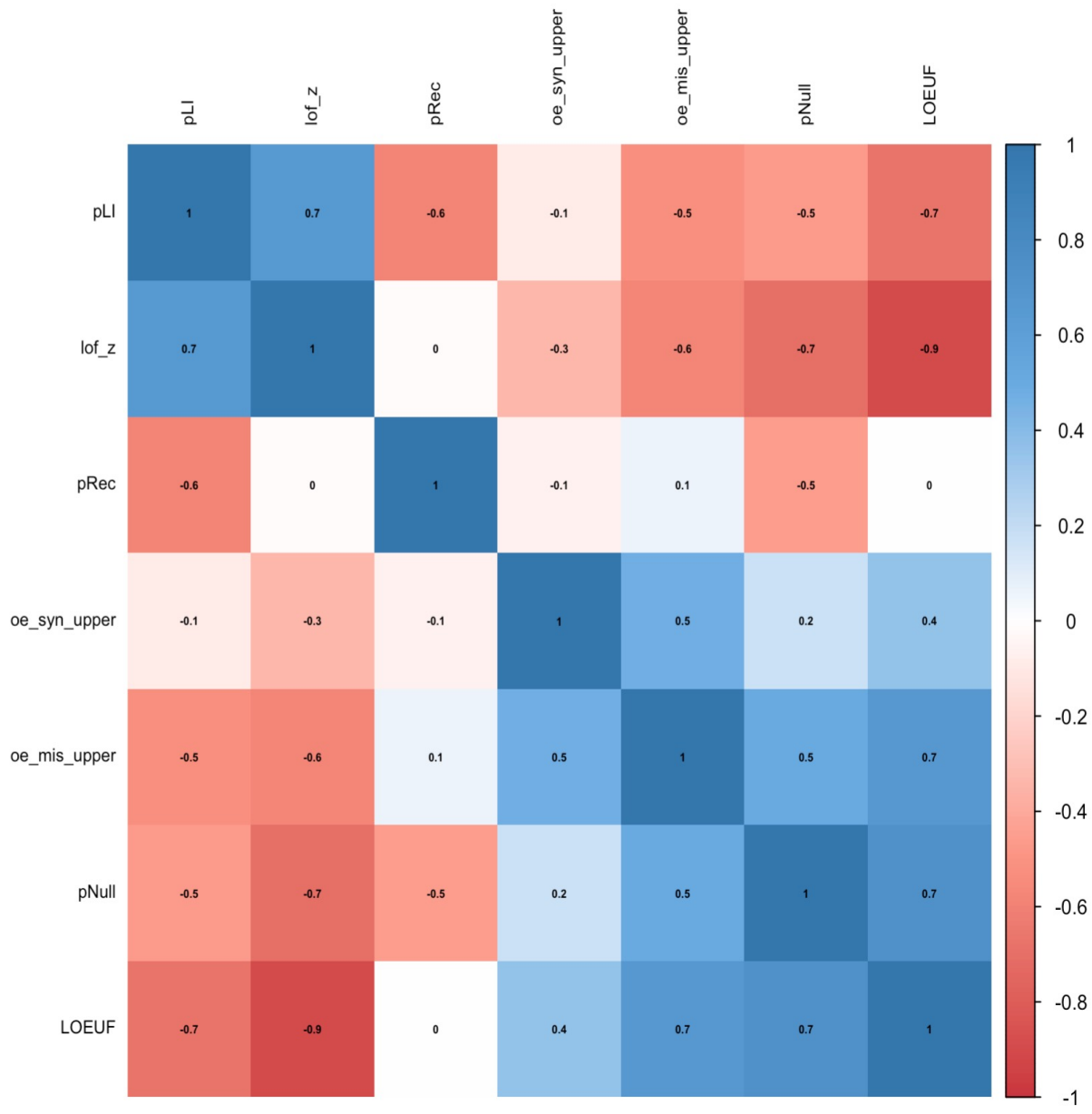


Fig. 6.3. Matrices de corrélation de Pearson pour la sélection de scores de contrainte génétique. En rouge nous avons les corrélations négatives et en bleu celles positives. Plus la valeur est proche de 1, plus ces scores sont corrélés.

Sur le Figure 6.4, nous avons les scores de spécificité d'expression chez Homo Sapiens (HS_Dumas), Néanderthal (NH_Dumas) et Denisovans (DS_Dumas) qui sont très fortement corrélés. Cependant nous retenons le HS_Dumas (le gène est-il très conservé chez l'homo sapiens ou non).

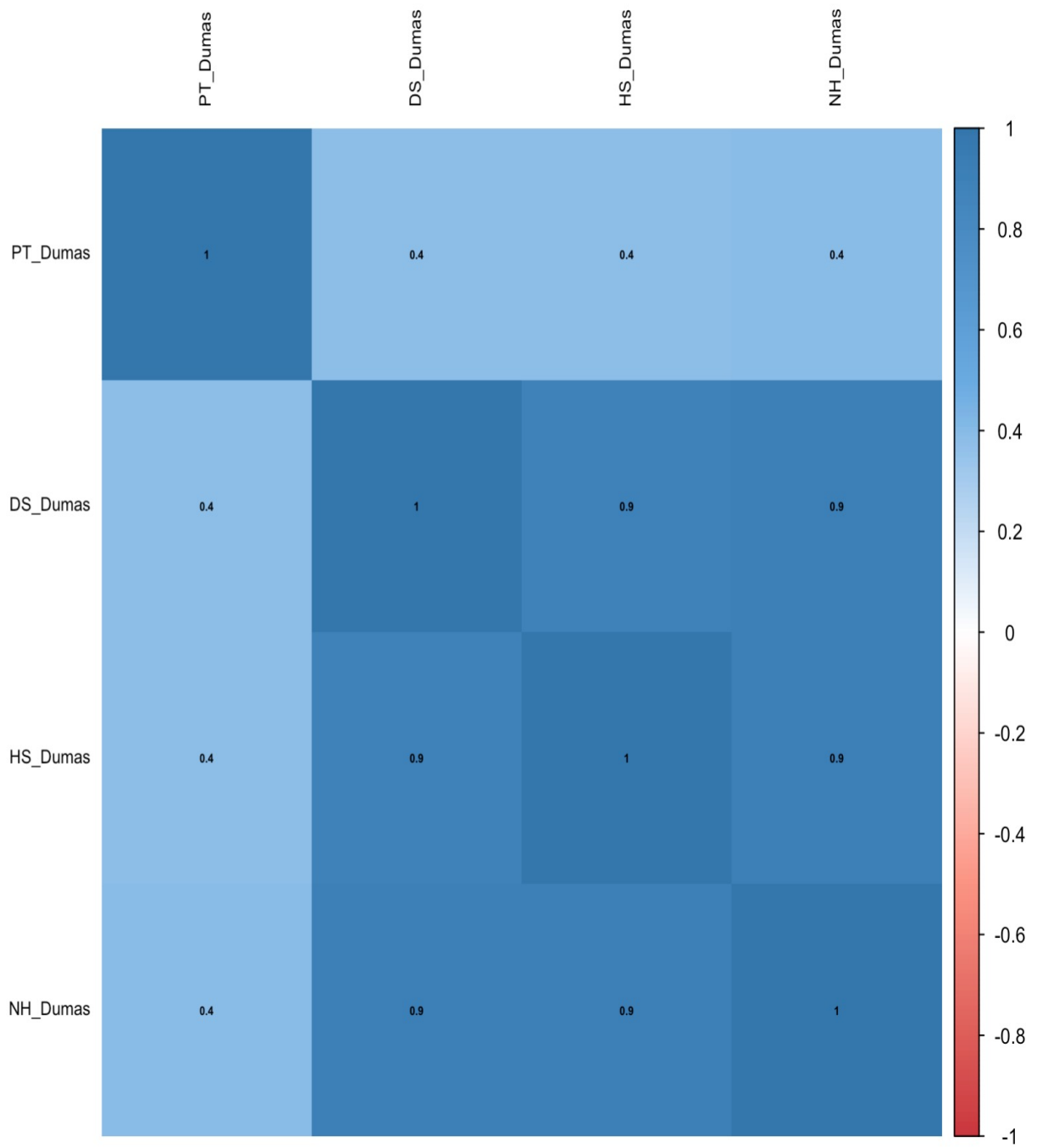


Fig. 6.4. Matrices de corrélation de Pearson pour la sélection de scores d'évolution. En rouge nous avons les corrélations négatives et en bleu celles positives. Plus la valeur est proche de 1, plus ces scores sont corrélés.

Après la sélection des scores génétiques pour la mise en place du score composite, nous nous retrouvons avec 15 scores de gènes dans chaque jeu de données. La Figure 6.5 montre la nouvelle classification sur ces 15 scores. Nous obtenons trois regroupements.

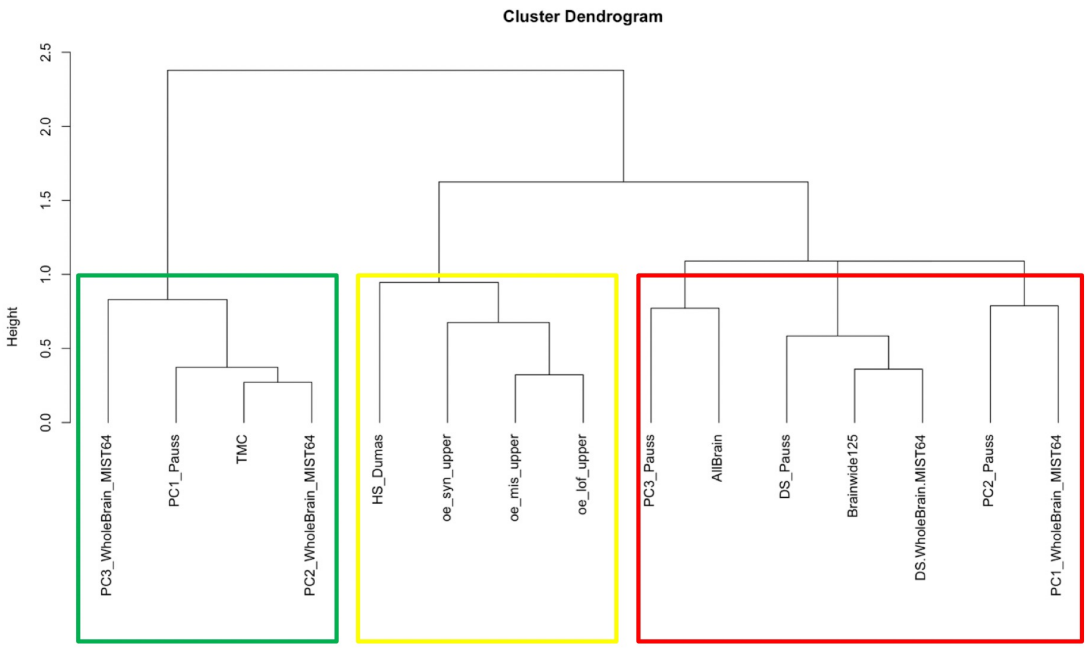


Fig. 6.5. Classification hiérarchique des scores sélectionnés. Nous avons 3 regroupements: En vert le regroupement 1 avec 4 scores, en jaune le regroupement 2 avec 4 scores et en rouge le regroupement 3 avec 7 scores.

6.2. Sélection du jeu de données pour le score composite

Nous avons vu que le score composite nécessite un jeu de données complet. De ce fait, nous avons établi une approche de gestion des données manquantes. Nous nous retrouvons avec deux jeux de données complets (un jeu de données supprimées et un jeu de données imputées) et un jeu de données avec données manquantes (remplacées par une valeur neutre (zéro)). Ce dernier va nous permettre de répliquer l'étude effectuée par Huguet et al. [37] dans le but de comparer les résultats avec ceux de nos deux jeux de données complets. L'étude effectuée avec la méthode de remplacement par zéro est alors la référence qui va nous permettre de savoir si c'est une bonne idée de supprimer ou d'imputer les données et si oui, lequel des deux est le meilleur modèle. Le pLI étant utilisé dans l'étude [37], nous allons l'intégrer dans nos scores pour cette phase de comparaison. Nous passons alors de 15 scores à 16 scores. Il est important de noter que nous avons 18 individus de moins que l'étude [37] et que la carte du génome a été mise à jour depuis les publications.

Les sections 6.2.1 et 6.2.2 sont appliquées à chacun de nos trois jeux de données. Pour plus de clarté, nous allons juste parler de jeux de données. Avant de procéder à l'annotation décrite dans la section 6.2.1, nous transformons le score de la stabilité différentielle par tissus, le score d'expression temporelle, les scores de l'expression spatiale des gènes au niveau cérébral, le score d'évolution de Dumas et les scores d'expression du cortex de Pauss et al. [49] (voir

Tableau 6.1) dans le but de séparer les scores positifs des scores négatifs. En effet, ces scores ont des valeurs positives ou négatives en fonction de ce qu'ils représentent. Par exemple un score du cerveau peut avoir des valeurs positives et négatives dépendant du fait qu'il soit un score "d'association au cortex préfrontal" ou qu'il soit "sensory-moteur". Le signe est pour leur distinction ce qui fait que nous n'allons garder que leur valeur absolue et les renommer en précédant chaque nom de score par Abs et en le faisant suivre de pos ou neg dépendant du signe qu'il avait avant la transformation. Nous effectuons aussi une transformation pour tous les scores de contrainte génétique en les inversant (1 divisé par le score de contrainte génétique). Les résultats du premier article [36] sont basés sur le score pLI qui a des valeurs allant de 0 à 1 et est délétère entre 0.9 et 1 tandis que le score LOEUF par exemple, a des valeurs allant de 0.03 à 2 et est délétère entre 0.03 et 0.35 (voir Figure 6.6) . Donc pLI est délétère quand ses valeurs sont hautes (0.9-1) et LOEUF quand ses valeurs sont basses (0.03-0.35). En transformant (1 divisé par LOEUF: 1/LOEUF), nous les ramenons tous les deux sur la même direction afin de faciliter l'interprétation. Dans la suite de ce mémoire, nous allons parler de LOEUF tout en faisant allusion à 1/LOEUF notamment dans les résultats. Nous rappelons que nous ne faisons pas encore de normalisation vu que les scores dans ce cas-ci sont étudiés de façon individuelle.

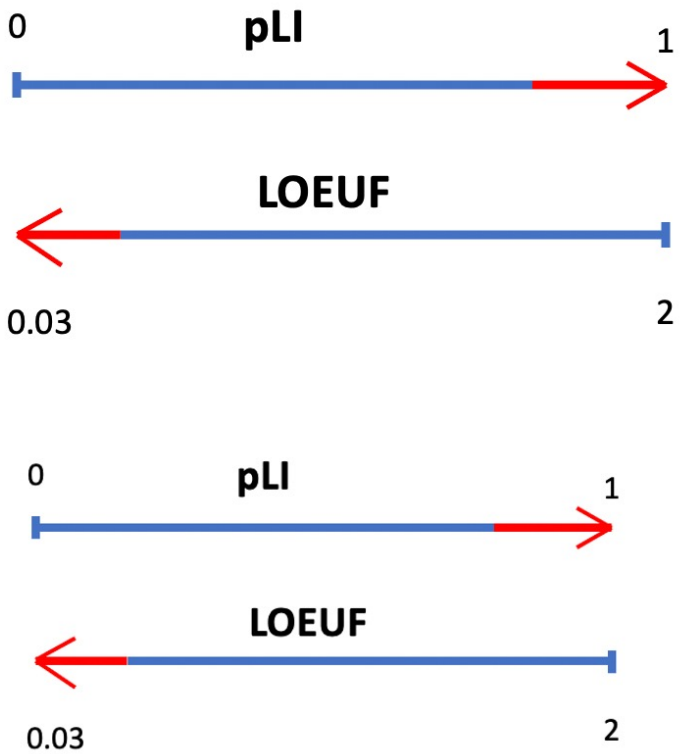


Fig. 6.6. Représentation des scores pLI et LOEUF. Le sens des fléches et la couleur rouge expliquent le fait que le pLI est délétère pour ses valeurs hautes (vers 1) et LOEUF est délétère pour ses valeurs basses (vers 0.03).

