Optimisation de la détection et de l'interpretation des variations génomiques issues de données d'exomes pour les études cas-contrôles

Olivier Quenez

Normandie Université

UNIVERSITÉ DE ROUEN NORMANDIE

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité **SCIENCES DE LA VIE ET DE LA SANTE**

Préparée au sein de l'**Université de Rouen Normandie**

Optimisation de la détection et de l'interpretation des variations génomiques issues de données d'exomes pour les études cas-contrôles

Présentée et soutenue par

OLIVIER QUENEZ

Thèse soutenue le 19/12/2023

devant le jury composé de :

MME MARIE DE TAYRAC	Professeur des Univ - Prat Hospitalier - UNIVERSITE RENNES 1	Rapporteur du jury
M. SEBASTIEN JACQUEMONT	Professeur des Universités - Université de Montréal (UdeM)	Rapporteur du jury
MME CELINE BELLENGUEZ	Chargé de Recherche - UNIVERSITE DE LILLE	Membre du jury
MME EMMANUELLE GENIN	Directeur de Recherche - UNIVERSITE BRET. OCCIDENTALE UBO	Membre du jury
M. JULIEN THEVENON	Professeur des Univ - Prat Hospitalier - UNIVERSITE GRENOBLE ALPES	Président du jury
M. GAEL NICOLAS	Professeur des Univ - Prat Hospitalier - Université de Rouen Normandie	Directeur de thèse

Thèse dirigée par **GAEL NICOLAS** (CANCER AND BRAIN GENOMICS)

ED 497 NBISE
Normande de Biologie Intégrative, Santé, Environnement

CBG CANCER & BRAIN GENOMICS

Remerciements

Je remercie très sincèrement les membres de mon jury, **Madame la Docteure Céline Bellenguez**, **Madame la Professeure Marie de Tayrac**, **Madame la Docteure Emmanuelle Génin**, **Monsieur le Professeur Sébastien Jacquemont** et **Monsieur le Professeur Julien Thévenon** qui me font l'honneur de juger mon travail.

Monsieur le Professeur Gaël Nicolas, merci d'avoir accepté de m'encadrer et de m'accompagner sur cette thèse, pour tes conseils et ta disponibilité malgré toutes tes responsabilités, c'est un réel plaisir et une chance de pouvoir travailler à tes côtés, en espérant que cela dure encore longtemps.

Monsieur le Docteur Dominique Campion, je vous remercie sincèrement pour tout ce que vous avez fait depuis mon arrivée il y a 10 ans. J'ai toujours été impressionné par votre vision de la génétique. Et vous voyez que finalement, il reste encore des choses à trouver !

Une pensée pour le **Professeur Thierry Frébourg**, qui a laissé sa marque sur chacun d'entre nous. J'ai appris énormément pendant les années au cours desquelles j'ai eu l'occasion de travailler avec lui, et merci à lui de m'avoir fait confiance dans les différentes missions qu'il m'a confié au fil des années.

Tous les membres de l'équipe neuro,

Catherine, merci pour tout, c'est un plaisir de travailler à tes côtés. Et désolé de t'avoir fait relancer toutes ces analyses des dizaines de fois suite à des réajustements ou des petites erreurs qui se sont glissées dans mes lignes de code. Sans toi tout ce travail aurait été impossible.

Anne, merci pour tout ce que tu m'as appris au cours de ces années, tu as joué un rôle très important dans tout ce travail, et bien sûr pour ta bonne humeur et l'ambiance chaleureuse de notre bureau. Mon seul regret sera de devoir quitter notre bureau en Janvier prochain.

Anne-Claire et Stéphane, merci beaucoup pour tout le travail que vous avez abattu toujours dans la bonne humeur, et ce malgré mes demandes très régulières. Sans tout ce travail de validation, ces travaux n'auraient pas la même qualité.

Kévin, merci pour toute l'aide et l'expertise que tu as apporté à ce projet, tu es toujours prêt à rendre service et j'espère pouvoir encore travailler avec toi dans le futur.

Camille, nous sommes arrivés en même temps dans le laboratoire et nous avons eu la chance de travailler ensemble depuis. Merci pour ton aide précieuse dans ce travail !

Magalie, **Laetitia**, **Romain**, **Sébastien,** merci de remettre de la biologie dans toutes ces données numériques ! Merci aussi pour votre sympathie, votre bonne humeur et nos échanges fréquents, et ce malgré le fait de ne pas toujours parler la même langue !

L'équipe du service de neurologie,

David, **Aline**, **Lou**, **Morgane**, et bien sûr **Didier**, vous nous rappelez que derrière toutes ces données se trouvent des patients et des familles.

Tous les membres du service de génétique moléculaire du Professeur Claude Houdayer, techniciens, ingénieurs, secrétaires et biologistes, en particulier :

Stéphanie et **Pascale**, merci pour votre appui dans ces travaux et pour nos échanges réguliers,

Sophie, avec qui j'ai la chance de travailler depuis longtemps, toujours de bonne humeur et toujours prête à aider en cas de besoin, on se retrouve au 2ème !

Et bien sûr tous les membres de l'unité U1245 que je n'ai pas cité ici mais qui ont participé chacun à leur manière à ce travail grâce à la bonne ambiance au laboratoire.

Des remerciements particuliers à tous ceux qui m'ont permis d'arriver où je suis aujourd'hui :

Gilles, tu as été un formidable maître d'apprentissage et j'ai beaucoup appris à tes côtés, de la place du bioinformaticien dans un laboratoire à la recherche de solutions innovantes pour répondre aux problèmes, en passant par la chasse au rapace mais ça, c'est une autre histoire...,

Une pensée pour **Christophe**, qui a eu confiance en le jeune débutant que j'étais et m'a remis le pied à l'étrier après une période difficile,

Olivier, travailler à tes côtés a été un plaisir, merci pour la liberté offerte au sein de ton équipe sans quoi je ne serai jamais retourné vers la recherche, et merci d'avoir fait de moi un "Roger Mayer" au service d'autres Jimi,

Et **Anne**, pour mon premier vrai retour à la recherche, merci pour ces discussions que nous avons eu sur la thèse qui m'ont motivé et qui ont finalement abouti, et merci encore d'avoir accepté de participer à mon suivi de thèse.

Merci à tous mes amis qui me supportent tous depuis tant d'années (dans tous les sens du terme !) : **Anaïs**, **Baptiste**, **Linda**, **Leslie**, **Quentin**, **Stéphane**, **Emilie**, **Valérie**, **Mara**, **Renan**, **Marjo** et tant d'autres.

Ma famille, toujours à mes côtés, en espérant vous rendre fier.

Ma femme **Alice**, merci pour ton soutien et ton amour permanent,

Et enfin mon fils **Augustin**, qui ne cesse de m'émerveiller jour après jours.

Résumé

Au cours des 20 dernières années, l'évolution des nouvelles technologies a révélé la grande variabilité de notre génome depuis la simple substitution jusqu'aux réarrangements chromosomiques. Les technologies de séquençage à haut débit ont particulièrement amélioré l'identification et l'interprétation des variations de petite taille tout en offrant l'opportunité d'explorer les variations de structure avec une résolution supérieure à celle disponible grâce aux analyses pangénomiques sur puces. Néanmoins, l'identification des variations de structure, et plus particulièrement des variations du nombre de copies (CNV) à partir de données de séquençage par capture, a été sous exploitée et peu évaluée. Notre objectif principal était de mettre en place un pipeline bioinformatique basé sur la profondeur de lecture pour l'identification des CNV, puis de l'appliquer à une études cas-témoins d'exome dans le cadre de la recherche sur la maladie d'Alzheimer.

La maladie d'Alzheimer (MA) est la maladie neurodégénérative la plus fréquente. Les facteurs génétiques individuels jouent un rôle important dans son déterminisme et de multiples facteurs de risque ont été identifiés, essentiellement des substitutions et petites insertions/délétions. Pourtant, des variations de structure ont déjà été identifiées dans des formes monogéniques de MA, comme les duplications complètes du gène *APP*. Les CNV restent très peu étudiés dans la MA et nous avons souhaité appliquer une approche cas-témoins à partir de données massives d'exomes pour détecter des CNV contribuant au risque de MA.

Dans un premier temps, nous avons établi une stratégie d'analyse basée sur le logiciel CANOES afin de détecter les CNV à partir de données de NGS issues d'une capture (panel, exomes). Cette approche a été validée à travers deux grands jeux de données de panels et d'exomes comparés à des techniques indépendantes. Dans le premier jeu de données (panels), la sensibilité et la spécificité étaient de 100% et nous obtenons une sensibilité de 87,25 % et une valeur prédictive positive de 88,5% sur la détection de CNV sur les données de séquençage d'exomes.

Par la suite, nous avons appliqué cette approche aux données d'exomes issues des consortium ADES (Alzheimer Disease Exome Sequencing) et ADSP (Alzheimer Disease Sequencing Project), regroupant, après un contrôle qualité extensif développé dans le cadre de ces travaux, 22 094 individus répartis entre 4077 formes précoces de MA, 8458 formes tardives et 9559 témoins. Nous avons mis au point des analyses au niveau des transcrits et appliqué une méthode statistique basée sur les dosages appliquée aux formes précoces et aux témoins. Nous avons pu identifier plusieurs potentiels nouveaux facteurs de risque dont la région du chr22q11.21, déjà impliquée dans les troubles du neurodéveloppement (p=3,8x10-4). De plus, nous avons identifié des délétions très rares dans les gènes *ABCA1* et *ABCA7* dont les variations perte de fonction sont connues comme facteurs de risque de MA depuis peu, et nous avons réalisé une analyse conjointe des délétions et des variations perte de fonction de petite taille.

En conclusion, nous avons montré que la détection de CNV issus de données d'exome est fiable et nous en avons mesuré les performances et les limites avant de les appliquer à un grand jeu de données afin d'identifier de nouveaux mécanismes contribuant au développement de la maladie d'Alzheimer.

Mots clef : Maladie d'Alzheimer, CNV, études d'association, séquence d'exome

Summary

Over the past 20 years, the evolution of new technologies has revealed the great variability of our genome, from simple substitutions to chromosomal rearrangements. High-throughput sequencing has particularly improved the identification and interpretation of small variations, while offering the opportunity to explore structural variations with a higher resolution than that available with genome-wide microarray analyses. Nevertheless, the identification of structural variations and more specifically copy number variations (CNVs) from capture sequencing data, has been under exploited and under evaluated. Our main objective was to develop a read depth based bioinformatics pipeline for CNV identification, and then apply it to a case-control exome study in Alzheimer's disease research.

Alzheimer's disease (AD) is the most common neurodegenerative disorder. Individual genetic factors play an important role in its determinism, and multiple risk factors have been identified, mainly substitutions and small insertions/deletions. However, structural variations have already been identified in monogenic forms of AD, such as complete duplication of *APP* gene. CNVs remain largely unstudied in AD, we set out to apply a case-control approach using massive exome data to detect CNVs contributing to AD risk.

As a first step, we established an analysis strategy based on CANOES software to detect CNVs from NGS data derived from a capture (gene panels, exomes). This approach was validated using 2 large gene panels and exome datasets, compared with independent targeted techniques. In the first dataset (gene panels), sensitivity and specificity were 100%, and we obtained a sensitivity of 87.25% and a predictive positive value of 88.5% for CNV detection in whole exome sequencing data.

We then applied this approach to whole exome data from the ADES (Alzheimer Disease Exome Sequencing) and ADSP (Alzheimer Disease Sequencing Project) consortia, grouping, after extensive quality control developed as part of this work, 22,094 samples divided between 4077 early onset cases, 8458 late onset and 9559 controls. We developed transcript-level analyses and applied a statistical method based on dosage applied on early onset cases and controls. We were able to identify several potential new risk factors, including the 22q11.21 regions, already implicated in neurodevelopmental disorders (p=3,8x10-4). In addition, we identified rare deletions in *ABCA1* and *ABCA7* genes, whose loss-of-function variations have recently been identified as risk factors for AD, and carried out a joint analysis of deletions and small loss-of-function variations.

In conclusion, we have shown that CNV detection from exome data is reliable, and we have measured its performance and limitations before applying it to a large dataset to identify new mechanisms contributing to the development of Alzheimer's disease.

Keywords: Alzheimer disease, CNV, association study, whole exome sequencing

Liste des abréviations

ACPA : Analyse Chromosomique par Puce à ADN
ADES : Alzheimer Disease European Sequencing project
ADSP : Alzheimer Disease Sequencing Project
CMT : Charcot-Marie Tooth
CNV : Copy Number Variation
CGH : Comparative Genomic Hybridization
DGV : Database of Genomic Variant
DDPCR : Digital Droplet PCR
EOAD : Early Onset Alzheimer Disease
ExAC : Exome Aggregation Consortium
FISH : Fluorescence In Situ Hybridization
FoSTeS : Fork Stalling and Template Switching
gnomAD : Genome Aggregation Database
IGAP : International Genomics of Alzheimer disease Project
LOAD : Late Onset Alzheimer Disease
LCR : Low Copy Repeat
LINE : Long Interspersed Elements
MLPA : Multiplex Ligation-dependent Probe Amplification
NAHR : Non-Allelic Homologous Rearrangement
NGS: Next Generation Sequencing
NHEJ: Non-Homologous End Joint
PCR : Polymerase Chain Reaction
PIEV : Pénétrance Incomplète et Expressivité Variable
QMPSF : Quantitative Multiplex PCR of Short fluorescent Fragments
SINE : Short Interspersed Elements
SNV : Single Nucleotide Variation
SV : Structural Variation
SVA : SINE-VNTR-Alu
T2T : consortium Telomere-2-Telomere
VNTR : Variable Number Tandem Repeat
WES : Whole Exome Sequencing
WGS : Whole Genome Sequencing
UTR : UnTranslated Regions

Table des matières

Liste des figures

Liste des tableaux

Avant-propos

Depuis la première publication de la séquence du génome humain au début des années 2000, les techniques permettant d'explorer notre ADN n'ont cessé d'évoluer. L'apport des nouvelles technologies, depuis les puces pangénomiques jusqu'aux technologies de séquençage à haut débit, a permis l'exploration de la diversité interindividuelle jusqu'alors sous-estimée. Tout ceci a permis de mettre en lumière la grande complexité de notre génome, sa mutabilité, mais aussi de développer nos connaissances sur les gènes et leurs fonctions et bien sûr leur lien avec les différentes pathologies. En associant toutes ces connaissances, il a été possible d'identifier l'impact des variations rares et fréquentes dans des maladies complexes et de découvrir un nombre exponentiel de gènes causant des maladies mendéliennes qui ne pouvaient être diagnostiquées jusqu'à récemment et restaient ainsi orphelines.

Parmi les différents types de variations, les événements impactant de grandes portions du génome, aussi appelés variants de structure, restent complexes à identifier et à interpréter. Lors de mes travaux de thèse, je me suis concentré sur l'identification des variations du nombre de copies (ou CNV – Copy Number Variation) rares à partir de données de séquençage à haut débit, approche souvent sous-exploitée. Ces CNVs rares ont potentiellement un impact fort dans certaines pathologies. L'objectif de mon travail a consisté à la mise en place d'un pipeline de détection et de filtration de CNV à appliquer sur un jeu de données important et hétérogène d'exomes dans le cadre d'une étude cas-témoins portant sur la maladie d'Alzheimer.

Nous commencerons par décrire la structure de notre génome et sa complexité, en rappelant que tout ceci a été possible grâce à des grands projets de séquençage de population et aux bases de données qui y sont associés. Nous essayerons ensuite de comprendre comment se produisent ces différents événements et leur impact à la fois évolutif et pathologique. Pour conclure cette première partie je présenterai les méthodes qui nous permettent d'identifier et de valider les différents variants de structure, à la fois par les approches moléculaires et bioinformatiques. Par la suite je présenterai la maladie d'Alzheimer, son origine génétique et sa complexité. Je terminerai en présentant les travaux effectués au cours de ma thèse et les différents résultats.

1. Introduction

1.1. Structure et variabilité du génome humain

1.1.1. Le séquençage du génome humain et sa structure

C'est en 2001 que les consortiums du Human Genome Project (Lander et al. 2001) et Celera Genomics (Venter et al. 2001) ont publié les premières séquences brutes du génome humain, qui conduiront à l'obtention en 2004 d'une première version de sa séquence (International Human Genome Sequencing Consortium 2004; Istrail et al. 2004). Cette première version, bien qu'incomplète, a permis de mettre en évidence la taille du génome (environ 3 Milliards de paires de bases - pb) et la composition de ce dernier, dévoilant qu'une faible proportion du génome correspond à des régions codantes. En effet, si les gènes couvrent environ 25% de notre génome, seul 1,5% de ce dernier correspond à des régions codantes, le reste étant composé de régions exoniques non traduites (UnTranslated Regions - UTR) et des introns. Les 75% restants de notre génome correspondent à des régions intergéniques.

Une caractéristique marquante du génome humain est sa composition à plus de 50% par des séquences répétées (de Koning et al. 2011). Ces séquences peuvent varier selon le type de répétition, leur quantité ou leur localisation et peuvent médier des remaniements de structure, qui font l'objet de ces travaux de thèse.

Le premier type de répétition correspond aux séquences répétées en tandem pour environ 6,5% de notre génome. Ces répétitions sont classées en fonction de la taille de la séquence répétée. On retrouve l'ADN micro-satellitaire constitué de répétitions entre une et 5 paires de base (pb) pouvant être répétées jusqu'à 5000 fois. Ces microsatellites ont été localisés dans toutes les régions du génome que ce soit dans les régions codantes ou non codantes. On retrouve ensuite l'ADN mini-satellitaire constitué de séquences entre 10 et 25 pb répétées entre 1000 et 2000 fois. Ces minisatellites sont retrouvés majoritairement dans les régions télomériques. Enfin, l'ADN satellitaire est constitué de séquences de 100 et 500 pb formant des régions pouvant aller jusqu'à plusieurs millions de pb. Ces ADN satellites constituent majoritairement les régions centromériques et para centromériques. Les 40 à 45% de régions répétées restantes correspondent à des éléments répétés dispersés aussi appelés éléments mobiles. Les séquences composant ces éléments mobiles peuvent être classées en plusieurs catégories. On y retrouve les SINE et LINE (Short/Long Interspersed Elements) (Liang et Fu 2021; Richardson et al. 2015) avec une limite de taille entre les deux types d'éléments situés à 500 pb. Parmi les SINE, la famille la plus répandue est celle des séquences Alu (Batzer et Deininger 2002). D'une taille

moyenne d'environ 300 pb, ces séquences comptent plus d'un million de copies dans notre génome, représentant 10% de ce dernier. Il est possible de retrouver ces différents éléments répétés dans des structures plus complexes, tel que les SVA (SINE-VNTR-Alu) (Gianfrancesco et al. 2019; H. Wang et al. 2005) correspondant à des rétrotransposons.

Lorsque ces éléments répétés, qu'ils soient en tandem ou dispersés, sont organisés dans des structures de plus de 1 kb et avec un niveau d'identité de plus de 90%, on parle alors de régions de duplication segmentaire (ou LCR - Low Copy Repeats). Ces structures complexes ont un rôle très important dans la variabilité et l'évolution de notre génome en étant impliquées dans certains mécanismes de recombinaison comme nous le verrons par la suite.

Tous ces éléments répétés ont pendant longtemps limité la possibilité d'obtenir une séquence complète de notre génome. La majorité des techniques permettant le séquençage complet de notre génome utilisaient jusqu'à maintenant des fragments courts (entre 75 et 150pb) qui ne permettaient pas d'obtenir une séquence précise des éléments de faible complexité tel que les répétitions en tandem longues, mais aussi de certaines régions de duplication segmentaire. Ce n'est que très récemment qu'une version entière du génome a été obtenue dans le cadre du Consortium T2T (Telomere To Telomere) (Nurk et al. 2022). Ceci a été rendu possible par l'utilisation des nouvelles technologies de séquençage de longues molécules permettant de séquencer des molécules pouvant atteindre plusieurs centaines de kb. Grâce à cela, il a été possible de séquencer entièrement des régions jusqu'alors inaccessibles tels que les télomères et les centromères, mais aussi les bras courts des chromosomes acrocentriques. Toutes ces régions, qui étaient restées inexploitées jusqu'à maintenant, correspondent à environ 8% du génome humain.

1.1.2. Variabilité du génome humain

L'apparition des technologies de séquençage massif en parallèle ou séquençage de seconde ou de « nouvelle » génération (Next Generation Sequencing - NGS), permettant la lecture simultanée de plusieurs millions de courts fragments d'ADN (ou reads) en une seule opération, et ce à un coût réduit, a permis la multiplication du séquençage d'individus dès les années 2010. La technologie produite par Illumina s'est progressivement imposée en tant que référence pour le séquençage de fragments courts (ou "Short read") entre 75 et 150 pb. Aujourd'hui, de nouvelles technologies émergent et viennent concurrencer Illumina sur les approches short reads, telles que la société BGI/MGI (MGI tech Co., Ltd.)

se basant sur une technologie similaire à celle d'Illumina pour ses séquenceurs, ou encore la société PacBIO et son nouveau séquenceur Onso, parmi d'autres concurrents qui arrivent actuellement sur le marché et contribuent à réduire les coûts de séquençage.

L'avènement de ces technologies a permis le séquençage d'un grand nombre d'individus, qui lui-même a mis en évidence un large spectre de type de variations génétiques depuis la simple substitution nucléotidique au réarrangement de grande taille. Les variations sont dichotomisées en deux grandes classes en fonction de la taille de l'événement.

D'une part, on retrouve les variations de petite taille, constituées à la fois des substitutions d'un nucléotide (ou SNV – Single Nucleotide Variant) et les petites insertions / délétions (ou indels), généralement avec une limite de taille fixée à 50 pb. Leur nombre est estimé entre 3 et 5 millions par individu en comparaison à un génome de référence du même sexe. Ces événements peuvent survenir dans toutes les régions du génome et représentent la plus grande variabilité en termes de nombres d'événements. Pour exemple, la base de données gnomAD v3.1 contient actuellement 76 156 génomes d'individus non apparentés et a identifié plus de 640 millions de variants différents de haute qualité dont plus de 60% sont présents à moins de 1%(Karczewski et al. 2020). Il est important d'observer que, plus on séquence d'individus, plus le nombre de singletons (événement unique à un individu au sein d'un jeu de données) augmente lui aussi, permettant de mettre en lumière une variabilité inter-individus très importante (Figure 1).

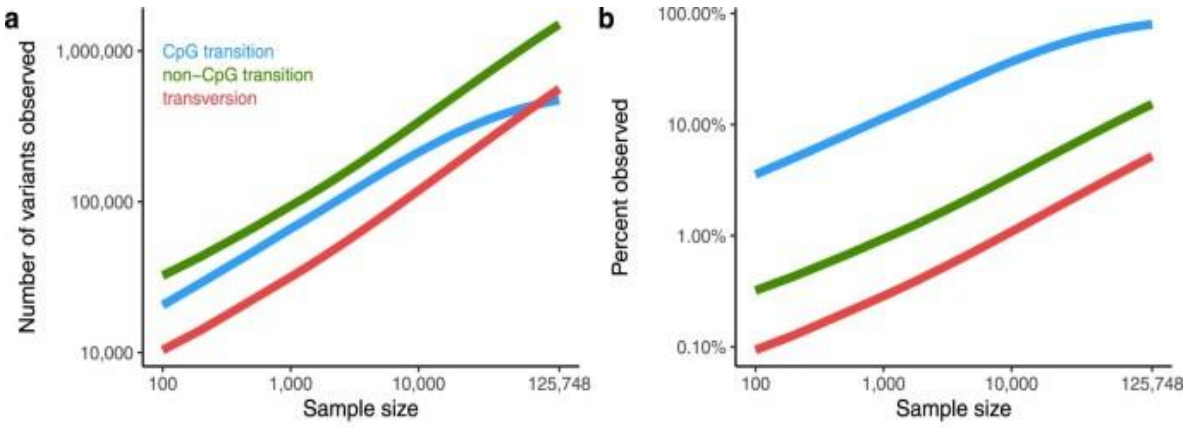


Figure 1 : Evolution du nombre de transition et de transversion en fonction du nombre d'individus
A: Nombre total de variants différents en fonction de la taille de l'échantillon. B: Pourcentage du nombre de variants totaux observés en fonction de la taille de l'échantillon. Figure issue de Karczewski et al., 2020.

D'autres part, on retrouve les événements de grande taille qui constituent les variations de structure (ou SV – Structural Variant). La limite de 50 paires de bases entre les indels et les variations de structure provient principalement du fait que les logiciels permettant d'identifier l'un ou l'autre des événements ne sont pas les mêmes. En effet, les petites insertions/délétions sont généralement identifiées par les mêmes logiciels et technologies que les substitutions, alors que les variations de structures sont identifiées par d'autres logiciels ou technologies. Ces événements sont beaucoup plus rares que les SNV et indels, mais affectent de grandes proportions du génome.

Les variations de structure sont classées en deux catégories en fonction de l'impact sur la quantité d'ADN (Figure 2). Dans le cas où la quantité d'ADN reste constante, on parlera de variation équilibrée. On retrouvera dans cette catégorie 3 types d'événements : les inversions d'une portion d'un chromosome, les translocations et les insertions de nouvelles séquences (tels que les rétrovirus).

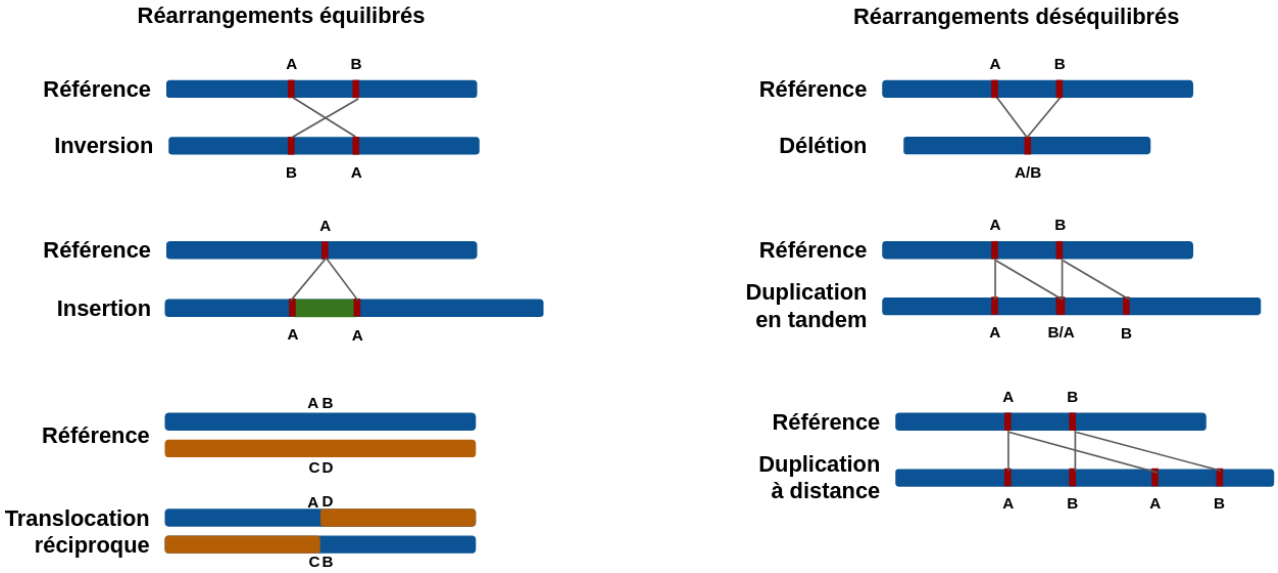


Figure 2 : Représentation des différents types de variation de structure
Les variations de structure sont catégorisées entre réarrangements équilibrés et déséquilibrés. Sur chaque schéma est représenté dans la partie supérieure la séquence de référence et dans la partie inférieure l'événement chez un individu.

Si la quantité d'ADN génomique est modifiée par la variation, on parlera alors d'événements déséquilibrés, que l'on retrouve aussi sous le nom de variations du nombre de copies ou CNV (Copy Number Variation). Ces variations regroupent les délétions et les duplications, ces dernières pouvant être soit en tandem, soit à distance.

Les chiffres les plus récents issus de la base de données gnomAD (R. L. Collins et al. 2020) révèlent un nombre médian de 7 439 variants de structure par génome nucléaire humain, dont 49,8% sont des singletons dans la base de données parmi 14237 individus (Figure 3.A). Parmi les événements les plus fréquemment identifiés, on retrouvera les insertions d'éléments répétés tels que les séquences Alu, les SVA et les LiINE1 (Figure 3.B). Toutes ces informations sur la grande variabilité de notre génome et la complexité de ce dernier n'auraient pas pu être mises en évidence sans la mise en place de projets de séquençage massif de populations.

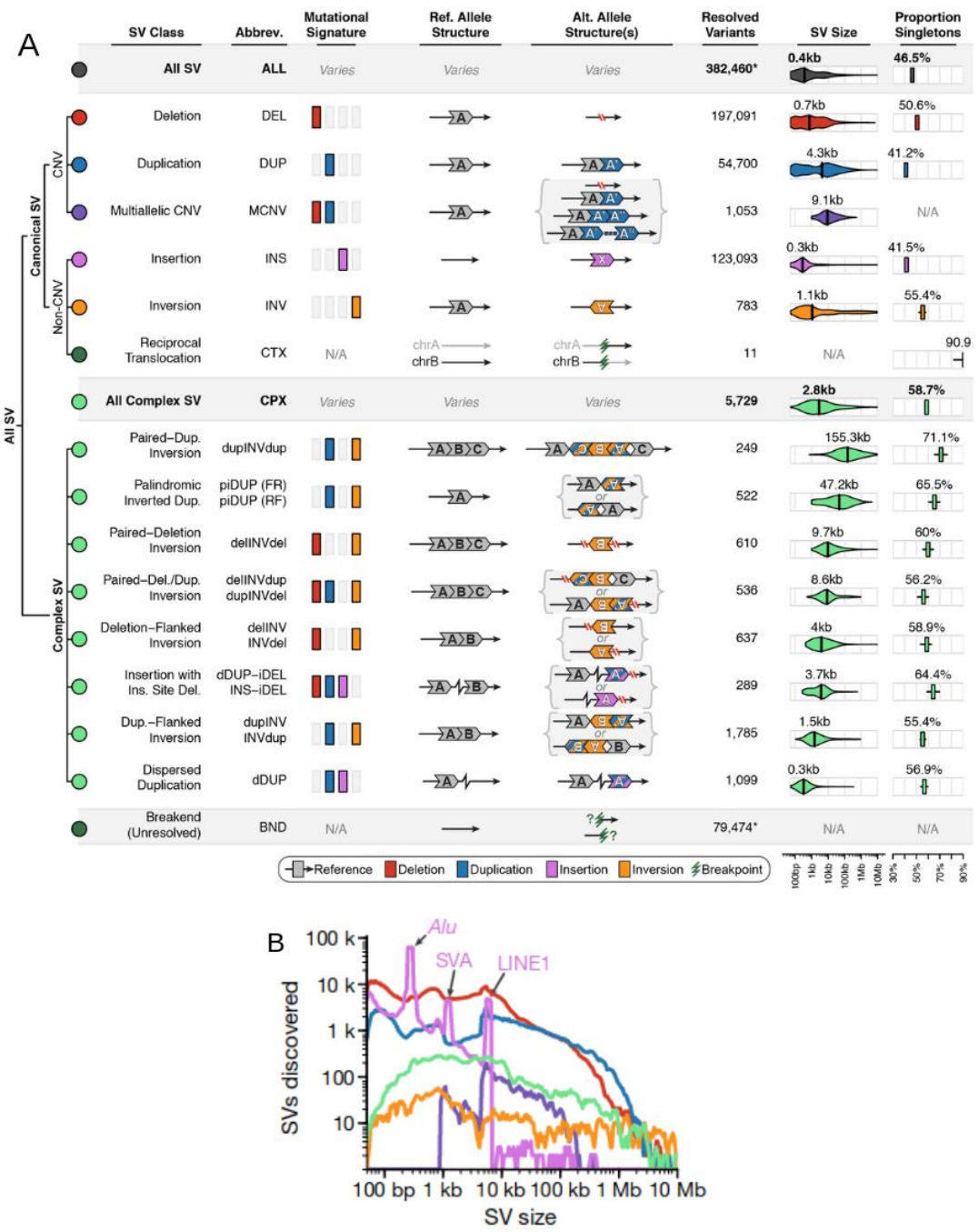


Figure 3 : Les variations de structure dans la base de données gnomAD

A - Classification des différents types de variations de structure. Dans la partie supérieure sont représentés les événements de base, et dans la partie inférieure les différentes compositions complexes possibles. Sont associés à chaque événement le nombre et la distribution de la taille chez 12 549 individus séquencés en génome issus de la base de données gnomAD. B – Distribution de la taille des événements en fonction du type de variation. Figure extraite de Collins et al., 2020.

1.1.3. Les grands projets de séquençage

Après le premier séquençage du génome humain, plusieurs grands projets ont été menés afin de caractériser génétiquement de grandes populations. Certains projets se concentrent sur la production de données de séquençage en effectuant une sélection des individus sur différents critères (population, pathologie) et vont associer aux données de séquençage des données phénotypiques. D'autres projets visent à agréger les données issues d'autres projets afin de pouvoir combiner les différentes informations et ainsi apporter une plus-value supplémentaire.

Les deux premiers projets historiques se sont basés sur deux approches légèrement différentes pour l'analyse des populations. Le premier consortium du "1000 Genomes Project" a visé à séquencer plus de 1000 individus issus de 14 populations différentes dans le monde (1000 Genomes Project Consortium et al. 2012; 2010) en combinant à la fois du séquençage de génome peu profond et du séquençage d'exome, complétant l'analyse à l'aide de puces de génotypage. Ce projet a évolué pour séquencer au final plus de 2500 individus, permettant d'avoir une cartographie mondiale de 26 populations, à la fois pour les variations ponctuelles (1000 Genomes Project Consortium et al. 2015) mais aussi pour les variations de structure (Sudmant et al. 2015).

Le second projet est celui de l'Exome Sequencing Project (ESP) mené par le National Heart, Lung, and Blood Institute (NHLBI). Celui-ci s'est concentré uniquement sur le séquençage d'exomes de personnes nord-américaines, mais en atteignant un effectif de plus de 6500 individus (Fu et al. 2013) classés en European-Americans et African-Americans.

En parallèle de ces deux premiers projets, le National Institute of Health (NIH) a mis en place une base de données visant à regrouper l'ensemble des variations connues et d'y associer un identifiant unique commençant par « rs », afin de faciliter les échanges. Cette base de données de référence dbSNP a évolué afin de devenir le projet ALFA (ALlele Frequency Aggregator) qui conserve la même idée d'agrégation de données et d'identifiant unique.

Parmi les projets majeurs d'agrégation de données figurent la base de données ExAC (Exome Aggregation Consortium et al. 2016; Karczewski et al. 2020) et son évolution gnomAD (Genome Aggregation Database) (S. Chen et al. 2022; Gudmundsson et al. 2022). Au-delà de la simple agrégation, ces deux bases ont entrepris d'uniformiser au maximum les différents projets, une réanalyse bio-informatique complète a été mise en place et associée à un contrôle qualité très strict. Grâce à cela, ExAC puis gnomAD ont pu associer les données de plus de 130 000 individus (v2.1). Il est à noter que

dans sa version la plus récente (v3), gnomAD se limite maintenant uniquement à des données issues de séquençage de génome, là encore afin d'uniformiser au maximum les données.

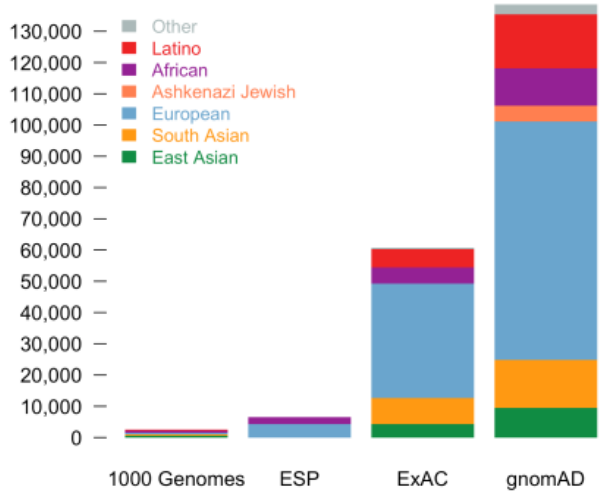


Figure 4 : Evolution des bases de données publiques

Comme on peut le voir sur la Figure 4, il existe un biais important dans l'origine ethnique des populations séquencées. Afin de compenser cela, de nombreux projets nationaux et internationaux ont émergé afin de caractériser différentes populations. On peut citer entre autres le projet chinois de la Chinese Millionome Database (CMDB) (Li et al. 2023), le GenomeAsia 100K(GenomeAsia100K Consortium et al. 2019), The African Genome Variation Project (Gurdasani et al. 2015) ou encore le projet Qatari(Zayed 2016). En Europe, un projet de coordination international, le 1+million genomes project, a pour vocation de coordonner les différents projets nationaux afin d'harmoniser et de partager les technologies et les procédures. Ces projets ont aussi pour vocation de fournir des populations témoin plus proches des patients qui sont nécessaires afin de mieux identifier des polymorphismes de la population générale et de pouvoir mettre en avant des variations rares potentiellement impliquées dans des pathologies.

Le plus grand projet de séquençage national est celui de la UK Biobank(R. Collins 2012; Littlejohns et al. 2020) du Royaume-Uni et sa base de données Decaf(Halldorsson et al. 2022). Cette dernière associe différentes données omiques (génomique, protéomique, puces) à des données biomédicales (imagerie, prélèvements sanguins, autres) pour plus de 500 000 individus. Cette base de données, en plus d'une quantité très importante de données variées, offre aussi un suivi longitudinal des patients, permettant de voir leur évolution au cours du temps.

En France, le projet FREX (French Exome Project) a permis d'établir une première cartographie nationale en séquençant l'exome de 600 individus, issus de 6 régions françaises (Genin 2017). Ces données de séquençage ont été complétées par des données de puces de génotypage (ou SNP-array). Plus récemment, et dans la continuité du projet FREX, le projet POPGEN du plan France Médecin Génomique 2025 a débuté, avec pour objectif le séquençage de 10250 individus.

Bases de données de variations de structure

Il existe un certain nombre de bases de données spécialisées dans les variations de structures. La plus ancienne d'entre elles est la Database of Genomic Variants(DGV) (MacDonald et al. 2014). Elle héberge actuellement environ 1 million de CNV distincts issus de 75 études et elle est limitée uniquement à des individus sains. Cette base de données contient majoritairement des données issues d'analyses de puces de CGH (Comparative Genomic Hybridation) ou de puces à SNP (voir le chapitre "Méthodes de détection des réarrangements génomiques" pour plus de détailw sur ces approches). Les données dans cette base étant très hétérogènes et afin de pouvoir fournir une information fiable, une version "Gold standard" existe, celle-ci se limitant aux événements détectés au moins 2 fois dans 2 études différentes. Ceci permet de limiter au maximum les artefacts. Ceci a permis à la DGV d'être utilisée comme base de référence de fréquence des variations de structure dans la population.

Lors des différentes étapes de réanalyse du projet gnomAD, des étapes d'identification des variations de structure ont été intégrées. Ceci a permis d'avoir sur une même base à la fois les variations ponctuelles et les variations de structure. L'intérêt principal pour notre projet est que la détection des variations de structure est effectuée à partir de données de séquençage et va donc se rapprocher au plus près de ce que nous allons obtenir. Ceci aura un impact important par la suite, car toutes les approches, qu'elles soient bio-informatiques ou moléculaires, ne sont pas capables d'identifier les mêmes événements et, pour un même événement donné, ne présenteront pas les mêmes bornes.

De la même manière que pour les variations ponctuelles avec dbSNP et ALFA, le NIH a mis en place une base de données spécialisée dans les variations de structure : dbVar. Chaque variant de structure aura un identifiant unique, le ssv.

Enfin il existe des bases de données spécialisées dans un type de variation de structure, on peut citer par exemple dbRIP (database of Retrotransposon Insertion Polymorphisms in humans)(J. Wang et al. 2006) qui est spécialisé dans la localisation des différents éléments répétés tels que les éléments Alu, LINE1 ou SVA.

Les bases de données cliniques

En parallèle des bases de données et des projets ayant pour vocation de déterminer la variabilité du génome humain dans la population générale, il existe des bases de données construites autour de données de séquençage de patients, incluant potentiellement des apparentés sains. L'objectif de ces bases de données est de répertorier des variations identifiées afin de faciliter l'interprétation de variations identifiées par la suite. De la même manière que pour les projets de séquençage en population générale, on retrouve à la fois des projets de séquençage et des bases de données d'agrégation de données.

On peut citer comme exemple la base TopMed (Trans-Omics for Precision Medicine) regroupant des cohortes de patients atteints de pathologies cardiaques ou pulmonaires. Cette base de données regroupe plus de 180 000 individus et contient non seulement des données de séquençage, mais aussi d'autres données omiques, tel des données de RNA-Seq ou de méthylation, ainsi que des données phénotypiques (Taliun et al. 2021). Une partie de TopMed est incluse dans gnomAD. Un autre exemple est la Simons Simplex Collection conduite par le SFARI (Simons Foundation Autism Research Initiative) visant à séquencer plus de 2600 cas index atteints d'un trouble du spectre autistique ainsi que les 2 parents et au moins un apparenté sain (Levy et al. 2011; O'Roak et al. 2011).

En lien avec notre étude, nous pouvons citer le projet ADSP (Alzheimer Disease Sequencing Project), un projet américain visant à identifier de potentiels facteurs de risque ou protecteurs en lien avec la maladie d'Alzheimer et d'autres démences associées (Crane et al. 2017; Raghavan et al. 2018; Bis et al. 2020). Ce projet contient actuellement plus de 20 500 individus séquencés en exome, et plus de 36 000 génomes (bientôt 70 000), incluant de potentiels apparentés sains. Nous reviendrons plus en détails sur ce projet par la suite, celui-ci étant intégré dans notre étude.

Concernant les bases de données spécialisées dans l'agrégation des variants, l'une des bases les plus importantes pour les variations de structure est DECIPHER (DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources) (Firth et al. 2009). Celle-ci contient les données issues de plus de 46 000 patients atteints de maladies rares, chacun renseigné par une équipe référente. Ceci permet d'avoir un niveau de confiance élevé dans les données mises à disposition qui comprennent à la fois les variations identifiées, mais aussi de nombreuses informations phénotypiques.

Leçons tirées des bases de données pour l'interprétation des variants

Les informations contenues dans ces différentes bases de données ont permis de mettre en lumière certaines caractéristiques de notre génome. Contrairement à ce qui serait attendu, les différents types de variations ne sont pas répartis de manière aléatoire. Si l'on prend les données de gnomAD, on obtient des valeurs théoriques d'une variation toutes les 4 pb et une variation rare toutes les 8 pb (S. Chen et al. 2022; Gudmundsson et al. 2022). Or, si l'on observe précisément la répartition, on peut observer que certains gènes ne présentent pas ou peu de variations. De la même manière, des variations potentiellement perte de fonction sont présentes de manière plus ou moins importante en fonction des gènes. A partir de ce constat, les équipes travaillant sur les données de gnomAD ont déterminé des scores de tolérance pour chaque gène à chaque type de variation : en fonction de la taille et de la composition en bases du gène, ils ont calculé une quantité théorique attendue des différents types de variations (synonyme, faux sens, perte de fonction, etc). Un score est alors calculé en fonction des données observées et attendues, ce score étant corrigé en fonction de la taille de l'échantillon observé. Plus le nombre de variants observés est faible par rapport au nombre attendu, plus le gène a de forte chance d'être soumis à une contrainte. Pour les variations perte de fonction, le score défini est la pLI (Probability of Loss-of-function Intolerance) qui varie entre 0 et 1, et on considère classiquement un transcrit comme fortement contraint lorsque le score est supérieur à 0,9. Ce score de pLI est défini sur les données d'ExAC, et avec la mise en place de la base gnomAD et ses effectifs plus importants, un nouveau score de tolérance aux différents types de variations a été mis en place. Basé sur le même principe du nombre d'événements observés et attendus, ce nouveau score présente une probabilité d'intolérance associé à un intervalle de confiance. Au final, plus que le score en lui-même, c'est la borne supérieure de l'intervalle de confiance (ou LOEUF – LoF Observed/Expected Upper Bound) qui est maintenant employée.

Le même constat peut être observé pour les variations du nombre de copies et la sensibilité au dosage des gènes. A partir des données de gnomAD, le groupe du Pr McArthur a observé une absence de délétion ou de duplication affectant spécifiquement certains gènes (Karczewski et al. 2020). Grâce à cela, ils ont pu établir un score d'haploinsuffisance (c'est à dire l'incapacité d'un gène à produire un phénotype sauvage (normal) en l'absence d'une copie) ou de triplosensibilité dans le cas où un gain de copies induit une perte du phénotype sauvage. Grâce à cela, il est possible de mieux identifier de potentiels événements d'intérêt diagnostique, ou au contraire d'exclure des réarrangements fréquents. La base de données ClinGen (Riggs et al. 2012) met elle aussi à disposition des informations

de sensibilité au dosage de gènes ou de régions génomiques en lien avec certaines pathologies. Cette base de données ne se base pas sur des données de fréquence dans la population mais sur une revue exhaustive de la littérature pour fournir ces informations d'haploinsuffisance et de triplosensibilité.

1.1.4. Mécanisme de génération des variations de structures

Les différents types de réarrangements, qu'ils soient équilibrés ou non, proviennent de ruptures dans la séquence chromosomique puis d'une réorganisation d'un ou plusieurs chromosomes. L'organisation et la taille d'un réarrangement dépendent du mécanisme impliqué dans ce dernier (Gu, Zhang, et Lupski 2008). Il existe trois grands mécanismes identifiés comme conduisant à ces réarrangements.

Recombinaison Homologues Non Allélique

Les recombinaisons entre régions homologues non alléliques (Non-Allelic Homologous Rearrangement - NAHR) sont des réarrangements structuraux qui surviennent via les régions de duplication segmentaire. Ces régions correspondent à des blocs d'au moins 1 kb ayant une similarité d'au moins 95% entre elles. Lors de la méiose, les régions homologues alléliques sont alignées l'une par rapport à l'autre, permettant des recombinaisons homologues. Ces dernières permettent le brassage de matériel entre les chromosomes paternels et maternels sans altérer la structure globale. Il arrive que les régions associées correspondent, non pas à des régions homologues, mais à des duplications segmentaires non alléliques. Dans ce cas, il peut survenir un réarrangement qui va induire un changement important dans la structure des deux chromosomes en sortie. Selon l'orientation des deux régions de duplication segmentaire l'une par rapport à l'autre, plusieurs cas de figure peuvent se présenter (voir Figure 5). Si les deux régions sont situées sur le même chromosome et sont orientées dans le même sens, il va alors y avoir génération d'une délétion et d'une duplication, chacune sur un des deux chromosomes. Si les deux régions de duplication segmentaire sont situées sur le même chromosome mais dans une orientation différente, il y aura inversion de la région située entre ces deux régions. Si les deux régions menant au réarrangement sont situées sur des chromosomes différents nous aurons alors affaire à une translocation inter chromosomique avec un échange de matériel entre les deux chromosomes allant des régions de duplication segmentaire jusqu'au télomère. Il peut aussi arriver que les réarrangements se produisent entre les deux chromatides d'un même chromosome. Dans ce cas nous aurons, de la même manière que pour deux chromosomes de la même paire, apparition d'une duplication et d'une délétion sur chacune des deux chromatides. Enfin, si le

réarrangement se produit entre deux régions sur la même chromatide, il va se produire un événement de circularisation produisant une délétion sur le chromosome et la formation d'un ADN circulaire.

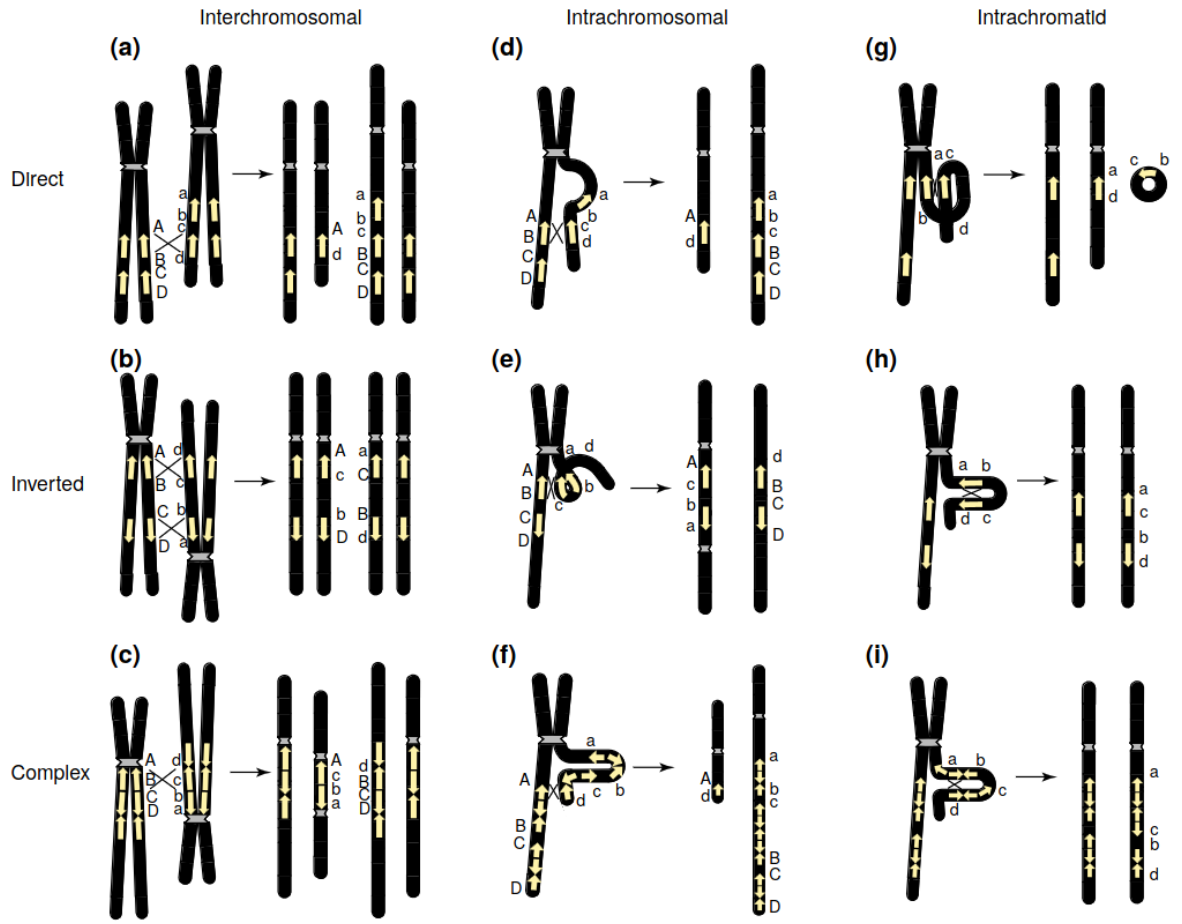


Figure 5: Représentation schématique des recombinaisons basées sur les mécanismes de NAHR.
Les flèches jaunes représentent les régions de LCR. Les réarrangements sont organisés en fonction des mécanismes (inter chromosomique, intra chromosomique ou intra chromatidique) et de l'orientation des LCR (direct, indirect, organisation complexe). Figure issue de Stankiewicz et Lupski, 2002

Il est à noter que toutes les régions de duplication segmentaire n'ont pas la même implication dans les mécanismes de réarrangements. Plusieurs critères influencent l'implication des régions de duplication segmentaire dans la fréquence d'apparition des réarrangements par NAHR. Trois critères ont été identifiés comme facteurs d'apparition de recombinaison. Le premier est basé sur la distance entre les régions de duplication segmentaire, considérant que plus deux régions sont proches, plus les possibilités de recombinaison sont importantes (Sharp et al. 2005; Stankiewicz et Lupski 2002). Le second critère à prendre en compte est la taille de la séquence homologue et plus particulièrement la présence de segments présentant un haut niveau d'identité entre les régions de duplication segmentaire (Rubnitz et Subramani 1984; Reiter et al. 1998). Ces segments composés de séquences de

haute similarité, aussi appelés Minimal Efficient Processing Segments (MEPS), présentent des tailles variant généralement de 300 à 500 pb en longueur, même si des segments de plus petite taille ont été identifiés dans certains remaniements. Le dernier critère correspond à l'environnement direct situé aux alentours des régions de duplication segmentaire avec la présence de structures de séquences spécifiques capable d'induire des cassures double brin tels que des palindromes, des séquences mini satellites ou des transposons. L'ensemble de ces critères définit le caractère récurrent des réarrangements chromosomiques médiés par ces régions répétées, et par conséquent la présence de duplications ou de délétions récurrentes dans la population, possédant des bornes similaires.

Ce mécanisme médie certaines pathologies, tel que le syndrome de Williams-Beuren qui, dans environ 90% des patients, est dû à une délétion d'environ 1.6 Mb dans la région du 7q11.23, cette délétion étant médiée par deux séquences LCR flanquantes (Osborne 1999).

Cassure double brin et jonction des extrémités non-homologues

La jonction des extrémités non-homologues (NHEJ - non homologous end joining) (Carvalho et Lupski 2016) est le second mécanisme de réarrangement chromosomique. Il s'agit d'un des mécanismes utilisés par les cellules eucaryotes pour réparer les cassures double brin de l'ADN qui peuvent apparaître suite à des radiations ionisantes par exemple. En fonction de la phase cellulaire à laquelle se produit la cassure double-brin, deux mécanismes différents pourront être impliqués, soit une recombinaison homologue basée sur la chromatide sœur, soit le mécanisme de NHEJ. La recombinaison homologue permet une réparation de l'ADN sans altération de sa séquence, mais elle nécessite que la chromatide sœur soit proche au moment de la réparation et implique aussi un temps beaucoup plus important (plusieurs heures contre quelques minutes). Ces différentes raisons conduisent à l'utilisation préférentielle du mécanisme de NHEJ malgré le risque d'altération de la séquence ADN.

Le mécanisme de réparation par les NHEJ implique 4 étapes principales (Figure 6) (Lieber 2008). Le complexe protéique Ku va identifier les cassures double brin et venir se fixer sur cette dernière. Un dimère de protéines va ensuite s'associer à la structure afin de rapprocher les 2 brins séparés. Il s'en suit une étape importante pour comprendre la structure des NHEJ : une modification des terminaisons par des étapes de délétions et d'insertions de nouvelles bases de façon à rendre les terminaisons compatibles. Enfin, l'étape finale est celle de la ligation des 2 brins. C'est l'étape de modification des terminaisons qui va laisser une empreinte connue sous le nom de "cicatrices informatives" qui va aider à la détection de ce type d'événement.

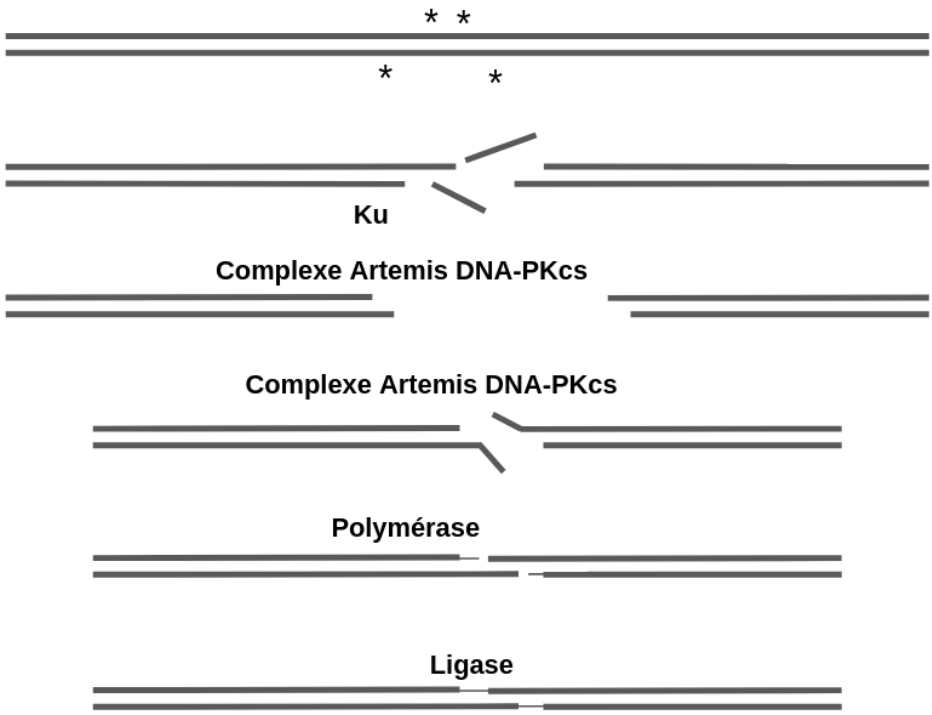


Figure 6 : Cassure double brin et jonction des extrémités non homologues
Suite à la génération de radicaux libres (*), perte de petit fragment d'ADN et apparition de cassure double brin. La protéine Ku va venir se fixer sur la cassure double brin. Le complexe Artemis-DNA-PKcs va éliminer quelques bases de part et d'autre de la cassure puis rapprocher les deux sections. Une ligase peut ensuite compléter le segment manquant avant que la ligase vienne joindre les deux sections. Adapté de Lieber et al., 2008

Ces cicatrices informatives sont généralement constituées de séquences de micro-homologie (entre 2 et 5 pb), ainsi que de petites insertions et duplications allant jusqu'à 25 bp (Nobile et al. 2002; Toffolatti et al. 2002; Hussain et al. 2021). De par leur nature, les cassures doubles brins ne nécessitent pas de régions répétées pour être provoquées, néanmoins dans environ 40% des cas présentés dans les travaux des équipes de Nobile et al. et de Toffolatti et al., les points de cassure ont été identifiés dans des régions répétées. Lorsque nous allons chercher à identifier de potentiels points de cassure, l'identification de ce type d'homologie pourra permettre de confirmer, ou du moins d'apporter du poids dans la validité de la variation détectée.

Enfin, il est à noter que les NHEJ sont impliqués dans les réparations des cassures double-brin somatiques générées durant le développement des lymphocytes B et T, et plus spécifiquement dans les recombinaisons V(D)J impliquées dans la génération de récepteurs antigéniques (Schatz et Swanson

2011), les séquences variables induites par l'étape de reconstruction du point de jonction apportant de la diversité supplémentaire aux récepteurs antigéniques.

FoSTeS (fork stalling and template switching)

Le dernier mécanisme basé sur la réplication de l'ADN a été identifié en 2007 par l'étude de Lee et al., (J. A. Lee, Carvalho, et Lupski 2007). Cette équipe travaillait sur la maladie de Pelizaeus Merzbacher (PMD), une maladie récessive liée à l'X majoritairement induite par la duplication non récurrente du gène *PLP1*, ainsi que des délétions non récurrentes et des variations ponctuelles perte de fonctions du même gène *PLP1*. L'analyse des différents points de jonction ne permettait pas d'expliquer par les mécanismes de NAHR ou de NHEJ l'origine de ces événements, et ce malgré la présence de micro-homologie au point de jonction rappelant les mécanismes de NHEJ. Le mécanisme suggéré est celui d'un décrochement de la fourche de réplication du brin original suite à un arrêt ou un ralentissement de la réplication. L'extrémité 3' change alors de matrice en utilisant une fourche de réplication proche (Figure 7).

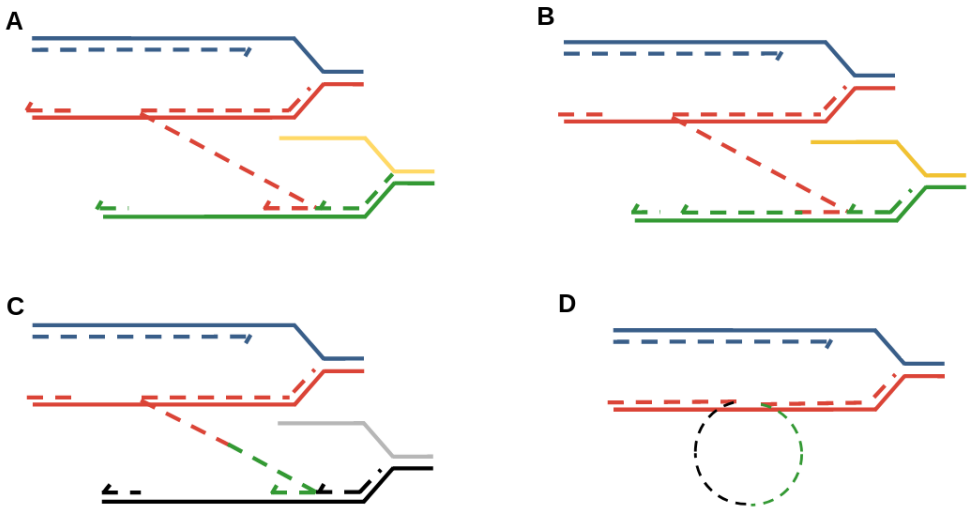


Figure 7 : Représentation schématique du mécanisme de FoSTeS
A. Suite à une liaison de l'ADN, une fourche de réplication (bleu et rouge) avec un brin ralenti (ligne rouge pointillée) va s'intercaler dans une seconde fourche (jaune et verte). B. Synthèse d'un segment complémentaire au niveau de la seconde fourche (ligne verte pointillée). C. Après le désengagement de la fourche, le brin ralenti de la fourche de réplication d'origine (bleu et rouge) peut s'intercaler dans une autre fourche de réplication (noire et grise), ce phénomène pouvant se produire plusieurs fois. D. Reprise de la réplication normale avec intégration de la séquence supplémentaire.

Cela nécessite des séquences identiques entre les 2 fourches, des séquences de micro-homologie étant suffisantes. Selon la complexité de la région et la possibilité d'avoir plusieurs arrêts ou pauses, il est

possible d'avoir plusieurs changements de brin et donc formation de réarrangements complexes. Comme visible sur les 2 exemples issus du papier de Lee et al. (Figure 8), ceci va conduire à des structures complexes incluant des duplications, des inversions et des insertions. Lorsque nous allons étudier des réarrangements à partir de données de séquençage, l'identification de séquence de micro-homologie pourra, là aussi, renforcer la confiance que l'on a dans les événements détectés.

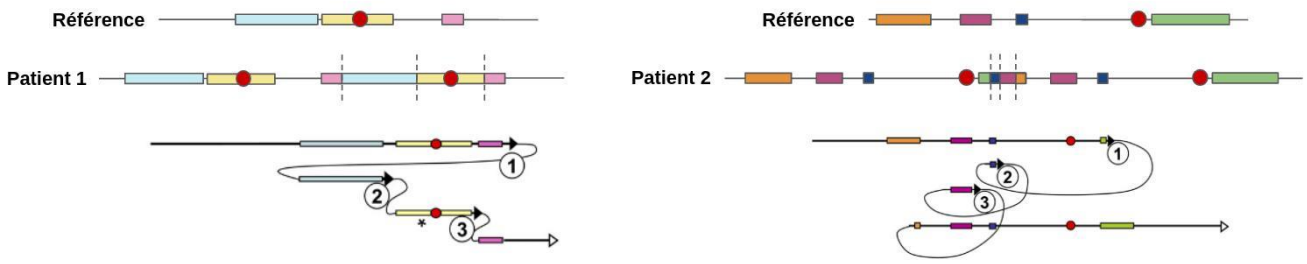


Figure 8: Exemple de réarrangement induit par les mécanismes de FoSTeS.
Pour chaque patient est représenté la séquence de référence avec les différents blocs de duplication, la séquence identifiée chez le patient. Les traits verticaux indiquent les points de cassure identifiés grâce aux séquences d'homologie entre les blocs. Le point rouge représente le gène *PLP1*. La partie inférieure représente l'ordre des événements. L'astérisque indique une orientation inconnue du bloc. Figure adaptée de Lee et al., 2007.

1.1.5. Impact des variations de structure pour l'homme

Les variations de structure dans l'évolution

Les variations de structure, et plus particulièrement les duplications de gènes ou de segments complets du génome ont eu un impact très important dans l'évolution du génome et l'adaptation des espèces à leur environnement (Dennis et Eichler 2016). L'apparition de nouvelles copies d'un gène permet une modulation directe par augmentation d'expression de la quantité de protéines présentes dans l'organisme. C'est le cas par exemple du gène *AMY1* codant pour l'amylase (Perry et al. 2007). Il a été démontré que le nombre de copies du gène *AMY1* est corrélé à la quantité de protéines d'amylase salivaire. Une pression de sélection est observée dans les populations ayant une alimentation riche en amidon (Figure 9), avec sélection préférentielle des individus ayant plus de copies du gène *AMY1* et ainsi pouvant plus facilement s'alimenter.