6.2.1. Annotation fonctionnelle

Les CNV que nous utilisons proviennent des travaux de Huguet et al. [36, 37] et de Douard et al. [38]. Ils ont annoté les CNV à l'aide de l'annotation Gencode V19 (la version de référence de la version hg19 du génome humain) avec les noms de gène de la base de données génomiques ENSEMBL (https://grch37.ensembl.org/index.html).
L'annotation se fait de façon hiérarchique. Nous avons les scores de gènes; les gènes qui sont contenus dans une CNV; une CNV qui est portée par un individu dont le QI est connu.
Nous utilisons notre jeu de données de scores pour annoter les gènes. Il n'y a que les gènes codants qui sont annotés. L'annotation des gènes consiste donc à attribuer les 16 scores aux gènes inclus dans les CNV identifiées chez chaque individu. Dans le cas où pour un gène donné le score n'a pas de valeur, on lui attribue un zéro (méthode de remplacement par une valeur neutre). L'annotation des CNV consiste à attribuer un score à une CNV. Pour ce faire, nous utilisons un modèle d'interaction génique additif. Pour chaque score, nous effectuons la somme des scores des gènes contenus dans cette CNV. Cependant nous avons des CNV résultant d'une délétion et d'autres d'une duplication. Nous nous retrouvons alors avec des scores des CNV pour les délétions et des scores des CNV pour les duplications. Si un individu a plusieurs CNV, pour calculer son score nous allons faire, pour chaque score, la somme des scores des CNV que porte cet individu. Encore une fois les délétions étant différentes des duplications, nous effectuons la somme des scores des CNV de délétion et la somme des scores des CNV de duplication. L'individu se retrouve avec deux scores: un score de l'individu pour les délétions et un autre pour les duplications. Le processus de cette annotation est illustré par la figure 6.7.

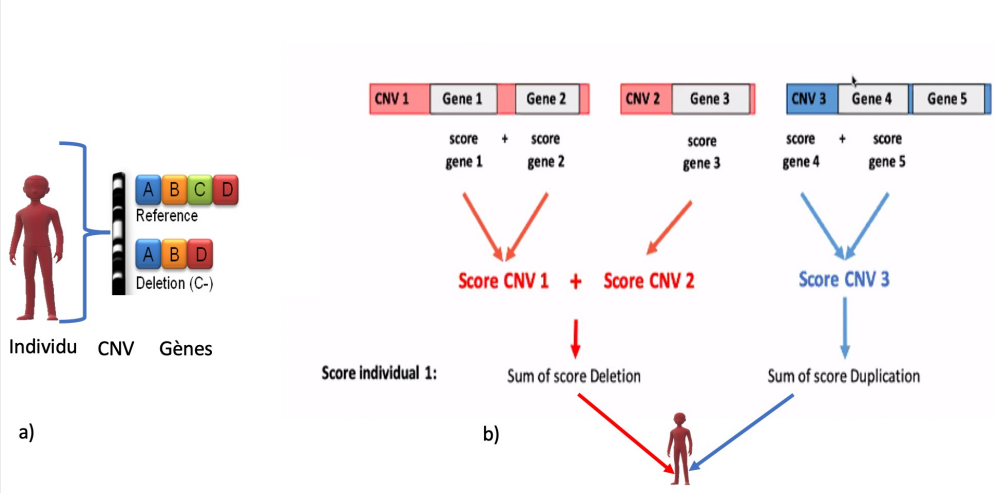


Fig. 6.7. Annotation du score de l'individu. Adapté de [36].
La Figure a) nous montre un individu qui porte une CNV de type délétion. En effet le gène C présent dans la séquence de référence ABCD est supprimé.
La figure b) montre le processus d'annotation du score de l'individu. Nous avons en rouge les délétions et en bleu les duplications.

Comme nous l'avons stipulé plus haut, ce processus est appliqué à nos trois jeux de données.

6.2.2. Analyse de données

Après avoir annoté nos scores de gènes, puis de CNV, nous avons à présent l'information à l'échelle de l'individu (le score de l'individu). Pour estimer les effets des CNV sur l'intelligence générale, nous adaptons le modèle développé par Huguet et al. [37]. En effet, après l'annotation de 10 scores de gènes, ils ont utilisé un modèle linéaire à effet mixte pour leur analyse; un modèle pour chaque score ; soit 10 modèles. Cependant les scores sont catégorisés en scores de délétion et scores de duplication. Nous avons alors 10 modèles pour les scores de délétion et 10 pour les scores de duplication. Ce modèle explique la variable d'intérêt $ZScore\_IQ$, le score Z de l'intelligence générale vu dans la section 4.2, en fonction d'un seul score. Le modèle linéaire mixte est ajusté en utilisant le type de test de QI des cohortes ($typeTest\_cohort$) comme effet fixe et le code de la famille comme un effet aléatoire. Vu que dans certaines cohortes nous avons plusieurs membres d'une même famille (par exemple la SSC), nous ajustons avec le code de famille comme un effet aléatoire (random effect).
Nous avons 16 scores qui ont été annotés dans notre jeu de données. Nous adoptons la même approche pour notre analyse. Nous utilisons toujours R avec le package nlme et la fonction lme(linear mixed effect). Cependant, dans le jeu de données de scores que nous utilisons, seuls deux scores (pLI, LOEUF) sont identiques à ceux utilisés dans l'étude de Huguet et al. [37]. Vu que nous effectuons une étude comparative, nous nous limiterons à la présentation des résultats pour les scores pLI et LOEUF.
Ayant trois jeux de données différents, nous calculons l'AIC(Akaike information criterion) pour chaque modèle de chaque type de données. L'AIC [73] est une métrique qui permet d'estimer la qualité du modèle. En statistiques, c'est l'une des métriques les plus utilisé pour la sélection de modèle. Il décrit dans quelle mesure un modèle donné ajuste les données. Plus il est bas, meilleur est le modèle.
Les Figures 6.8, 6.9 et 6.10 montrent les résultats. Le seuil de significativité désiré est de 0.05. Dans l'analyse, les p-value présentes sur les figures sont corrigées en utilisant la méthode de Benjamini et Hochberg afin de contrôler le taux de fausses découvertes (FDR). En effet, le FDR ajuste les p-value de façon à limiter le nombre de faux positifs qui étaient considérés comme significatifs avant la correction. Les p-value corrigées sont alors comparées au seuil de significativité et si elles sont inférieures, les résultats sont significatifs. L'estimé donne une indication sur le nombre de points de $Zscore\_IQ$ perdus par unité de score. Pour les résultats du QI, vu que cette dernière a été normalisée en utilisant le score Z, pour connaitre la perte de QI estimée, nous multiplions l'estimé (Est.) par 15 qui représente l'écart type du

QI (voir section 4.2). À chaque fois que nous parlons de LOEUF, nous faisons allusion au LOEUF inversé.

DÉLÉTION

	Est.	Lower	Upper	P-value	AIC
pLI	-0.1672704	-0.1996184	-0.1349224	5.121845e-24	71374
1/LOEUF	-0.0261180	-0.03081814	-0.02141785	1.843843e-27	71362.05

DUPLICATION

	Est.	Lower	Upper	P-value	AIC
pLI	-0.0457891	-0.06319161	-0.02838660	2.558773e-07	71451.14
1/LOEUF	-0.006770512	-0.009193114	-0.004347910	4.420418e-08	71451.68

Fig. 6.8. Résultats du QI pour le pLI et LOEUF inversé basés sur le jeu de données suivant la méthode de remplacement par zéro. Est.= estimé, p-value, AIC= Akaike information criterion, Lower= borne inférieure de l'intervalle de confiance, Upper= borne supérieure de l'intervalle de confiance. Nous avons un niveau de confiance 95% ce qui signifie que nous sommes sûr à 95% que la valeur de l'estimée (Est.) est comprise entre la valeur du Lower et la valeur du Upper. Nous avons d'abord les résultats des délétions puis ceux des duplications.

La Figure 6.8 présente nos résultats sur le jeu de données suivant la méthode de remplacement par zéro. Ces tests sont effectués de façon simultanée. Nos résultats sont très similaires à ceux de Huguet et al. [37]. La diminution du $ZScore\_IQ$ est exprimée par la colonne Est. de la Figure 6.8 que nous arrondissons. Toutes les p-values fournies entre parenthèse sont des p-value corrigées par FDR après l'analyse. Le seuil de significativité désiré est de 0.05.
Pour le score pLI nous avons des résultats significatifs, nous trouvons que chaque point de pLI réduit correspond à une diminution du $ZScore\_IQ$ de 0.165 points ce qui signifie que chaque point de pLI réduit correspond à une diminution du QI de 2.475 points. Les résultats montrent une p-value très significative ($p = 1{,}02 \times 10^{-23}$). Chaque point de pLI dupliqué, correspond à une diminution du $ZScore\_IQ$ de 0.045 points ce qui correspond à une diminution du QI de 0.675 points avec une p-value significative ($p = 2.5 \times 10^{-7}$). Pour le score LOEUF les résultats sont aussi significatifs, chaque point de LOEUF réduit correspond à une diminution du $ZScore\_IQ$ de 0.025 points; ce qui correspond à une diminution du QI de 0.375 points avec une p-value significative ($p = 7.37 \times 10^{-27}$). Chaque point de LOEUF dupliqué correspond à une diminution du $ZScore\_IQ$ de 0.007 points; ce qui correspond à

une diminution du QI de 0.105 points avec une p-value significative ($p = 5.8 \times 10^{-8}$).
Tout comme l'étude précédente [37], sur la base de l'AIC, le LOEUF est sélectionné comme étant le meilleur score pour les délétions. Cependant pour les duplications, les valeurs de l'AIC pour le pLI et le LOEUF sont très similaires et ne permettent pas de conclure le quel des 2 scores est le meilleur.

DELETION

	Est.	Lower	Upper	P-value	AIC
pLI	-0.1648745	-0.1968859	-0.1328632	7.679734e-24	71374.83
1/LOEUF	-0.02533252	-0.02992624	-0.02073879	4.512516e-27	71363.88

DUPLICATION

	Est.	Lower	Upper	P-value	AIC
pLI	-0.04534255	-0.06257912	-0.02810599	2.574186e-07	71451.17
1/LOEUF	-0.006611485	-0.008983471	-0.004239500	4.796480e-08	71451.88

Fig. 6.9. Résultats du QI pour le pLI et LOEUF inversé basés sur le jeu de données imputées. Est.= estimé, p-value, AIC= Akaike information criterion, Lower= borne inférieure de l'intervalle de confiance, Upper= borne supérieure de l'intervalle de confiance. Nous avons un niveau de confiance 95% ce qui signifie que nous sommes sûr à 95% que la valeur de l'estimée (Est.) est comprise entre la valeur du Lower et la valeur du Upper. Nous avons d'abord les résultats des délétions puis ceux des duplications.

La Figure 6.9 présente nos résultats sur le jeu de données imputées. La diminution du $ZScore\_IQ$ est exprimée par la colonne Est. de la Figure 6.9 que nous arrondissons. Toutes les p-values fournies entre parenthèse sont des p-value corrigées par FDR après l'analyse. Le seuil de significativité désiré est de 0.05.
Les résultats du pLI pour les délétions sont significatifs ($p = 1.53 \times 10^{-23}$) et nous montrent qu'un point de pLI réduit correspond à une diminution de 0.165 points du $ZScore\_IQ$; ce qui correspond à une diminution de 2.475 points de QI. Les résultats pour les duplications restent aussi significatifs ($p = 2.6 \times 10^{-7}$); chaque point de pLI dupliqué correspond à une diminution du $ZScore\_IQ$ de 0.045 points et du QI de 0.675 points. Pour le LOEUF, nous avons des résultats significatifs pour les délétions ($p = 1.8 \times 10^{-26}$). En effet, un point de LOEUF réduit correspond à une diminution de 0.025 points du $ZScore\_IQ$ donc une diminution de 0.375 points de QI. Les résultats du LOEUF pour les duplications sont également significatifs ($p = 6.3 \times 10^{-8}$). Un point de LOEUF dupliqué correspond à une diminution du $ZScore\_IQ$ de 0.0066 points et du QI de 0.099 points .

Nous constatons que les résultats sont très similaires avec ceux obtenus avec la méthode de remplacement par zéro.
Les résultats de l'AIC sélectionnent le LOEUF comme étant le meilleur score pour les délétions. Pour les duplications, les valeurs de l'AIC pour le pLI et le LOEUF sont presque pareils.

DELETION

	Est.	Lower	Upper	P-value	AIC
pLI	-0.2240620	-0.2794125	-0.1687115	2.359681e-15	71412.56
1/LOEUF	-0.03770782	-0.04616275	-0.02925288	2.697629e-18	71402.91

DUPLICATION

	Est.	Lower	Upper	P-value	AIC
pLI	-0.05497394	-0.0840215420	-0.0259263350	2.089501e-04	71462.94
1/LOEUF	-0.01037799	-0.01456093	-0.006195046	1.175981e-06	71456.94

Fig. 6.10. Résultats du QI pour le pLI et LOEUF inversé basés sur le jeu de données supprimées. Est.= estimé, p-value, AIC= Akaike information criterion, Lower= borne inférieure de l'intervalle de confiance, Upper= borne supérieure de l'intervalle de confiance. Nous avons un niveau de confiance 95% ce qui signifie que nous sommes sûr à 95% que la valeur de l'estimée (Est.) est comprise entre la valeur du Lower et la valeur du Upper. Nous avons d'abord les résultats des délétions puis ceux des duplications.

La Figure 6.10 présente nos résultats sur le jeu de données manquantes supprimées. La diminution du $ZScore\_IQ$ est exprimée par la colonne Est. de la Figure 6.10 que nous arrondissons. Toutes les p-values fournies entre parenthèse sont des p-value corrigées par FDR après l'analyse. Le seuil de significativité désiré est de 0.05.
Un point de pLI réduit correspond à une diminution du $ZScore\_IQ$ de 0.224 points; ce qui correspond à une diminution du QI de 3.36 points avec une p-value $p = 4.7 \times 10^{-15}$. Un point de pLI dupliqué correspond à une diminution du $ZScore\_IQ$ de 0.055 points ce qui correspond à une diminution de 0.825 du QI avec une p-value $p = 2 \times 10^{-4}$. Les résultats pour le LOEUF sont aussi significatifs avec une p-value $p = 1.07 \times 10^{-17}$ pour les délétions et $p = 1.15 times 10^{-6}$ pour les duplications. Un point de LOEUF réduit correspond à une diminution du $ZScore\_IQ$ de 0.038 points; ce qui correspond à une diminution du QI de 0.57 points. Un point de LOEUF dupliqué correspond à une diminution du $ZScore\_IQ$ de 0.01 points; ce qui correspond à une diminution du QI de 0.15 points.
Nous constatons que les résultats sont moins significatifs et que les estimations sont plus

sévères comparés aux résultats obtenus avec la méthode de remplacement par zéro.
Les résultats de l'AIC sélectionnent le LOEUF comme étant le meilleur score pour les délétions et le pLI comme étant le meilleur score pour les duplications.

JEU DE DONNÉES	DELETION		DUPLICATION	
	1/ LOEUF	pLI	1/ LOEUF	pLI
DM SUPPRIMÉES	71402.91	71412.56	71456.94	71462.94
MÉTHODE LAB	71362.05	71374	71451.68	71451.14
DM IMPUTÉES	71363.88	71374.83	71451.88	71451.17

Fig. 6.11. Comparaison des AIC selon le type de jeu de données pour la sélection du meilleur modèle. Nous avons mis en évidence la méthode de référence (en rouge) qui est la méthode de remplacement par zéro et qui va nous permettre de choisir le meilleur entre les deux jeux de données complets.