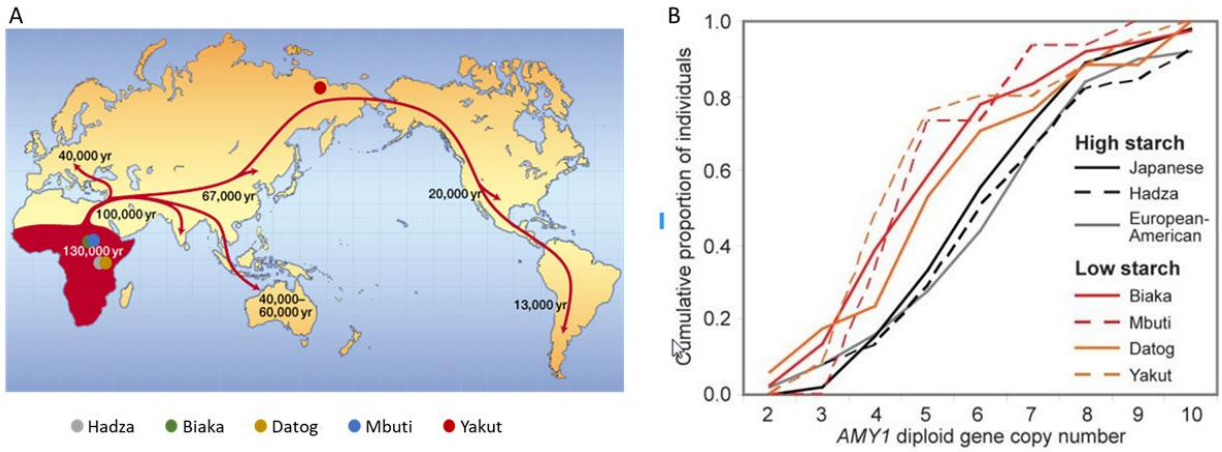


Figure 9 : Répartition du nombre de copies d'AMY1 dans différentes populations.
(A) Migration de la population humaine et localisation des différentes populations étudiées. (B) Proportion cumulative en fonction du nombre de copies d'AMY1 dans des populations avec des régions plus ou moins riche en amidon. Le nombre de copies dans les populations n'est pas corrélé avec la migration de l'espèce humaine, indiquant une sélection locale. Adapté de Perry et al., 2007

La multiplication du nombre de copies pourrait aussi être une des réponses au paradoxe de Peto (Caulin et Maley 2011), du nom de l'épidémiologiste Richard Peto. Le point de départ est le suivant : si toutes les cellules vivantes ont la même chance de devenir cancéreuses, les grands animaux possédant un plus grand nombre de cellules devraient développer plus de cancers que les humains. Or, il n'y a aucune corrélation entre le nombre de cellules d'un organisme et le taux d'apparition du cancer (Peto et al., 1975). Une partie de la solution à ce problème pourrait provenir de la multiplication du nombre de copies du gène suppresseur de tumeurs *TP53* dans les grands organismes. C'est le cas de l'éléphant qui possède une copie de *TP53*, ainsi que 19 rétrocopies de *TP53* (ou TP53 retrogene - TP53RTG) dont les trois quarts codent pour des protéines (Sulak et al. 2016). Ceci expliquerait en partie le fait que l'éléphant développe un nombre plus faible de cancer (Callaway 2015).

Par ailleurs, l'apparition de duplications de gènes complets permet de créer une redondance et ainsi libérer les gènes des contraintes sélectives. Plusieurs événements peuvent alors se produire sur cette nouvelle copie, tout d'abord l'intégration de cette copie à un autre endroit du génome peut interagir avec de nouveaux éléments régulateurs, modifiant l'expression du gène. Il est aussi possible que de nouvelles mutations dans les nouvelles copies puissent apporter de nouvelles fonctions (Lynch et Conery 2000). On peut citer en exemple le gène *SRGAP2* (Slit-Robo Rho GTPase activating protein 2), codant pour une protéine impliquée dans le développement cortical dans les phases de développement précoce en agissant comme un régulateur sur la migration et la différenciation neuronale (Dennis et al. 2012). La copie ancestrale *SRGAP2A* est partagée entre tous les primates. Une

première duplication partielle des exons 1 à 9 a conduit à l'apparition du gène *SRGAP2B* partagé avec certains primates puis un second événement conduisant à l'apparition du gène *SRGAP2C* contenant les mêmes exons 1 à 9, cette fois spécifique à l'homme. La copie ancestrale *SRGAP2A* est toujours sous pression de sélection, et une perte de copie fonctionnelle du gène par mutations perte de fonction ou par délétion d'une copie vont induire une pathologie du développement (Saitsu et al., 2012). Le gène *SRGAP2B* a dégénéré au fil de l'évolution et n'est maintenant que faiblement exprimé dans le cerveau, sans contrainte évolutive, ceci étant confirmé entre autres par la présence de gains et pertes de copie du gène, dont une délétion homozygote, chez des témoins sains. Enfin, le gène *SRGAP2C* est lui aussi toujours sous pression de sélection, et possède toujours un rôle biologique. On a donc ici avec un exemple avec un gène dupliqué en plusieurs copies, un gène d'origine toujours actif, une copie ayant développé une nouvelle fonction et enfin une copie ayant dégénéré pour devenir un pseudogène inactif.

Variants de structure et pathologie

Les réarrangements chromosomiques jouent un rôle important dans notre évolution, mais ces derniers sont aussi à l'origine de nombreuses pathologies, telles que les maladies du développement ou les troubles du spectre autistique.

Les plus grands événements rapportés correspondent aux maladies chromosomiques et impliquent la duplication ou la délétion d'un chromosome ou d'un bras complet de chromosome (Figure 10). Parmi les maladies chromosomiques les plus connues on peut bien sûr citer le syndrome de Down (trisomie 21), ou bien le syndrome d'Edwards (trisomie 18) (Outtaleb et al. 2020). Ces grandes duplications affectent de très nombreux gènes et vont par conséquent induire une altération très importante du fonctionnement et du développement cellulaire, induisant dans le cadre de ces deux trisomies des dysmorphies crânio-faciales caractéristiques ainsi que des troubles neurodéveloppementaux. Dans la cadre du syndrome d'Edwards, il y a une mortalité de plus de 95% dans la première année de vie. Ces événements concernant généralement tout un chromosome sont appelés aneuploïdies et ne sont pas nécessairement classés parmi les CNV car elles ne résultent pas des mêmes mécanismes.

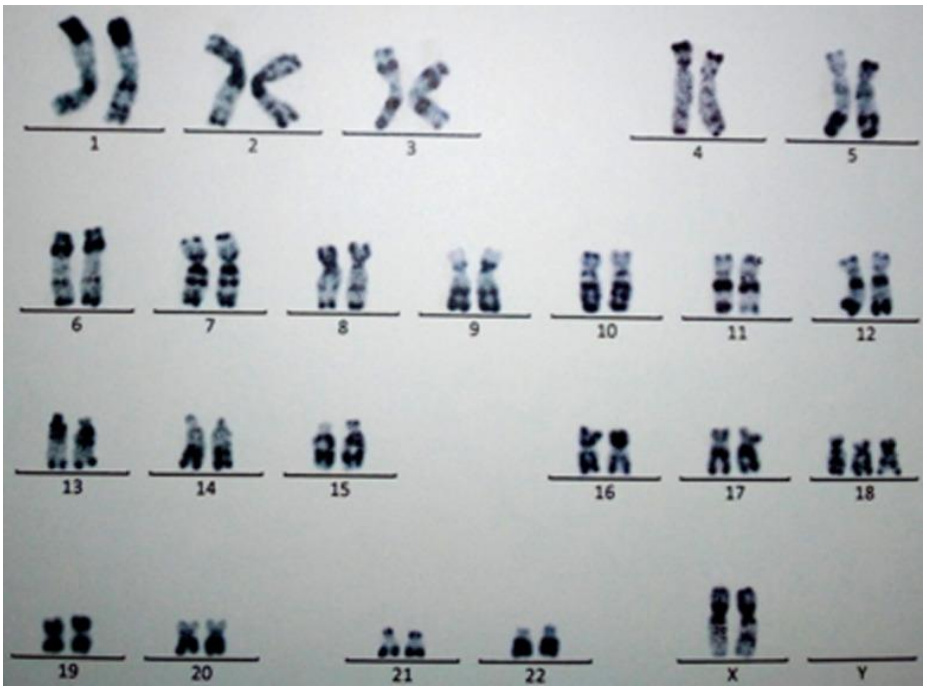


Figure 10: Etude cytogénétique en bandes RHG
Mise en évidence d'un caryotype anormal avec présence d'un chromosome 18 surnuméraire. Figure issue de Outtaleb et al., 2020.

Il existe aussi de nombreux exemples de variations du nombre de copies, par définition étant de plus petite taille, dans différentes pathologies telles que des déficiences intellectuelles, des troubles du spectre autistique ou bien des anomalies du développement. J'ai cité plus tôt la région du chr15q11.2 BP1-BP2 qui présente des délétions et duplications récurrentes de 443kb; les délétions de cette région étant actuellement la variation du nombre de copies la plus fréquemment rencontrée dans le trouble du neurodéveloppement (Rafi and Butler, 2020).

Un CNV n'a néanmoins pas besoin d'affecter une grande portion chromosomique pour induire un effet pathogène. C'est par exemple le cas des duplications emportant le gène *APP*, cette duplication étant responsable de formes autosomique dominante de formes précoces de la maladie d'Alzheimer (Anne Rovelet-Lecrux et al. 2006), comme nous le verrons dans la partie dédiée, mais aussi de délétions monoexoniques entrainant la perte de fonction d'une protéine.

Un autre exemple de CNV emportant un nombre restreint de gènes est celui de la maladie de Charcot-Marie-Tooth, une neuropathie périphérique affectant aussi bien les nerfs sensitifs que moteurs. Il a été identifié en 1991 la présence d'une duplication sur la région chromosomique 17p12 associée à la forme de type 1A (CMT1A) autosomique dominante de la maladie (Lupski et al. 1991) avant de définir

plus précisément le gène *PMP22* comme étant responsable de la pathologie (Lupski et Garcia 1992). Ce gène code pour une protéine membranaire présente au niveau de la myéline du système nerveux périphérique. De manière intéressante, l'apparition de mutations ponctuelles affectant *PMP22* peuvent induire une autre forme de la maladie de Charcot-Marie-Tooth (forme CMT1E). Il existe, en miroir de cette duplication (Figure 11), une délétion de *PMP22* qui est responsable d'une autre pathologie, la neuropathie héréditaire avec hypersensibilité à la pression (HNPP) (Chrestian 1993). Ceci indique une haplo-sensibilité du gène, c'est à dire que toute variation du nombre de copies va induire un changement phénotypique. Ces duplications et délétions de *PMP22* apparaissent avec des bornes identiques et dites en miroir. Ceci s'explique par les mécanismes qui permettent l'apparition de ces événements, médiés par le mécanisme de NAHR et la présence de régions de duplication segmentaire de part et d'autre de la région d'intérêt (Lupski 1999).

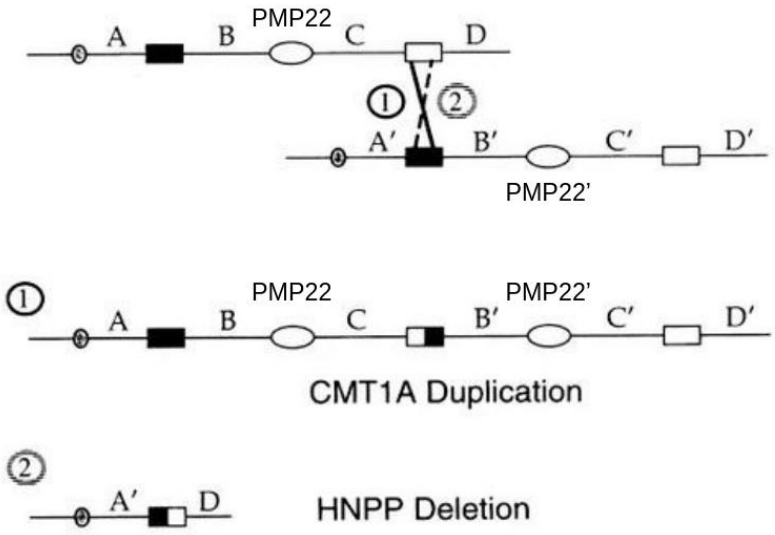


Figure 11 : Mécanisme de formation de la duplication CMT1A et de la délétion HNPP réciproque.
Les lettres A, B, C, D correspondent à des régions uniques flanquantes de *PMP22*, et les lettres A', B', C', D' aux mêmes sections sur le second chromosome. Les deux rectangles représentent les régions de duplication segmentaire médiant la recombinaison. Figure adaptée de Lupski, 1999

Différentes approches sont mises en œuvre pour déterminer la pathogénicité de ces variations génétiques. Outre les validations biologiques qui identifient l'impact de l'événement sur le fonctionnement cellulaire ou de l'organisme, des recommandations sont établies par diverses sociétés savantes à l'échelle nationale et internationale. C'est le cas de l'ACMG-ClinGen (Riggs et al. 2012) ou encore du réseau AChro-Puce en France (Dupont et al. 2022). Ces recommandations sont basées sur

l'ensemble des critères suivants : (i) la présence de régions codantes pour des protéines ou des éléments fonctionnels importants, (ii) la présence de gènes haploinsuffisants (pour les délétions) ou triplosensibles (pour les duplications), (iii) le nombre de gènes impactés, (iv) les connaissances sur les gènes impactés (e.g. un des gènes est déjà connu comme responsable d'une pathologie), (v) la fréquence de l'événement dans des bases de données et (vi) l'information de ségrégation de la variation dans une famille entre individus malades et sains lorsque l'information est disponible. En se basant sur ces critères, les variations sont ensuite classées dans des catégories allant de 'bénigne' à 'pathogénique'."

1.2. Méthodes de détection des réarrangements génomiques

Pour mettre au point une méthode fiable de détection des CNV à partir des données de séquençage à haut débit, il est crucial de la comparer avec un "Gold Standard" établi en utilisant des techniques de validation bien établies et reconnues. De plus, une connaissance approfondie des méthodes historiques et actuelles de détection des variants de structure est essentielle pour interpréter correctement les résultats et les comparer à ce qui existe déjà dans les bases de données. En effet, historiquement, les bases de données des variants de structure étaient principalement alimentées par les techniques de caryotypie, d'hybridation génomique comparative (ou CGH - Comparative Genomic Hybridization) et d'autres techniques ciblées. Cependant, avec l'avènement du NGS, de nombreux variants de structure à petite échelle ont été ajoutés à ces bases de données.

Il existe de nombreuses approches moléculaires pour détecter ou valider des variations de structure qui vont varier en fonction des régions étudiées (analyse pangénomique contre analyse ciblée) et de leur résolution.

1.2.1. Approches moléculaires pangénomiques

Le caryotype

La première approche employée pour l'identification des variations de structure est la technique du caryotype, pour la première fois mis au point en 1956 (Tjio 1978) par l'équipe de Levan et Tjio. Cette approche consiste à mettre en culture des cellules en présence de colchicine qui va bloquer les cellules en métaphase lors de la mitose. Ces cellules vont ensuite être incubées dans un milieu hypotonique

de façon à faire gonfler et éclater ces dernières puis on va venir les fixer sur une lame de verre. On colorera ensuite les cellules à l'aide de giemsa afin de révéler les bandes chromosomiques, puis on effectuera une observation au microscope des chromosomes ainsi marqués (Figure 12). Grâce à cette approche, il est possible d'identifier les paires de chromosomes ainsi que les possibles anomalies de grandes tailles : gain ou perte de chromosomes, gain ou perte de bras de chromosome, pour une résolution moyenne entre 5 et 10Mb.

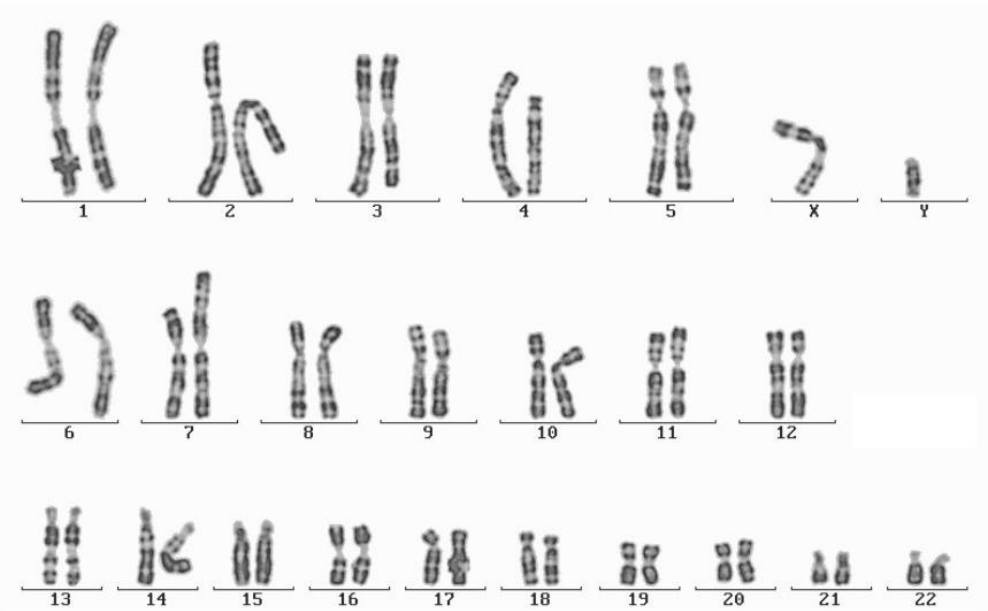


Figure 12 : Caryotype standard

La CGH (Comparative Genomic Hybridization) et les puces de CGH

En 1992, une nouvelle technique d'hybridation génomique comparative (comparative genomic hybridization) a permis de détecter des variations du nombre de copies entre deux individus (Kallioniemi et al. 1992). Le principe consiste à marquer l'ADN dénaturé (simple brin) de deux individus chacun avec un fluorochrome différent (Figure 13). L'un des ADN servira de référence tandis que l'autre sera celui que l'on cherche à tester. On va ensuite venir co-hybrider ces ADN, de la même manière que pour la technique de FISH, sur des chromosomes en métaphase d'un 3e individu. La quantité d'ADN venant s'hybrider étant dépendante du nombre de copies, on observera une coloration différente en fonction du nombre de copies relatives de la référence et de l'individu testé. En cas de délétion, l'ADN de référence marquera plus fortement et en cas de duplication ce sera celui testé.

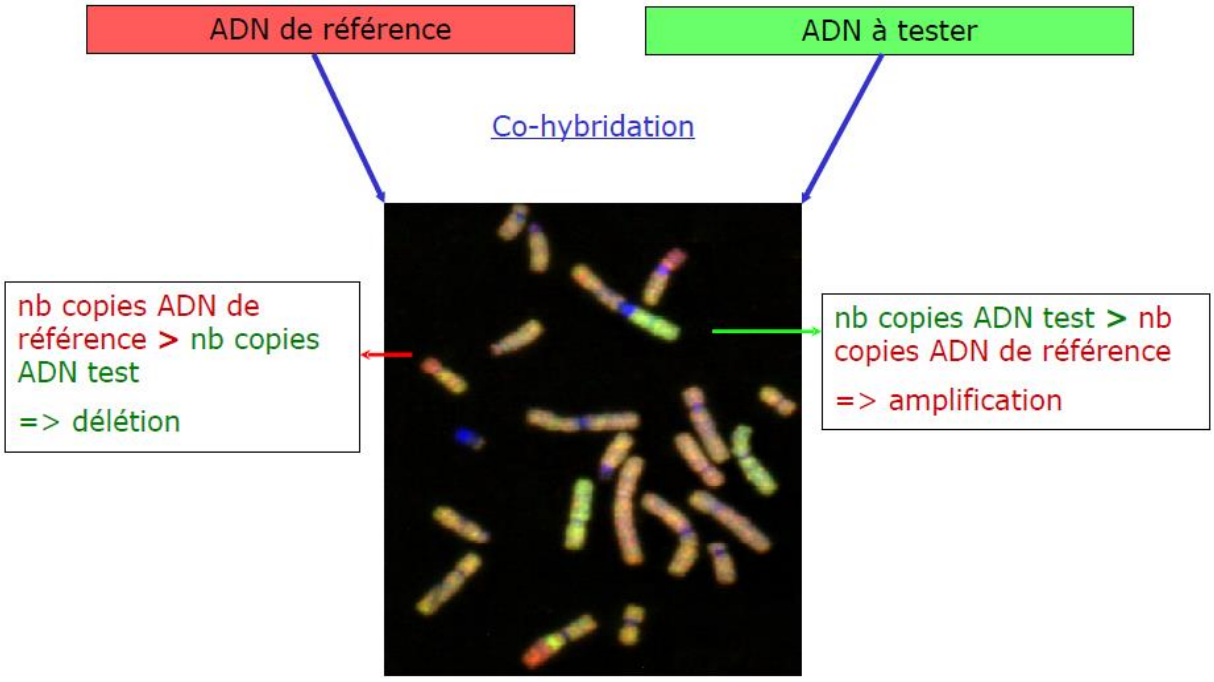


Figure 13 : Principe de la CGH.
L'ADN de l'individu contrôle est marqué par un fluorochrome rouge et celui de l'individu testé en vert. En cas de duplication chez l'individu testé, la fluorescence verte sera plus forte. En cas de duplication, c'est la fluorescence rouge du contrôle qui sera visible.

Cette technique, basée ici aussi sur une observation au microscope, ne permet pas d'avoir une résolution très élevée (à peu près 5 Mb) mais permet d'avoir une vue globale des délétions et des duplications. Une évolution de cette technique se base sur des puces à ADN (ou CGH-array). Utilisant toujours l'idée d'une hybridation compétitive entre un cas et un ou plusieurs contrôles, on vient déposer l'ADN marqué et fragmenté non plus sur des chromosomes en métaphase, mais sur des lames de verre sur lesquelles ont été fixées des sondes correspondant à des positions spécifiques et connues du génome. Chaque sonde est en réalité un spot sur la puce contenant plusieurs oligomères identiques de 60 nucléotides simple brin. A la différence de la CGH classique, on utilise souvent un pool de plusieurs individus en tant que témoins dans le but de lisser les CNVs des différents témoins, tous marqués avec le même fluorochrome. On va ensuite lire la fluorescence sur la puce à l'aide d'un scanner spécifique et identifier de potentielles délétions et duplications si plusieurs sondes consécutives émettent la même fluorescence (Figure 14).

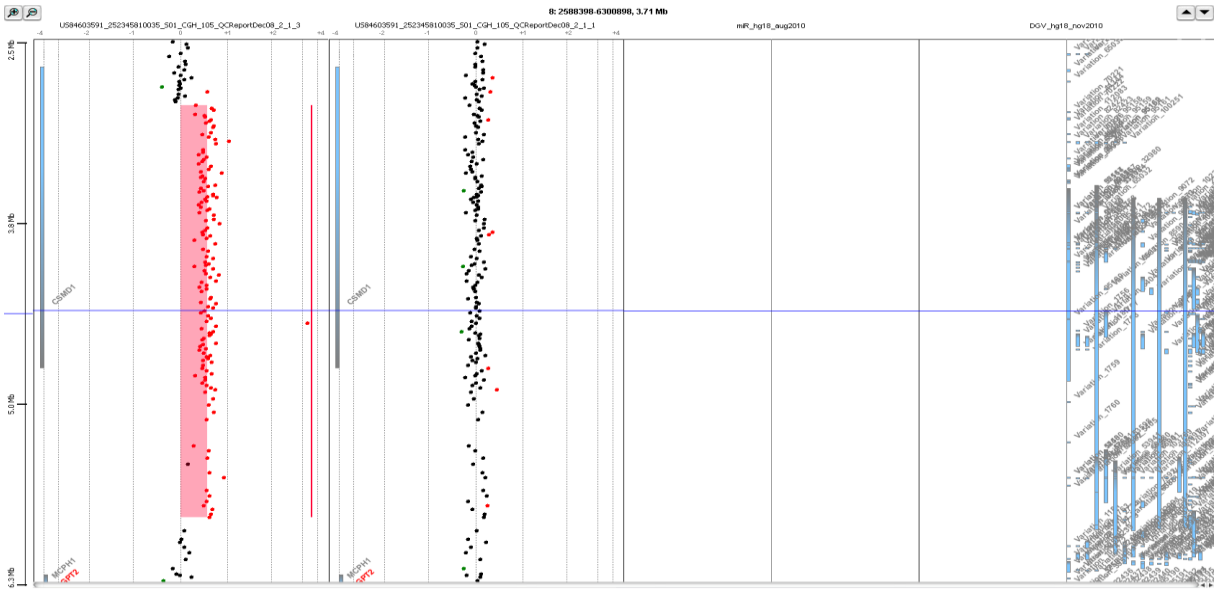


Figure 14 : Analyse de puce à CGH dans la région du gène CSMD1.
L'individu 1 présente une délétion partielle du gène *CSMD1*, contrairement à l'individu 2.

Plusieurs points sont à prendre en compte lorsqu'on utilise cette technique. Premièrement, la résolution est dépendante du nombre de sondes sur la puce. Les puces pangénomiques présentant le plus de sondes (1millions) permettent d'obtenir une résolution de 8 kb environ. Il existe des puces présentant un nombre inférieur de sondes permettant d'analyser plus d'individus en une seule analyse, mais avec une résolution plus faible. Il est donc nécessaire de faire un compromis entre la résolution souhaitée et la quantité de patients analysés en parallèle. Ensuite la répartition des différentes sondes n'est pas distribuée de manière égale sur l'entièreté du génome. Certaines régions complexes ne présentent pas de sonde, telles que les régions de duplication segmentaire, et certains gènes ne sont pas ou peu ciblés par les puces. C'est pourquoi il existe des puces dites "à façon" sur lesquelles des régions spécifiques ont été ciblées et des sondes supplémentaires ont été définies (Figure 15). De cette manière il est possible d'avoir une analyse précise d'une ou plusieurs régions d'intérêt spécifique à la pathologie étudiée.

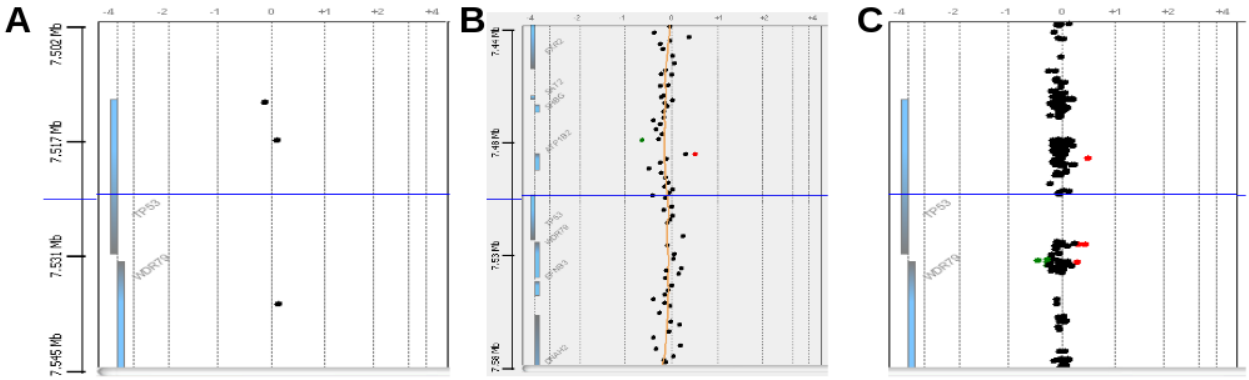


Figure 15 : Exemple de résolution entre différentes puces, région du gène TP53.
A. Puce 180 000 sondes, 4 échantillons B. Puce 1 million de sondes, 1 échantillon C. Puce custom 180 000 sondes “p53 PATHWAY”, 4 échantillons

Les puces à SNP ou puces de génotypage

Une première approche bioinformatique pour la détection des CNVs est basée sur l'utilisation des données de puces à SNP (ou SNP-array).

Cet outil a été développé initialement pour analyser une grande quantité d’allèles(D. G. Wang et al. 1998) en parallèle de façon à pouvoir génotyper des polymorphismes fréquents (ou SNP - Single Nucleotide Polymorphism) dans la population, c'est-à-dire ayant une fréquence dans la population supérieure à 1%. Ces puces sont utilisées dans des études d'association pangénomiques (ou GWAS – Genome-Wide Association Study) regroupant de grandes quantités de patients et de témoins. L'objectif de ces études est de pouvoir associer des variants fréquents en tant que facteur de risque ou protecteur pour des phénotypes particuliers. De nombreuses pathologies ont été étudiées en se basant sur des études de GWAS, tel que le diabète (Bradfield et al. 2011), l’obésité, les risques cardiovasculaires (Smith et Newton-Cheh 2015), mais aussi dans le cas de la maladie d'Alzheimer, comme nous le verrons par la suite.

Une utilisation détournée de ces puces a été mise en évidence par l’équipe de Bignell (Bignell et al. 2004) puis développée de manière plus large par le logiciel PennCNV (K. Wang et al. 2007) et permet d’identifier des variations du nombre de copies. Ils ont observé que l’intensité du signal total et le rapport d’intensité entre les deux allèles variaient en fonction du nombre de copies de la région. En combinant l’intensité totale du signal et la balance allélique des SNPs dans une région (Figure 16), il est possible d’identifier des gains ou des pertes de régions par cette approche, avec une résolution dépendant de la densité de la puce, comme c’est le cas pour les analyses par CGH array.

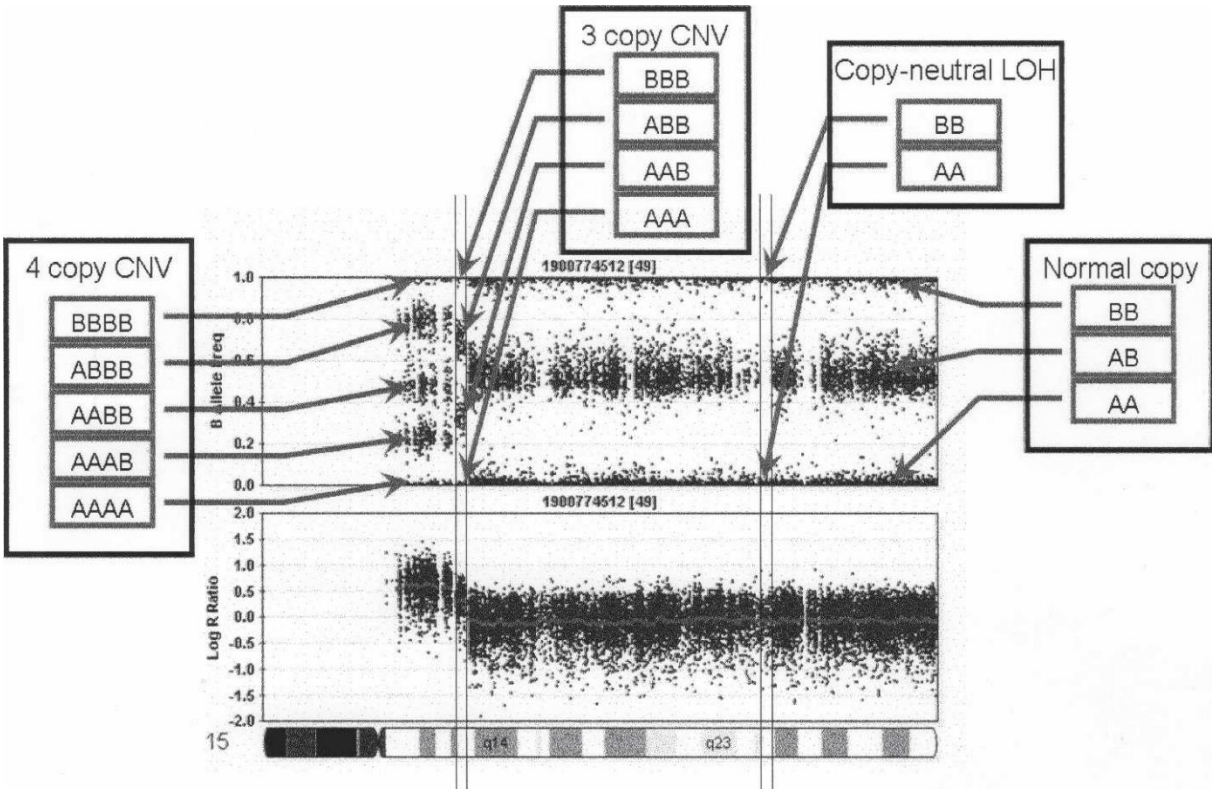


Figure 16 : Lecture du signal des puces de génotypage pour la détection de CNV.
Représentation de la balance allélique (partie supérieure) et de l'intensité du signal normalisée (partie inférieure) issue d'une puce de génotypage. Les balances alléliques attendues sont normalement de 0 (homozygote sauvage), 1/2 (hétérozygote) ou 1 (homozygote mutant). En cas de délétion ou de duplication, ces balances vont évoluer : 0 ou 1 dans le cas d'une délétion 0, 1/3, 2/3, 1 pour une duplication. Figure issue de Wang et al., 2007.