Nous rappelons que ces résultats sont ceux de la sélection du meilleur jeu de données pour notre score composite qui requiert un jeu de données complet. Pour ce faire, nous avons les résultats issus de deux jeu de données complets (DM supprimées et DM imputées) que nous allons comparer avec ceux de la méthode de remplacement par zéro. Les valeurs de la Figure 6.11 proviennent des AIC des Figures 6.8, 6.9 et 6.10.
Les AIC pour les délétions et les duplications du modèle avec les données manquantes supprimées sont assez éloignés du modèle de référence (le modèle de remplacement par zéro). Cependant les AIC du modèle avec les données manquantes imputées et ceux du modèle de

remplacement par zéro sont très similaires. Les figures 6.12 et 6.13 nous permettent de mieux apprécier cette comparaison entre les deux jeux de données complets et le jeu de données suivant la méthode de remplacement par zéro.
Le jeu de données avec DM imputé est donc de loin le meilleur pour notre score composite comparé à celui où les DM sont supprimées car il a toutes les valeurs et il reproduit les mêmes observations et conclusions que les observations obtenus avec la méthode de remplacement par zéro.

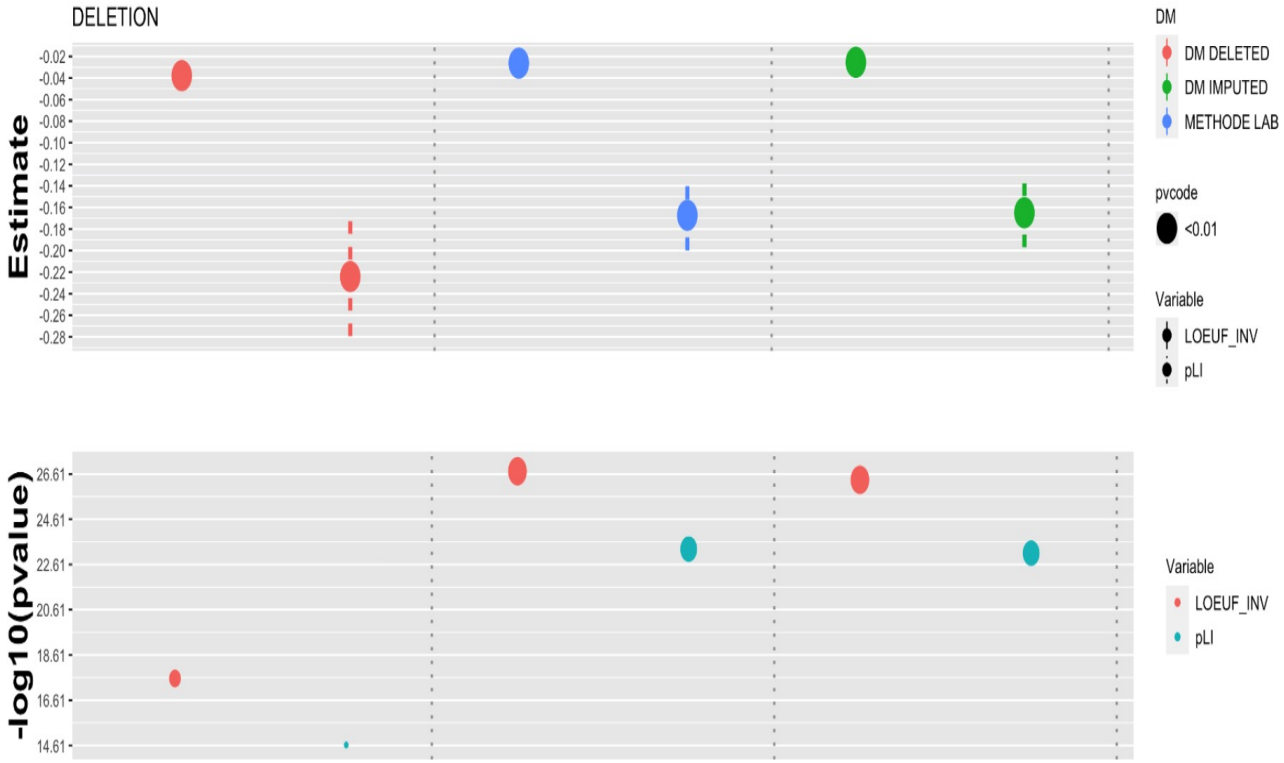


Fig. 6.12. Comparaison des résultats selon le type de jeu de données pour la sélection du meilleur modèle. En rouge nous avons les résultats avec le jeu de données supprimées, en vert ceux avec les données imputées et en bleu nous avons les résultats issues des données avec la méthode de remplacement par zéro. pvcode= p-value; En trait plein nous avons le LOEUF inversé et en tirets nous avons le pLI. En ordonnée nous avons les estimées et aussi le -log10(Pvalue) qui nous permet de bien voir la p-value par rapport à la figure avec les estimées. Cette figure est pour les délétions.

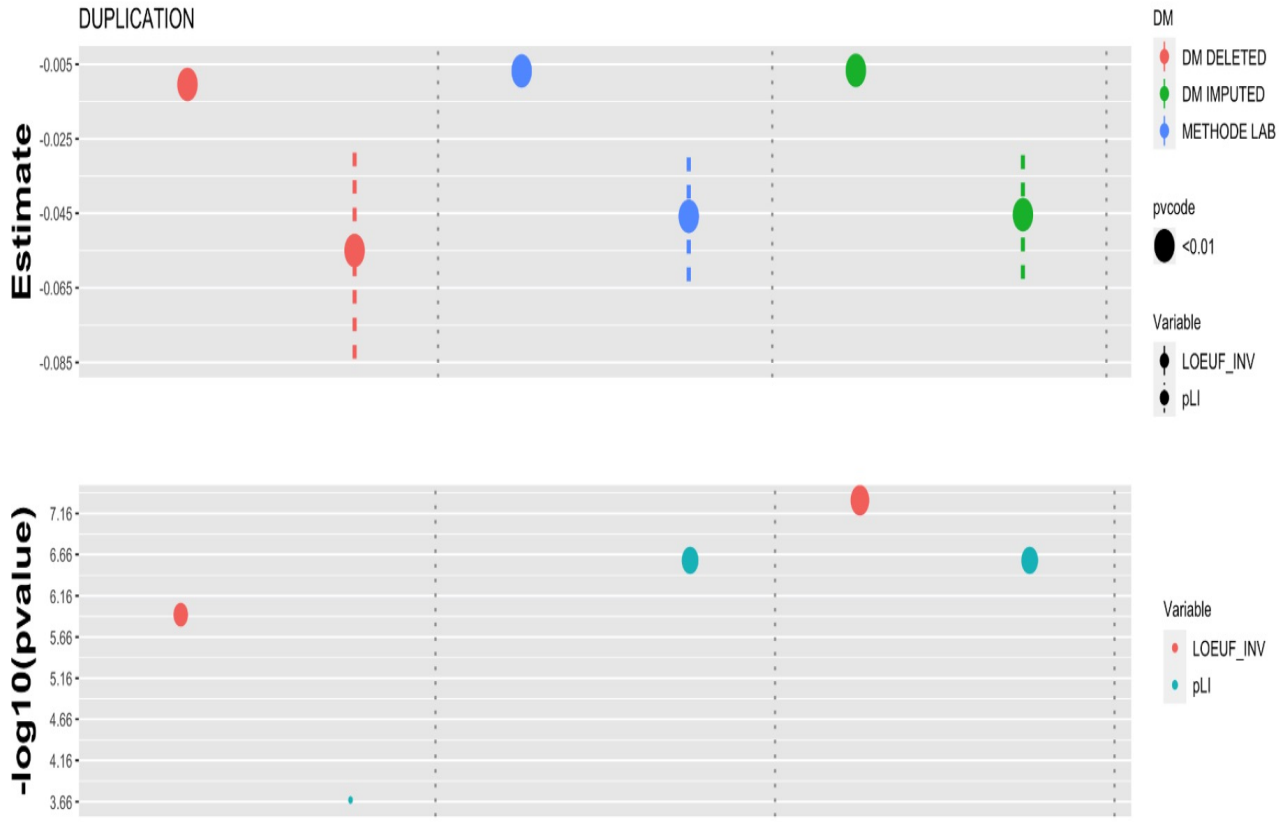


Fig. 6.13. Comparaison des résultats selon le type de jeu de données pour la sélection du meilleur modèle. En rouge nous avons les résultats avec le jeu de données supprimées, en vert ceux avec les données imputées et en bleu nous avons les résultats issues des données avec la méthode de remplacement par zéro. pvcode= p-value; En trait plein nous avons le LOEUF inversé et en tirets nous avons le pLI. En ordonnée nous avons les estimées et aussi le -log10(Pvalue) qui nous permet de bien voir la p-value par rapport à la figure avec les estimées. Cette figure est pour les duplications..

6.2.3. Concordance avec la littérature

Nous allons maintenant comparer nos résultats avec ceux de la littérature et de la cohorte UKBB (voir Huguet et al. [37], table 17). Cependant pour simplifier, nous allons utiliser le terme littérature pour les deux. Les résultats du QI que nous avons de la littérature proviennent de 47 CNV récurrents. Nous allons d'abord tester la concordance du modèle suivant le jeu de données de la méthode de remplacement par zéro avec les CNV de la littérature et ensuite la concordance du modèle avec le jeu de données imputées et les CNV de la

littérature. Cette concordance est mesurée par le coefficient de corrélation intraclasse (ICC) qui est une métrique qui mesure à quel point des observations peuvent être similaires.
Les 47 CNV récurrents sont annotées suivant la méthode d'annotation présentée à la section 6.2.1. Pour finir, nous utilisons notre modèle linéaire mis en place à la section 6.2.2 pour prédire le QI des personnes avec ces 47 CNV récurrents. Cette prédiction est effectuée grâce à la fonction *predict* de R. Les résultats de QI de cette analyse sont ensuite comparés aux résultats des QI trouvés dans la littérature. Nous utilisons les résultats de QI obtenus avec la variable LOEUF pour effectuer cette comparaison. La Figure 6.14 illustre cette comparaison. Nous obtenons une concordance de 66% (0.6589) avec la littérature pour les délétions et une concordance de 83% (0.828) pour les duplications.
Pour la concordance du modèle avec le jeu de données imputées, nous allons fusionner les scores de gènes provenant du jeu de données imputées avec les gènes des 47 CNV récurrents afin d'avoir des données imputées pour ces CNV récurrents. La même approche pour prédire les CNV récurrents suivant la méthode de remplacement par zéro est utilisée. La Figure 6.15 illustre cette comparaison. Nous obtenons une concordance de 66% (0.6618) avec la littérature pour les délétions et une concordance de 83% (0.8279) pour les duplications. Ces concordances sont similaires avec celles obtenues précédemment.

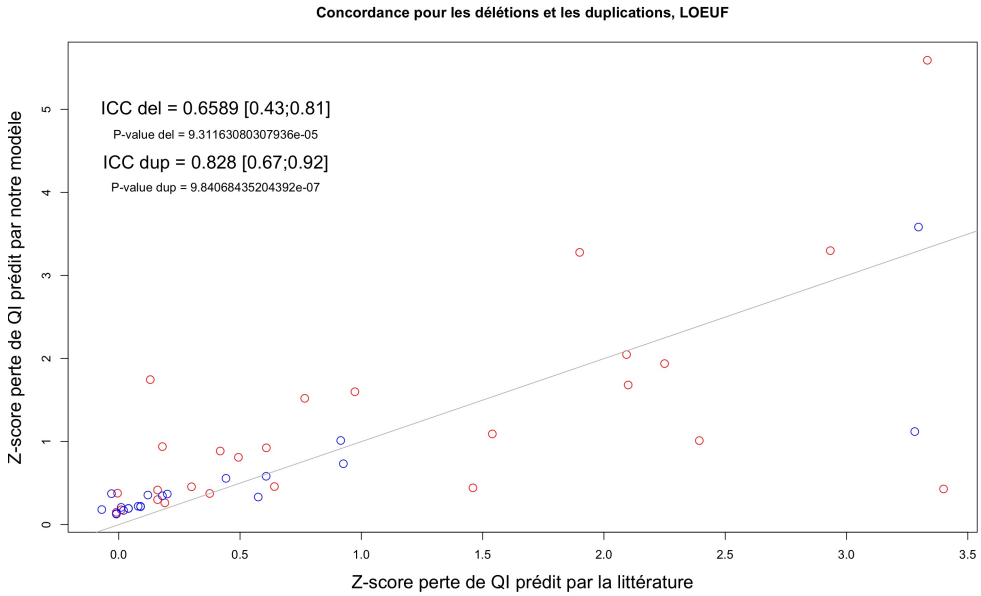


Fig. 6.14. Concordance des résultats du score LOEUF du jeu de données suivant la méthode de remplacement par zéro avec les résultats de la littérature. En rouge nous avons les délétions et en bleu les duplications. En abscisse nous avons le score Z de la perte de QI provenant de la littérature et en ordonné celui estimé par nos modèles. P-value del représente la P-value pour les délétions et P-value dup celle des duplications. ICC=Coefficient de corrélation intraclasse.

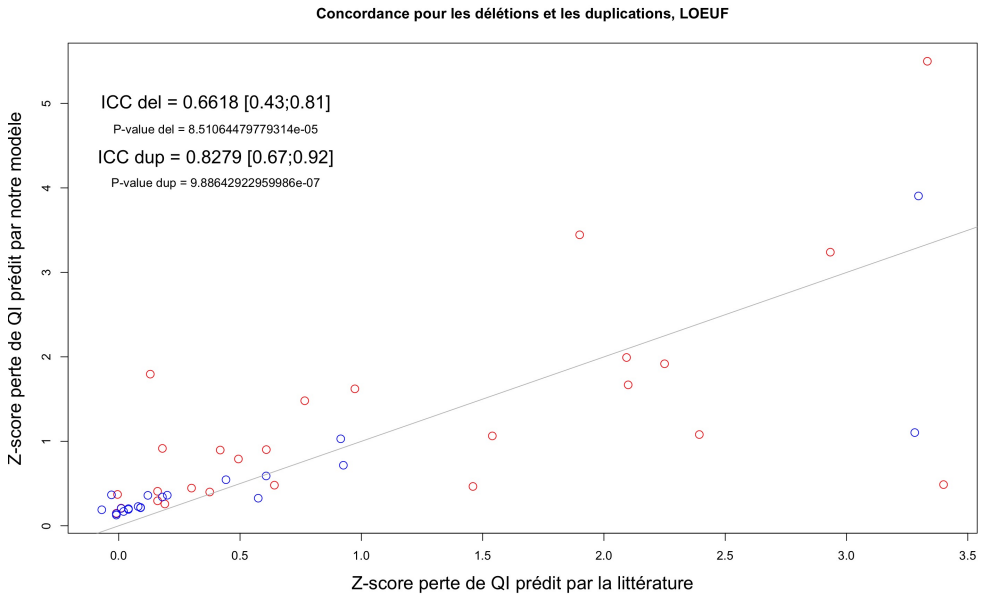


Fig. 6.15. Concordance des résultats du score LOEUF du jeu de données imputées avec les résultats de la littérature. En rouge nous avons les délétions et en bleu les duplications. En abscisse nous avons le score Z de la perte de QI provenant de la littérature et en ordonné celui estimé par nos modèles. P-value del représente la P-value pour les délétions et P-value dup celle des duplications. ICC=Coefficient de corrélation intraclasse.