Cette approche est maintenant incluse dans les méthodes d'Analyse Chromosomique par Puce à ADN (ACPA) au même titre que la CGH-array, et permet, par une seule méthode d'obtenir à la fois le signal des SNP fréquents et les variations du nombre de copies. Avec cette approche, il est uniquement possible de détecter les événements déséquilibrés (délétions et duplications) avec une résolution dépendant de la densité de puce utilisée, généralement entre 30 à 50 kb.

Cartographie optique

La technique la plus récente d'analyse de structure de notre génome est basée sur la linéarisation de l'ADN et sur une cartographie optique par la plateforme Saphyr de la société Bionano Genomics (Figure 17). On va pour cela extraire l'ADN de manière classique, le dénaturer puis isoler les molécules à haut poids moléculaire, c'est à dire des molécules peu fragmentées. On marque ensuite l'ADN avec des sondes spécifiques sur tout le génome, ces sondes correspondant à des positions identifiées sur le génome. L'échantillon ainsi marqué est déposé dans une cartouche qui va permettre l'analyse. Cette

cartouche est constituée d'un ensemble de piliers puis de canaux qui vont permettre, lors de la migration de l'ADN, d'allonger et de linéariser la molécule. Une fois la molécule linéarisée elle passe devant un capteur qui va scanner la molécule et générer des images. Après que de nombreuses molécules ont traversé le système, les images vont être agrégées puis analysées par un algorithme qui va générer une séquence consensus des marqueurs. Cette séquence consensus va ensuite être confrontée au génome de référence de façon à identifier de potentiels événements : selon l'alignement des sondes de l'échantillon sur le génome de référence, on détecte de potentiels gains ou pertes de copies, mais aussi les insertions, les inversions et les translocations. En fonction de la distance des différents marqueurs il est possible d'atteindre une résolution de 500 pb. L'apport très important de cette approche est de permettre d'avoir en une seule analyse une vue complète de la structure d'un génome avec une résolution relativement élevée. Cette technique obtient de très bons résultats pour les détections des variations de structure mais présente comme principal défaut de ne pas fournir d'informations sur la séquence en nucléotides elle-même, contrairement aux puces à SNP (fournissant des informations sur les polymorphismes fréquents) ou les données de séquençage à haut débit, en particulier celles obtenues à partir des séquençages de longs fragments.

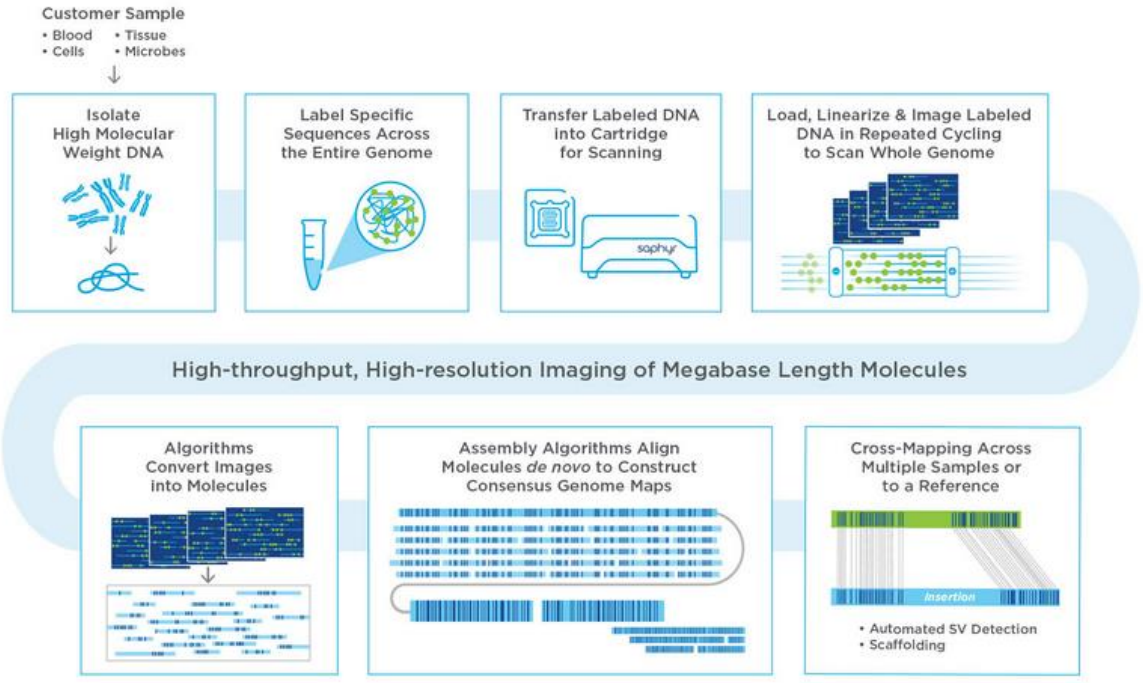


Figure 17 : Principe général de la cartographie optique.
Pipeline d'analyse de données utilisant la technologie Bionano représentant les différentes étapes du processus

1.2.2. Approches moléculaires ciblées

Technique FISH (Fluorescence In Situ Hybridization)

La technique du FISH permet de cibler spécifiquement une région chromosomique relativement grande. Elle consiste à fixer une sonde marquée à l'aide d'un fluorochrome qui va reconnaître une région spécifique de l'ADN (Figure 18). Pour ce faire, on effectue une dénaturation de l'ADN permettant d'obtenir de l'ADN simple brin, rendant possible l'hybridation avec la sonde. On effectue ensuite une observation au microscope, de manière similaire au caryotype. Cette technique présente plusieurs limites. Premièrement le fait que l'on ne puisse marquer qu'une ou 2 séquences à chaque expérience. Ensuite, la résolution reste limitée, à partir de 100 - 300 kb. Cette technique permet de détecter à la fois des gains ou des pertes de copies, mais aussi des translocations.

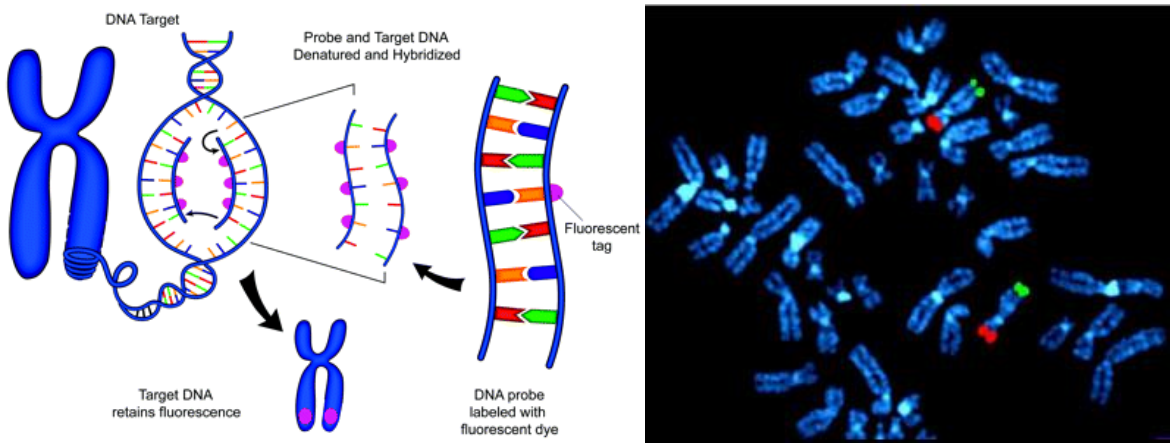


Figure 18 : Principe général de la technique FISH.
Principe général de la technique de Fish (à gauche) et visualisation de deux sondes au microscope à fluorescence (à droite) sans anomalie détectée.

Jusqu'à maintenant toutes les techniques présentées ne permettaient pas de descendre à une résolution inférieure à quelques dizaines de kb. Les approches suivantes permettent d'obtenir une résolution au niveau de l'exon.

PCR en temps réel

La plus ancienne technique permettant de quantifier un nombre de copies est la PCR quantitative en temps réel (ou Real-Time-qPCR) (Svanvik et al. 2000). L'amplification d'un fragment d'ADN par PCR passe par 3 étapes : une phase d'initiation, une phase exponentielle linéaire au cours de laquelle le nombre de copies est multiplié par 2 à chaque cycle et enfin une phase de plateau qui représente la quantité maximale d'amplicons générés. Le temps nécessaire pour atteindre la phase exponentielle sera d'autant plus rapide que la quantité d'ADN initiale que l'on cherche à amplifier est importante. On quantifie la quantité d'ADN à chaque cycle en utilisant soit un intercalant de l'ADN de type SYBRGreen (signal aspécifique qui permet de détecter l'ADN double brin), soit une sonde d'hydrolyse spécifique de la région. En utilisant 2 sondes différentes correspondant à 2 fragments distincts, il est possible d'établir une quantification relative de la quantité initiale des 2 fragments. Un seuil de fluorescence est choisi et on calcule le cycle de PCR auquel la fluorescence atteint ce seuil, que l'on appelle la valeur de Ct (Cycle threshold). Deux valeurs de Ct identiques indique que la quantité initiale d'ADN était équivalente. En revanche, des valeurs de Ct différentes indiquent une disparité dans le nombre initial de copies des deux fragments. Etant donné que la quantité d'ADN double entre chaque cible, il est même possible de déterminer le nombre de copies relatives entre les deux régions (Figure 19).

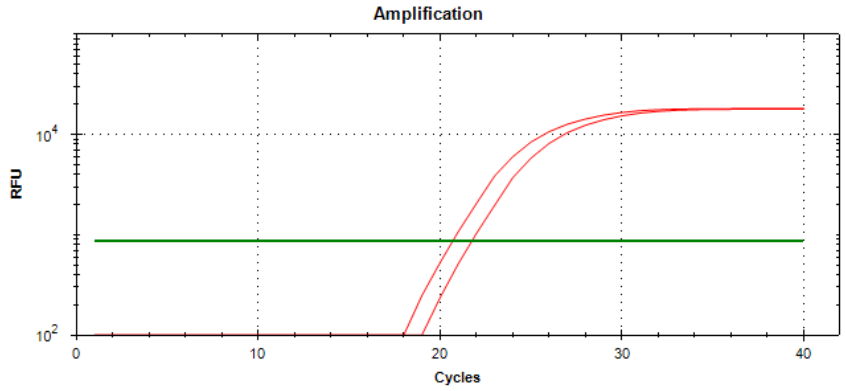


Figure 19 : Courbe d'amplification PCR en temps réel (ou RT-PCR).
Le seuil de fluorescence est défini manuellement de façon à se situer pendant la phase exponentielle (ligne verte). On mesure le nombre de cycle (ou cT) pour atteindre le seuil de chaque échantillon, et on compare les CT entre deux individus pour identifier de potentielles délétions ou duplications.

En plaçant une sonde dans une région où l'on a détecté un CNV et en la comparant avec une sonde placée dans un gène de ménage (c'est-à-dire un gène ne supportant pas une variation du nombre de

copies), il est possible de confirmer ou d'infirmer l'événement. Une limite importante de cette approche est que chaque expérience ne peut tester qu'une seule sonde, et dans le cadre d'un CNV emportant un ou plusieurs gènes, il sera nécessaire de faire plusieurs expériences pour valider les différents exons ou gènes emportés.

MLPA

La MLPA (Multiplex Ligation-dependent Probe Amplification) est une analyse basée elle aussi sur la technique de PCR et sur le fait que la quantité d'ADN amplifié lors de la phase exponentielle est proportionnelle à sa quantité initiale (Figure 20). Contrairement à l'approche par PCR en temps réel, l'idée est de définir plusieurs couples de sondes de façon à analyser en parallèle plusieurs exons et/ou gènes (Schouten et al. 2002). Pour cela, on définit une sonde complémentaire d'une région cible, et une seconde sonde en reverse distante d'une base, sur le même brin. Ces deux sondes comportent des couples d'amorces de PCR universelle séparés de la sonde par des séquences de tailles différentes (séquences stuffer). On utilise ensuite une ligase qui permet de lier les deux sondes si elles se sont hybridées sur la région d'intérêt. Les fragments ainsi formés sont ensuite amplifiés par PCR universelle, en parallèle, en arrêtant l'amplification pendant la phase exponentielle. L'intérêt d'utiliser des séquences stuffer de longueur variable est de pouvoir séparer les fragments par électrophorèse en fonction de leur taille sur un séquenceur capillaire. On obtient un signal pour chaque amplicon, et en comparant les résultats à un échantillon contrôle, il est possible de confirmer ou d'infirmer une variation du nombre de copies. L'intérêt de cette approche est de pouvoir travailler avec plusieurs amplicons en parallèle et donc de pouvoir analyser plusieurs exons et/ou gènes en parallèle.

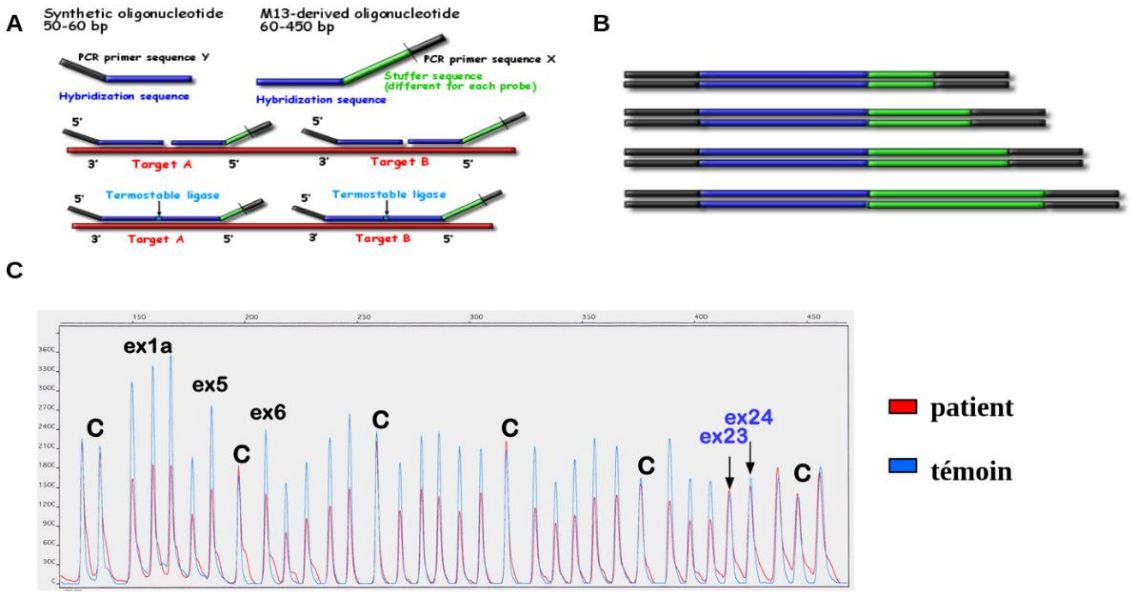


Figure 20 : Principe général de la technique MLPA.
A. Construction des couples d'amorces PCR sur deux cibles différentes. B. Obtention des amplicons de différentes tailles obtenus grâce à l'ajout de Stuffer de taille variable. C. Visualisation des résultats révélant une délétion partielle de BRCA1Δ1-22. Les pics "C" (contrôle) correspondent à un gène de ménage et permet la calibration des deux profils entre eux. Chaque pic correspond à un exon du gène, la différence d'intensité de signal indique ici une perte de copie chez le patient.

QMPSF

La QMPSF (Quantitative Multiplex PCR of Short Fluorescent Fragments) est aussi basée sur l'utilisation de la PCR en phase exponentielle et permet l'analyse en parallèle de plusieurs fragments (Charbonnier et al. 2000). Cette technique, conçue et développée au sein de notre laboratoire, est différente de la MLPA par la construction des fragments amplifiés : en QMPSF, on marque directement l'une des deux amorces de PCR pour chaque amplicon, et il n'y a pas de construction avec des séquences stuffer, ni de ligation, ni de PCR universelle (Figure 21). Il faut donc, au moment de la construction des couples d'amorce, sélectionner des tailles d'amplicons différents de façon à pouvoir multiplexer les amorces et diminuer le nombre de réactions. Le reste du processus sera le même, avec une comparaison d'un échantillon d'un individu contre un contrôle, et la séparation des différents amplicons par électrophorèse sur séquenceur.

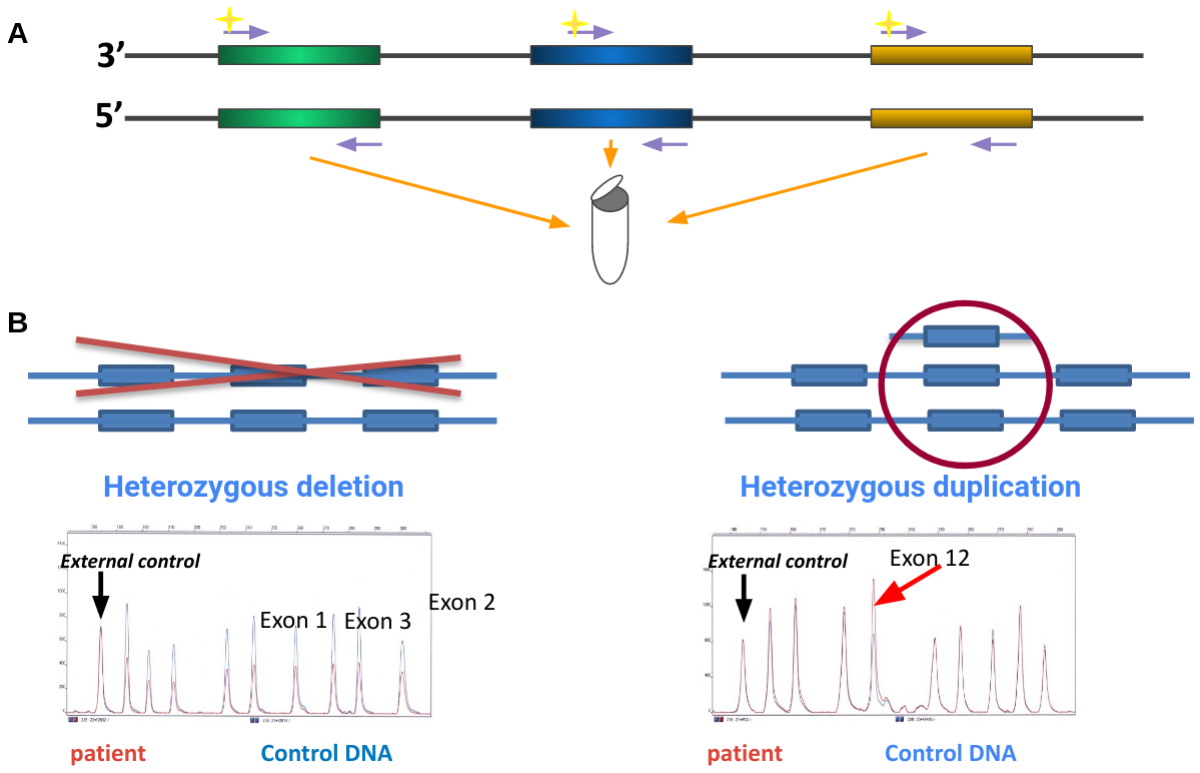


Figure 21 : Principe général de la technique de QMPSF.
A. Construction des différents couples d'amorces, chaque couple produisant un fragment de taille différente de façon à multiplexer les résultats. B. lecture des résultats et représentation schématique d'une délétion (à gauche) et d'une duplication (à droite) et "nom du graphe" de résultat. Ces graphes se lisent de la même manière que pour la MLPA.

PCR digitale en émulsion (Digital Droplet PCR)

Une des dernières approches en date permettant la détection et la validation des variations du nombre de copies est basée sur les techniques de PCR digitale. Cette dernière a été mise en place suite à la nécessité d'avoir à disposition une méthode permettant de quantifier des variations à faible fraction allélique dans le cadre de recherche de mosaïque, d'ADN tumoral circulant ou d'identification d'ADN fœtal circulant dans le sang maternel. Le principe repose sur l'idée de fractionner l'échantillon jusqu'à obtenir une dilution limite, c'est à dire de façon à obtenir 0 ou 1 molécule d'ADN par réaction. La technique initiale consistait à effectuer des dilutions successives (Sykes et al. 1992) pour effectuer la séparation des fragments, puis la technique a évolué avec l'apparition de la PCR digitale en émulsion (ou Digital Droplet PCR – DDPCR) (Vogelstein et Kinzler 1999). Celle-ci consiste à fractionner le mélange de PCR et d'ADN sous forme de microgouttelettes émulsionnées dans de l'huile, chaque gouttelette agissant comme des unités réactionnelles indépendantes (Figure 22). Contrairement aux autres approches (RT-PCR / MLPA / QMPSF), on effectue une PCR en point final, la quantification s'effectuant

par le nombre de gouttelettes contenant l'un des fragments d'intérêt. La ddPCR a été initialement mise en place pour identifier des variations ponctuelles mais elle a été adaptée pour être appliquée à l'identification des CNVs (Cassinari et al. 2019).

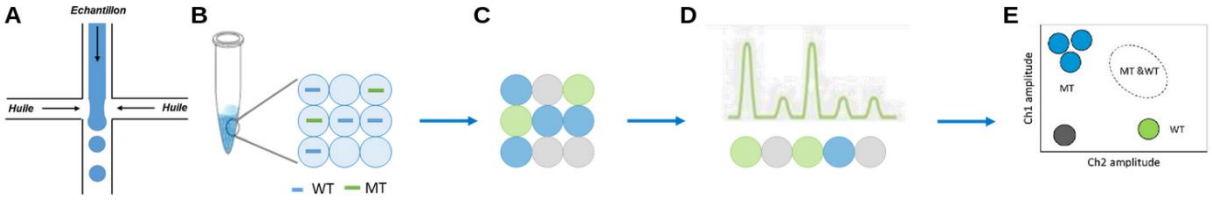


Figure 22 : Principe général de la ddPCR
A. Génération des gouttelettes par émulsion B. Après émulsion, chaque gouttelette contient entre 0 et 1 fragment C. PCR en point final D. Lecture de la fluorescence pour chaque gouttelette E. Génération du graphe de résultat.

Afin d'identifier les fragments, on utilise des sondes d'hydrolyse libérant deux fluorochromes différents, l'un pour une référence (généralement un gène de ménage, comme pour les autres techniques) et un second pour la séquence ciblée. Chaque gouttelette est ensuite analysée afin de déterminer quel fragment est amplifié. On calcule ensuite le ratio entre le nombre de gouttelettes positives pour la région cible et celles pour la région de référence afin d'obtenir le nombre de copies estimé de la région d'intérêt.

Comme indiqué, l'un des points forts de cette technique est de pouvoir définir de manière précise le nombre de copies et de détecter des événements ayant une faible fraction allélique. La contrepartie est, comme pour la qPCR, de ne pouvoir valider qu'un seul événement à la fois.

1.2.3. Variants de structure et séquençage massif en parallèle

L'évolution des technologies de séquençage

Le séquençage Sanger (Rizzo et Buck 2012) est la première technologie permettant une lecture automatisée de la séquence d'ADN. Elle consiste à effectuer une première amplification par PCR du fragment à séquencer, puis à effectuer une seconde PCR à partir d'une seule des deux amorces tout en ajoutant des didesoxynucleotides triphosphate (ddNTP) fluorescent dans le milieu de réaction. Ces ddNTP sont dits terminateurs, c'est à dire que leur incorporation va arrêter la synthèse du fragment en empêchant l'incorporation d'autres dNTP. L'intégration des ddNTP s'effectuant de manière

aléatoire, on obtient des fragments de taille variable. Il suffit ensuite d'utiliser un séquenceur capillaire qui va séparer par électrophorèse les fragments pour enfin produire la séquence. Ce séquençage de première génération permet d'obtenir des fragments de 500 à 1000 pb, ce qui est parfait pour effectuer le séquençage ciblé d'un exon. Cette approche est donc utilisée en routine pour des tests ciblés. Malgré les capacités actuelles des séquenceurs capillaires multicanaux (jusqu'à 96 en parallèle), il est impossible d'imaginer le séquençage de l'ensemble de l'exome ou du génome par cette approche, que cela soit au niveau matériel ou humain. On estimait dans le milieu des années 2000 un cout entre 5 et 30 Millions de dollars et environ 60 ans d'analyse sur un seul séquenceur capillaire (Bennett et al. 2005; Hert, Fredlake, et Barron 2008). Malgré l'évolution des séquenceurs capillaires et l'évolution des coûts depuis cette estimation, il est évident qu'il a été nécessaire de voir émerger de nouvelles technologies pour produire des analyses d'exomes et de génomes en très grands nombres.

En conclusion, le facteur limitant principal justifiant le passage à haut débit est la possibilité de lire en parallèle de grandes quantités de séquences. C'est dans ce contexte qu'ont été mises au point les technologies de deuxième génération, aussi appelées séquençage massif en parallèle. Il existe plusieurs types de séquenceurs possédant des chimies différentes, mais le principe général va rester le même pour toutes les approches.

Il est tout d'abord nécessaire de convertir l'ADN ciblé sous forme de librairie. Cela consiste à fragmenter l'ADN, sélectionner les fragments selon leur taille puis faire la ligation d'adaptateur. Ce sont ces adaptateurs qui vont permettre l'amplification (de la même manière que des amorces de PCR). Ces adaptateurs peuvent aussi jouer un rôle dans la fixation de la librairie sur une surface d'analyse, qu'il s'agisse de la surface d'une flowcell ou de solution en billes, ceci afin de permettre la parallélisation de multiples réactions de séquençage. Les fragments d'ADN qui constituent la librairie ne sont généralement pas lus dans leur intégralité. Au lieu de cela, il est effectué une lecture des extrémités de chaque fragment, ce qui donnera au final deux lectures par fragment, que l'on appelle des données pairées (paired-end sequencing). Ceci va jouer un rôle très important dans la recherche de variations de structure, comme nous le verrons par la suite (Figure 23).

Figure 23: Représentation schématique d'une paire de reads issus d'un fragment d'une librairie d'ADN. Chaque read sens est associé à son read antisen., La distance entre les deux reads est la taille d'insert, cette dernière est définie par la fragmentation de l'ADN et est normalement proche pout l'ensemble d'une librairie.

La technologie Illumina est actuellement la plus répandue dans les laboratoires de recherche et de diagnostic, et c'est aussi la technologie utilisée pour la production des exomes et des panels de gènes qui seront décrits dans la partie résultat. Cette technologie est basée sur l'utilisation de dNTP terminateurs réversibles fluorescents. Le principe général est présenté en Figure 24. La librairie d'ADN va être distribuée de manière aléatoire sur une flowcell sur laquelle sont disposés des primers d'accroche qui permettent la fixation des fragments préparés. Une première série d'amplifications en pont va permettre l'amplification des fragments fixés pour former des clusters, chaque cluster étant constitué, en théorie, du résultat de l'amplification d'un seul fragment d'ADN. La seconde étape consiste en une série de cycles définis comme suit : (i) fixation d'un dNTP terminal fluorescent, qui est, contrairement aux ddNTP employés dans la technologie Sanger, réversible, (ii) lecture de la fluorescence de chaque cluster directement sur la flowcell (iii) clivage de la partie terminale et du marquage avant le début d'un nouveau cycle.

Grâce à cette nouvelle génération de séquenceurs, il est possible d'accéder de manière précise à la séquence d'un individu, et de pouvoir identifier avec précision les variations ponctuelles ou les insertions et délétions de petites tailles. Néanmoins, il est difficile par ces approches d'obtenir une vue plus globale de l'organisation du génome. Ceci est d'autant plus vrai que la parallélisation du séquençage s'est accompagnée d'une réduction de la taille des fragments ; en comparaison du séquençage Sanger permettant d'obtenir des séquences de 1000pb avec une qualité relativement constante, la seconde génération de séquenceurs obtient des fragments allant de 50 à 250pb, avec une diminution de la qualité significative avec l'augmentation de la taille des séquences. Il est donc complexe d'obtenir la structure du génome au sens large et d'identifier de potentielles variations de structure avec des fragments courts même si des approches bioinformatiques tentent de répondre à la question, comme nous le verrons ensuite.

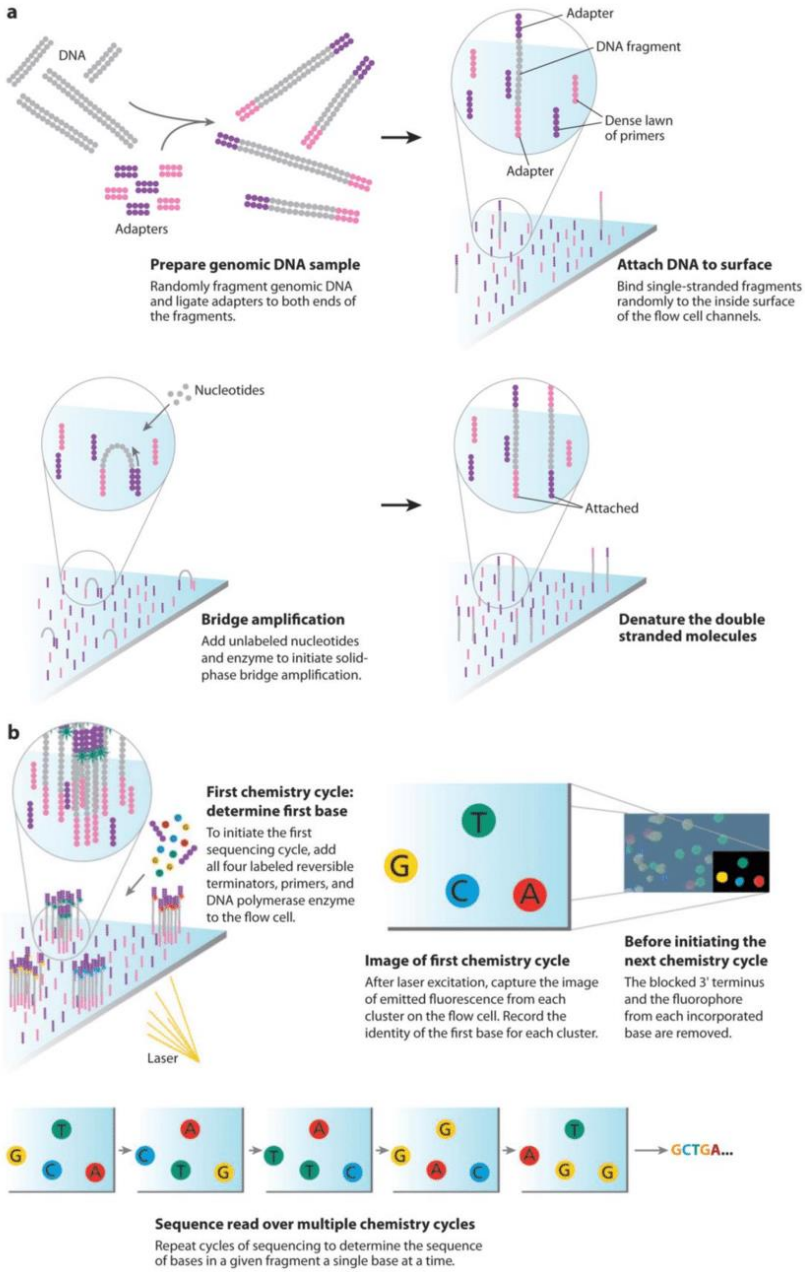


Figure 24: Présentation de la technologie de séquençage Illumina.

a : Préparation des clusters : Préparation de la librairie par fragmentation aléatoire de l'ADN puis ajout des adapteurs spécifiques au séquençage. La fragmentation peut être mécanique (généralement par sonication) ou chimique (enzyme de restriction). Les fragments sont ensuite fixés sur la flowcell par l'intermédiaire des adaptateurs. Après formation d'un pont entre deux points de fixation sur la flowcell, on effectue une première phase d'amplification sans fluorescence. Enfin, il y a dénaturation des molécules double brin et répétition de l'amplification en pont jusqu'à formation de cluster.

b : Phase de séquençage : (i) exécution de cycle de séquençage : ajout des dNTP fluorescent avec terminateur réversible, (ii) lavage pour retirer les dNTP non fixés (iii) lecture de la fluorescence pour déterminer la base (iv) clivage du groupement terminateur afin de permettre la fixation du dNTP lors du prochain cycle. Figure issue de www.illumina.com

Les séquenceurs capables d'analyser les longues molécules uniques (ou séquenceurs LongRead) constituent la 3ème génération. L'objectif premier de ces plateformes est avant tout l'identification de variants de structure, au détriment de la qualité de séquençage au niveau de la base. Les deux plateformes principales actuelles proviennent des entreprises PacBio et Oxford Nanopore (ONT) et fonctionnent sur deux approches complètement différentes.

La technologie d'ONT n'est pas basée sur une technologie de synthèse, mais sur une lecture directe de l'ADN. La flowcell utilisée est constituée de milliers de nanopores présents dans une membrane électro-résistante. L'ADN à séquencer est présenté sur une des surfaces de la flowcell, un complexe enzymatique va capturer et dénaturer la molécule d'ADN avant de la présenter à l'un des canaux. Lorsque la molécule d'ADN passe à travers le nanopore, cela induit un changement de pH qui est mesuré et interprété. Cette technique de séquençage permet d'obtenir des fragments pouvant atteindre 100 kb, mais avec une qualité de séquençage qui se révèle plus faible que les autres techniques, rendant difficile l'identification des variations ponctuelles, surtout à faible profondeur. Néanmoins, une nouvelle chimie, très récente, réduirait cette limitation.

Les séquenceurs de la société PacBio fonctionnent sur un système de séquençage par synthèse, nommé SMRT pour Single Molecule Real Time. La première étape consiste à fragmenter de façon limitée l'ADN, l'objectif étant d'obtenir des fragments de grande taille. Ensuite, des adaptateurs "en épingle" vont être ligués sur les fragments double brin, et vont venir former une boucle, permettant la circulation du fragment. Enfin, chaque fragment sera amplifié dans un puits présent sur la flowcell, chaque puits ne contenant, en théorie, qu'une seule molécule. De cette manière, chaque puits est un milieu réactionnel indépendant, et la fluorescence peut être suivie de manière indépendante dans chaque réaction. Enfin, la circularisation du fragment permet de lire plusieurs fois chaque séquence, permettant de corriger les erreurs de séquençage. Au final, ces reads corrigés après plusieurs relectures (ou reads HiFi pour High Fidelity) atteignent une qualité similaire voire supérieure aux technologies de seconde génération tout en produisant des reads entre 10 et 20 Kb.

Nous reviendrons plus en détail sur l'impact de ces séquenceurs de 3ème génération dans la partie Discussion de ce document.

Les approches bio-informatiques pour la détection de variation de structure

Il existe un très grand nombre d'outils bio-informatiques permettant la détection des variations de structure à partir de données de séquençage à haut débit de type Short read (Zhao et al. 2013; Tattini, D'Aurizio, et Magi 2015), chacun ayant plus ou moins de spécificité en fonction de l'approche employée ou des données biologiques ciblées. D'un point de vue bio-informatique, on peut classer les outils en 5 catégories en fonction de l'approche employée, ces approches s'appuyant sur un des aspects du séquençage et/ou des données paired-end. L'idée derrière tous ces outils et approches est de réussir à identifier les points de cassure et les associer pour définir et caractériser les événements. De la même manière que l'on recherchera les différences entre les données de séquençage et la référence pour identifier les variations ponctuelles, on va chercher à identifier les reads incohérents avec la référence.

Anomalie d'alignement des paires de reads

Les techniques de Paired-end mapping ou ReadPaired, se basent sur la position relative des deux reads d'une même paire. En théorie, les deux reads d'une même paire vont être en antisens et face à face l'un par rapport à l'autre, et à distance moyenne définie par la taille des fragments séquencés.

Avec cette approche, on va rechercher toutes les paires de reads présentant soit une orientation anormale, soit une taille d'insert incohérente avec le reste des données, et l'on pourra définir tout ou partie de l'événement en fonction de cette information. Si l'on observe une taille d'insert anormalement grande entre nos paires de reads, ceci indiquera une délétion d'une partie de la région intermédiaire. Dans de rares cas, si la taille de l'événement est inférieure à la taille de l'insert, il est possible de détecter l'insertion d'un élément. Si l'orientation relative des reads pairés vient à changer, ceci indiquera une possible inversion ou translocation, ces dernières pouvant être inter ou intra chromosomiques. Plusieurs exemples sont représentés sur la Figure 25. Le défaut de cette approche est de ne pas définir précisément les points de cassure des événements, ces derniers étant localisés dans la région intermédiaire entre les deux reads sans le localiser précisément. Enfin, pour que cette approche puisse fonctionner correctement, il est nécessaire que les reads ne soient pas chevauchants, et donc que la fragmentation lors de la préparation de l'ADN ne soit pas trop importante.

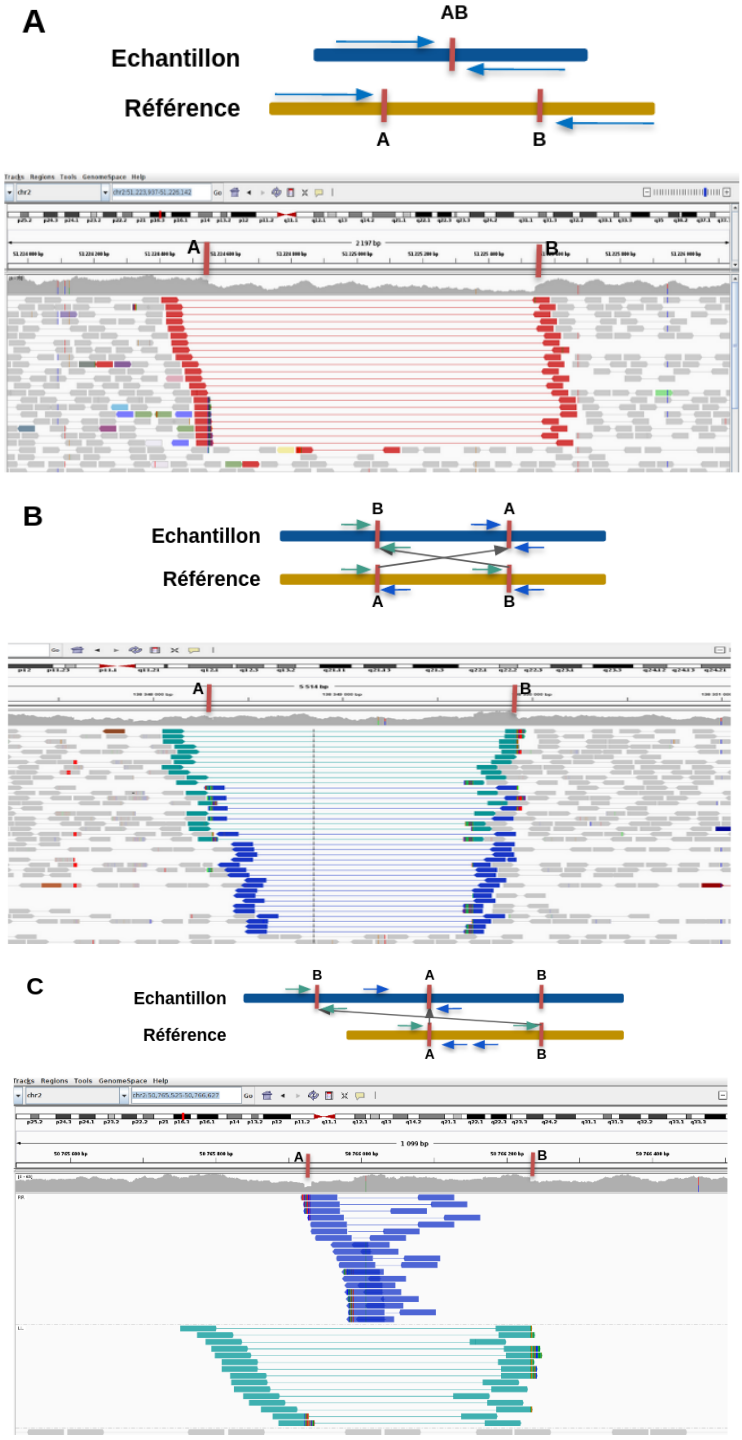


Figure 25 : Détection d'événements par utilisation des paires de read

Visualisation provenant du logiciel IGV. A. Délétion de 1kb. Les reads sont représentés par paires, apparaissant en rouge si la taille de l'insert est anormalement élevée par rapport aux restes des données. B. Inversion de 1,3kb. Les reads sont organisés par orientation des fragments par rapport au génome de référence : orientation normale des paires sens/anti-sens en gris, sens/sens en turquoise, et anti-sens/anti-sens en bleu. La profondeur de lecture et les fins de reads mal alignés semblent indiquer une duplication autour du point de cassure B. C. Duplication en tandem inversé de 250pb.

Alignements multiples et reads chimériques

La seconde approche est basée sur l'alignement multiple d'un read sur le génome (ou split Reads) ; il est possible d'aligner tout ou partie d'un read à plusieurs endroits d'un génome : si un read est entièrement aligné à plusieurs positions sur le génome, cela met en avant la présence de régions répétées, sans pour autant indiquer un possible événement. Ce sont donc les reads partiellement alignés qui vont être informatifs pour l'identification d'un événement (parfois appelés reads chimériques). Les deux alignements indiquent une jonction entre deux régions du génome normalement distantes. On interprète ensuite l'événement en fonction de la position des deux fragments : par exemple si les deux fragments s'alignent sur le même chromosome, il peut s'agir d'une délétion, sur deux chromosomes différents, d'une translocation. Cette approche permet aussi de détecter les points d'insertion d'éléments mobiles, en séquençant à la fois le point d'insertion et le début de la séquence insérée (Figure 26).

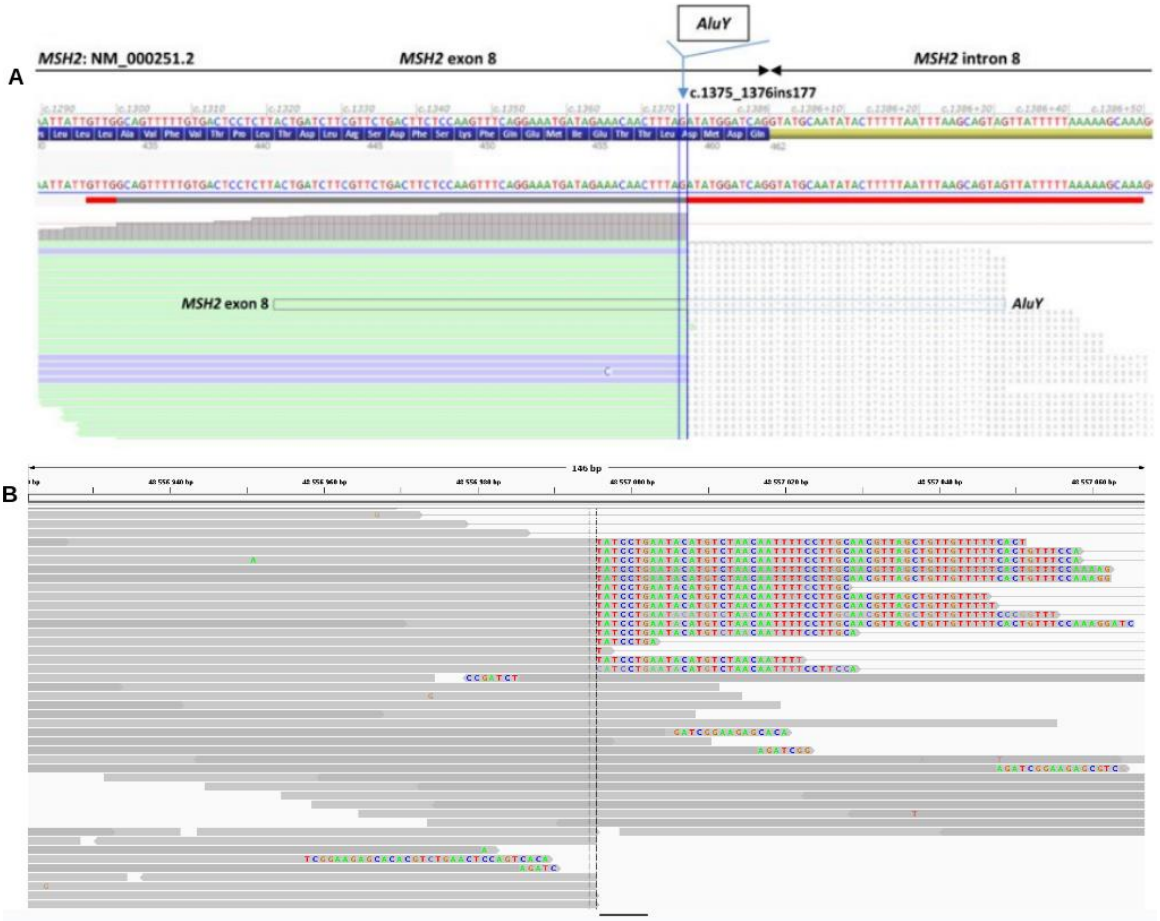


Figure 26 : Détection de point de cassure par les splitReads.
A. Insertion d'une séquence Alu dans l'exon 8 du gène *MSH2* (visualisation sur le logiciel Alamut Visual).
B. Délétion située au niveau de l'exon 1 du gène *SMAD4* (visualisation sur le logiciel IGV).

L'un des principaux avantages de cette technique est sa capacité à déterminer avec précision les points de cassure, permettant de lier les deux segments du génome concernés, quel que soit le type d'événement observé. Toutefois, elle présente des limitations. D'abord, pour que la technique soit efficace, il est impératif de pouvoir séquencer le point de cassure avec précision, ce qui, en fonction du type de séquençage utilisé (panel de gènes, exome, génome), n'est pas toujours possible. De plus, la séquence doit être suffisamment longue pour s'aligner correctement en différents endroits. Plus la séquence est courte, moins la localisation est précise. Enfin, étant donné que les réarrangements sont principalement induits par des régions répétées, identifier des reads chevauchant ces régions et ne s'alignant pas en plusieurs positions peut être complexe. Par conséquent, cette approche est surtout pertinente pour les insertions et les événements n'impliquant pas de régions répétées.

Profondeur de lecture

La troisième approche se base sur la distribution des reads le long du génome et la profondeur de lecture. En effet, l'idée du séquençage haut débit est de couvrir plusieurs fois les régions que l'on souhaite séquencer, qu'il s'agisse de capture ciblée, de séquençage d'exome ou de génome, afin de déterminer efficacement les deux allèles d'un individu et ce, malgré les possibles erreurs de séquençage. La quantité de reads s'alignant sur une partie du génome étant corrélée au nombre de copies de cette dernière, il est possible, en comparant la profondeur le long du génome, de déterminer le nombre de copies de chaque région (Figure 27). Si l'on souhaite appliquer cette approche, il est donc nécessaire de découper le génome en segments et d'effectuer le comptage de reads sur chaque élément, puis de les comparer entre deux. Il sera nécessaire pour cette approche de considérer différemment le séquençage avec ou sans capture. Dans le cas du séquençage de génome sans capture, les reads seront alignés de manière aléatoire et équilibrés le long du génome. Il est possible de directement segmenter ce dernier en bloc et de compter les reads sur ces derniers. Puis, par une approche statistique, on peut identifier les régions présentant une couverture supérieure ou inférieure au reste du génome, révélant des duplications ou des délétions. Dans le cas de séquençage incluant une étape de capture, il faudra considérer les données différemment. En effet, la capture nécessite l'utilisation de sondes nucléotidiques qui, en fonction de la composition de la séquence, vont s'hybrider de manière non uniforme en fonction du taux de GC (Benjamini et Speed 2012) des séquences ciblées et de la longueur des sondes utilisées. De plus, la taille de la région capturée varie, induisant encore de la variabilité entre ces dernières. Il est donc impossible de comparer directement

les différentes couvertures de chaque région ciblée entre elles chez le même individu et il est nécessaire de comparer plusieurs individus entre eux. On établit une référence basée sur un pool d'individu, et on compare ensuite l'individu testé au modèle de façon à identifier de potentielles délétions et/ou duplications. L'intérêt de cette approche est qu'elle est adaptée aux données de séquençage ciblé qui sont déjà décomposées en régions pour le comptage. Cette approche est par contre limitée en deux points : tout d'abord, elle ne peut être utilisée que pour la détection de variations du nombre de copies (délétion et duplication) et dans le cas des duplications, il n'y aucune idée concernant la localisation de la séquence dupliquée. Dans le cas de ces travaux, nous avons sélectionné le logiciel CANOES (Backenroth et al. 2014) comme outil pour construire notre pipeline.

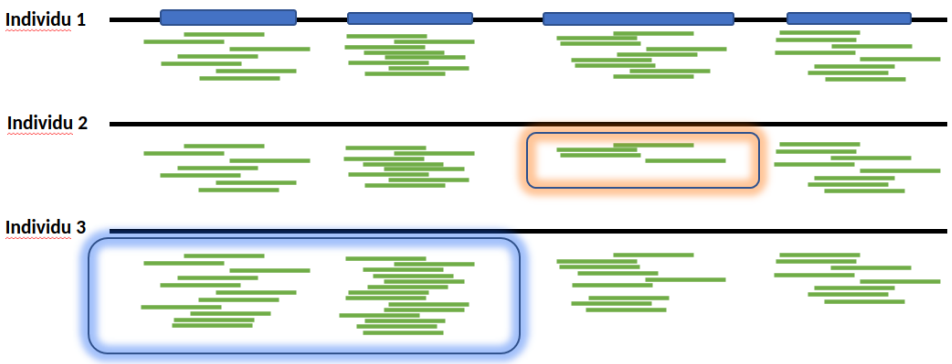


Figure 27 : Représentation schématique de l'utilisation de la profondeur de lecture
Les rectangles bleus correspondent aux cibles du kit de capture, les lignes vertes aux reads alignés pour chacun des 3 individus. En appliquant un modèle statistique et en comparant les différents individus entre eux, on détecte des délétions (encadrement rouge) ou des duplications (encadrement bleu)

Assemblage de novo

Enfin la dernière approche consiste à ne pas aligner les données séquencées sur un génome de référence, mais à effectuer un assemblage "de novo" sans a priori. Une fois le nouvel assemblage terminé, celui-ci est comparé à une référence qui peut être soit le génome de référence, soit d'autres données plus spécifiques au projet. En prenant l'exemple d'une analyse en trio, on peut considérer d'effectuer l'assemblage de novo des 3 individus séparément (les deux parents sains et le cas index malade), puis comparer les données de l'enfant à ses parents. De cette manière, il est possible d'identifier les haplotypes provenant du père ou de la mère, puis d'identifier toutes les structures aberrantes chez l'enfant correspondant à un événement de novo.

L'intérêt de cette technique est de s'affranchir de la notion de référence unique figée qui ne prend pas en compte la grande variabilité des haplotypes existant dans la population. Le défaut de cette

approche est la nécessité de faire un assemblage de novo, qui nécessite de séquencer le génome à une forte profondeur. De plus, les temps de calcul informatique pour l'assemblage sont beaucoup plus importants que pour un simple alignement. L'assemblage de novo peut aussi être utilisé pour définir de manière plus précise les points de cassure en sélectionnant uniquement les reads présentant des caractéristiques spécifiques (reads multi alignés, régions spécifiques, etc). En se limitant uniquement à une partie des données, il est possible de s'affranchir d'une grande partie des calculs d'assemblage. La limitation de cette approche sera la même que l'utilisation des reads chimériques s'appuyant sur les alignements multiples : si l'événement est médié par des régions répétées, les reads utilisés pour l'assemblage de novo pourraient ne pas être assez long pour identifier les deux extrémités du point de cassure.

Approches combinatoires

Il faut bien sûr considérer que, pour toutes ces approches, c'est l'accumulation de preuve, c'est à dire de reads informatifs, qui vont confirmer ou non un événement : de la même manière que des erreurs peuvent se produire lors du séquençage, il peut se produire des erreurs lors de l'alignement des reads conduisant à des faux positifs. Ces erreurs lors de l'alignement seront d'autant plus grandes que les régions sont complexes à séquencer : plus une région est répétée, plus il est difficile de correctement aligner les reads sur cette dernière, conduisant à de potentiels faux positifs. Chacune des quatre approches ayant plus ou moins de réussite en fonction de l'événement, il peut être intéressant d'utiliser une combinatoire de ces 4 approches lorsque l'on va essayer de détecter des événements. La majorité des outils exploitant les données de 3ème génération, tel que Sniffles (Smolka et al. 2022), pbsv (Töpfer 2022) ou encore svim (Heller et Vingron 2019), utilisent ce type d'approche combinatoire afin d'exploiter au mieux les données disponibles mais c'est aussi le cas pour certains outils exploitant les données de 2ème génération. On peut prendre par exemple le logiciel GRIDSS (Cameron et al. 2017) qui utilise 3 des 4 approches : à partir des fichiers d'alignement BAM, GRIDSS extrait tous les reads présentant des anomalies, c'est-à-dire les paires de reads mal orientés ainsi que les reads chimériques. À partir de ces séquences il va reconstruire, comme le ferait un assemblage de novo, les séquences des points de cassure sous forme de séquence consensus. Ces dernières seront réalignées sur le génome pour améliorer la localisation des points de cassure.

Afin de combiner les différentes approches, une autre possibilité est d'utiliser plusieurs outils se complétant les uns les autres. On pourra par exemple utiliser un logiciel se basant sur la profondeur

de lecture afin d'identifier les délétions et les duplications, puis de le compléter avec un logiciel travaillant sur les données chimériques pour localiser de manière précise les points de cassure.

Quelles approches pour quelles données ?

Chacune des approches que nous venons de décrire ne sera pas applicable à tous les types de données de séquençage. En effet, nous avons vu jusqu'à maintenant que la majorité des événements sont médiés par des régions répétées et/ou complexes, signifiant que les points de cassure que l'on cherche à identifier sont localisés dans ces régions. Or, ces séquences répétées sont généralement situées dans des régions non codantes (intergéniques ou introniques). Cela signifie que si l'on utilise des données de séquençage obtenues après capture d'exons, nous aurons peu de chances d'avoir les points de cassure précis des réarrangements. Au final, si l'on travaille sur des données de séquençage avec capture, seule la technique basée sur la comparaison des profondeurs de lecture sera vraiment efficace. Malheureusement, cela signifie que l'on devra se limiter à la détection des variations du nombre de copies, la détection des réarrangements équilibrés étant impossible par cette approche. Bien sûr, l'utilisation d'autres outils reste possible (voire souhaitable, dans le cadre du diagnostic), pour ne pas manquer certaines insertions d'éléments mobiles ou points de cassure séquencés par exemple, mais dans le cadre d'une large étude visant un point de vue pangénomique, les outils n'utilisant pas la profondeur de lecture sur des données issues de capture d'exons ne seront pas exploitables. En ce qui concerne le séquençage de génome en revanche, la combinaison d'outils semble être l'approche la plus efficace, sachant les intérêts et les limites de chacune des approches. Les outils utilisés génèrent pour certains de nombreux artefacts ou points de cassure isolés ne définissant pas un évènement, les approches combinatoires, bien que très demandeuses en calcul, s'imposent comme la règle pour le séquençage de génome (R. L. Collins et al. 2019; W.-P. Lee et al. 2021).

Tableau 1 : Liste non exhaustive de logiciel bioinformatique de détection de variation de structure

Outil	Approche bioinformatique	Type de données	Référence
CANOES	RD	SR capture	(Backenroth et al. 2014)
XHMM	RD	SR capture	(Fromer et al. 2012)
ExomeDepth	RD	SR capture	(Plagnol et al. 2012)
CoNIFER	RD	SR capture	(Krumm et al. 2012)
cn.MOPS	RD	SR WGS	(Klambauer et al. 2012)
CANVAS	RD	SR WGS	(Roller et al. 2016)
CNVnator	RD	SR WGS	(Abyzov et al. 2011)
BreakDancer	RP	SR WGS	(K. Chen et al. 2009)
PEMer	RP	SR WGS	(Korbel et al. 2009)
Pindel	SR	SR WGS	(Ye et al. 2009)
Cortex	AS	SR WGS	(Iqbal et al. 2012)
Magonlya	AS	SR WGS	(Nijkamp et al. 2012)
Manta	RP / SR	SR WGS	(X. Chen et al. 2016)
Lumpy	RP / SR	SR WGS	(Layer et al. 2014)
Delly	RP / SR	SR WGS	(Rausch et al. 2012)
Hydra	RP / AS	SR WGS	(Quinlan et al. 2010)
GenomeStrip	RP / SR/ RD	SR WGS	(Handsaker et al. 2011)
SVDetect	RP / RD	SR WGS	(Zeitouni et al. 2010)
GRIDSS	RP / SR / AS	SR WGS	(Cameron et al. 2017)
Sniffles2	RD / SR / AS	LR	(Smolka et al. 2022)
Picky	SR	LR	(Gong et al. 2018)
PBHoney	SR	LR	(English, Salerno, et Reid 2014)
svim	SR	LR	(Heller et Vingron 2019)
pbsv	SR / AS	LR	(Töpfer 2022)

RD : Read Depth. **SR** : Split Read. **RP** : Read Pair. **AS** : assemblage Denovo. **SR** : Short Read. **LR** : Long Read. **WGS** : Whole Genome Sequencing

Au total, de nombreux outils bioinformatiques existent pour détecter les variations du nombre de copies et autres variations structurales à partir de données de NGS. Les performances des outils, seuls ou en combinaison, sont rarement évaluées et dépendent à la fois des données et des procédures d'analyse. Il n'existe néanmoins aucune solution parfaite et pas de réel gold standard. Dans le cadre de l'étude de remaniements rares individuels pour le diagnostic de maladies monogéniques, il reste impératif de confirmer l'existence des variations structurales par une technique indépendante, même si la visualisation des alignements dans les logiciels dédiés (comme IGV, par exemple) permet, lorsqu'il existe des arguments concordants, et en particulier dans des données de génome complet, d'avoir une forte probabilité de l'existence du remaniement. Les laboratoires du plan France Médecine Génomique 2025 rendent par exemple des résultats de remaniements sur des approches bioinformatiques à partir des données de séquençage de génome, lorsque le faisceau d'arguments en faveur du remaniement est suffisant. Néanmoins, l'utilisation pour le conseil génétique rend nécessaire la confirmation par une technique ciblée. En approche pangénomique comme dans les études cas témoins avec des données massives, les ADN des individus inclus sont rarement disponibles pour une confirmation orthogonale des remaniements d'intérêt. Il est donc nécessaire de connaître les performances et les limites de détection des outils employés. Ce travail d'évaluation a constitué la première étape de mon travail, avant d'appliquer notre approche sur un grand jeu de données dans un second temps.

1.3. Génétique de la maladie d'Alzheimer

1.3.1 Clinique et étiologie de la maladie

La maladie d'Alzheimer est la maladie neurodégénérative la plus fréquente au monde. Elle entraîne une neurodégénérescence progressive responsable de troubles cognitifs multi-domaines, commençant le plus souvent par une altération de la mémoire épisodique, s'étendant ensuite à d'autres domaines cognitifs et responsable d'une altération de la vie quotidienne entraînant un trouble neurocognitif majeur également appelé démence. Le facteur de risque le plus associé à la maladie d'Alzheimer reste le vieillissement, si bien que cette pathologie est un problème de santé publique majeur croissant dans de nombreuses populations.

La maladie d'Alzheimer est définie sur le plan neuropathologique par la présence de deux types de lésions : les plaques amyloïdes ou plaques séniles constituées en leur cœur de peptides amyloïde beta

(Aβ) agrégés, intra parenchymateuses, et d'autre part des dégénérescences neurofibrillaires, intra neuronales, constituées de protéines Tau hyper et anormalement phosphorylées.

La maladie d'Alzheimer est une maladie complexe chez la plupart des malades, elle est alors causée par une agrégation de facteurs génétiques et de facteurs non génétiques, dits environnementaux. La composante génétique dans l'étiologie de la maladie d'Alzheimer est considérée comme élevée, suite aux études de jumeaux qui ont permis de comparer la concordance de diagnostic de maladie d'Alzheimer chez des jumeaux monozygotes (19-41%) par rapport à des jumeaux dizygotes (41-60%), l'hypothèse étant que la plupart des jumeaux partagent leur environnement mais que les dizygotes et monozygotes diffèrent par leurs variations génétiques (Gatz et al., 2006). Ce type d'étude a également conduit à mesurer l'héritabilité, qui correspond à la part de variance d'un trait phénotypique liée à des facteurs génétiques dans une population donnée. Nous n'aborderons pas en détails ce concept dont l'utilité est controversée. Ici, nous retiendrons simplement que les études de jumeaux permettent d'affirmer l'importance des facteurs génétiques dans la maladie d'Alzheimer.

1.3.2. Formes autosomiques dominantes

Les formes précoces de la maladie d'Alzheimer (survenue des premiers symptômes avant l'âge de 65 ans, EOAD - Early-Onset Alzheimer Disease) concernent environ 5% des malades (Sirkis et al. 2022). Bien que les études de jumeaux n'estiment pas précisément la part des facteurs génétiques dans la maladie d'Alzheimer à début précoce, il est communément admis que la part des facteurs génétiques est très largement majoritaire parmi ces patients. Il existe des formes purement génétiques également appelés formes héréditaires de la maladie d'Alzheimer, de transmission autosomique dominante (ADEOAD – Autosomal Dominant EOAD). Ces formes héréditaires sont causées par des variations pathogènes dans un des 3 gènes suivants : *APP*, *PSEN1* ou *PSEN2*. Les formes héréditaires représentent probablement moins de 15% des malades avec une forme précoce (Schramm et al. 2022). La pénétrance est habituellement totale à l'âge de 65 ans pour l'écrasante majorité des mutations, bien qu'il existe quelques variations de pénétrance réduite à 65 ans. L'âge de début moyen est en fait bien inférieur à 65 ans pour la plupart des mutations, on retrouvera des âges moyens par gène de 51, 44 et 54 ans, respectivement pour *APP*, *PSEN1* et *PSEN2* (Lanoiselée et al. 2017).

Ces mutations ont contribué à formuler puis à valider l'hypothèse de la cascade amyloïde (Hardy et Selkoe 2002), qui place le peptide Aβ au centre de la physiopathologie de la maladie d'Alzheimer. En

effet, le gène *APP* code pour le précurseur du peptide bêta-amyloïde, et les gènes présénilines 1 et 2 (*PSEN1* et *PSEN2*) codent pour la sous-unité catalytique du complexe gamma secretase qui clive le précurseur du peptide bêta-amyloïde permettant de produire du peptide Aβ. Toutes ces mutations mènent soit à une augmentation de la production de peptide Aβ, soit une augmentation du ratio entre les formes longues par rapport aux formes courtes, augmentant ainsi l'agrégabilité. Certaines mutations au centre de la séquence codante pour Aβ dans *APP* entraînent une augmentation de l'agrégabilité de ce peptide et ont une tendance plus importante à s'agréger dans les vaisseaux du cerveau, causant une angiopathie amyloïde cérébrale (Grangeon et al. 2021; Voigt et al. 2022).