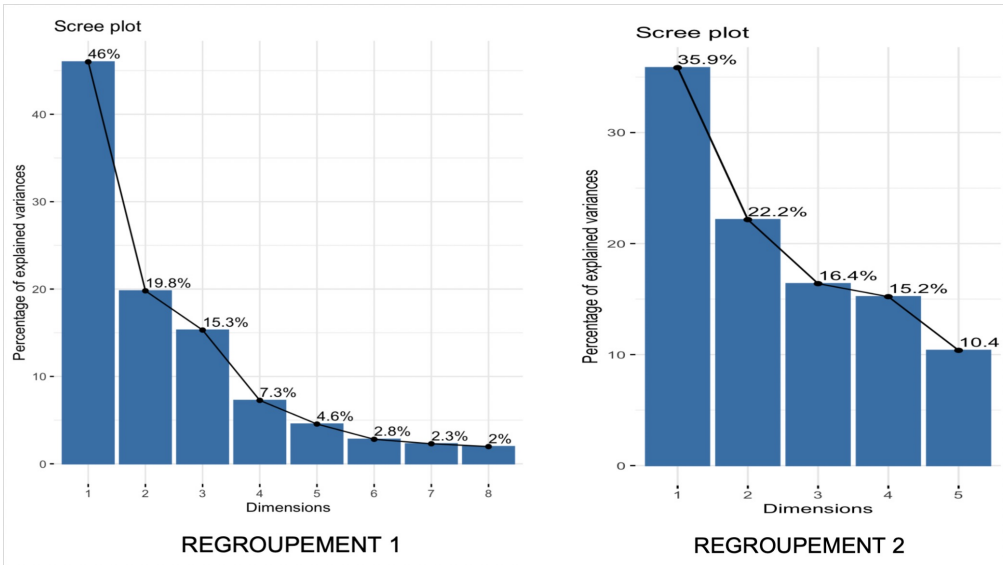
6.3. Calcul du score composite et résultats

Comme vu dans le chapitre 2.2.1, le score composite consiste à regrouper des mesures individuelles en une mesure globale visant à déterminer différents aspects d'un modèle conceptuel. Dans ce sens, il permet de réduire le nombre de variables à étudier. Il sera appliqué sur le jeu de données imputées vu que nous avons obtenu de meilleurs résultats avec celui-ci. Nous allons adopter une des approches (construction formative ou construction réflexive) permettant de construire un score composite (voir section 2.2.1). Étant donné que nos scores sont corrélés entre eux, nous ne pouvons pas utiliser la construction formative qui nécessite des mesures (dans notre études ces mesures sont les scores) non corrélées. Nous allons donc utiliser l'approche de la construction réflexive. Cette dernière utilise la méthode d'analyse par composante principale et requiert que les scores soient normalisés.

6.3.1. Analyse par composante principale

L'analyse par composantes principales (ACP) est un algorithme mathématique qui réduit la dimensionnalité des données tout en conservant la majeure partie de la variation dans l'ensemble de données [62]. Elle nous donne comme résultat de nouvelles variables appelées composantes principales ou facteurs ou encore dimensions. Ces dernières sont des combinaisons linéaires des variables initiales. L'ACP nous permet donc de réduire nos variables en

perdant le moins d'informations possible.
L'ACP est appliquée sur le jeu de données retenu comme étant meilleur après la sélection: le jeu des données imputées. Pour ce faire, nous utilisons les packages "FactoMineR", "Factoextra" de R. Cette analyse par composante principale se fera sur les différents regroupements de la Figure 6.5. Sur chaque regroupement, on applique l'ACP. Les scores étant sur différentes échelles, nous allons normaliser les données en utilisant la fonction scale.unit du même package. Dans la section 6.1, les scores n'étaient pas normalisés car ils sont étudiés de façon individuelle alors qu'ici, ils sont combinés. Pour appliquer l'ACP, nous utilisons la fonction PCA de ces packages sur chaque regroupement. Le résultat de cette fonction nous renvoie plusieurs informations parmi lesquelles le pourcentage de la variance expliqué par chaque composante. Afin de mieux visualiser ce pourcentage, nous allons mettre en place le graphique des valeurs propres appelé scree plot (Figure 6.16) en anglais en utilisant la fonction R $fviz\_eig()$. Nous allons également accéder aux variables de ces résultats avec la fonction $get\_var()$ qui nous renvoie entre autres le cos2 ou cosinus carré qui nous permet de voir la qualité de la représentation des variables dans chaque composante (Figure 6.18). Ces deux graphiques mis en place avec la fonction corrplot de R, nous permettent de connaitre le nombre de composantes principales que nous allons retenir et qui expliquent le plus la variance.



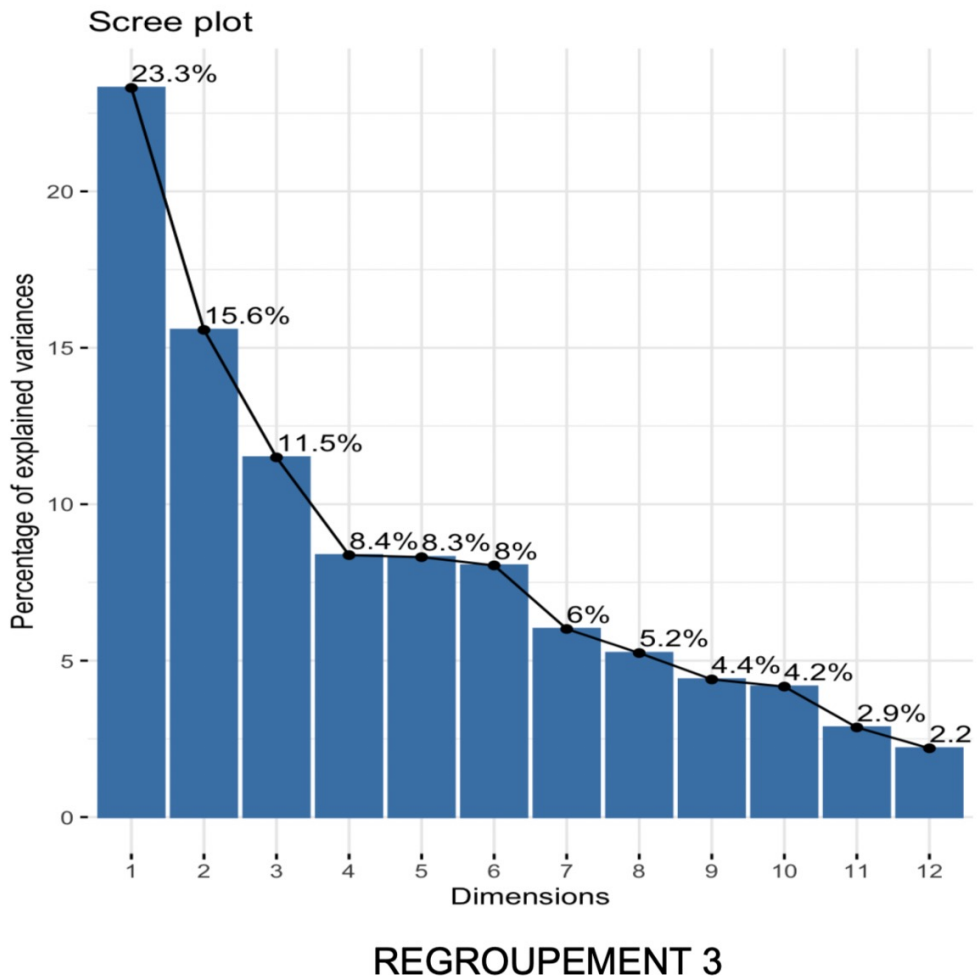
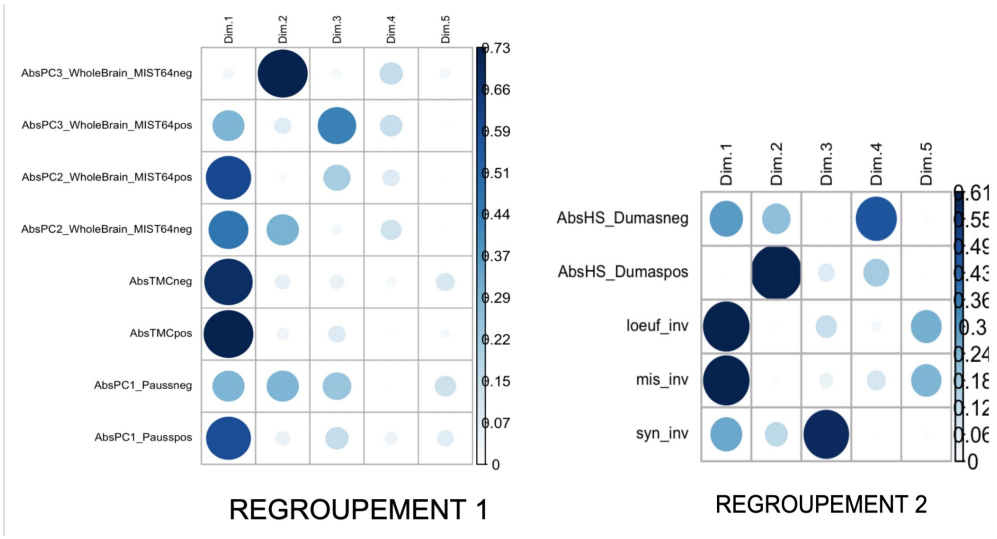


Fig. 6.16. Résultats du scree plot pour les trois regroupements. En ordonnée, nous avons le pourcentage de la variance présent dans chaque composante principale et en abscisse nous avons les différentes composantes principales.

Les résultats de la Figure 6.16 nous montrent les scree plot des trois regroupements. Il faut noter que les dimensions sont appelées composantes principales.

Pour le regroupement 1, nous avons 8 dimensions et dans chaque dimension, nous avons le pourcentage de la variance; 46% de la variance est expliqué par la dimension 1 et nous remarquons que plus de la moitié de la variance (65%) est expliquée par les deux premières composantes. Pour le regroupement 2, nous avons 5 dimensions et tout comme le regroupement 1, plus de la moitié de la variance est expliquée par la première et la deuxième dimension (57.9%). Le regroupement 3 nous montre 12 dimensions. Cependant contrairement aux deux autres regroupement, 38.9% de la variance est expliqué par les dimensions 1 et 2.

Les résultats du scree plot suggèrent que pour les regroupements 1 et 2, les deux premières dimensions expliquent la majorité de la variance. Nous allons donc les retenir. Le regroupement 3, la moitié de la variance est expliquée en sélectionnant les trois premières dimensions. Cependant, nous allons nous baser sur la qualité de la représentation des scores dans ces différentes dimensions pour valider ces résultats.



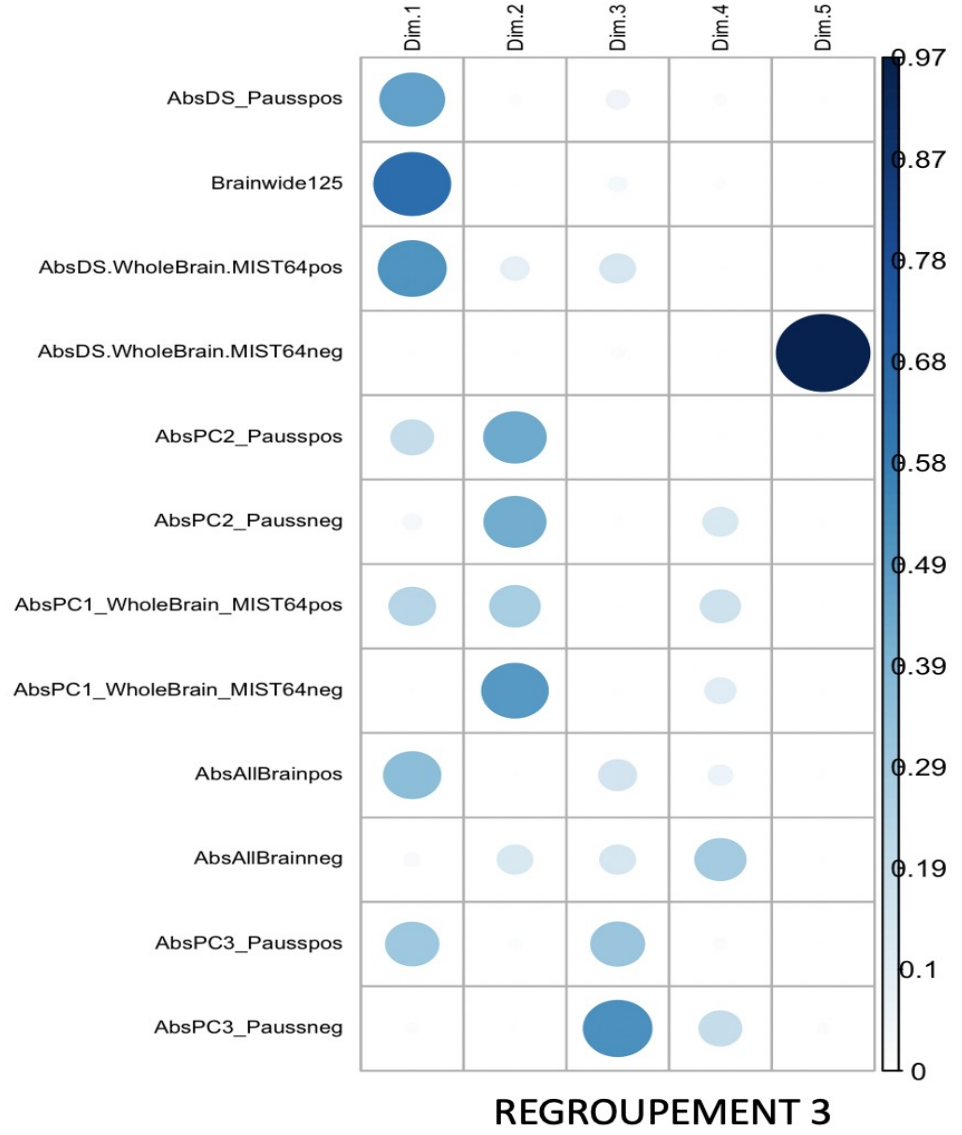


Fig. 6.17. Matrice expliquant la corrélation entre les dimensions et les scores individuels c'est-à-dire la qualité de la représentation des scores dans chaque dimension. En abscisse nous avons les dimensions, en ordonné à gauche nous avons les scores et en ordonnée à droite nous avons le gradient de couleur qui explique la qualité de la représentation.

La Figure 6.18 nous donne un aperçu sur la qualité de la représentation de nos scores dans chaque composante. Plus le cercle est foncé, plus la qualité de la représentation est meilleure. C'est une corrélation entre chaque dimension et les variables de notre jeu de données.
Dans le regroupement 1, nous notons que les variables sont très bien représentées dans les dimensions 1 et 2. Dans le regroupement 2, les scores négatifs de HS_Dumas sont mieux représentés dans la dimension 4 puis la dimension 1 et le score oe_syn_upper inversé est mieux représenté dans la dimension 3 puis la dimension 1. Cependant, le reste des scores du regroupement 2 est bien représenté dans les dimensions 1 et 2. Ces résultats pour les regroupements 1 et 2 renforcent la décision de garder les deux premières dimensions.
Tous les scores du regroupement 3 sont mieux représentés dans les dimensions 1 et 2 sauf le $PC3\_Pauss$ négatif et le $DS\_WholeBrain$ négatif qui sont successivement mieux représentés dans les dimensions 3 et 5. Nous savons que pour ce regroupement, la moitiée de la variance est représentée par les regroupements 1,2 et 3. Cependant, les scores sont très faiblement représentés dans la dimension 3 et le fait de garder cette dernière pour nos analyses pourrait insérer du biais dans nos résultats.
D'après les résultats des Figures 6.16 et 6.18, nous retenons les composantes principales 1 et 2 de chaque regroupement.
N'oublions pas que dans le jeu de données sur lequel l'ACP est appliqué, les colonnes sont des gènes et les variables sont les scores. Nous allons récupérer les informations des gènes pour chaque composante principale (CP ou dimension) retenue. Pour ce faire, nous allons utiliser la fonction $get\_pca\_ind()$ qui nous renvoie des résultats dont ceux des coordonnées des gènes pour chaque composante principale. Nous nous retrouvons avec trois jeux de données issus des trois regroupements dont les lignes sont des gènes et les colonnes (variables) sont les composantes principales. En effet, chaque composante principale (CP) est considérée comme un score. Nous les combinons afin d'obtenir un seul jeu de données dénoté "jeu de données du score composite" et nous renommons les variables suivant le regroupement. Nous avons alors le jeu de données du score composite avec comme variables (scores composites) PC1CL1, PC2CL1, PC1CL2, PC2CL2, PC1CL3, PC2CL3.

6.3.2. Annotation fonctionnelle

Nous performons la même annotation que celle effectuée dans la section 6.2.1 sauf que cette fois ci, au lieu d'être appliquée sur les scores de façon individuelle, nous l'appliquons sur le jeu de données du score composite. Nous avons maintenant 6 scores composites au lieu de 15 scores individuels et ces 6 scores composites expliquent les informations contenues dans les 15 scores que nous avions. Par exemple pour annoter le score composite PC1CL1, nous faisons d'abord l'annotation des gènes en réconciliant chaque score composite avec son

gène, sa position dans la CNV (début, fin) et son type. Maintenant que nous avons les scores composites de gènes, nous allons calculer les scores composites de délétion et de duplication des CNV de PC1CL1 en faisant la somme des scores composites de gène de PC1CL1 pour les délétions et la somme pour les duplications. Le score composite PC1CL1 de l'individu pour les délétions et le score composite PC1CL1 de l'individu pour les duplications sont calculés en faisant la somme des scores composites CNV de PC1CL1 pour les délétions d'une part et d'autre part pour les duplications.

6.3.3. Analyse de données

Après avoir annoté les scores composites nous adoptons le même modèle linéaire que celui utilisé dans la section 6.2.2. La variable d'intérêt $ZScore\_IQ$ est expliquée en fonction d'un score composite. Pour chaque score composite, nous avons un modèle pour les délétions et un modèle pour les duplications. À côté de ce modèle nous avons mis en place un modèle qui regroupe tous les scores composites de délétion d'une part et un modèle qui regroupe tous les scores composites de duplication d'autre part.
Les résultats fournis par l'AIC (voir annexe A.5) pour les délétions sont meilleurs pour ce modèle par rapport au modèle que nous présentons à la section (6.19).
Nous avons aussi mis en place un modèle qui regroupe tous les scores composites de délétion et de duplication. Les résultats fournis par l'AIC pour ce modèle (voir annexe A.6) montrent que notre modèle est meilleur (6.19 et 6.20).
La figure illustre de façon très simple la méthodologique adoptée pour mettre en place le score composite et quantifier les effets des CNV sur le QI.
En effet, nous avons d'abord caractérisé les scores génétiques en étudiant leur distribution et en faisant leur classification hiérarchique. Nous avons ensuite effectué la gestion des données manquantes des scores génétiques en utilisant trois méthodes. Les données issues de ces méthodes sont par la suite utilisées pour répliquer l'étude de Huguet et al [37]. Cette étude utilise une méthode d'annotion décrite à la section 6.3.2 et un modèle linéaire pour l'analyse de données. Une analyse comparative a ensuite été faite pour sélectionner le meilleur jeu de données pour construire notre score composite.
Avant de construire le score composite, une classification hiérarchique a été effectuée sur les scores sélectionnés. L'approche de la construction formative a été utilisée avec la méthode d'analyse par composante principale. Les résultats de ces analyses nous fournissent des composantes principales qui sont à présent considérées comme nos nouveaux scores génétiques. Ces derniers sont une combinaison des scores génétiques de notre jeu de données. Pour finir, les mêmes méthodes d'annotation et d'analyse que celles utilisées pour répliquer l'étude de Huguet et al [38] ont été utilisées.

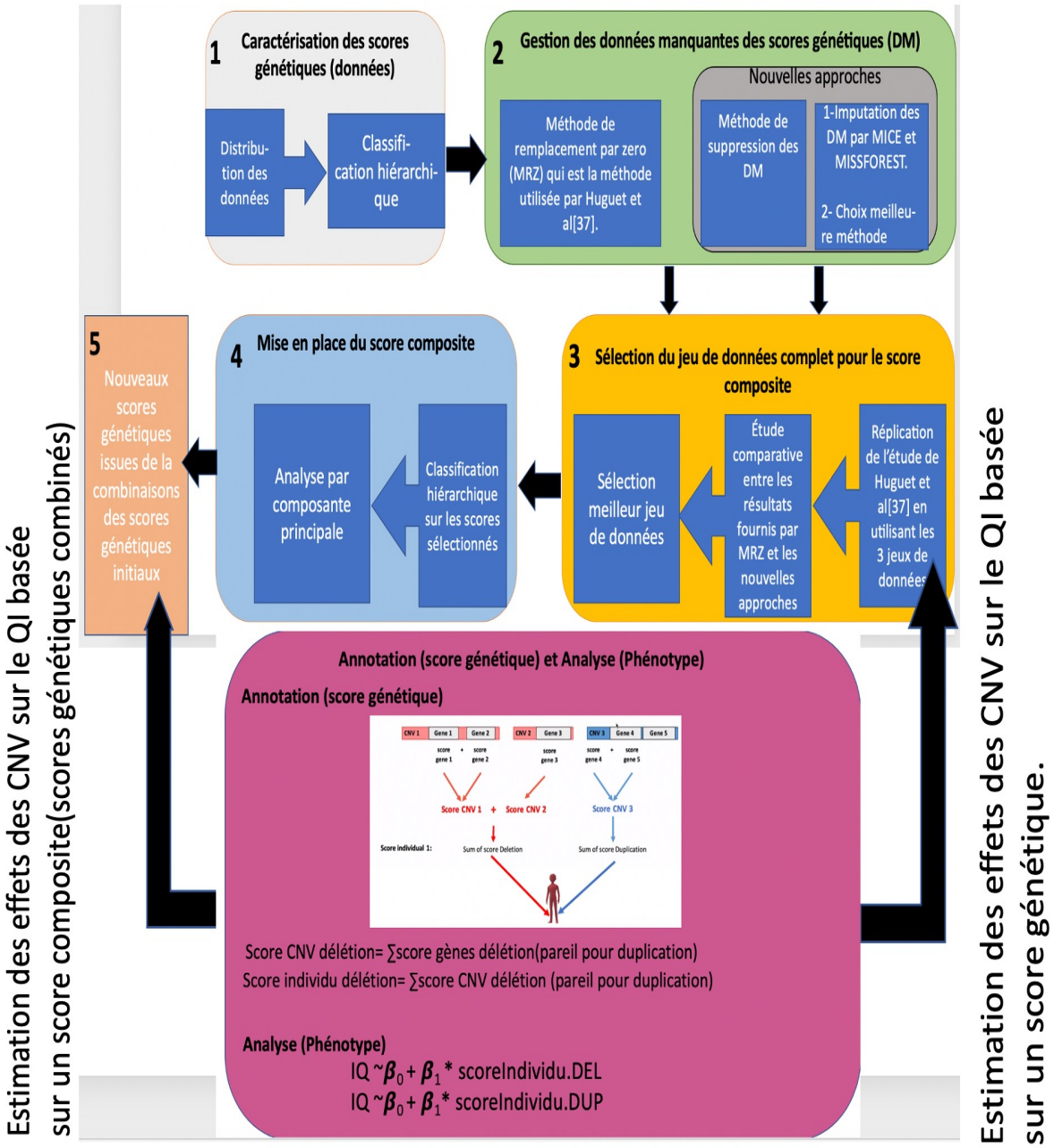


Fig. 6.18. Flux de la méthodologie. Les chiffres représentent le sens de déplacement. Les grandes étapes sont différenciées par des couleurs. La phase d'annotation et d'analyse statistique est utilisée lors de la réplication de l'étude de Huguet et al (estimation des effets des CNV sur le QI avec un score génétique) et lors de l'estimation des effets des CNV sur le QI avec un score composite (plusieurs scores génétiques combinés).ß0,ß1 sont des coefficients de régression.

Nous allons présenter les résultats de l'estimation du QI obtenu en utilisant le score composite sur les Figures 6.19 et 6.20.

	Est.	Lower	Upper	P-value	AIC
DEL.PC1CL1	-0.02221316	-0.02879296	-0.01563336	3.8676e-11	71435.94
DEL.PC2CL1	-0.02956530	-0.03845502	-0.02067558	7.4625e-11	71436.63
DEL.PC1CL2	-0.03578360	-0.04647099	-0.02509620	5.5699e-11	71435.67
DEL.PC2CL2	0.00627938	-0.00594428	0.01850305	3.1402e-01	71477.41
DEL.PC1CL3	-0.02548327	-0.03222831	-0.01873824	1.4234e-13	71424.86
DEL.PC2CL3	-0.01624269	-0.02592679	-0.00655859	1.0148e-03	71468.10

Fig. 6.19. Résultats du QI pour les composantes principales pour les délétions. Est.= estimé, p-value, AIC= Akaike information criterion, Lower= borne inférieure de l'intervalle de confiance, Upper= borne supérieure de l'intervalle de confiance. Nous avons un niveau de confiance 95% ce qui signifie que nous sommes sûr à 95% que la valeur de l'estimée (Est.) est comprise entre la valeur du Lower et la valeur du Upper.

	Est.	Lower	Upper	P-value	AIC
DUP.PC1CL1	-0.00378633	-0.00761389	4.12198e-05	5.254420e-02	71476.99
DUP.PC2CL1	-0.01415784	-0.01935927	-0.00895641	9.77818e-08	71451.69
DUP.PC1CL2	-0.00971698	-0.01495004	-0.00448391	2.7472e-04	71466.88
DUP.PC2CL2	-0.00338986	-0.00975566	0.00297594	2.96641e-01	71478.64
DUP.PC1CL3	-0.0078872	-0.01184698	-0.00392747	9.52669e-05	71465.45
DUP.PC2CL3	-0.00575814	-0.01061701	-0.00089927	2.02135e-02	71474.88

Fig. 6.20. Résultats du QI pour les composantes principales pour les duplications. Est.= estimé, p-value, AIC= Akaike information criterion, Lower= borne inférieure de l'intervalle de confiance, Upper= borne supérieure de l'intervalle de confiance. Nous avons un niveau de confiance 95% ce qui signifie que nous sommes sûr à 95% que la valeur de l'estimée (Est.) est comprise entre la valeur du Lower et la valeur du Upper.

Toutes les p-values fournies entre parenthèse sont des p-value corrigées par FDR après l'analyse. Le seuil de significativité désiré est de 0.05.

Tous nos résultats sont significatifs pour les délétions et pour les duplications sauf pour la dimension 2 du regroupement 2 (DEL.PC2CL2 et DUP.PC2CL2). La diminution du $ZScore\_IQ$ est exprimée par la colonne Est. des Figures 6.19 et 6.20 que nous arrondissons.
La réduction d'un point de la dimension 1 du regroupement 1 (DEL.PC1CL1) correspond à un score $ZScore\_IQ$ réduit de 0.022 points ce qui correspond à une réduction du QI de 0.33 points ($p = 2.22 \times 10^{-10}$) tandis que la duplication d'un point de la dimension 1 du regroupement 1 (DUP.PC1CL1) correspond à un score $ZScore\_IQ$ réduit de 0.004 et un QI réduit de 0.06 points ($p = 6.3 \times 10^{-2}$).
La réduction d'un point de la dimension 2 du regroupement 1 (DEL.PC2CL1) correspond à un score $ZScore\_IQ$ réduit de 0.029 points ce qui correspond à une réduction du QI de 0.435 points ($p = 2.23 \times 10^{-10}$) tandis que la duplication d'un point de la dimension 2 du regroupement 1 (DUP.PC2CL1) correspond à un score $ZScore\_IQ$ réduit de 0.014 et un QI réduit de 0.21 points ($p = 2.34 \times 10^{-7}$).
La réduction d'un point de la dimension 1 du regroupement 2 (DEL.PC1CL2) correspond à un score $ZScore\_IQ$ réduit de 0.036 points ce qui correspond à une réduction du QI de 0.54 points ($p = 2.22 \times 10^{-10}$) tandis que la duplication d'un point de la dimension 1 du regroupement 2 (DUP.PC1CL2) correspond à un score $ZScore\_IQ$ réduit de 0.01 et un QI réduit de 0.15 points ($p = 4.7 \times 10^{-4}$).
La réduction d'un point de la dimension 2 du regroupement 2 (DEL.PC2CL2) correspond à un score $ZScore\_IQ$ réduit de 0.006 points ce qui correspond à une réduction du QI de 0.09 points ($p = 3.14 \times 10^{-1}$) tandis que la duplication d'un point de la dimension 2 du regroupement 2 (DUP.PC2CL2) correspond à un score $ZScore\_IQ$ réduit de 0.003 et un QI réduit de 0.045 points ($p = 3.14 \times 10^{-1}$).
La réduction d'un point de la dimension 1 du regroupement 3 (DEL.PC1CL3) correspond à un score $ZScore\_IQ$ de 0.025 points ce qui correspond à une réduction du QI de 0.375 points ($p = 1.7 \times 10^{-12}$) tandis que la duplication d'un point de la dimension 1 du regroupement 3 (DUP.PC1CL3) correspond à un score $ZScore\_IQ$ réduit de 0.008 et un QI réduit de 0.12 points ($p = 1.9 \times 10^{-4}$).
La réduction d'un point de la dimension 2 du regroupement 3 (DEL.PC2CL3) correspond à un score $ZScore\_IQ$ réduit de 0.016 points ce qui correspond à une réduction du QI de 0.24 points ($p = 1.5 \times 10^{-3}$) tandis que la duplication d'un point de la dimension 2 du regroupement 3 (DUP.PC2CL3) correspond à un score $ZScore\_IQ$ réduit de 0.006 et un QI réduit de 0.09 points ($p = 2.6 \times 10^{-2}$).
Nous remarquons aussi qu'à part la dimension 2 du regroupement 2, les effets (Est.) vont dans le même sens au sein du même regroupement que ce soit pour les délétions ou pour les duplications.
Les résultats de l'AIC sélectionnent la variable PC2CL1 comme étant la meilleure pour les délétions et PC1CL3 comme étant la meilleure pour les duplications.

Chapitre 7

Discussion

Nous rappelons que cette étude vise à rassembler l'ensemble des informations connues sur l'expression spatio- temporelle des gènes dans le cerveau afin d'en extraire un score composite pour quantifier les effets des variations du nombre de copies de gènes sur la cognition (IQ). Pour ce faire, nous combinons différents scores fonctionnels. Pour mettre en place le score composite, nous avons d'abord cherché à obtenir un jeu de données complet. Nous allons d'abord sélectionner un jeu de données complet et ensuite nous allons mettre en place un score composite.