1.3.3. Facteurs de risques

Il est maintenant considéré que la très grande majorité des malades, en dehors des formes précoces à transmission autosomique dominante dont la cause a été établie, sont des formes complexes avec une part probablement variable de facteurs génétiques. De nombreux facteurs génétiques ont été identifiés ces dernières années par des techniques variées. Le facteur de risque principal, tant en termes d'effet que de fréquence, reste l'allèle ε4 du gène *APOE*. Ce facteur de risque a d'abord été identifié par des études de liaison dans des familles avec plusieurs cas de maladie d'Alzheimer (Schellenberg et al. 1987), puis il a été abordé sous la forme d'études d'association cas-témoins, permettant d'exprimer son effet sous la forme d'odds ratios (OR), utilisant l'allèle le plus fréquent, ε3, en référence (Campion et al. 1999; Genin et al. 2011). Avec une fréquence de 15,6 % en population caucasienne et des OR respectivement de 3.2 et de plus de 10 pour les porteurs hétérozygotes et pour les porteurs homozygotes, l'allèle ε4 concerne les formes précoces et les formes tardives et joue un rôle central dans la physiopathologie. Un autre allèle minoritaire du gène *APOE*, appelé ε2, joue quant à lui un rôle protecteur, puisqu'il est retrouvé plus fréquemment chez les témoins. Ce gène code pour une apolipoprotéine et joue de nombreux rôles dans le cerveau. Dans le contexte de la maladie d'Alzheimer, il a également été étudié sous de multiples aspects. On retiendra que l'allèle ε4 favorise l'agrégation du peptide Aβ et réduit sa clairance par rapport à l'allèle ε3 et par rapport à l'allèle ε2.

L'identification des autres facteurs de risque fréquents a été essentiellement abordés grâce à des études d'association pangénomique basées sur des puces à SNP (ou GWAS - Genome-wide Association Study). Les premières grandes études de GWAS sur Alzheimer ont été menées par le consortium International Genomics of Alzheimer disease Project (IGAP) et ont identifié plus de 40 facteurs de

risques (Baker et al. 2019) entre 2013 et 2019. C'est ensuite le consortium européen “European Alzheimer and Dementia Biobank” (EADB) qui a publié une étude permettant de doubler le nombre de loci facteurs de risque connus (Bellenguez et al. 2022). Ces études étant basées sur l'identification de loci à partir de variants fréquents, certains d'entre eux sont localisés dans des régions intergéniques. Dans ce cas, on associera le signal identifié au gène le plus proche, dont certains sont impliqués dans le métabolisme du peptide Aβ.

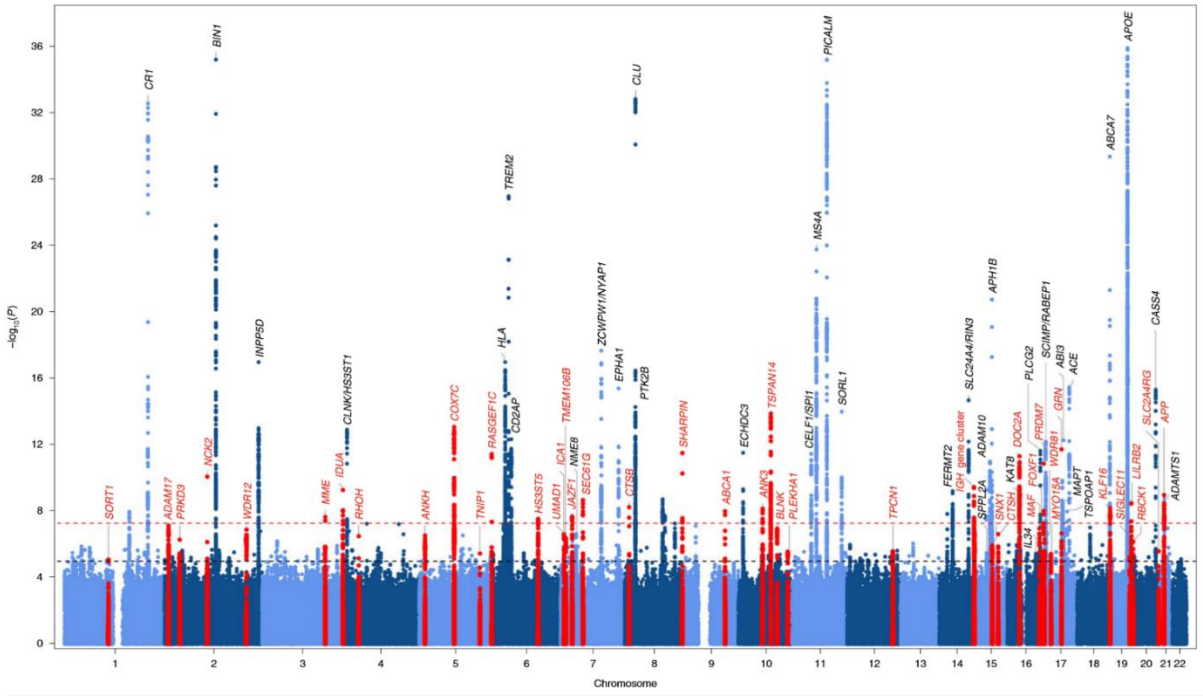


Figure 28: Manhattan-plot des différents loci identifiés par les études de GWAS.
Chaque point correspond un SNP génotypé (ou imputé) le long du génome avec la p-valeur associée. Les variants en rouge représentent des nouveaux loci par rapport aux études précédentes. Figure issue de Bellenguez et al., 2022

Ces différentes études travaillant à partir de variants fréquents, elles vont généralement permettre d'identifier des haplotypes portant un allèle associé à la pathologie, et rarement le SNP directement responsable de l’effet biologique. Il s’agit le plus souvent d’allèles non codants à effet très modeste. La combinaison de ces petits effets dans des scores, appelés score de risque polygénique, tente de faire l’addition de ces petits effets en un seul score, qui manque néanmoins une grande partie du risque génétique et ne tient pas compte du risque non génétique.

Une grande avancée, plus récente, a consisté à s’intéresser aux variants rares (fréquence inférieure à 1% dans la population), ce qui a permis d’associer des variants avec un effet fort sur le risque de

développer la maladie. Afin d'identifier ces variants rares, il est nécessaire d'effectuer du séquençage en exome ou en génome. Grâce à cela et au séquençage de centaines puis de plusieurs milliers de cas et de contrôles, il a été possible d'identifier des variants rares facteur de risque de la maladie. C'est le cas par exemple du variant p.Arg47His du gène *TREM2* (fréquence entre 0,3 et 0,6%) identifié par une étude islandaise. Ce variant augmente le risque de 2,26 à 4,56 fois. Néanmoins et par définition, Il est très difficile d'obtenir une récurrence d'un variant rare dans une population de cas ou de témoins, et identifier un seul variant rare associé à une maladie est très difficile. Il n'y a que peu d'exemple dans la maladie d'Alzheimer (Sims et al. 2017). En fonction de la taille des cohortes de cas et de témoins, la majorité des variants ne sont identifiés que sous la forme de singleton ou très peu de récurrences Et il n'est pas possible de les étudier de manière individuelle. La stratégie généralement mise en place est alors d'agréger des variants rares pour chaque gène, et d'identifier des agrégations dans certains gènes. C'est le travail qui a été effectué au sein du Centre National de Référence pour les Malades Alzheimer Jeunes (CNR-MAJ) qui a identifié en 2015 une accumulation de variants très rares dans le gène *SORL1* dans une cohorte de patients EOAD avec antécédent familiaux (Nicolas et al. 2016), des variants rares de ce gène avaient auparavant été identifiés chez des cas index de familles ressemblant à des formes autosomiques dominantes (C. Pottier et al. 2012). Cette étude d'association basé sur 484 cas jeunes et 498 témoins a permis de mesurer l'effet des variations dans le gène *SORL1*, pour une fréquence cumulée de 2,9% chez les patients et de 0,6% chez les témoins avec un OR de 5,03[0,02-14,99] (Nicolas et al. 2016). Avec l'accumulation de plus en plus importante de cas et de témoins, et la mise en place de collaborations internationales dans des consortiums, il a été possible d'identifier avec la même méthode 3(Bellenguez et al. 2017) puis 5(Holstege et al. 2022) gènes associés en tant que facteur de risque avec la maladie d'Alzheimer : *SORL1*, *TREM2*, *ABCA7* puis *ABCA1* et *ATP8B4*. Il a pour cela été nécessaire de regrouper 32 558 individus, répartis entre 16 036 cas et 16 522 témoins. Une représentation des différents signaux identifiés en fonction de la fréquence et de l'effet est disponible Figure 29.L'effet des variants rares a été évalué dans plusieurs études et répliqué. Il est maintenant clair que les variants rares actuellement détectés et détectables ont un effet modéré à fort, et ces variants sont clairement enrichis dans les formes précoces. L'étude préférentielle de malades jeunes apporte donc une puissance supplémentaire.

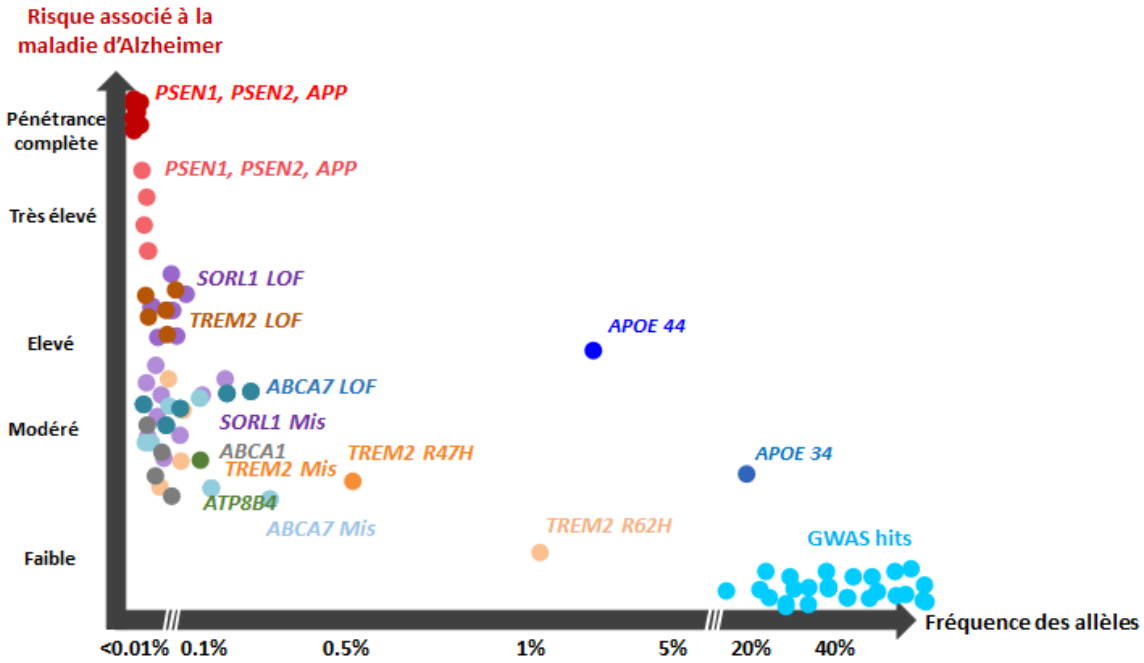


Figure 29 : Distribution des variants à risque en fonction de leur fréquence et de leur impact
Figure adaptée de Manolio et al., 2009

1.3.4. Impact des variations de structure

Les formes autosomiques dominantes

Les variations de structure jouent également un rôle dans la maladie d'Alzheimer. Tout d'abord dans les formes autosomiques dominantes de maladie d'Alzheimer précoce, des duplications du gène *APP*, seules ou avec plusieurs gènes de cette région du chromosome 21, ont été retrouvées chez des patients avec maladie d'Alzheimer autosomique dominante de survenue précoce avec ou sans angiopathie amyloïde cérébrale (Anne Rovelet-Lecrux et al. 2006; Sleegers et al. 2006). Plus récemment, nous avons également identifié une triplication (4 copies du gènes *APP*) chez un patient avec un phénotype similaire (Grangeon et al., 2021). Ces duplications sont associées à une augmentation de l'expression du gène avec des quantités relatives d'ARN messager mesurées aux alentours de 1,5 fois plus que les contrôles pour les duplications d'*APP* et environ 2 fois plus pour le seul cas disponible avec triplication (Grangeon et al. 2021; Cyril Pottier et al. 2012). Par ailleurs, une délétion en phase de l'exon 9 ou une délétion en phase de l'exon 9 et de l'exon 10 du gène *PSEN1* ont déjà été retrouvées comme responsables de formes autosomiques dominantes (Brooks et al. 2003;

Kilan Le Guennec et al. 2017; Prihar et al. 1999). Il ne s'agit pas de délétions entraînant une haploinsuffisance puisque ces exons sont en phase, mais un mécanisme plus complexe lié à un changement faux-sens se produisant au niveau de la jonction aberrante entre l'exon 8 et l'exon 10 ou entre l'exon 8 et l'exon 11. Une variation ponctuelle affectant l'épissage de l'exon 10, sans altérer le nombre de copies, a des conséquences similaires, aboutissant à un saut d'exon 10 conduisant ainsi à ce changement faux-sens aberrant (Doğan et al. 2022; Vidal et al. 1996).

Analyse par les techniques ACPA

En parallèle de l'identification de formes autosomiques dominantes, plusieurs études ont cherché à identifier des CNVs impliqués dans la maladie d'Alzheimer, que cela soit des familles compatibles avec une transmission autosomique dominante, non expliquée, ou bien des facteurs de risque. Ces dernières sont majoritairement basées sur l'utilisation des données de puces de génotypage, comme décrit dans la section 1.2.1 de ce document. Pour rappel, le principe est d'utiliser un logiciel bioinformatique (le plus connu étant PennCNV) afin d'analyser l'intensité du signal et la balance allélique de chaque variant afin d'identifier de potentiels CNVs, avec une résolution entre 50 et 100 kb minimum. Ces différentes études possèdent des approches relativement similaires, avec trois grands axes principaux d'analyse : soit une étude de cas et de témoins à la recherche d'un enrichissement, une analyse centrée sur une liste de gènes préétablie, ou bien la réplication de signaux déjà identifiés dans d'autres études. Une grande partie de ces études, présenté dans le tableau 2, sont très proches et vont soit partager une partie des cas et/ou témoins, soit servir de réplication les unes par rapport aux autres.

Tableau 2 : Etudes cas témoins conduites par technique ACPA

Études	Nombre de cas	Nombre de contrôles	Cohortes	Méthodes	taille minimum des CI	Liste de gènes	Note
Heinzen et al., 2010	331	899	Duke University	puce de génotypage PennCNV			
Swaminathan et al., 2011	136 MCI 222 AD	143	ADNI	puce de génotypage PennCNV	100 kb	AlzGene (317) Bertram et al., 2007	
Swaminathan et al., 2012	711	171	ADNI, NIA-LOAD ("étude familiale")	puce de génotypage PennCNV	100 kb	AlzGene (317) Bertram et al., 2007	Pas d'études de ségrégation
Swaminathan et al., 2012	728	438	ADNI, cohorte complémentaire TGen	puce de génotypage PennCNV	100 kb	AlzGene (317) Bertram et al., 2007	
Rovelet-Lecrux et al., 2012	33	1078 + 912 LOAD	CNR-MAJ	Puces de CGH QMPSF	15 kb		
Chapman et al., 2012	3260	1290	GERAD (UK)	puce de génotypage PennCNV	100 kb		
Li et al., 2012 Szigeti et al., 2013	375	192	TARCC	puce de génotypage PennCNV	100 kb		réplication dans NIA-LOAD
Ghani et al., 2013	392	357	Carribean Hispanic	puce de génotypage PennCNV	100 kb		

AD : Alzheimer Disease. ADNI : Alzheimer Disease Neuroimaging Initiative. CNR-MAJ : Centre National de Référence pour les Malades Alzheimer Jeunes. LOAD : Late-Onset Alzheimer Disease. MCI : Mild Cognitive Impairement. NIA-LOAD : National Institute of Aging : Genetics Initiative for Late-Onset Alzheimer's Disease. TARCC : Texas Alzheimer Research and Care Consortium.

Au final, de nombreux signaux ont été identifiés soit dans des gènes connus comme facteurs de risques par ailleurs en GWAS tel que les gènes *CR1* (Brouwers et al. 2012), *BIN1* (Szigeti et al. 2013), ou *CHRNA7* (Heinzen et al. 2010), soit de nouveaux gènes non associés. Ces études présentent des défauts importants : tout d'abord la résolution des puces de génotypage qui ne descendent pas, en général, en dessous de 100 kb de résolution. Ensuite et selon les études, les délétions et les duplications ne sont pas différenciées, considérant que les deux types d'événements induisent le même impact biologique. Enfin, le niveau de significativité des signaux identifiés est souvent faible : plusieurs de ces études se contentent d'une p-valeur inférieure à 0,05, sans appliquer de correction en fonction du nombre de tests considérés. Au final, aucun des gènes ou des régions chromosomiques identifiés par ces études usant de puces de génotypage n'ont été répliqués dans des études indépendantes.

Parmi les SV identifiés par ces différentes études, deux facteurs de risque ressortent. Tout d'abord, le gène *MAPT* (Microtubule-Associated Protein Tau), codant pour la protéine Tau et qui est impliqué directement dans la pathologie Alzheimer. La région codant ce gène présente deux haplotypes majeurs différents, H1 et H2, l'haplotype H1 présentant un risque plus important de développer la maladie (Sánchez-Juan et al. 2019)avec un OR de 1.12[1.04–1.20]. Les haplotypes H1 et H2 sont en inversion l'un par rapport à l'autre (Bowles et al. 2022)Figure 30 et il est possible de les différentier soit grâce à des variants fréquents en déséquilibre de liaison, soit par identification des points de cassure par séquençage, ou encore par une différence dans le nombre de copies des gènes bordant l'inversion.

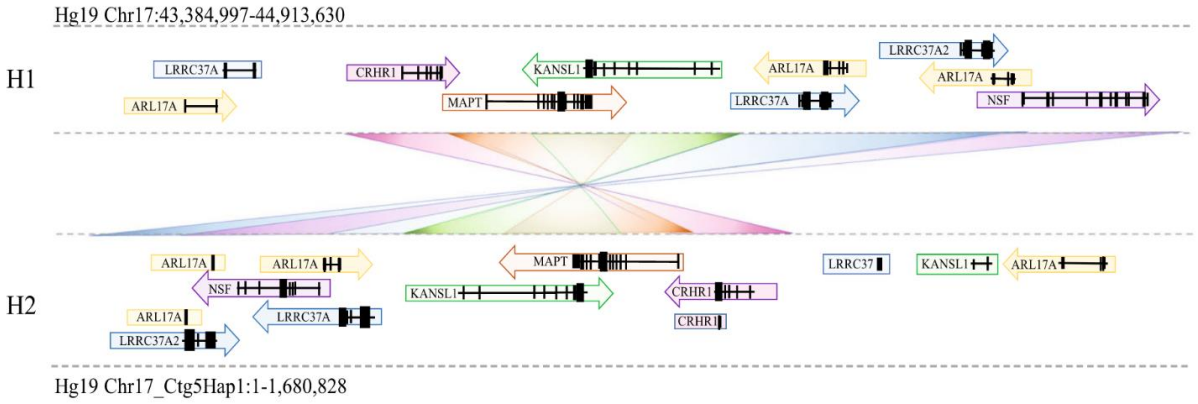


Figure 30 : Structure des haplotype H1/H2.
Figure issue de Bowles et al., 2022

Le second facteur de risque est la présence d'un LCR dans le gène *CR1* (OR = 1.32; 95% CI: 1.10–1.59, *P*=0.0025 non corrigé), la duplication de ce dernier conduisant à l'apparition de l'isoforme CR1-S (Brouwers et al. 2012) (Figure 31). Cette isoforme, conduit à l'inactivation du fragment C3b/C4b qui a

un rôle dans la capture et la dégradation du peptide Aβ. La réduction de la dégradation du peptide conduisant à une augmentation de sa toxicité.

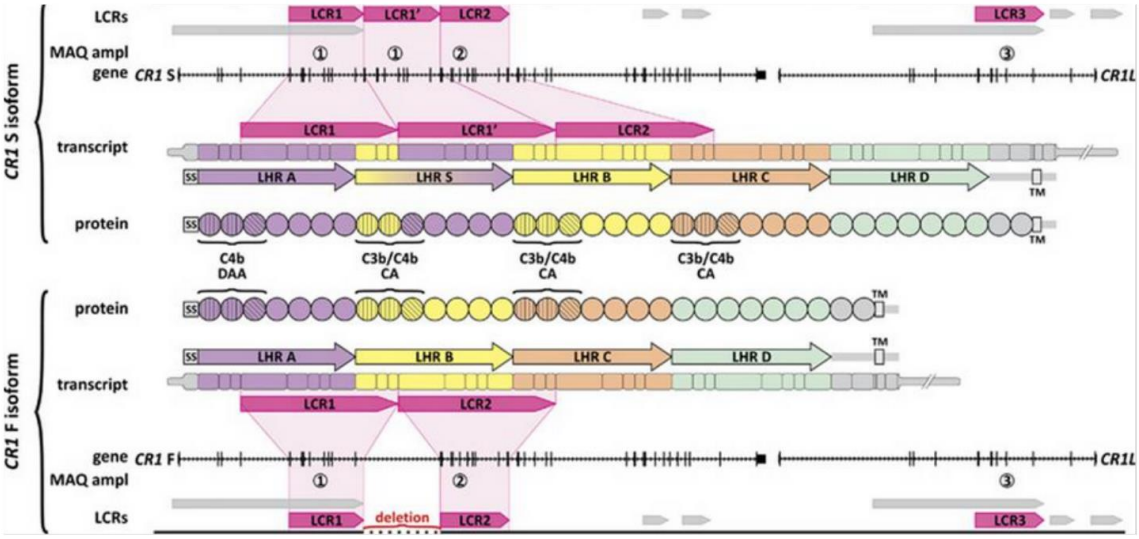


Figure 31 : Isoforme du gène CR1 médié par la duplication du LCR1.
L'isoforme majoritaire CR1-F présente une seule copie du LCR1. L'apparition de la duplication LCR1 dans l'isoforme CR1-S conduit à une altération du fragment C3b/C4b qui perd sa capacité à lier le peptide Aβ. Figure issue de Brouwers et al.,2012.

En parallèle de ces études basées sur des puces de génotypage, une étude, conduite par notre laboratoire (Anne Rovelet-Lecrux et al. 2012), a été conduite en utilisant des techniques de puces CGH. Cette étude était concentrée sur deux groupes de patients, d'une part 12 patients ayant déclaré la maladie de manière très précoce (début avant 55 ans) et d'autres part 23 cas index issus de familles avec une transmission compatible avec une forme autosomique dominante, mais sans avoir identifié de variations dans les gènes responsable des formes mendéliennes *APP*, *PSEN1* et *PSEN2*. Au final, 7 CNVs singletons ont été identifiés et n'ont pas été retrouvés dans une cohorte de 1078 témoins et de 912 patients atteints de forme tardive de la maladie d'Alzheimer. Parmi ces 7 CNVs, 4 affectent des gènes impliqués dans le métabolisme d'Aβ : *KLK6*, *MEOX2*, *SLC30A3* et *FPR2*.

En dehors de ces études exploitant des cohortes de cas et de témoins, deux études se sont basées sur des familles (

). La première étude, issue de l'équipe de R Tanzi (Hooli et al. 2014), a identifié à l'aide de puces de génotypage et parmi 261 familles, 10 CNVs ségrégant dans 10 familles différentes sans concerner les apparentés sains et n'apparaissant chez aucun autre individu de l'étude. Parmi ces 10 CNVs d'intérêt,

on peut noter la présence de CNV affectant *CHMP2B* et *MAPT*, deux gènes impliqués dans les démences frontotemporales. La seconde étude, issue de notre laboratoire, est basée sur une analyse de 14 trios de forme précoce de la maladie d'Alzheimer (A. Rovelet-Lecrux et al. 2015). Dans celle-ci, nous avons aussi pu identifier, toujours par puce CGH, une délétion de novo intronique dans le gène *BACE2*, un gène codant pour la β-secretase 2, une enzyme ayant un rôle dans le métabolisme du peptide Aβ.

Tableau 3 : Etudes familiales conduitent sur la maladie d'Alzheimer par technique ACPA

Études	Nombre de fan	Nombre d'individus	Méthodes	Taille minimale des C	Liste de gène	Commentaire
Hooli et al., 2013	261	109	puce de génotypage PennCNV	100 kb		Exclusion des LOAD dans les études de ségrégation
Rovelet-Lecrux et al., 2015	14	42	Puces de CGH QMPSF	10 kb	Réseau Aβ (355) Campion et al., 2016	

Les CNVs ultra rares identifiés par ces trois dernières études n'ont pas pu être répliqués, mais l'absence d'occurrence parmi les témoins, leur ségrégation parmi les cas dans les familles et leur lien biologique potentiel avec le peptide Aβ en font des candidats potentiels pour rechercher des réplications.

Utilisation des données de séquençage

Peu d'études ont été actuellement menés concernant l'identification de CNVs par des approches de séquençage à haut débit dans le cadre de la maladie d'Alzheimer. Une seule étude actuellement a cherché à exploiter les données de capture en exome pour identifier des CNVs dans le cadre de la maladie d'Alzheimer. Cette étude, issue de notre laboratoire (K Le Guennec et al. 2017), inclut 546 cas EOAD et 584 témoins issus de la cohorte FREX. Cette étude a identifié 4 duplications complètes du gène *MAPT* (locus comprenant 4 gènes en 17p21.31) chez des patients présentant tous les signes cliniques et des biomarqueurs compatibles avec une maladie d'Alzheimer, y compris un taux de peptide Aβ bas dans le liquide Céphalo-Rachidien. Ceci était inattendu car la surexpression de MAPT, démontrée dans le sang (K Le Guennec et al. 2017) puis dans des cellules neuronales induites issues de cellules souches pluripotentes dérivées d'un porteur d'une duplication *MAPT* (Miguel et al. 2022)ne serait pas compatible avec la cascade amyloïde, si cet évènement est suffisant pour causer la maladie. Une analyse complémentaire avait permis de montrer que les patients n'avaient pas d'agrégats amyloïdes détectables en PETscan et une analyse neuropathologique a montré l'absence de dépôts amyloïdes, confirmant qu'il s'agissait finalement d'une tauopathie primaire, phénocopie mimant la maladie d'Alzheimer cliniquement et par de nombreux autres aspects, en dehors d'Aβ (K Le Guennec et al. 2017). Par la suite, des duplications similaires ont été identifiées chez des patients avec paralysie supranucléaire progressive (PSP), une tauopathie primaire (Z. Chen et al. 2019). Enfin, une étude plus

poussée de notre laboratoire a regroupé 10 cas explorés de manière complète, dont 4 cas autopsiés, montrant que cette duplication est responsable de tauopathies primaires variées, conférant un spectre de lésions neuropathologiques de la protéine Tau dont les caractéristiques de composition et de localisation corrèlent avec la clinique, de formes intéressant d'abord les noyaux gris centraux et conférant une PSP, à des formes concernant le cortex et responsables d'une maladie de Pick (forme de démence fronto-temporale liée à Tau) ou des formes intermédiaires mimant la maladie d'Alzheimer, mais sans dépôts d'Aβ (Wallon et al. 2021).

En dehors de ce résultat fournissant un diagnostic différentiel pour certains cas, nous rapportions un enrichissement en CNVs rares dans une liste de gènes préétablie basé sur le réseau Aβ (Campion et al. 2016) dans notre étude cas-témoins, tel que des délétions et des duplications partielles affectant les gènes *ABCA1*, *ABCA7*, *KLK6* ou *CTSB* (K Le Guennec et al. 2017). Néanmoins, comme dans les études en CGH array, les données de ségrégation étaient absentes ou insuffisantes et les données d'enrichissement n'étaient pas significatives, du fait de la rareté des évènements observés.

Depuis cette étude, le consortium américain ADSP a construit une cohorte visant à réunir et séquencer plus de 70 000 individus, encore en cours de séquençage. Au fur et à mesure de la production et de la mise à disposition des données, des études à la fois sur les variations ponctuelles et sur les variations de structure sont produites. Dans ce contexte, le consortium ADSP a déjà publié deux études basées sur 3800 (W.-P. Lee et al. 2021) puis 16900 individus dans une étude actuellement non encore reviewée par des pairs (H. Wang et al. 2023). Il est important de noter que, dans les premières étapes du projet, les données étaient produites à la fois en WES et en WGS, mais les études sur les variations de structure se sont limitées à analyser les données de WGS. Dans cette étude, ils se sont concentrés à la fois sur les CNVs ultra rares (fréquence inférieure à 0.1%) en effectuant deux analyses. Ils ont tout d'abord identifié les événements affectant des gènes précédemment rapportés en tant que facteur de risque. Ils ont par exemple identifié des CNVs dans les gènes *SORL1* et *ABCA7*, et des CNV non codants, mais aucun n'était suffisamment récurrent et enrichi chez les patients ou les contrôles pour montrer une différence significative, à l'échelle du CNV ou à l'échelle du gène. De manière intéressante, la charge en CNV rares était légèrement plus importante chez les cas, et la charge en CNV affectant une liste de gènes de maladie d'Alzheimer était également plus importante chez les cas. Il est à noter tout de même que les patients ont majoritairement un début tardif (74 ans d'âge de début moyen, écart type de 10 ans), et que l'étude a porté sur environ 6600 cas et 6900 contrôles, qui sont des effectifs importants comparé aux études précédentes, mais inférieurs à l'étude que nous avons menée dans le cadre du consortium européen et dont les résultats sont rapportés ici (Holstege et al. 2022).

Même si plusieurs de ces résultats sont prometteurs, aucun ne permet d'obtenir un niveau de significativité statistique suffisant pour une analyse au niveau du génome, c'est à dire en corrigeant la p-valeur pour le nombre de gènes testés au total. Ceci met en avant le besoin de constituer des cohortes plus importantes de façon à pouvoir identifier des récurrences dans les gènes affectés par les CNVs.

1.4. Objectifs

L'objectif de mes travaux a été d'identifier, à partir d'un jeu de données de plus de 20 000 exomes, de potentiels nouveaux facteur de risque associé aux formes précoces de la maladie d'Alzheimer. Une des contraintes liées à l'utilisation de ce jeu de données a été l'impossibilité de valider par une technique indépendante les CNVs détectés par notre approche. Il était donc nécessaire de bien caractériser notre approche afin de connaitre précisément les performances et les limites de nos résultats.

Mon projet a été constitué de deux étapes. Premièrement, la mise en place d'un pipeline bioinformatique permettant la détection des CNVs rares à partir de données de séquençage à haut débit en capture et l'évaluation de ce pipeline. Nous avions pour cela à disposition des données obtenues à la fois dans le cadre de la recherche et du diagnostic, avec des captures en exomes ou en panels, les résultats obtenus pouvant être évalués par différents outils moléculaires : CGH array, MLPA ou QMPSF, l'objectif étant d'établir de manière précise les valeurs prédictives positives, sensibilité et spécificité de l'approche.

Dans un second temps, j'ai cherché à l'appliquer à notre étude cas-témoins constituée de plus de 20 000 individus en s'intéressant plus particulièrement aux formes précoces de la maladie d'Alzheimer. La problématique principale de cette étude a été de réussir à adapter notre approche à des jeux de données hétérogènes, tout en mettant en place une méthode d'analyse des résultats robuste et pertinente.

2. Résultats

Les données produites dans notre laboratoire, que ce soit dans le cadre de la recherche ou du diagnostic, sont majoritairement des données de séquençage par capture, que ce soit du panel de gènes ou bien du séquençage d'exome complet. Nous avons vu que la détection de CNV en utilisant ces approches doit avant tout s'appuyer sur des outils utilisant l'information de profondeur de lecture. Il a donc été nécessaire de mettre en place un pipeline d'analyse se basant sur un outil utilisant la profondeur de lecture pour détecter les variations du nombre de copies. Malheureusement, les évaluations comparatives des différents outils sont souvent limitées : de nombreux articles vont énoncer les différents outils et leurs spécificités (approche bioinformatique utilisée, sur quel type de données les employer, outils dédiés aux données constitutionnelles ou somatiques, ...) sous forme de catalogues (Zhao et al. 2013) mais sans définir précisément la sensibilité et la spécificité de ces derniers. Parfois, les informations proviennent de l'équipe ayant développé un nouvel outil qui va prouver son efficacité par rapport à un outil de référence en se limitant à un petit jeu de données. Dans notre cas, nous avons choisi d'évaluer notre solution à des jeux de données "gold standard" produit au sein du laboratoire selon différentes approches (chapitre 1 des résultats) : des données issues de CGH array combiné à du séquençage d'exome, des confirmations exhaustives par technique QMPSF combiné à du séquençage ciblé sur panel de gènes, et une validation de CNVs détecté sur des données de séquençage d'exome par des approches ciblées de type QMPSF et ddPCR. Toutes ces étapes de validation ont été effectuées sur des données contrôlées et standardisées de notre laboratoire. Ensuite, nous avons appliqué cette nouvelle approche à des données très hétérogènes d'exomes issus des consortiums ADES (Alzheimer Disease Exome Sequencing project) et ADSP afin de pouvoir mener notre étude cas/témoins portant sur les formes précoces de la maladie d'Alzheimer (chapitre 2 des résultats).