Pour mettre en place le score composite, nous avons besoin d'un jeu de données complet concernant nos scores génétiques. Étant donné que nos scores contiennent un gros volume de données manquantes, il nous faut les gérer. Nous avons deux solutions, supprimer les données manquantes ou les imputer avec des méthodes plus élaborées. Nous avons alors opté en premier pour la suppression, ce qui nous a donné un jeu de données de scores avec données manquantes supprimées. Comme seconde alternative, nous avons imputé nos données avec MICE et Missforest et effectué une analyse comparative des deux. Ayant obtenu de meilleurs résultats avec MissForest, nous l'avons utilisé pour imputer nos données. Nous avons à présent deux jeux de données complets: un jeu de données avec données manquantes supprimées et un autre avec données manquantes imputées. Cependant, ces deux méthodes de gestion des données manquantes peuvent biaiser les analyses. En effet, en présence d'un gros volume de données manquantes, le fait de supprimer les données manquantes peut biaiser les résultats car les analyses sont effectuées sur des observations incomplètes; dans ce cas-ci, les analyses sont effectuées que sur une partie du génome. Dans le contexte où les données manquantes sont imputées, le biais est causé par le fait que la valeur qui remplace la donnée manquante n'est pas réelle mais prédite. Ne sachant pas laquelle est la meilleure pour nos analyses (causer moins de biais), nous avons adopté les deux méthodes et avons effectué une analyse comparative avec comme référence une analyse basée sur la méthodologie utilisée au laboratoire du Dr Jacquemont qui utilise le jeu de

données avec la méthode de remplacement par zéro.
Pour ce faire nous répliquons l'étude de Huguet et al. [37], qui est la méthode de référence pour l'étude des scores de façon individuelle dans le laboratoire du Dr Jacquemont. Cette étude utilise le jeu de données avec la méthode de remplacement par zéro. Après avoir annoté ce jeu de données, nous l'analysons en utilisant un modèle linéaire à effet mixte. Après correction par FDR, tous les résultats obtenus sont significatifs. Ils sont très proches des résultats de Huguet et al. [37] sachant que nous avons 18 individus de moins qu'eux (voir section 6.1) et que la carte du génome a été mise à jour depuis les publications. En effet, nous trouvons qu'un point de pLI réduit correspond à une diminution du QI de 2.55 points et un point de pLI dupliqué correspond à une diminution du QI de 0.75 points. Huguet et al.[37] trouvent qu'un point de pLI réduit correspond à une diminution du QI de 2.64 points et un point de pLI dupliqué correspond à une diminution du QI de 0.81 points. Pour le LOEUF, nous trouvons une diminution du QI de 0.39 points pour chaque point de LOEUF perdu et une diminution de 0.105 points pour chaque point de LOEUF dupliqué. L'étude de Huguet et al. [37] ont trouvé qu'un point de LOEUF réduit correspond à une diminution de 0.43 points du QI et un point de LOEUF dupliqué correspond à une diminution du QI de 0.135 points. Ces résultats sont utilisés comme référence pour sélectionner le meilleur jeu de données complet (jeu de données supprimées ou imputées).
Les deux jeux de données complets sont annotés puis analysés suivant la méthode d'annotation et d'analyse de Huguet et al. [37]. Tous les résultats obtenus sont significatifs (p-value < 0.05). Nous comparons alors ces résultats avec ceux obtenus avec la méthode de remplacement par zéro et nous utilisons le critère d'information de Akaike qui permet d'évaluer les modèles pour sélectionner le meilleur modèle. Nous notons qu'entre les données supprimées et les données imputées, ce sont ces dernières qui ont des résultats très proches de ceux des données avec la méthode de remplacement par zéro; que ce soit pour le LOEUF ou le pLI, pour les délétions ou les duplications. En effet, nous avons presque les mêmes estimées, les mêmes p-value et les critères d'information de Akaike (Akaike information criterion) sont similaires à une virgule près.
L'approche avec données manquantes supprimées a fourni des résultats avec un gap assez conséquent par rapport à la méthode de remplacement par zéro. De plus concernant la concordance avec la littérature, nous obtenons une bonne concordance et les résultats de concordance de la méthode avec imputation et ceux de la méthode de remplacement par zéro sont très similaires. Ce qui laisse penser que notre méthode d'imputation n'introduit pas de biais par rapport à l'étude précédente [37].
Le modèle avec les données imputées est sélectionné comme étant le meilleur modèle car il ne rajoute pas du bruit à notre modèle (comparé au modèle avec la méthode de remplacement par zéro) contrairement au modèle avec données manquantes supprimées. Avec cette dernière, nous avons perdus beaucoup de gènes alors qu'avec le modèle avec les

données imputées, nous arrivons à garder toutes nos informations génétiques sans rajouter du bruit à notre modèle (vu que nous obtenons des résultats similaires à ceux de l'étude précédente).
En outre, étant dans le cadre d'intéractions géniques additives, la méthode de remplacement par zéro sous-estime le score d'un gène (en remplaçant la donnée manquante d'un score par zéro) contrairement à la méthode avec les données imputées. Nous notons alors qu' avec la méthode de remplacement par zéro, nous passons à côté d'informations importantes. Par ailleurs, le modèle avec les données manquantes supprimées nous a permis de constater que la perte des gènes sur-estime la perte de QI. En effet, les résultats des analyses montrent une plus grande perte de QI pour le jeu de données avec données manquantes supprimées ensuite pour le jeu de données avec la méthode de remplacement par zéro. L'approche par imputation et par suppression de données nous amène à croire que la méthode de remplacement par zéro qui est la méthode utilisée dans toutes les études effectuées au laboratoire du Dr Jacquemont sur-estime le QI car elle n'utilise pas la totalité du génome et que le poids des gènes manquants est alloué aux autres gènes présents. Il serait intéressant de le vérifier en faisant des études supplémentaires.
L'imputation des données manquantes nous permet alors d'avoir un jeu de données complet en gardant tous les gènes sans rajouter du bruit à notre modèle. En d'autres termes, nous arrivons à estimer la perte de QI en utilisant la totalité du génome.

Notre objectif est de mettre en place un score composite. Pour ce faire, nous utilisons le meilleur modèle fournit par l'analyse précédente c'est-à-dire le modèle avec le jeu de données imputées. Nous effectuons une approche de score composite avec la construction réflexive (voir section 2.2.2.2.2) en utilisant l'analyse par composante principale. Nous séparons nos scores en regroupements et nous effectuons l' analyse par composante principale sur chaque regroupement. Les résultats de cette analyse nous ont permis de réduire les 15 scores que nous avions, dans le jeu de données imputé, en 6 scores qui expliquent le plus la variance contenue dans les 15 scores. Nous obtenons ainsi les scores PC1CL1, PC2CL1, PC1CL2, PC2CL2, PC1CL3, PC2CL3.
Sur ces scores, nous effectuons une annotation fonctionnelle puis une analyse en utilisant le même modèle linéaire à effet mixte (voir section 6.3.3). Nos résultats sont significatifs sauf pour la dimension 2 du regroupement 2. Ces résultats nous permettent d'étudier les effets des CNV sur le QI en tenant en compte plusieurs scores contrairement à l'étude précédente qui le faisait de façon individuelle. De plus, comme dans les études précédentes [36,37], pour les délétions par rapport aux duplications, nous avons un effet différent et la taille de l'effet pour les délétions est supérieure à celle des duplications. Pour PC1CL1 et PC2CL2 qui regroupent les scores de l'expression spatiale des gènes au niveau cérébral, les scores d'expression du cortex de Paus et al. et de Burt et al. [49,50], nous voyons que ces

scores combinés au niveau de PC1CL1 correspondent à une réduction du QI de 0.33 points pour chaque point réduit et de 0.06 points pour chaque point dupliqué, et au niveau de PC2CL1 ils correspondent à une réduction du QI de 0.435 points pour chaque point réduit et de 0.21 points pour chaque point dupliqué. Pour le PC1CL2 et PC2CL2 qui regroupent le score d'évolution de Dumas et les scores de contraintes génétiques, nous remarquons que ces scores combinés au niveau de PC1CL2 correspondent à une réduction du QI de 0.54 points pour chaque point réduit et de 0.15 points pour chaque point dupliqué. Pour la PC2CL2, nos résultats ne sont pas significatifs. Enfin nous avons le groupe de scores PC1CL3 et PC2CL3 qui regroupe les scores de la stabilité différentielle par tissus et les scores d'expressions temporelles. Dans ce groupe, nous voyons que pour PC1CL3, un point réduit correspond à une réduction du QI de 0.375 points et un point dupliqué correspond à une réduction du QI de 0.12 points tandis que pour PC2CL3, un point réduit correspond à une réduction du QI de 0.24 points et un point dupliqué correspond à une réduction du QI de 0.09 points. Ceci est très intéressant car ça nous permet de voir comment les scores ont été combinés pour mesurer les effets des CNV sur le QI. Nous voyons qu'au lieu de tenir en compte qu'une dimension biologique, nous en combinons plusieurs.
En effet, chaque score rempli une fonction biologique particulière et a sa propre nature, en les combinant au lieu de les étudier de façon individuelle, nous prenons en compte la fonction biologique et la nature de chaque score. L'étude précédente de Huguet et al. [37] ne prenait en compte qu' une seule fonction en étudiant les scores individuellement. Nos résultats sont donc basés sur l'impact de plusieurs scores au lieu d'un score. Nous remarquons un effet différentiel dans les résultats de la perte de QI (voir 6.19 et 6.20) par rapport à l'étude des scores de façon individuelle. En effet, comparé aux résultats de l'étude des scores imputés(voir 6.9) étudiés de façon individuelle, nous remarquons une plus importante réduction du QI. Ces résultats suggèrent que le fait de prendre en compte qu'une seule fonction biologique sur-estime la perte de QI.

Interprétation dans un contexte clinique

Dans le contexte clinique, cette étude peut aider à estimer la dangerosité du variant CNV. Est-ce que le variant est suffisamment pathogène au point d'expliquer le phénotype observé (dans cet étude le QI)? La quantification des effets des CNV sur le cognitif permettra aux cliniciens d'estimer la contribution de ces variants aux symptômes neuro-développementaux chez un patient donné. Si les résultats ne sont pas significatifs et qu'il n'y a pas d'association entre le variant et le phénotype, le clinicien devrait penser à investiguer plus sur la génétique de l'individu. Cependant, ce n'est pas un outil diagnostique mais un outil d'aide à la prise de décision.
Dans le laboratoire du Dr Jacquemont, les cliniciens ont une application basée sur les études précédentes qui permet au clinicien d'estimer la dangerosité du variant CNV par rapport au

phénotype observé. Cette étude est une continuité de ces études précédentes.

Limites de l'étude

Dans cette étude, nous avons six scores composites et chaque score composite est lié à un groupe de fonctions de gènes. Notre étude aurait été plus complète si on était parvenu à combiner tous les 6 scores en un seul score en utilisant des pondérations. En outre la méthode d'analyse par composante principale (ACP) cause du biais. En effet, le fait de choisir que trois composantes alors que nous en avons plusieurs nous fait perdre de l'information. Dans l'étude future, l'idée est de trouver des coefficients de pondération optimaux ce qui correspond à la méthode de construction formative (voir 2.2). Si nous n'utilisons pas la méthode de construction réflexive, nous pourrons nous passer des ACP. Trouver les bons coefficients de pondérations reste donc un défi.

Le modèle actuel utilise à la fois les variants *de novo* et les variants hérités. Dans la littérature [37], on sait que l'effet des variants *de novo* est plus fort que l'effet des CNV héritées. Ce qui fait qu'on sous-estime l'effet des variants *de novo* et on sur estime l'effet des variants hérités. Dans des travaux futurs, il serait intéressant de séparer ces deux variants afin de mieux quantifier les effets des CNV sur la cognition.

Recommandations pour les chercheurs futurs

L'idée dans cette étude est d'améliorer le score composite pour qu'elle fasse l'objet d'une publication. Dans le cas d'une éventuelle publication, le score sera mis à la disposition des chercheurs. Cependant tout n'est pas à refaire à chaque fois. Ce sont les estimations obtenues et validées par un comité qui seront utilisées. Par exemple dans l'étude de Huguet et al, un point de pLI réduit est associé à une réduction du QI de 2.74 points [36] ; ceci est validé par un comité et donc utilisé dans chaque étude (nous n'avons plus besoin de refaire les analyses pour trouver cette valeur).

Chapitre 8

Conclusion et perspectives

En définitive, cette étude nous a permis de voir que les scores génétiques peuvent être combinés et que cette combinaison permet de prendre en compte plusieurs dimensions biologiques dans l'estimation des effets des CNV sur la cognition. En utilisant un modèle linéaire sur cette combinaison des scores génétiques, nos résultats suggèrent que l'étude de la quantification des effets des CNV sur le QI, grâce aux scores génétiques utilisés de façon individuelle, sur-estime la perte de QI. En effet, en prenant en compte plusieurs fonctions biologiques, nous remarquons un effet différentiel dans les résultats de la perte de QI par rapport à l'étude effectuée lorsqu'on prend en compte qu'une seule fonction biologique.

Nous prévoyons continuer cette étude afin d'obtenir un seul score qui combine tous les autres grâce à des coefficients de pondérations. Nous prévoyons également répliquer cette étude sur les variants SNV (single nucleotide variant) qui correspondent à la substitution d'un nucléotide par un autre. En effet, nous avons déjà commencé à étudier ces derniers en mettant en place des filtres afin de sélectionner les SNV rares conduisant à une perte de fonction. Une fois ces aspects pris en compte, cet effort à grande échelle pour caractériser les CNV pourra aider le clinicien pour l'interprétation des tests de diagnostic génétique et éclairer la prise de décision dans les cliniques de neuro-développement. La quantification des effets des CNV sur le cognitif permettra aux cliniciens d'estimer la contribution de ces variants aux symptômes neuro-développementaux chez un patient donné.

Annexe A

Imputation

A.1. Simulation pour le choix des paramètres des méthodes d'imputations

Cette section représente un aperçu des différents tests effectués pour choisir les paramètres pour MICE et MissForest. Les tests sont effectués avec un jeu de données contenant 31960 gènes et 41 scores dont 29.7% des données sont manquantes. La métrique RMSE donnant d'énormes erreurs par rapport à la métrique MAE comme nous l'avons vu dans le chapitre 6.3.3, nous allons nous baser sur cette métrique et sur le temps d'exécution. Nous avons ordonné les tableaux suivant la combinaison qui donne de meilleurs résultats.

A.1.1. Simulation pour le choix des paramètres de MICE

Ces simulations sont effectuées pour sélectionner les paramètres m et maxit. Nous rappelons que la méthode utilisée est PMM (Predictive Mean Matching). Nous avons m=5,10,20,25,50,100. Pour chaque m, nous avons effectué des combinaisons avec maxit=5 puis 10 20, 25, 50,100. Nous nous retrouvons avec 36 résultats. Cependant nous ne retenons que les combinaisons qui ont donné de meilleurs résultats.
En rouge nous avons la combinaison qui a donné de meilleurs résultats pour le RMSE et en bleu celui qui a de meilleurs résultats pour le temps d'exécution. Nous avons retenu la combinaison m=5 et maxit=20 car pour le RMSE il vient en deuxième et est très proche de celle qui a donnée de meilleurs résultats (m=50, maxit=10). Étant donné que les résultats pour le RMSE sont très proches et que m=50, maxit= 10 donne un temps de computation assez élevé, nous avons choisi m=5, maxit=20.

Method=pmm;

- m=5,10,25,50,100
  - ✓ maxit=5,10,20,25,50,100

m	maxit	RMSE Global
50	10	18,66451
5	20	18,69802
50	25	18,8535
50	20	19,076
25	50	19,17101
5	25	19,84056

m	maxit	TIME
5	20	7,54
5	25	10,43
50	10	34,97
50	20	67,62
50	25	83,54
25	50	86,06

Fig. A.1. Simulation avec MICE. m=nombre d'imputations multiples, maxit=nombre maximal d'itérations pour chaque imputation multiple, RMSE=root mean square error, time= temps d'exécution. En rouge nous avons la combinaison qui donne le meilleur pour RMSE et en bleu celui qui donne le meilleur pour le temps.

A.1.2. Simulation pour le choix des paramètres de MissForest

Ces simulations sont effectuées pour sélectionner les paramètres ntree et maxit. Nous avons ntree=100, 150, 200, ...