2.1. Définition d'un pipeline de détection des CNVs à partir de données de séquençage d'exome

Afin de détecter des CNV issus de données de NGS obtenus après capture par une approche de comparaison de données de profondeurs, deux options se sont offertes : soit combiner des outils existants mais utilisant l'information de profondeur, soit optimiser un outil existant. Nous avons choisi la seconde option. Le choix a été fait de choisir l'outil CANOES qui fonctionne sur la même approche que le logiciel de référence XHMM mais qui présente de meilleurs résultats (Backenroth et al. 2014).

Nous avons cherché à optimiser la détection en se concentrant uniquement sur cet outil. L'idée a été de préparer les données en amont de façon à avoir le moins de variabilité possible et les données les plus informatives possibles puis, après la détection, d'appliquer des filtres de façon à ne conserver que les variations probablement vraies.

Afin de valider le pipeline, nous avons appliqué notre analyse sur 137 individus pour lesquels nous possédions à la fois les données de séquençage en exome et les données de CGH array. Il a été nécessaire de sélectionner les événements à la fois détectables par CANOES et par des approches de CGH array. Pour ce faire, nous avons regardé parmi les données de CANOES quels étaient les événements qui contenaient au moins 5 sondes de CGH array que nous avions utilisées. Puis à partir des données de CGH array nous avons sélectionné uniquement les événements qui emportaient au moins un exon capturé par nos kits de capture. Afin de compléter cette première comparaison, nous avons fait une analyse plus poussée de chaque événement détecté par l'une des 2 méthodes afin de comprendre pourquoi il n'était pas détecté par l'autre.

Afin de compléter cette évaluation de notre approche, nous avons travaillé avec les données issues du soin dans le laboratoire de génétique moléculaire du CHU de Rouen afin de tester notre outil sur les données produites en routine dans le cadre du diagnostic. Nous avons eu accès à 3776 individus répartis en 3 panels de gènes différents, dont 461 individus avaient eu une QMPSF exhaustive pour chaque exon codant de 4 gènes.

Enfin, une dernière étape a consisté à utiliser notre pipeline afin d'analyser notre cohorte de patients atteints de forme précoce de la maladie d'Alzheimer et les témoins issus du projet de séquençage national FREX. Dans cette cohorte nous avons évalué notre outil en effectuant une confirmation soit par QMPSF soit par ddPCR des événements détectés sur une liste de gènes d'intérêt de manière systématique.

Au total, nous avons établi une approche permettant la détection de variations du nombre de copies à partir de données de séquençage obtenues après capture. Ce pipeline est maintenant appliqué de manière systématique sur les données de séquençage en exome et sur les données de panel à la fois dans le cadre de la recherche et du diagnostic dans nos laboratoires. L'utilisation systématique de la QMPSF pour l'identification des variations sur les gènes *APP*, *PSEN1* et *MAPT* n'est plus nécessaire pour les patients ayant un exome dans le cadre de l'exploration d'une maladie d'Alzheimer précoce, et seuls les CNVs d'intérêts sont maintenant confirmés par une seconde technique. Dans le cadre du diagnostic des maladies rares du développement et en oncogénétique, le pipeline est lui aussi utilisé de manière

systématique sur les exomes et les différents panels et a même remplacé la MLPA systématique de 4 gènes en oncogénétique. Cette approche a également été utilisée dans plusieurs études, comme par exemple une étude menée par notre équipe qui incluait notre cohorte de patient atteint de forme précoce de la maladie d'Alzheimer et la cohorte de contrôle FREX et qui a permis de mettre en évidence les duplications du gène *MAPT* comme étant responsable d'une forme de tauopathie primaire (K Le Guennec et al. 2017), une seconde étude, également de notre équipe, qui a permis de mettre en évidence des délétions du gène *SLC20A2* dans la maladie rare « calcifications cérébrales primaires » (David et al. 2016) , ou encore une étude menée par l'équipe du Pr Tournier-Lasserve et qui portait sur l'analyse d'une cohorte de patient atteints de la maladie de moya moya (Aloui et al. 2020).

Les résultats de notre étude décrivant la mise en place du workflow centré sur l'outil CANOES et décrivant ses performances comparatives fait l'objet de ce premier chapitre de résultats. L'article, présenté ci-dessous, a été publié dans la revue European Journal of Human Genetics. J'ai également été amené à présenter les résultats oralement lors des journées satellites de l'ESHG à Goteborg en 2019 et lors des assises de génétique de Tours en 2020.

European Journal of Human Genetics (2021) 29:99–109
https://doi.org/10.1038/s41431-020-0672-2

ESHG

ARTICLE

Detection of copy-number variations from NGS data using read depth information: a diagnostic performance evaluation

Olivier Quenez[1] · Kevin Cassinari[1] · Sophie Coutant[2] · François Lecoquierre[2] · Kilan Le Guennec[1] · Stéphane Rousseau[1] · Anne-Claire Richard[1] · Stéphanie Vasseur[2] · Emilie Bouvignies[2] · Jacqueline Bou[2] · Gwendoline Lienard[2] · Sandrine Manase[2] · Steeve Fourneaux[2] · Nathalie Drouot[2] · Virginie Nguyen-Viet[2] · Myriam Vezain[2] · Pascal Chambon[2] · Géraldine Joly-Helas[2] · Nathalie Le Meur[2] · Mathieu Castelain[2] · Anne Boland[3] · Jean-François Deleuze[3] · FREX Consortium · Isabelle Tournier[2] · Françoise Charbonnier[2] · Edwige Kasper[2] · Gaëlle Bougeard[2] · Thierry Frebourg[2] · Pascale Saugier-Veber[2] · Stéphanie Baert-Desurmont[2] · Dominique Campion[1,4] · Anne Rovelet-Lecrux[1] · Gaël Nicolas[1]

Received: 22 November 2019 / Revised: 20 May 2020 / Accepted: 9 June 2020 / Published online: 26 June 2020

Abstract

The detection of copy-number variations (CNVs) from NGS data is underexploited as chip-based or targeted techniques are still commonly used. We assessed the performances of a workflow centered on CANOES, a bioinformatics tool based on read depth information. We applied our workflow to gene panel (GP) and whole-exome sequencing (WES) data, and compared CNV calls to quantitative multiplex PCR of short fluorescent fragments (QMSPF) or array comparative genomic hybridization (aCGH) results. From GP data of 3776 samples, we reached an overall positive predictive value (PPV) of 87.8%. This dataset included a complete comprehensive QMPSF comparison of four genes (60 exons) on which we obtained 100% sensitivity and specificity. From WES data, we first compared 137 samples with aCGH and filtered comparable events (exonic CNVs encompassing enough aCGH probes) and obtained an 87.25% sensitivity. The overall PPV was 86.4% following the targeted confirmation of candidate CNVs from 1056 additional WES. In addition, our CANOES-centered workflow on WES data allowed the detection of CNVs with a resolution of single exons, allowing the detection of CNVs that were missed by aCGH. Overall, switching to an NGS-only approach should be cost-effective as it allows a reduction in overall costs together with likely stable diagnostic yields. Our bioinformatics pipeline is available at: https://gitlab.bioinfo-diag.fr/nc4gpm/canoes-centered-workflow.

Members of the FREX Consortium are listed below Acknowledgements.

Supplementary information The online version of this article (https://doi.org/10.1038/s41431-020-0672-2) contains supplementary material, which is available to authorized users.

✉ Gaël Nicolas
gaelnicolas@hotmail.com

[1] Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Department of Genetics and CNR-MAJ, Normandy Center for Genomic and Personalized Medicine, Rouen, France

[2] Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Department of Genetics, Normandy Center for Genomic and Personalized Medicine, Rouen, France

[3] Centre National de Recherche en Génomique Humaine, Institut de Génomique, CEA and Fondation Jean Dausset-CEPH, Evry, France

[4] Department of research, Centre hospitalier du Rouvray, Sotteville-lès-Rouen, France

Introduction

Copy-number variations (CNVs) are a major cause of Mendelian disorders [1] as well as risk factors for common diseases [2]. With the advent of next-generation sequencing (NGS), a number of software tools have been developed to detect CNVs [3–5]. Whole-genome sequencing (WGS) is often presented as an almost universal technique allowing the assessment of almost any type of variation, including CNVs and other structural variations [6]. WGS may eventually be used as a first-tier diagnostic tool in the context of genetically highly heterogeneous disorders. However, the detection of structural variations from data generated using the technology of short read sequencing is still associated with a number of false positives. Such events can be detected using a plethora of bioinformatics tools based on different principles, including depth of coverage (DOC)

Fig. 1 Principles of depth of coverage (DOC) comparison. Schematic distribution of reads among three different samples over five sequenced exons. **a** The absence of any CNV. **b** Duplication of two exons (2 and 3). **c** Deletion of exon 4. In order to call those CNVs, software tools have to establish a reference. Some tools compare paired data from the same patient, e.g., tumor tissue against germline, while others build their reference from a pool of samples and then compare a given sample to this reference, as the CANOES tool used in our workflow.

information, relative position of paired reads, split reads and De Novo Assembly [7]. Besides the development of WGS, targeted sequencing of gene panel, and whole-exome sequencing (WES) remain of primary use in many diagnostic and research laboratories. They are indeed still considered as more affordable and of easier access as they can be processed using usual informatics facilities accessible to most laboratories. Moreover, the input of WGS is questioning in disorders with low genetic heterogeneity and high phenotypic specificity. Hence, gene panels and WES remain largely used.

The detection of CNVs from exonic capture-based targeted sequencing solutions primarily relies on DOC information [8, 9]. Tools based on DOC information compare one sample with a reference, and predict deletions or duplications depending on the increase or decrease of the DOC as compared with the reference (Fig. 1). As each tool was set up and trained on a specific dataset, one of the main challenges is to evaluate the specificity and sensitivity of a given software tool on large datasets. Studies evaluating the diagnostic performances of CNV detection pipelines are scarce although they appear to be critical for their use in routine procedures [10–12]. In order to optimize CNV detection from NGS data, a classical approach consists in running multiple tools in parallel and then aggregate the results to keep a CNV as candidate only if multiple tools called it [13]. As it is more effective to do so with tools using different types of bioinformatics methods (DOC, split reads, etc.), this combinatory approach is most adapted when working on WGS, or at least if most of the intergenic or intronic regions—where breakends are more frequently found—are captured. Here, we decided to focus on one tool using the DOC approach as it still remains the most adapted one for exonic capture. In a *precision workflow* approach, we developed a workflow based on the already existing software tool CANOES [14]. To select this tool, we previously compared features of each tool that are related to the definition of a reference for CNV detection. Indeed, defining the reference is critical [15], as calling candidate deletions or duplications requires the comparison of DOC data of each sample to the reference. Our main criteria concerning the definition of reference were that (i) the tool should take into account information from multiple samples and that (ii) it should associate a Hidden Markov Model (HMM) with a distribution model to represent the variability of coverage between samples and between each target. CANOES appeared as the best candidate as it adopts a pooling strategy to build its reference model and it uses an HMM associated to a binomial negative distribution. In addition, CANOES defines the reference independently for each sample, by selecting samples with the closest mean and variance.

We performed a diagnostic performance evaluation of this workflow regarding gene panel and WES data, in two steps. First, we compared CNV calls with a reference technique, namely a comprehensive assessment by quantitative multiplex PCR of short fluorescent fragments (QMPSF) [16] or array comparative genomic hybridization (aCGH), regarding targeted gene panel and WES data, respectively. Second, we implemented our workflow in our

routine procedures and performed an additional evaluation of the positive predictive value of our CANOES-centered workflow using targeted confirmation of CNVs using an independent targeted technique.

Material and methods

Gene panel sequencing

In order to evaluate our workflow, we analyzed data from three gene panels (for detailed information, see Supplementary Table 1). Patients provided informed written consent for genetic analyses in a diagnostic setting.

Panel 1 was set up to focus on genes involved in predisposition to colorectal cancer and digestive polyposis or Li-Fraumeni syndrome [17]. This panel was implemented in two successive versions. V1 was used to sequence 11 genes in 2771 samples. V2 was used to sequence 15 genes (same 11 genes plus 4) in 549 samples. In both versions and for all genes, exons, and introns outside repeated sequences were captured.

Panel 2 also has two successive versions and was designed to focus on two clinical indications: (i) hydrocephaly (3 genes) and (ii) Cornelia de Lange syndrome and differential diagnoses (24 genes in v1, 30 in v2). In total, 320 samples were sequenced using this panel (240 with v1, 80 with v2). For this panel, introns outside repeated sequences were captured only for two genes, namely *L1CAM* and *NIPBL*.

Panel 3 was designed to focus on genes involved in nonspecific intellectual disability. It has been used to analyze 220 samples and is composed of 48 genes (coding regions only). The list of genes is available upon request.

Assessment of CNV calls from gene panel data: step 1

For the comparison with a reference technique, we used data obtained from samples for which both NGS (panel 1, v1) and comprehensive QMPSF screening data were available ($n = 465$). This QMSPF assessment included all 60 exons of 4 genes from this panel (*APC*, *MSH2*, *MSH6*, and *MLH1*) and was applied to all 465 samples.

Assessment of CNV calls from gene panel data: step 2

Following step 1, we implemented our CANOES-centered workflow in our routine diagnostic procedures on NGS data from all three panels ($n = 3311$ additional samples in total). We performed confirmations of candidate CNVs using QMPSF or multiplex ligation-dependent probe amplification (MLPA) only in samples with a CANOES call. Primers used for QMPSF screening and validation are available upon request.

Whole-exome sequencing

Patients provided informed written consent for genetic analyses either in a diagnostic or in a research setting, following the approval by respective ethics committees.

Whole exomes were sequenced in the context of diverse research and diagnostic purposes (Supplementary Table 1). Exomes were captured using Agilent SureSelect Human All Exon kits (V1, V2 V4 + UTR, V5, V5 + UTR, and V6) (Agilent technologies, Santa Clara, CA, USA). Final libraries were sequenced on an Illumina Genome Analyzer GAIIX (corresponding to exomes captured with the V1, V2, or V4UTR kit, $n = 10$), or on an Illumina HiSeq2000, 2500, or 4000 with paired ends, 76 or 100 bp reads (Illumina, San Diego, Ca, USA). Exome sequencing was performed in three sequencing centers: Integragen (Evry, France) ($n = 6$), the French National Center in Human Genomics Research (CNRGH, Evry, France) ($n = 1065$) and the Genome Quebec Innovation Center (Montreal, Canada) ($n = 128$) [18]. Exomes were all processed through the same bioinformatics pipeline following the Broad Institute Best Practices recommendations [19]. Reads were mapped to the 1000 Genomes GRCh37 build using BWA 0.7.5a. [20]. Picard Tools 1.101 (http://broadinstitute.github.io/picard/) was used to flag duplicate reads. We applied GATK [21] for short insertion and deletions (indel) realignment and base quality score recalibration. All quality checks were processed as previously described [18].

Assessment of CNV calls from WES data: step 1

For the comparison with a reference technique, we analyzed data from 147 unrelated individuals with both WES and aCGH data available.

Array CGH analysis

Oligonucleotide aCGH was performed as previously described [22]. Briefly, high-resolution aCGH analysis was performed using the 1 × 1M Human High-Resolution Discovery Microarray Kit or the 4 × 180k SurePrint G3 Human CGH Microarray kit (Agilent Technologies, Santa Clara, CA, USA), using standard recommended protocols. An in-house and sex-matched genomic DNA pool of at least ten control individuals was used as reference sample. Hybridization results were analyzed with the Agilent's DNA-Analytics software (version 4.0.81, Agilent Technologies) or the Agilent Genomic Workbench (version 7.0, Agilent Technologies). Data were processed using the ADM-2 algorithm, with threshold set at 6.0 SD or 5.0 SD.

WES/aCGH comparison

Array CGH enables the detection of genome-wide rearrangements thanks to the measurement of the deviation of the fluorescent signal of the patient as compared with a control DNA. The number of probes depends of the type of chip that is used (here, Agilent 1M or 180k). The threshold to consider a deletion or a duplication was set to the deviation of five or three consecutive probes, respectively. This restricts the detection to CNVs of 8 or 20 kb for Agilent 1M or Agilent 180k chips, respectively, on average. On the contrary, as CANOES analysis is based on WES data, it is strictly restricted to CNVs covering exonic sequences, but it can detect CNVs as small as one single exon.

In order to combine these approaches to evaluate the sensitivity of our workflow, we filtered out CNVs located in intronic and intergenic regions exclusively from the aCGH data (and on X and Y chromosomes for the samples processed without sex-chromosome CNV calling). Moreover, as CANOES analysis is based on the calculation of a mean and variance of coverage on a given genomic region, the detection of polymorphic rearrangements is very uncertain. For that reason, we also filtered out all polymorphic CNVs from aCGH data. We defined as polymorphic a CNV that overlaps at least at 70% with CNVs reported in the Gold Standard section of the Database of Genomic Variants with a frequency superior to 1% [23].

Regarding the evaluation of the positive predictive value of our workflow, we restricted our analysis to candidate non-polymorphic CNVs detected from WES data (i) that are theoretically detectable by aCGH as they encompass at least three or five probes, depending on the chip used and (ii) that overlap with segmental duplication regions <50% of the CANOES target regions. The segmental duplication regions have been extracted from the UCSC Table browser [24] (https://genome-euro.ucsc.edu/cgi-bin/hgTables).

As most aCGH data were processed using the hg18 genome as reference, we used the lift over tool from UCSC (https://genome.ucsc.edu/cgi-bin/hgLiftOver) to establish the correspondence to hg19. If there were no lift over possibility, we manually checked genes encompassing CNVs.

Assessment of CNV calls from WES data: step 2

Following step 1, we implemented our workflow in our routine procedures. From additional 1056 WES (Supplementary Table 1), we performed targeted confirmations following the detection of candidate CNVs by CANOES using QMPSF or ddPCR [25]. We focused our confirmations on a list of 350 genes that belong to the so-called Aβ network [26], as all the samples used at this step were sequenced in the context of Alzheimer disease research. This list of genes was built thanks to literature curation on Alzheimer pathophysiology, independently of any genomic information. Candidate CNVs were selected for targeted confirmation if (i) they encompassed genes belonging to this network, and (ii) they were not polymorphic i.e., with a frequency below 1% in our dataset.

Primers used for QMPSF or ddPCR validation are available upon request.

CNV calling from NGS data using CANOES

The CANOES software tool implements an algorithm dedicated to the detection of quantitative genomic variations based on DOC information. Basically, CANOES requires DOC data for each target of the capture kit used for each of the sample that are analyzed together. It also integrates the GC content information of each target to reduce the background variability observed in high-throughput sequencing data [27]. The read depth was calculated using BEDtools [28], and the GC content was determined using the GATK suite.

CANOES builds its statistical reference model from a subset of the samples included in the same analysis (at least 30 samples are recommended). To obtain the best possible fit, CANOES selects the samples that are the most correlated to the currently analyzed sample. This allows the detection of small CNVs, but also reduces the detection susceptibility of recurrent events. CANOES uses a Hidden Markov Model to represent the variability of the DOC distribution built from the selected samples. Then, it uses the Viterbi algorithm to assign deletions, duplications or normal regions. After the calling step, a "Not Applicable" (NA) score is attributed to all CNVs from samples carrying more than 50 rearrangements. Such samples are usually characterized by higher or lower average read depth and cannot be compared with the reference model. All CNVs assigned with an NA score were thus removed from further analyses. As CANOES used the capture kit definition to detect CNVs, boundaries of events were defined by the start position of the first target and the end position of the last target detected as deviated in comparison with the model.

A CANOES-centered workflow

To optimize CANOES performances, we focused on two different approaches, a methodological approach in sample selection and a bioinformatics approach (Fig. 2).

As previously described, CANOES defines a statistical model for a particular sample from a judicious selection of other samples included in the analysis. The first step of our workflow consisted in the implementation of rules to select the samples that should better be analyzed together. In order to get enough material to build an efficient statistical model and following the CANOES recommendations, we always worked with at least 30 samples. Importantly, we analyzed

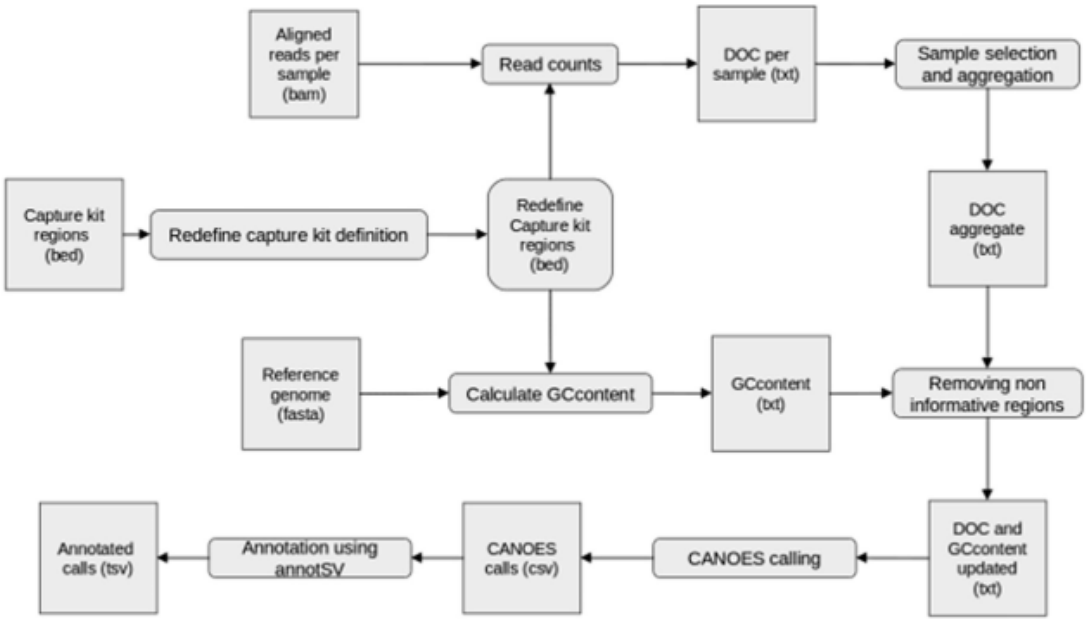


Fig. 2 CANOES-centered workflow. File (square) with their format in parenthesis, and process (rounded) constituting the workflow. From the original capture kit definition, we merge closed target from the same exon, then do in parallel the DOC and the GC content estimation. We gather DOC individual files depending on the project, sequencing batch, unrelated samples, and remove non-informative regions. The last steps consist in CNV calling using CANOES and annotation with annotSV.

samples with the less technical variability from each other. Practically, this consists in analyzing samples from the same run, and not to merge multiple runs if not necessary. When merging multiple runs was inevitable (e.g., sequencing of <30 samples per run), we combined sequencing runs from the same platform and processed using the same capture kit and technical conditions, including the same number of samples per lane in order to reduce read depth variability from each sample. Of note, CANOES is not originally set up for the analysis of CNVs on sex-chromosomes, but we implemented modifications in the original script in order to include sex-chromosomes in our analyses with a modification into the output file, the copy number is replaced by a GAIN or LOSS information. Hence, we ran our workflow after gathering either $N \geq 30$ males or $N \geq 30$ females for the analysis of gene panels 2 and 3 that contain X-linked genes and for WES data.

Bioinformatics optimization

The first step consisted in the modification of the target definition from the capture kit information. We decided to merge close targets (<30 bp) if they covered the same exon. The only exception to this rule is if a target is larger than 1 kb. In this particular case, targets are split. Concerning gene panels that include introns, we decided to split large targets that include both intronic and exonic regions. In this case, we split targets at the intron/exon junction.

In order to gain flexibility in our analysis and to be able to add or remove samples easily, we implemented a two-step strategy consisting in (i) performing the read count step for each sample separately, and then (ii) aggregating selected samples before running CANOES. Doing so allowed, for example, intrafamilial analyses including patient–parent trio approaches, where cases can be analyzed without taking related samples into account, preventing biasing the statistical model. Finally, we removed non-informative regions from our analyses. We considered a region as non-informative if more than 90% of the samples each had <10 reads on the target. Then, we called the CNVs using CANOES, and annotated the results using AnnotSV [29] in order to get additional information about the possible effect and populations frequencies.

Nextflow integration

In order to complete our optimization of processing and analysis time, we integrated our bioinformatics pipeline into Nextflow, a data-driven workflow manager [30]. This software tool allows a quick deployment of new pipelines on different kind of computational environments, from local computers to a cloud environment. Another interest of Nextflow is to increase the performance by distributing the different steps of the workflow in regards to the computational resources available. The complete workflow, including the specific adaption of CANOES to analyze sex-

chromosomes, is available on https://gitlab.bioinfo-diag.fr/nc4gpm/canoes-centered-workflow.

Interpretation of CNVs

The CNV detection workflow that is available on the above-mentioned reference finally includes the whole code required for both calling steps and annotation as a last step. Hence, the output file is a tab-delimited file that can be opened in a spreadsheet software to allow further filtration or sorting. For CNV interpretation in a Mendelian context, we prioritized CNVs based on (i) their frequencies in the DGV, potentially refined by frequencies in Exac [31], (ii) the inclusion or not of a gene from the OMIM morbid list [32] and (iii) probability of loss-of-function intolerance score (pLi) based on gnomAD database [33] and visualized CNVs in the UCSC genome browser [34].

Results

After building a workflow centered on the CANOES tool, we assessed its performances in the context of (i) gene panel NGS data and (ii) WES data, both generated following capture and Illumina short read sequencing.

Gene panel sequencing data

We first evaluated the performances of the CANOES tool using targeted sequencing data of a panel of 11 genes (panel 1, $n = 465$ samples). In parallel, all samples were assessed using custom comprehensive QMPSF assessing the presence or absence of a CNV encompassing any of the 60 coding exons of four of these genes. We identified 14 CNVs by QMPSF (12 deletions, 2 duplications, size range: [1,556 bp–97 kbp]). All of them were accurately detected by our CANOES-based workflow from NGS data (Table 1). In addition, no additional CNV was called by CANOES, allowing us to obtain a sensitivity and a specificity of 100% (95% CI: [73.24–100]) for those four genes. (see Supplementary Table 2).

To further assess the positive predictive value (PPV) of our workflow in the identification of CNVs from gene panels, we applied it to additional NGS data obtained from three gene panels (2222 samples from panel 1, 320 samples from panel 2, and 220 samples from panel 3). We detected 101 candidate CNVs in 98 samples and assessed their presence using either QMPSF or MLPA (Table 2). We validated 87/101 CNVs (86.13%, 95% CI: [77.50–91.94], false-positive rate: 13.9%). Overall, the PPV of our workflow applied to gene panel sequencing data was 87.83% (95% CI: [80.01–92.94]). True positive calls of our workflow were 71 deletions (size range: [391 bp–1.06 Mbp]) and 16 duplications (size range: [360 bp–39.4 kbp]) (see Supplementary Table 3). False positives were mainly deletions (10/14) and five of them were monoexonic.

Table 1 Summary of step 1 evaluation of CANOES-centered workflow.

	Gene panel	Whole exome
Gold standard	Comprehensive QMPSF/4 genes	aCGH data
Number of samples	465	147
Comparison to	GPS-CANOES calls (from panel 1)	WES-CANOES calls
Number of gold standard CNVs	14	102[a]
True positives	14	89
False negatives	0	13
Sensitivity	100% (CI: [73.24–100])	87.25% (CI: [78.84–82.77])
Number of CANOES calls	14	223[a]
True positives	14	190
False positives	0	33
Positive predictive value	100% (CI: [73.24–100])	85.2% (CI: [79.70–89.46])

aCGH array comparative genomic hybridization, *CNV* copy-number variation, *GPS* gene panel sequencing, *QMPSF* quantitative multiplex PCR of short fluorescent, *WES* whole-exome sequencing, *CI* confidence interval.

[a]The number of CNVs is different due to the selection of theoretically detectable events from one method in regards to the other.

Whole-exome sequencing data

We then evaluated the performances of our workflow for the detection of CNVs from WES data. We first applied our workflow to the data obtained from 147 samples with both WES (average DOC = 110×) and aCGH data available (50 samples assessed with the Agilent 1M chip and 97 samples with the Agilent 180k chip). Overall, ten samples were removed due to a high or low number of rearrangements detected by aCGH or exome, mostly due to low DNA quality or low coverage in WES.

From aCGH data, we detected 1873 CNVs over the 137 samples remaining, of which 102 were non-polymorphic exonic CNVs. Our workflow accurately detected 89 (87.2%) of them (Table 1 and Supplementary Table 4). Among the CNVs that were missed by our workflow, seven were large CNVs (from 14 to 80 kb) that encompassed only one ($n = 5$) or two ($n = 2$) targets defined by the capture kit (see Fig. 3).

In order to determine the PPV of our workflow from WES data, we selected 223 CNVs called by our workflow and (i) theoretically detectable by aCGH as encompassing at least three (180k chips) or five (1M chips) probes and (ii) which

did not overlap with segmental duplication regions for more than 50% of the CANOES targets. Of them, 190 (85.2%) CNVs were confirmed as true positives following aCGH data assessment (Table 1 and Supplementary Table 5).

Of note, an additional set of 519 candidate CNVs were detected by our CANOES-based workflow that overlapped <50% of segmental duplication regions but encompassed <3 (180k chips) or 5 aCGH probes (1M chips). Hence, they were not reported by the CGH analysis tool and would then have been overlooked following classical aCGH data analysis (see Fig. 4). We did not perform targeted confirmation of all these candidate CNVs. Instead, with the aim to further assess the PPV of our workflow regarding exonic non-polymorphic CNVs of any size, we applied it to 1056 additional WES performed in the context of Alzheimer disease research (with no corresponding aCGH data). We selected non-polymorphic CNVs targeting 355 genes belonging to the Aβ network involved in the pathophysiology of Alzheimer disease [26], whatever their size. We validated 111/125 candidate CNVs (88.8%, false-positive rate: 11.2%) by QMPSF [35] or ddPCR (Table 2 and Supplementary Table 6). True positive calls of our workflow were 39 deletions (size range: [165 bp–24.2 Mbp]) and 69 duplications (size range [166 bp–5.9 Mbp]). Interestingly, among the 125 candidate CNVs obtained from our workflow, 78 were considered to be theoretically detectable by aCGH 1M, and 47 were considered as not detectable by aCGH 1M. Among the ones theoretically detectable by aCGH, 74 were true positives (94.9%). Among the theoretically not detectable ones, 37 were true positives (78.7%).

Overall, the PPV of our CANOES-based workflow was 86.49% from WES data after taking into account results from step 1 and step 2 altogether.

Table 2 Summary of step 2 evaluation of CANOES-centered workflow.

	Gene Panel	Whole Exome
Data source	GPS-CANOES calls	WES-CANOES calls
Number of samples	3311	1056
Comparison to	QPMSF/MLPA	QMPSF/ddPCR
CANOES calls	101	125
True positives	87	111
False positives	14	14
Positive predictive value	86.13% (CI: [77.50–91.94])	88.8% (CI: [81.61–93.51])

aCGH array comparative genomic hybridization, *CNV* copy-number variation, *GPS* gene panel sequencing, *QMPSF* quantitative multiplex PCR of short fluorescent, *WES* whole-exome sequencing, *CI* confidence interval.

Discussion

Multiple tools have been developed to detect CNVs from NGS data. As long as such tools are being implemented in diagnostic laboratories, there is a critical need to evaluate their performances. Previous studies showed a large diversity of performances, while a number was performed using simulated datasets [5]. For example, from gene panel data, the DeCON tool [11] reached an overall sensitivity of 93% on a cancer gene panel sequencing of 94 genes, with a 100% sensitivity and 99% specificity on *BRCA1* and *2* genes and, with a 96% PPV on the complete gene panel after validation by MLPA. Another study on 60,000 samples focusing on a panel of 48 genes reached a 100% sensitivity with a PPV of 63.2% compared with array CGH and MLPA [36]. From WES data, a study analyzing 1017 samples with XHMM obtained 67% of sensitivity and 15.76% PPV on rare CNVs compared with SNP array [37], while a comparative study of three tools [15] on 861 WES samples revealed an important diversity of sensitivity (from 20 to 75%) and PPV (from 20 to near 100%).