Pour chaque ntree, nous avons effectué des combinaisons avec maxit=5 puis 10, 20, 25, 50,1 00. Nous avons effectué des combinaisons avec maxit et ntree; cependant nous nous sommes vite rendus compte qu'à partir de 100, ntree est stable c'est-à-dire que pour une valeur de maxit donnée, si on change le ntree après 100, le résultat ne change plus. Nous ne retenons que les combinaisons qui ont donné de meilleurs résultats.

En bleu nous avons la combinaison qui a donné de meilleurs résultats. En effet pour MissForest, la meilleure combinaison est celle à partir de laquelle les résultats sont stables; c'est-à-dire pour ntree= 100 le résultat reste le même si on augmente la valeur de maxit dans les autres combinaisons (maxit=5 et ntree=100). Cependant c'est la combinaison m=15 et ntree=100 qui donne de meilleurs résultats pour le temps. Étant donné que les résultats pour le temps sont très proches pour m=5, ntree=100 et n=15, ntree= 100, nous gardons la combinaison à partir de laquelle les résultats sont stables.

- ntree=100,150,200,...
  - ✓ maxit=5,10,15,25,50

maxit	ntree	RMSE Global
5	100,150,200,...	16,53818
10	100	16,38195
15	100	16,38195
20	100	16,38195
25	100	16,38195

maxit	ntree	TIME
5	100	5,65
10	100	5,61
15	100	5,47
20	100	5,83
25	100	5,65

Fig. A.2. Simulation avec MissForest. ntree=nombre d'arbres générés aléatoirement à chaque itération, maxit=nombre maximal d'itérations pour chaque imputation multiple, RMSE=root mean square error, time= temps d'exécution. En bleu nous avons la combinaison qui donne de meilleurs résultats.

A.2. Résultats des 2000 gènes pour les regroupement 2,4 et 5

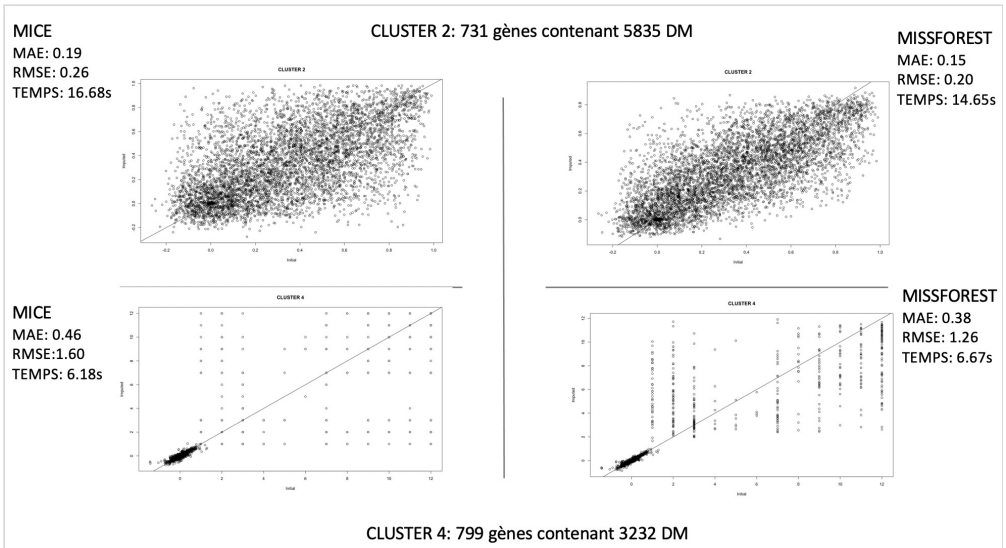


Fig. A.3. Résultats de l'imputation pour le jeu de données manquantes M. Aperçu des regroupements 2 et 4. Ici nous avons tracé la droite d'équation y=aX + b afin de faire une comparaison entre les valeurs imputées et les vraies valeurs issues du jeu de données C. En abscisse nous avons les vraies valeurs (initiales) et en ordonnée nous avons les valeurs imputées. Nous avons pour chaque méthode, le résultats des métriques temps d'exécution, RMSE et MAE. Dans le regroupement 2, nous avons 731 gènes et 16 scores ce qui nous donne 11696 valeurs. Parmi elles, nous avons 5835 données manquantes. Dans le regroupement 4, nous avons 799 gènes et 9 scores ce qui nous donne 7191 valeurs. Parmi elles, nous avons 3232 données manquantes.

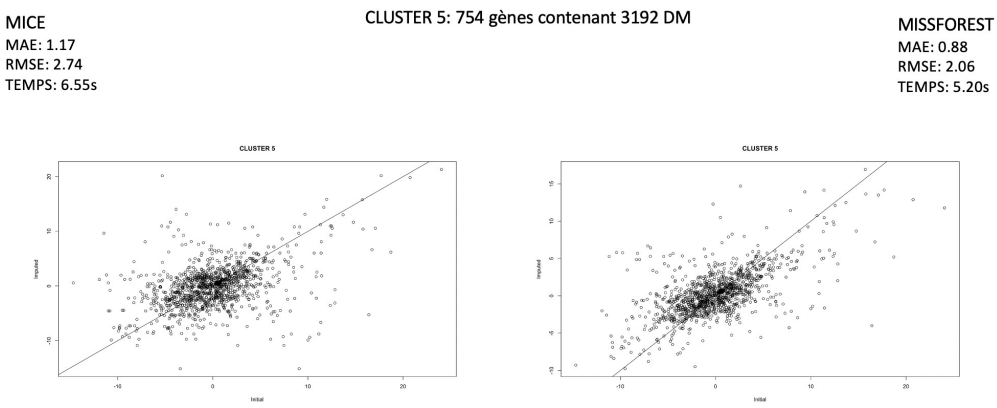


Fig. A.4. Résultats de l'imputation pour le jeu de données manquantes M. Aperçu du regroupement 5. Ici nous avons tracé la droite d'équation y=aX + b afin de faire une comparaison entre les valeurs imputées et les vraies valeurs issues du jeu de données C. En abscisse nous avons les vraies valeurs (initiales) et en ordonnée nous avons les valeurs imputées. Nous avons pour chaque méthode, le résultats des métriques temps d'exécution, RMSE et MAE. Dans le regroupement 5, nous avons 754 gènes et 9 scores ce qui nous donne 6786 valeurs. Parmi elles, nous avons 3192 données manquantes.

A.3. Annexe pour le score composite

Cette figure présente les résultats de la perte de QI si on utilise un modèle additif et qu'on addition tous les scores de délétions dans un modèle et tous les scores de duplications dans un autre contrairement au modèle que nous avons présenté qui utilise un score à la fois.

	Est.	Lower	Upper	Pvalue	AIC
DEL.PC1CL1	-0.021406321	-0.027985583	-0.014827059	1.889713e-10	71430.56
DEL.PC2CL1	-0.028457328	-0.037346116	-0.019568541	3.652015e-10	71430.56
DEL.PC1CL2	-0.036126520	-0.046825772	-0.025427268	3.838027e-11	71467.24
DEL.PC2CL2	0.008242870	-0.003983507	0.020469248	1.863976e-01	71467.24
DEL.PC1CL3	-0.024552568	-0.031335779	-0.017769357	1.392281e-12	71452.32
DEL.PC2CL3	-0.012474431	-0.022204493	-0.002744369	1.199360e-02	71452.32
DUP.PC1CL1	-0.002289263	-0.006157797	0.001579272	2.461356e-01	71485.96
DUP.PC2CL1	-0.013695233	-0.018955040	-0.008435426	3.401836e-07	71485.96
DUP.PC1CL2	-0.009514787	-0.014794105	-0.004235470	4.136770e-04	71501.17
DUP.PC2CL2	-0.001862507	-0.008282991	0.004557977	5.696594e-01	71501.17
DUP.PC1CL3	-0.008161448	-0.012126436	-0.004196460	5.516816e-05	71494.18
DUP.PC2CL3	-0.006293403	-0.011157688	-0.001429118	1.123364e-02	71494.18

Fig. A.5. Résultats du QI pour les composantes principales pour le modèle qui regroupe tous les scores de délétions d'une part et tous les scores de duplications d'autre part. Est.= estimé, p-value, AIC= Akaike information criterion, Lower= borne inférieure de l'intervalle de confiance, Upper= borne supérieure de l'intervalle de confiance. Nous avons un niveau de confiance 95% ce qui signifie que nous sommes sûr à 95% que la valeur de l'estimée (Est.) est comprise entre la valeur du Lower et la valeur du Upper.

Cette figure présente les résultats de la perte de QI si on utilise un modèle additif et qu'on additionne tous les scores de délétions et de duplication dans un seul modèle contrairement au modèle que nous avons présenté dans le chapitre 6.3 qui utilise un score à la fois.

	Est.	Lower	Upper	Pvalue	AIC
DEL.PC1CL1	-1.902104e-02	-0.026181632	-0.011860458	1.965402e-07	71560.36
DEL.PC2CL1	-1.733465e-02	-0.028192675	-0.006476633	1.758720e-03	71560.36
DEL.PC1CL2	-1.832352e-02	-0.030651248	-0.005995785	3.584967e-03	71560.36
DEL.PC2CL2	7.984198e-03	-0.004694173	0.020662570	2.171185e-01	71560.36
DEL.PC1CL3	-1.014664e-02	-0.018633063	-0.001660224	1.912745e-02	71560.36
DEL.PC2CL3	2.257823e-05	-0.011235747	0.011280904	9.968638e-01	71560.36
DUP.PC1CL1	-1.038475e-03	-0.005345446	0.003268496	6.365201e-01	71638.93
DUP.PC2CL1	-1.072138e-02	-0.016731792	-0.004710962	4.740198e-04	71638.93
DUP.PC1CL2	-4.377488e-03	-0.010407662	0.001652687	1.548203e-01	71638.93
DUP.PC2CL2	-1.283068e-03	-0.007820085	0.005253949	7.004657e-01	71638.93
DUP.PC1CL3	-3.114922e-03	-0.008000231	0.001770386	2.114343e-01	71638.93
DUP.PC2CL3	-1.480175e-03	-0.007414540	0.004454190	6.249439e-01	71638.93

Fig. A.6. Résultats du QI pour les composantes principales pour le modèle qui regroupe tous les scores de délétions et tous les scores de duplications dans un seul modèle. Est.= estimé, p-value, AIC= Akaike information criterion, Lower= borne inférieure de l'intervalle de confiance, Upper= borne supérieure de l'intervalle de confiance. Nous avons un niveau de confiance 95% ce qui signifie que nous sommes sûr à 95% que la valeur de l'estimée (Est.) est comprise entre la valeur du Lower et la valeur du Upper.

Bibliographie

1. Calier, M. et al. *Les sciences cognitives et l'école,* 9-54, 2003. Récupéré de https://www.cairn.info/les-sciences-cognitives-et-l-ecole–9782130534976-page-9.htm#

2. Dima D. C. et al. *Electrophysiological network Translational Psychiatry alterations in adults with copy number variants associated with high neurodevelopmental risk,* 10(1) - 324, 2020.

3. Dixneuf P. *Analyse de la performance de la méthode d'imputation de données manquantes missForest et application à des données environnementales,* Mémoire de maîtrise en génie de l'environnement, Université du Quebec, 2019. Récupéré de https://espace.etsmtl.ca/id/eprint/2360/2/$DIXNEUF\_Paul$-web.pdf

4. CPA Webmaster. *Série "La psychologie peut vous aider ": Troubles Cognitifs Et Démence, Société Canadienne de psychologie,* 2020. Récupéré de https://cpa.ca/fr/psychology-works-fact-sheet-cognitive-disorders-and-dementia/

5. George D. et al. *SPSS for Windows step by step: a simple guide and reference,* 11.0 update (14th ed.), Boston, MA: Allyn and Bacon, 2016.

6. Hehir-Kwa J. Y. et al. *Exome sequencing and whole genome sequencing for the detection of copy number variation,* Expert Review of Molecular Diagnostics 15(8) : 1023-1032, 2015.

7. Jiang, L., Huguet, G., Schramm, C., et al. *Estimating the effects of copy-number variants on intelligence using hierarchical Bayesian models,* Genet Epidemiol, 44(8), 825-840, 2020.

8. Jabir M. *Comparaison de méthodes d'imputation des données manquantes appliquées à la base nationale sur les collisions,* mémoire de maîtrise en sciences de gestion, HEC Montréal, 2018. Récupéré de https://biblos.hec.ca/biblio/memoires/m2018a609812.pdf

9. Jones R et al. *The potential of composite cognitive scores for tracking progression in Huntington's disease,* Journal of Huntington's Disease 3(2) : 197-207, 2014.

10. Kim S. H. et al. *Language characterization in 16p11.2 deletion and duplication syndromes,* American Journal of Medical Genetics Part B Neuropsychiatric Genetics 183(6) : 380-391, 2020.

11. Sanders S. J. et al. *Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci,* Neuron 87(6) : 1215-1233, 2015.

12. Morrow E. M. *Genomic copy number variation in disorders of cognitive development,*

American Academy of Child and Adolescent Psychiatry 49(11) : 1091-1104, 2010.
13. Lautenschlager N. T. et al. *Risk of dementia among relatives of Alzheimer's disease patients in the MIRAGE study: What is in store for the oldest old?,* Neurology 46(3) : 641-650, 1996.
14. Lee C. et al. *Copy number variations and clinical cytogenetic diagnosis of constitutional disorders,* Nature Genetics 39 : S48-S54, 2007.
15. Le syndrome de l'X fragile. *Qu'est-ce que le Syndrome de l'X fragile,* McGill. Récupéré le 04 novembre 2020 de https://www.mcgill.ca/buildinglinks/fr/troubles-et-syndromes/xfragile
16. Lupski J. R. *Cognitive phenotypes and genomic copy number variations,* JAMA 313(20) : 2029-2030, 2015.
17. Nassar H. M. et al. *Comparison of weighted and composite scores for pre-clinical dental learners,* European Journal of Dental Education 22(3) : 192-197, 2018.
18. Peretti C. S. et al. *Cognitive skill learning in healthy older adults after 2 months of double-blind treatment with piribedil,* Psychopharmacology (Berl) 176(2) : 175-181, 2004.
19. Potier M. *Cognitive deficits in Down syndrome, from birth to dementia: mechanisms and treatments,* Bulletin de l'Académie Nationale de Médecine, 200(8-9): 1543-1557, 2016.
20. Romana S. et al. *Cytogénétique moléculaire, Collége National des Enseignants et Praticiens de Génétique Médicale,* 2012. Récupéré de http://campus.cerimes.fr/genetique-medicale/enseignement/genetique19/site/html/cours.pdf
21. Nakatochi, M. et al. *Implications of germline copy-number variations in psychiatric disorders: review of large-scale genetic studies,* J Hum Genet, 66(1), 25-37, 2021.
22. Sénéchal G. *Rôle de la neuropsychologie dans les études génétiques : caractérisation détaillée d'une famille multiplexée avec épilepsie et troubles d'apprentissage,* Doctorat en psychologie, Université de Montréal, 2011. Récupéré de https://archipel.uqam.ca/4643/1/D2195.pdf
23 Valsesia, A. et al. *The Growing Importance of CNVs: New Insights for Detection and Clinical Interpretation,* Front Genet, 4, 92, 2013.
24. Société québécoise de l'autisme. *Vue sur les particularités neurobiologiques et cognitives des autistes,* 1999. Récupéré le 04 novembre 2020 de https://www.autisme.qc.ca/tsa/recherche/etiologie/particularites-neurobiologiques-et-cognitives-des-autistes.html
25. Meigs, J. B. *The Genetic Epidemiology of Type 2 Diabetes: Opportunities for Health Translation,* Curr Diab Rep, 19(8), 62, 2019.
26. Dang, V., Kassahn, K., Marcos, A. et al. *Identification of human haploinsufficient genes and their genomic proximity to segmental duplications,* Eur J Hum Genet 16, 1350–1357, 2008.
27. Mountford, H. S. et al. *Copy number variation burden does not predict severity of*

neurodevelopmental phenotype in children with a sex chromosome trisomy, Am. J. Med. Genet. Part C 184, 256–266, 2020.
28. Yamasaki, M. et al. *Sensitivity to gene dosage and gene expression affects genes with copy number variants observed among neuropsychiatric diseases,* BMC Med Genomics 13, 55, 2020.
29. Stefansson, H. et al. *CNVs conferring risk of autism or schizophrenia affect cognition in controls,* Nature, 505(7483), 361-366, 2014.
30. MacLeod, A. K. et al. *Genetic copy number variation and general cognitive ability,* PLoS One, 7(12), e37385, 2012.
31. Mannik, K. et al. *Copy number variations and cognitive phenotypes in unselected populations,* JAMA, 313(20): 2044-2054, 2015.
32. Thygesen, J. H. et al. *Genetic copy number variants, cognition and psychosis: a meta-analysis and a family study,* Mol Psychiatry, 2020.
33. Kendall, K. M. et al. *Cognitive performance and functional outcomes of carriers of pathogenic copy number variants: analysis of the UK Biobank,* Br J Psychiatry, 214(5), 297-304, 2019.
34. Guillaume Huguet, Catherine Schramm, Elise Douard et al. *SA47 - QUANTIFYING THE EFFECT OF COPY-NUMBER VARIANTS ON GENERAL INTELLIGENCE IN UNSELECTED POPULATIONS,* European Neuropsychopharmacology, Volume 29, Supplement 3, 2019.
35. Lappalainen, T. et al. *Genomic Analysis in the Age of Human Genome Sequencing,* Cell, 177(1), 70-84, 2019.
36. Huguet, G., Schramm, C., Douard, E. et al. *Measuring and Estimating the Effect Sizes of Copy Number Variants on General Intelligence in Community-Based Samples,* JAMA Psychiatry, 75(5), 447-457, 2018.
37. Huguet, G., Schramm, C., Douard, E. et al. *Genome-wide analysis of gene dosage in 24,092 individuals estimates that 10,000 genes modulate cognitive ability,* Mol Psychiatry, 2021.
38. Douard, E., Zeribi, A., Schramm, C. et al. *Effect Sizes of Deletions and Duplications on Autism Risk Across the Genome,* Am J Psychiatry, 178(1), 87-98, 2021.
39. Robert Nussbaum et al. *THOMPSON AND THOMPSON GENETICS IN MEDICINE,* Elsevier Health Sciences, 2015.
40. Zeribi A. *Contribution différentielle des variations du nombre de copies aux troubles du spectre autistique et aux traits cognitifs,* Mémoire de maîtrise en sciences biomédicales, Université de Montréal, 2018. Récupéré de https://papyrus.bib.umontreal.ca/xmlui/bitstream/handle/1866/21376/*Zeribi_Abderrahim*_2018_*mem*
41. Breiman, L. *Random Forests,* Machine Learning 45, 5–32, 2001.
42. Salgado C.M. et al. *Missing Data. In: Secondary Analysis of Electronic Health Records,*

Springer, Cham, 2016.
43. Do, K. T. et al. *Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies,* Metabolomics, 14(10), 128, 2018.
44. Shahla Faisal, Gerhard Tutz. *Multiple imputation using nearest neighbor methods,* Information Sciences, Volume 570, 2021.
45. Crane, P. K., Carle, A. et al. *Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI),* Brain Imaging Behav, 6(4), 502-516, 2012.
46. Langfelder, P., Zhang, B., et Horvath, S. *Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R,* Bioinformatics, 24(5), 719-720, 2008.
47. van Buuren, S., et Groothuis-Oudshoorn, K. *mice: Multivariate Imputation by Chained Equations in R,* Journal of Statistical Software, 45(3), 1 - 67, 2011.
48. Aguert, M., et Capel, A. *Mieux comprendre les scores z pour bien les utiliser,* Rééducation Orthophonique, 274, 61-85, 2018.
49. French, Leon, and Tomás Paus. *A FreeSurfer view of the cortical transcriptome generated from the Allen Human Brain Atlas,* Frontiers in neuroscience 9, 2015.
50. Burt, J. B. et al. *Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography,* Nature neuroscience, 21(9), 1251-1259, 2018.
51. Kang, Hyo Jung, et al. *Spatio-temporal transcriptome of the human brain,* Nature 478.7370 : 483-489, 2011.
52. Dumas, Guillaume, Simon Malesys, and Thomas Bourgeron. *Systematic detection of brain protein-coding genes under positive selection during primate evolution and their roles in cognition,* Genome Research, 2021.
53. Urchs, S., Armoza, J., Moreau, C. et al. *MIST: A multi-resolution parcellation of functional brain networks,* MNI Open Research, 1, 3, 2019.
54. Hawrylycz, Michael, et al. Canonical genetic signatures of the adult human brain, Nature neuroscience 18.12 : 1832, 2015.
55. Marc Guerrien. *L'intérêt de l'analyse en composantes principales (ACP) pour la recherche en sciences sociales,* Cahiers des Amériques latines [En ligne], 43 | 2003, mis en ligne le 10 août 2017, consulté le 29 août 2021. URL : http://journals.openedition.org/cal/7364 ; DOI : https://doi.org/10.4000/cal.7364
56. Chawner, S., Owen, M. J., Holmans, P., et al. *Genotype-phenotype associations in children with copy number variants associated with high neuropsychiatric risk in the UK (IMAGINE-ID): a case-control cohort study,* Lancet Psychiatry, 6(6), 493-505, 2019.
57. Vorstman, J. A. S., Parr, J. R., et al. *Autism genetics: opportunities and challenges for clinical translation,* Nat Rev Genet, 18(6), 362-376, 2017.
58. Kirov, G., Grozeva, D., Norton, N., Ivanov, D. et al. *Support for the involvement of*

large copy number variants in the pathogenesis of schizophrenia, Hum Mol Genet, 18(8), 1497-1503, 2009.
59. Fernandez-Blanco, A., et Dierssen, M. *Rethinking Intellectual Disability from Neuro- to Astro-Pathology,* Int J Mol Sci, 21(23), 2020.
60. Iwase, S., Berube, N. G., et al. *Epigenetic Etiology of Intellectual Disability,* J Neurosci, 37(45), 10773-10782, 2017.
61. Kendall, K. M., Bracher-Smith, M., Fitzpatrick, H. et al. *Cognitive performance and functional outcomes of carriers of pathogenic copy number variants: analysis of the UK Biobank,* Br J Psychiatry, 214(5), 297-304, 2019.
62. Ringnér, M. *What is principal component analysis?,* Nat Biotechnol 26, 303–304, 2008.
63. Pearson, K. *Determination of the Coefficient of Correlation.,* Science, 30(757), 23-25, 1909.
64. Lanchbury, J. S. *The Human Genome Project.,* Rheumatology, Volume 37, Issue 2, Pages 119–121, 1998.
65. Watson, J. D., et Crick, F. H. *The structure of DNA.,* Cold Spring Harb Symp Quant Biol, 18, 123-131, 1953.
66. Vidhya A. *Tutorial on 5 Powerful R Packages used for imputing missing values,* 2016. Récupéré de https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values.
67. Williams P. A. et al. *Communicating efficacy information based on composite scores in direct-to-consumer prescription drug advertising,* Patient Educ Couns 99(4) : 583-590, 2016.
68. Writing Committee for the Enigma et al. *Association of Copy Number Variation of the 15q11.2 BP1-BP2 Region With Cortical and Subcortical Morphology and Cognition,* JAMA Psychiatry 77(4) : 420-430, 2020.
69. Shwartz M. et al. *Composite Measures of Health Care Provider Performance : A Description of Approaches,* Milbank Q 93(4) : 788-825, 2015.
70. Miller DT, Adam MP, Aradhya S, et al. *Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies,* Am J Hum Genet, 86(5):749-764, 2010.
71. Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie et al. *Missing value estimation methods for DNA microarrays ,* Bioinformatics, Volume 17, Issue 6, Pages 520–525, June 2001,.
72. Daniel J. Stekhoven, Peter Bühlmann. *MissForest—non-parametric missing value imputation for mixed-type data,* Bioinformatics, Volume 28, Issue 1, Pages 112–118, January 2012.
73. Akaike H. *Akaike's Information Criterion,* In: Lovric M. (eds) International Encyclopedia of Statistical Science, Springer, Berlin, Heidelberg, 2011.
74. Morrill, S. A., er Amon, A. *Why haploinsufficiency persists?,* Proc Natl Acad Sci U S

A, 116(24), 11866-11871, 2019.
80. CDC, Findings : *Developmental Disabilities Prevalence Trends j CDC,* Feb. 2015.
81. Chung, B. H., Tao, V. Q. et al. *Copy number variation and autism: new insights and clinical implications,* J Formos Med Assoc, 113(7), 400-408, 2014.
82. Alkan, C., Coe, B. P.,et al. *Genome structural variation discovery and genotyping,* Nat Rev Genet, 12(5), 363-376, 2011.
83. Feuk, L., Carson, A. R., et Scherer, S. W. *Structural variation in the human genome,* Nat Rev Genet, 7(2), 85-97, 2006.
84. McRae, A. et al. *No Association Between General Cognitive Ability and Rare Copy Number Variation,* Behavior Genetics 43(3) : 202-207, 2013.
85. Girirajan, S. et al. *Phenotypic heterogeneity of genomic disorders and rare copy-number variants,* N Engl J Med 367(14) : 1321-1331, 2012.
86. Tyner C. E. et al. *Development of Composite Scores for the TBI-QOL,* Arch Phys Med Rehabil 101(1) : 43-53, 2020.
87. Pang, A.W., et al. *Towards a comprehensive structural variation map of an individual human genome,* Genome Biol 11, R52, 2010.
88. McGill *"Le syndrome de l'X fragile Qu'est-ce que le Syndrome de l'X fragile?",* consulté le 29 Aôut 2021.
89. Statistics solution, *Composite Scoring and Reliability*, récupéré le 04 novembre 2020 de https://www.statisticssolutions.com/composite-scoring-and-reliability/
90 Martins TD, Annichino-Bizzacchi JM, Romano AVC, Filho RM. *Principal Component Analysis on Recurrent Venous Thromboembolism,* Clinical and Applied Thrombosis/Hemostasis, January 2019.
91. Shuangge Ma, Ying Dai. *Principal component analysis based methods in bioinformatics studies,* Briefings in Bioinformatics, Volume 12, Issue 6, Pages 714–722, November 2011.
92. Tammimies K. et al. *Molecular Diagnostic Yield of Chromosomal Microarray Analysis and Whole-Exome Sequencing in Children With Autism Spectrum Disorder,* JAMA 314(9) : 895-903, 2015.
93. Kaufman AS, Flanagan DP, Alfonso VC, Mascolo JT. *Test Review: Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV),* J Psychoeduc Assess.,24:278–295, 2006.
94. Canivez GL, Watkins MW. *Long-Term Stability of the Wechsler Intelligence Scale for Children-Third Edition among Demographic Subgroups: Gender, Race/Ethnicity, and Age,* J Psychoeduc Assess.,17:300–313, 1999.
95. Ensor RCK. *The trend of Scottish intelligence: a comparison of the 1947 and 1932 surveys of the intelligence of eleven-year-old pupils,* Eugen Rev. ,41:196–197, 1950.
96. Deary IJ, Gow AJ, Taylor MD, Corley J, Brett C, Wilson V, et al. *The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and*

beyond, BMC Geriatr. ,7:28, 2007.
97. Akshoomoff N. *Use of the Mullen Scales of Early Learning for the Assessment of Young Children with Autism Spectrum Disorders,* Child Neuropsychol J Norm Abnorm Dev Child Adolesc.,12:269–277, 2006.
98. Leiter RG. *Leiter international performance scale,* 1979.
99. Gale H. Roid, Miller LJ. *Leiter international performance scale-revised,* 1997.
100. Raven JC, Court JH, Raven J. *Raven's Progressive Matrices,* 1998.
101. Coolican J, Bryson SE, Zwaigenbaum L. *Brief report: data on the Stanford-Binet Intelligence Scales (5th ed.) in children with autism spectrum disorder,* J Autism Dev Disord, 38:190–197,2008.
102. Baron IS. *Test review: Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV),* Child Neuropsychol J Norm Abnorm Dev Child Adolesc.,11:471–475, 2005.
103. Kaufman AS, Raiford SE, Coalson DL. *Intelligent Testing with the WISC-V,*John Wiley and Sons, 2016.
104. *Wechsler D. Wechsler Abbreviated Scale of Intelligence,* 1999.
105. *Wechsler D. Wechsler Abbreviated Scale of Intelligence - Second Edition,* 2011.
106. Ryan JJ, Carruthers CA, Miller LJ, Souheaver GT, Gontkovsky ST, Zehr MD. *Exploratory Factor Analysis of the Wechsler Abbreviated Scale of Intelligence (WASI) in Adult Standardization and Clinical Samples. Appl Neuropsychol,* 10:252–256, 2003
107. *Wechsler D. Wechsler Preschool and Primary Scale of Intelligence - Fourth Edition,* 2012.
108. Farmer C, Golden C, Thurm A. *Concurrent Validity of the Differential Ability Scales, Second Edition with the Mullen Scales of Early Learning in Young Children with and without Neurodevelopmental Disorders,* Child Neuropsychol J Norm Abnorm Dev Child Adolesc. ,22:556–569, 2016.
109. Schumann G, Loth E, Banaschewski T, Barbot A, Barker G, Büchel C, et al. *The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology,* Mol Psychiatry, 15:1128–1139, 2010.
110. Pausova Z, Paus T, Abrahamowicz M, Bernard M, Gaudet D, Leonard G, et al. *Cohort Profile: The Saguenay Youth Study (SYS),* Int J Epidemiol. https://doi.org/10.1093/ije/dyw023.
111. Awadalla P, Boileau C, Payette Y, Idaghdour Y, Goulet J-P, Knoppers B, et al. *Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics,* Int J Epidemiol, 42:1285–1299, 2013.
112. *Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness,* International Journal of Epidemiology | Oxford Academic, https://academic.oup.com/ije/article/42/3/689/909916 Consulté le 30 Août 2021.

113. Deary IJ, Gow AJ, Pattie A, Starr JM. *Cohort Profile: The Lothian Birth Cohorts of 1921 and 1936,* Int J Epidemiol, 41:1576–1584, 2012.
114. Fischbach GD, Lord C. *The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors,* Neuron, 68:192–195, 2010.
115. *Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder,* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5501701/. Consulté le 22 Janvier 2020.
116. Seiser EL, Innocenti F. *Hidden Markov Model-Based CNV Detection Algorithms for Illumina Genotyping Microarrays,* Cancer informatics, 2014.
117. Travaux de Catherine Proulx, étudiante au Laboratoire, 2020.
118. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, et al. *PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data,* Genome Res, 17:1665–1674,2007.
119. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, et al. *QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data*, Nucleic Acids Res, 35:2013–2025,2007.
120. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline*, Curr Protoc Bioinforma, 43:11.10.1-33, 2013.
121. Trost B, Walker S, Wang Z, Thiruvahindrapuram B, MacDonald JR, Sung WWL, et al. *A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data*, Am J Hum Genet, 102:142–155, 2018.
122. International Human Genome Sequencing Consortium. *Finishing the euchromatic sequence of the human genome*, Nature 431, 931–945, 2004.
123. *For Some Genes, Acetylation/Deacetylation Cycling Is the Real Turn-On*, PLoS Biol, 3(12):e431, 2005.
124. Marc Guerrien. *INPUT: GENE SCORES,* ermineJ, mis en ligne le 14 janvier 2019, consulté le 24 novembre 2021. URL : https://erminej.msl.ubc.ca/help/input-files/gene-scores/ "