After having defined a CANOES-centered workflow, we applied it to three different gene panels and WES data. Overall, we reached very high detection performances following the comparison with independent techniques.

From gene panel data, we obtained a 100% sensitivity among a set of four genes, the copy number of all coding exons of which having been assessed prior to NGS in 465 samples. In addition, we obtained a 87.83% PPV among all genes with a CANOES call. Such high

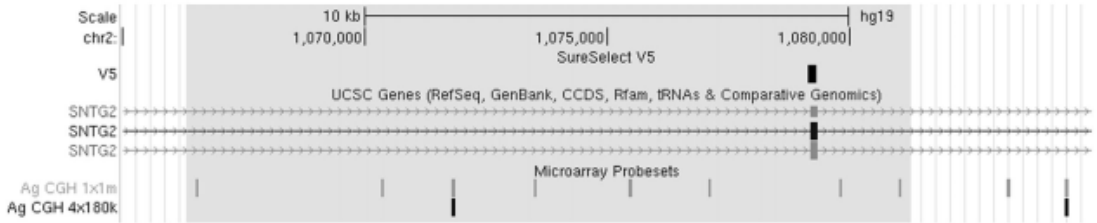


Fig. 3 Example of a CNV detected by aCGH but missed by the CANOES-centered workflow. A CNV (highlight region) detected by aCGH encompassing multiple CGH probes (1M probes array, in gray) but only one target from the SureSelect V5 capture kit. Of note, this deletion would have been missed by using a 180k probes array CGH (in black). View extract from UCSC genome Browser (https://genome-euro.ucsc.edu/cgi-bin/hgTracks).

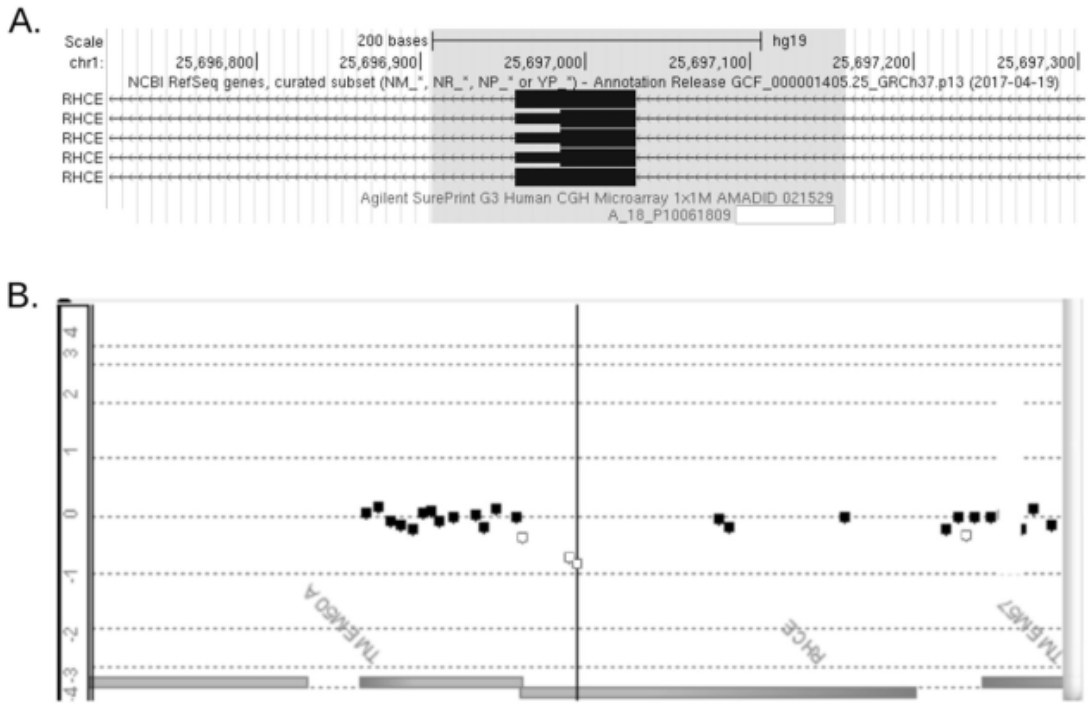


Fig. 4 Example of CNVs detected by the CANOES-centered workflow from WES data but missed by aCGH. a The highlighted region represents the CNV called by the CANOES-centered workflow, encompassing one exon of *RHCE*. **b** View of the same region from DNA-Analytics (aCGH data 1M) in the same patient. This deletion was not called following aCGH data analysis as the number of deviated probes did not reach the threshold for calling. However, as three probes (in white) were deviated, this allows the confirmation of the deletion of the region. View extract from UCSC genome Browser (https://genome-euro.ucsc.edu/cgi-bin/hgTracks).

performances have previously been reported for other tools applied to small NGS panels [11]. Among 14 false positives, we observed recurrent events, which can be easily reported as so and be ignored in further analyses. We also observed false positive CNVs in regions homologous to pseudogenes. In that case, it is possible to reduce false positive calls by improving the design of the capture to reduce the chance that probes target the homologous regions, or by optimizing the alignment.

Of note, for all genes of Panel 1 and two genes of Panel 2, introns were captured in addition to exons. This might have increased the chances to detect CNVs that can be considered as small from an exon-only point of view but that can actually be much larger at the genomic level. An advantage of capturing introns might indeed be a gain in statistical power for the normalization process: increasing the number of targets may increase the robustness of the model. Among 101 CNVs detected from NGS data from all three panels, 75 CNVs encompassed one of these genes with intronic-plus-exonic capture. Interestingly, only 18 of these 75 CNVs encompassed a single coding exon. Such a frequency of monoexonic CNVs is not unexpected regarding mutation screens in MMR genes, in which monoexonic deletions account for 26.92–46.27% of all pathogenic deletions [38–40], or other rare diseases [41–44], for example. We hypothesize that all other CNVs, encompassing multiple targets, would probably have been easily detected, had the introns been excluded from the capture design. Further analyses may be required to better assess the performances of our workflow from single exon CNVs and the effect of including introns or not in the capture design. The observed higher rate of false positives in CNV calls encompassing genes without introns captured (22.22%) may also require further assessments,

We used here a *precision workflow* approach, focusing on the optimization of one tool based on DOC. Interestingly, as some of our genes included noncoding sequences in gene panels, these specific exonic-plus-intronic captures could provide us the possibility to apply complementary tools using different approaches, like the ones developed for WGS. This can indeed increase both detection performances of CNVs and the spectrum of structural variants that can be detectable in these data.

Of note, all our panels included multiple genes. We do not expect that a design including a single gene, even with its intronic sequences, would reach the sufficient number of targets for CANOES to build a robust model.

We also applied our workflow to multiple WES datasets and reached an overall PPV of 86.49% (95% CI: [82.34–89.81]). As for gene panel CNV detection, a

confirmation by an independent technique is hence still required following the detection of a candidate CNV from WES data, although the low false-positive rate that we show here is expected to be associated with a limited number of molecular confirmations. One of the major features usually required to apply a new technique in a diagnostic workflow is a high sensitivity as compared with a reference technique. Here, we reached a sensitivity of 87.25% (95% CI: [78.84–82.77]). It should be noted that, after visualization on UCSC, five CNVs undetected by CANOES were located in polymorphic regions according to DGV or DGV gold standard tracks but not excluded from our comparison. This point can be explained by a different size between these CNVs in DGV and the one we called, so the criterion of 70% overlap required to consider them as identical was not fulfilled. Thus, this criterion led us to include some polymorphic CNVs in the comparison and hence could underestimate the sensitivity of our workflow. We still chose to keep this parameter as initially set up, because it is widely used in the literature and remains essential for a standardized analysis. Although the sensitivity was not 100%, it is important to notice that aCGH is considered as reference here although the spectrum of events that can be detected is still limited. When comparing our results with aCGH data, it appeared that we missed fewer events than the potential number of true positive CNVs that were missed by aCGH itself. Indeed, from aCGH data, we missed 13 CNVs, but our analyses called 519 candidate CNVs from corresponding WES data and which were theoretically undetectable by aCGH (i.e., either small CNVs or in regions with no aCGH probes coverage). Our PPVs suggest that the vast majority are eventually true. There is no reason to think that some of the CNVs detected by CANOES only might not be as or more deleterious than CNVs detected by both techniques or exclusively by aCGH. Knowing that aCGH misses many CNVs, even using the high-sensitivity chips such as the Agilent 1M one, and even if other chip designs might increase aCGH performances on coding regions, switching to a WES-only approach for CNV detection in a diagnostic setting should not reduce the overall diagnostic yield. Indeed, pathogenicity of CNVs cannot rely only on the size of CNVs as the deletion of a single coding exon in a gene can be sufficient to cause a Mendelian disorder. For example, we previously detected a single exon deletion that was not detectable by array CGH and was clearly pathogenic [42]. In addition, we expect that switching to a WES-only approach for CNV detection could be associated with reduced costs by skipping the CNV screen step by array technologies, although we did not perform cost-effectiveness analyses here.

As compared with aCGH, CANOES allowed the identification of CNVs of any size in regions not covered by probes but also for small CNVs including few exons. In addition, it is important to notice that the majority of CANOES false negatives were also CNVs with only few exons, which implies few targets for CANOES although noncoding probes may help detect some of them by aCGH. This decreased rate of detection of CNVs encompassing few targets has already been shown in other datasets [4, 12] and appears as a limitation inherent to DOC comparison methods.

Interestingly, CANOES allowed the detection of two mosaic rearrangements out of WES data: an *SLC30A3* duplication and a 24 Mb CNV corresponding to a chromosome 20-long arm deletion (Supplementary Table 6). QMPSF data indicated that both CNVs were indeed confirmed albeit with ratios outside the ranges expected for germline events. Those examples highlight the capacity of CANOES to detect mosaic rearrangements, although the tool does not indicate such a feature, which can only be identified following the use of a targeted technique. Of note, the chromosome 20-long arm deletion was detected in a healthy control. This kind of postzygotic rearrangement is not rare in aging people (0.1% after 50 years old) [45]. Those examples highlight the capacity of CANOES to detect mosaic rearrangement.

Beyond the above-mentioned limitations of CNV detection tools from NGS data, somatic CNVs remain a challenge, both for array-based technologies and for NGS-based tools [10]. Among the CNVs detected by our workflow, at least one was considered as likely somatic, as suggested by QMPSF data. However, the sensitivity of DOC tools might remain low in this context [10].

Of note, it is possible to increase the detection of small events or events in complex regions by using the "GenotypeCNV" function of CANOES. The aim of this function is to look precisely at specific regions and call the genotype of the sample for these specific regions, however it is associated with an increase in false positive calls [42], as well as an increase in time and computational resources needed. In particular cases, when known core genes have already been identified in a given disorder, it is possible to combine our approach to call CNVs at the exome level and focus on specific genes using the GenotypeCNV function applied to every exon of these genes to increase the detection performances in core genes at the same time.

In conclusion, we performed an evaluation of the performances of a CNV detection workflow based on read depth comparison from capture-prepared NGS data, one of the most popular methods for NGS in research and diagnostic settings. We highlighted very high sensitivity and positive predictive value, for both NGS gene panel and WES. Although the sensitivity was not perfect for WES data as compared with aCGH, a number of additional true calls were not detected by the so-called reference technique. This highlights the absence of a genuine gold standard up to now. Overall, we consider that switching to an NGS-only

approach is cost-effective as it allows a reduction in overall costs together with likely stable diagnostic yields.

Acknowledgements This study received fundings from Clinical Research Hospital Program from the French Ministry of Health (GMAJ, PHRC 2008/067), the JPND PERADES and France Génomique. This study was co-supported by the Centre National de Référence Malades Alzheimer Jeunes (CNR-MAJ), European Union and Région Normandie. Europe gets involved in Normandie with the European Regional Development Fund (ERDF).

Collaborators
FREX Consortium

Principal Investigators: Emmanuelle Génin[5], Dominique Campion[1,4], Jean-François Dartigues[6], Jean-François Deleuze[3], Jean-Charles Lambert[7], Richard Redon[8]

Bioinformatics group: Thomas Ludwig[5], Benjamin Grenier-Boley[7], Sébastien Letort[5], Pierre Lindenbaum[5], Vincent Meyer[3], Olivier Quenez[1]

Statistical genetics group: Christian Dina[8], Céline Bellenguez[7], Camille Charbonnier[1], Joanna Giemza[8]

Data collection: Stéphanie Chatel[7], Claude Férec[5], Hervé Le Marec[7], Luc Letenneur[6], Gaël Nicolas[1], Karen Rouault[5]

Sequencing: Delphine Bacq[3], Anne Boland[3], Doris Lechner[3]

[5]Inserm UMR 1078, CHRU, University Brest, Brest, France; [6]Inserm UMR 1219, University Bordeaux, Bordeaux, France; [7]Inserm UMR 1167, Institut Pasteur, Lille, France; [8]Inserm UMR 1087/CNRS UMR 6291, l'institut du thorax, Nantes, France

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, et al. De novo rates and selection of large copy number variation. Genome Res. 2010;20:1469–81.
2. Huguet G, Schramm C, Douard E, Jiang L, Labbe A, Tihy F, et al. Measuring and estimating the effect sizes of copy number variants on general intelligence in community-based samples. JAMA Psychiatry. 2018;75:447–57.
3. Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, Guo H, et al. An evaluation of copy number variation detection tools from whole-exome sequencing data. Hum Mutat. 2014;35:899–907.
4. Samarakoon PS, Sorte HS, Kristiansen BE, Skodje T, Sheng Y, Tjønnfjord GE, et al. Identification of copy number variants from exome sequence data. BMC Genomics. 2014;15:661.
5. Roca I, González-Castro L, Fernández H, Couce ML, Fernández-Marmiesse A. Free-access copy-number variant detection tools for targeted next-generation sequencing data. Mutat Res. 2019;779:114–25.
6. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019;20:246.
7. Hehir-Kwa JY, Pfundt R, Veltman JA. Exome sequencing and whole genome sequencing for the detection of copy number variation. Expert Rev Mol Diagn. 2015;15:1023–32.
8. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012;28:423–5.
9. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, et al. Copy number variation detection and genotyping from exome sequence data. Genome Res. 2012;22:1525–32.
10. Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC Bioinformatics. 2017;18:286.
11. Fowler A, Mahamdallie S, Ruark E, Seal S, Ramsay E, Clarke M, et al. Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. Wellcome Open Res. 2016;1:20.
12. Miyatake S, Koshimizu E, Fujita A, Fukai R, Imagawa E, Ohba C, et al. Detecting copy-number variations in whole-exome sequencing data using the eXome Hidden Markov Model: an 'exome-first' approach. J Hum Genet. 2015;60:175–82.
13. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Khera AV, et al. An open resource of structural variation for medical and population genetics. Genomics. 2019. https://doi.org/10.1101/578674.
14. Backenroth D, Homsy J, Murillo LR, Glessner J, Lin E, Brueckner M, et al. CANOES: detecting rare copy number variants from whole exome sequencing data. Nucleic Acids Res. 2014;42:e97.
15. Kuśmirek W, Szmurło A, Wiewiórka M, Nowak R, Gambin T. Comparison of kNN and k-means optimization methods of reference set selection for improved CNV callers performance. BMC Bioinformatics. 2019;20:266.
16. Charbonnier F, Raux G, Wang Q, Drouot N, Cordier F, Limacher JM, et al. Detection of exon deletions and duplications of the mismatch repair genes in hereditary nonpolyposis colorectal cancer families using multiplex polymerase chain reaction of short fluorescent fragments. Cancer Res. 2000;60:2760–3.
17. Baert-Desurmont S, Coutant S, Charbonnier F, Macquere P, Lecoquierre F, Schwartz M, et al. Optimization of the diagnosis of inherited colorectal cancer using NGS and capture of exonic and intronic sequences of panel genes. Eur J Hum Genet EJHG. 2018;26:1597–602.
18. Le Guennec K, Nicolas G, Quenez O, Charbonnier C, Wallon D, Bellenguez C, et al. ABCA7 rare variants and Alzheimer disease risk. Neurology. 2016;86:2134–7.
19. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
20. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
21. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
22. Rovelet-Lecrux A, Deramecourt V, Legallic S, Maurage C-A, Le Ber I, Brice A, et al. Deletion of the progranulin gene in patients with frontotemporal lobar degeneration or Parkinson disease. Neurobiol Dis. 2008;31:41–5.
23. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014;42:D986–92.
24. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32:D493–6.

25. Cassinari K, Quenez O, Joly-Hélas G, Beaussire L, Le Meur N, Castelain M, et al. A simple, universal, and cost-efficient digital PCR method for the targeted analysis of copy number variations. Clin Chem. 2019;65:1153–60.
26. Campion D, Pottier C, Nicolas G, Le Guennec K, Rovelet-Lecrux A. Alzheimer disease: modeling an Aβ-centered biological network. Mol Psychiatry. 2016;21:861–71.
27. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40:e72.
28. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
29. Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, et al. AnnotSV: an integrated tool for structural variations annotation. Berger B, editor. Bioinformatics. 2018. https://doi.org/10.1093/bioinformatics/bty304/4970516.
30. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9.
31. Exome Aggregation Consortium, Ruderfer DM, Hamamsy T, Lek M, Karczewski KJ, Kavanagh D, et al. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. Nat Genet. 2016;48:1107–11.
32. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43:D789–98.
33. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. Genomics. 2019. https://doi.org/10.1101/531210.
34. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.
35. Le Guennec K, Quenez O, Nicolas G, Wallon D, Rousseau S, Richard A-C, et al. 17q21.31 duplication causes prominent tau-related dementia with increased MAPT expression. Mol Psychiatry. 2017;22:1119–25.
36. Mu W, Li B, Wu S, Chen J, Sain D, Xu D, et al. Detection of structural variation using target captured next-generation sequencing data for genetic diagnostic testing. Genet Med Off J Am Coll Med Genet. 2019;21:1603–10.
37. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet. 2012;91:597–607.
38. Di Fiore F, Charbonnier F, Martin C, Frerot S, Olschwang S, Wang Q, et al. Screening for genomic rearrangements of the MMR genes must be included in the routine diagnosis of HNPCC. J Med Genet. 2004;41:18–20.
39. Taylor CF, Charlton RS, Burn J, Sheridan E, Taylor GR. Genomic deletions in MSH2 or MLH1 are a frequent cause of hereditary non-polyposis colorectal cancer: identification of novel and recurrent deletions by MLPA. Hum Mutat. 2003;22:428–33.
40. van der Klift H, Wijnen J, Wagner A, Verkuilen P, Tops C, Otway R, et al. Molecular characterization of the spectrum of genomic deletions in the mismatch repair genes MSH2, MLH1, MSH6, and PMS2 responsible for hereditary nonpolyposis colorectal cancer (HNPCC). Genes Chromosomes Cancer. 2005;44:123–38.
41. Baker M, Strongosky AJ, Sanchez-Contreras MY, Yang S, Ferguson W, Calne DB, et al. SLC20A2 and THAP1 deletion in familial basal ganglia calcification with dystonia. Neurogenetics. 2014;15:23–30.
42. David S, Ferreira J, Quenez O, Rovelet-Lecrux A, Richard A-C, Vérin M, et al. Identification of partial SLC20A2 deletions in primary brain calcification using whole-exome sequencing. Eur J Hum Genet EJHG. 2016;24:1630–4.
43. Guo X-X, Su H-Z, Zou X-H, Lai L-L, Lu Y-Q, Wang C, et al. Identification of SLC20A2 deletions in patients with primary familial brain calcification. Clin Genet. 2019;96:53–60.
44. Nicolas G, Rovelet-Lecrux A, Pottier C, Martinaud O, Wallon D, Vernier L, et al. PDGFB partial deletion: a new, rare mechanism causing brain calcification with leukoencephalopathy. J Mol Neurosci MN. 2014;53:171–5.
45. Machiela MJ, Zhou W, Caporaso N, Dean M, Gapstur SM, Goldin L, et al. Mosaic chromosome 20q deletions are more frequent in the aging population. Blood Adv. 2017;1:380–5.

Supplementary table 1.A — **Gene Panel Sequencing**

	Panel	number of samples	nb of genes	nb of exons	mean Depth of coverage	comprehensive comparison	targeted confirmation
Reference comparison dataset (step 1)	Panel 1 v1	465	11	137	250x	comprehensive QMPSF 4 genes (60 exons)	none
additional dataset (step 2)	Panel 1 v1	2222	11	137	600x	none	QMPSF
	Panel 1 v2	549	15	247	800x	none	QMPSF or MLPA
	Panel 2 v1	240	27	556	500x	none	targeted QMPSF
	Panel 2 v2	80	33	752	600x	none	targeted QMPSF
	Panel 3	220	48	815	600x	none	targeted QMPSF

Supplementary table 1.B — **Whole Exome Sequencing**

	capture Kit	number of samples	Project Origin	Sequencing center	whole genome comparison	targeted confirmation	mean Depth of Coverage
Reference comparison dataset (step 1)	v5UTR	70	ASD	CNRGH	aCGH 180k	none	120x
	v6	27	ID	CNRGH	aCGH 180k	none	100x
	v1	4	AD	Integragen	aCGH 1M	none	60x
	v4UTR	4	AD	Integragen	aCGH 1M	none	90x
	v5	40	AD	CNRGH	aCGH 1M	none	120x
	v2	2	AD	Integragen	aCGH 1M	none	140x
additional dataset (step 2)	v5	472	AD	CNRGH	none	QMPSF / ddPCR	120x
	v5UTR	584	control	CNRGH	none	QMPSF	120x

Supplementary table 1: Summary of datasets used for the evaluation of the CANOES-centered workflow.
Step 1 consisted in the comparison of reference datasets to the results of our CANOES-centered Workflow. Workflow from Gene Panel(1.A) or Whole Exome Sequencing(1.B).
Step 2 consisted in the validation by independent targeted techniques of candidate CNVs detected by our CANOES-centered
1.A : Summary of Gene Panel Sequencing datasets. QMPSF : quantitative multiplex PCR of short fluorescent fragments. MLPA : multiplex ligation-dependent probe amplification
1.B : Summary of Whole Exome Sequencing datasets. All capture kit used come from on Agilent SureSelect technology.
AD : Alzheimer Disease. ASD : Autism spectrum disorder. ID : Intellectual Disability. ddPCR : digital droplet PCR.

CHR	START	END	SAMPLE	CNV	HGVS_nomenclature	size(Kb)	MLCN	Q_SOME	GENE
5	112172694	112174250	10911-001	DEL	NC_000005.9:g.112172694_112174250del	1.556	1	99	APC
5	112079082	112176250	11665-006	DEL	NC_000005.9:g.112079082_112176250del	97.168	1	99	APC
2	48008928	48021943	13753-002	DEL	NC_000002.11:g.48008928_48021943del	13.015	1	99	MSH6
3	37050093	37054402	16308-002	DUP	NC_000003.11:g.37050093_37054402dup	4.309	3	99	MLH1
3	37052503	37060105	17832-001	DEL	NC_000003.11:g.37052503_37060105del	7.602	1	99	MLH1
2	47611056	47635862	21201-001	DEL	NC_000002.11:g.47611056_47635862del	24.806	1	99	MSH2
2	47625841	47658831	24026-001	DEL	NC_000002.11:g.47625841_47658831del	32.99	1	99	MSH2
2	47671472	47679355	25933-001	DUP	NC_000002.11:g.47671472_47679355dup	7.883	3	99	MSH2
2	47611056	47635862	27535-001	DEL	NC_000002.11:g.47611056_47635862del	24.806	1	99	MSH2
2	47629720	47650246	27541-001	DEL	NC_000002.11:g.47629720_47650246del	20.526	1	99	MSH2
2	47611056	47631850	24999-001	DEL	NC_000002.11:g.47611056_47631850del	20.794	1	99	MSH2
2	47637119	47677358	24431-001	DEL	NC_000002.11:g.47637119_47677358del	40.239	1	99	MSH2
3	37044183	37047271	23143-001	DEL	NC_000002.11:g.37044183_37047271del	3.088	1	99	MLH1
2	48023663	48032326	22574-001	DEL	NC_000002.11:g.48023663_48032326del	8.663	1	99	MSH6

Supplementary table 2 : Summary of CNVs used in step 1 (comparison to reference technique) analysis from Gene Panel Sequencing
CNVs were initially detected by a comprehensive QMPSF techniques, and all of them were properly detected by our CANOES-centered workflow from Gene Panel Sequencing data

Chr	Start	End	Sample	CNVType	HGVS_nomenclature	Kb	CopyNbr Estimated	Quality Score	Gene	QMPSF / MLPA Confirmation
1	45780462	45819892	17-01859	DUP	NC_000001.10:g.45780462_45819892dup	39.43	3	99	MUTYH	1
1	153785686	153792286	32223-001	DEL	NC_000001.10:g.153785686_153792286d	6.6	1	99	GATAD2B	1
2	47625842	47637867	04-03911	DUP	NC_000002.11:g.47625842_47637867dup	12.025	3	99	MSH2	1
2	47687044	47696053	06-05306	DUP	NC_000002.11:g.47687044_47696053dup	9.009	3	99	MSH2	1
2	47662983	47688699	10-05453	DUP	NC_000002.11:g.47662983_47688699dup	25.716	3	99	MSH2	1
2	48008928	48021943	13753-002	DEL	NC_000002.11:g.48008928_48021943del	13.015	1	99	MSH6	1
2	47612182	47631802	15-04820	DEL	NC_000002.11:g.47612182_47631802del	19.62	1	99	MSH2,EPCAM	1
2	47629720	47666006	16-11333	DEL	NC_000002.11:g.47629720_47666006del	36.286	1	99	MSH2	1
2	47675213	47694296	17-04290	DEL	NC_000002.11:g.47675213_47694296del	19.083	1	99	MSH2	1
2	47611093	47611213	17-06105	DEL	NC_000002.11:g.47611093_47611213del	0.12	1	3	EPCAM	0
2	47629721	47637867	17-06424	DEL	NC_000002.11:g.47629721_47637867del	8.146	1	99	MSH2	1
2	47611093	47739675	17-09186	DEL	NC_000002.11:g.47611093_47739675del	128.582	1	99	MSH2,EPCAM	1
2	48009642	48011569	17-10896	DEL	NC_000002.11:g.48009642_48011569del	1.927	1	99	MSH6	1
2	47611093	47631802	17-14175	DEL	NC_000002.11:g.47611093_47631802del	20.709	1	99	MSH2,EPCAM	1
2	47635436	47637867	18-01858	DEL	NC_000002.11:g.47635436_47637867del	2.431	1	99	MSH2	1
2	47625842	47658771	18-08533	DEL	NC_000002.11:g.47625842_47658771del	32.929	1	99	MSH2	1
2	47611056	47635862	21201-001	DEL	NC_000002.11:g.47611056_47635862del	24.806	1	99	MSH2	1
2	47625841	47658831	24026-001	DEL	NC_000002.11:g.47625841_47658831del	32.99	1	99	MSH2	1
2	47671472	47679355	25933-001	DUP	NC_000002.11:g.47671472_47679355dup	7.883	3	99	MSH2	1
2	47611056	47635862	27535-001	DEL	NC_000002.11:g.47611056_47635862del	24.806	1	99	MSH2	1
2	47629720	47650246	27541-001	DEL	NC_000002.11:g.47629720_47650246del	20.526	1	99	MSH2	1
2	47667732	47696101	28088-001	DEL	NC_000002.11:g.47667732_47696101del	28.369	1	66	MSH2	1
2	48029073	48032326	28314-001	DEL	NC_000002.11:g.48029073_48032326del	3.253	1	62	MSH6	1
2	47635471	47635862	28316-001	DEL	NC_000002.11:g.47635471_47635862del	0.391	1	69	MSH2	1
2	47635471	47635862	28470-001	DEL	NC_000002.11:g.47635471_47635862del	0.391	1	99	MSH2	1
2	48030351	48032326	28743-001	DEL	NC_000002.11:g.48030351_48032326del	1.975	1	99	MSH6	1
2	47692011	47695115	29016-001	DEL	NC_000002.11:g.47692011_47695115del	3.104	1	99	MSH2	1
2	47637119	47677358	29023-001	DEL	NC_000002.11:g.47637119_47677358del	40.239	1	99	MSH2	1
2	47611056	47631850	30973-001	DEL	NC_000002.11:g.47611056_47631850del	20.794	1	99	MSH2,EPCAM	1
2	47697780	47705898	31168-001	DEL	NC_000002.11:g.47697780_47705898del	8.118	1	99	MSH2	1
2	48004883	48034112	31227-001	DEL	NC_000002.11:g.48004883_48034112del	29.229	1	99	MSH6	1
2	47629720	47631850	33298-001	DEL	NC_000002.11:g.47629720_47631850del	2.13	1	99	MSH2	1
2	47637119	47639026	33355-001	DEL	NC_000002.11:g.47637119_47639026del	1.907	1	99	MSH2	1
2	47637119	47725127	33496-001	DEL	NC_000003.11:g.47637119_47725127del	88.008	1	99	MSH2	1
3	37050136	37054360	11-05279	DUP	NC_000003.11:g.37050136_37054360dup	4.224	3	99	MLH1	1
3	37054017	37059498	16-07141	DEL	NC_000003.11:g.37054017_37059498del	5.481	1	99	MLH1	1
3	37050093	37054402	16308-002	DUP	NC_000003.11:g.37050093_37054402dup	4.309	3	99	MLH1	1
3	37032820	37093414	17-02455	DEL	NC_000003.11:g.37032820_37093414del	60.594	1	99	MLH1	1
3	37044691	37046147	17-04308	DUP	NC_000003.11:g.37044691_37046147dup	1.456	3	11	MLH1	1
3	37052503	37060105	17832-001	DEL	NC_000003.11:g.37052503_37060105del	7.602	1	99	MLH1	1
3	37090248	37094765	31712-001	DEL	NC_000003.11:g.37090248_37094765del	4.517	1	99	MLH1	1
3	37045138	37046147	32099-001	DUP	NC_000003.11:g.37045138_37046147dup	1.009	3	99	MLH1	1
3	37023622	37049251	7623-001	DEL	NC_000003.11:g.37023622_37049251del	25.629	1	99	MLH1	1
5	112172694	112174250	10911-001	DEL	NC_000005.9:g.112172694_112174250del	1.556	1	99	APC	1
5	112079082	112176250	11665-006	DEL	NC_000005.9:g.112079082_112176250del	97.168	1	99	APC	1
5	112158630	112196935	16-07754	DEL	NC_000005.9:g.112158630_112196935del	38.305	1	99	APC	1
5	112170274	112177689	17-10219	DEL	NC_000005.9:g.112170274_112177689del	7.415	1	99	APC	1
5	112032510	112061176	180-004	DEL	NC_000005.9:g.112032510_112061176del	28.666	1	99	APC	1
5	112028285	112196935	30956-001	DEL	NC_000005.9:g.112028285_112196935del	168.65	1	99	APC	1
5	112151589	112196935	32712-001	DEL	NC_000005.9:g.112151589_112196935del	45.346	1	99	APC	1
5	36995238	37000692	23912-001	DEL	NC_000005.9:g.36995238_37000692del	5.454	1	4	NIPBL	1
5	37002967	37065173	31262-001	DEL	NC_000005.9:g.37002967_37065173del	62.206	1	99	NIPBL	1
6	157469721	157470121	8205-001	DEL	NC_000006.11:g.157469721_157470121d	0.4	1	99	ARID1B	1
7	6029053	6029652	11781-001	DEL	NC_000007.13:g.6029053_6029652del	0.599	1	99	PMS2	1
7	6020403	6038966	25189-001	DEL	NC_000007.13:g.6020403_6038966del	18.563	1	99	PMS2	1
7	6016973	6022754	26162-001	DUP	NC_000007.13:g.6016973_6022754dup	5.781	3	8	PMS2	0
7	6020403	6025843	27898-001	DEL	NC_000007.13:g.6020403_6025843del	5.44	1	99	PMS2	0
7	6016973	6027324	29135-001	DUP	NC_000007.13:g.6016973_6027324dup	10.351	3	99	PMS2	1
7	6005065	6020110	31357-001	DEL	NC_000007.13:g.6005065_6020110del	15.045	1	52	PMS2	0
9	3247928	3301730	14682-001	DEL	NC_000009.11:g.3247928_3301730del	53.802	1	99	RFX3	0
9	13106807	13279672	32202-001	DEL	NC_000009.11:g.13106807_13279672del	172.865	1	99	MPDZ	1
10	89684804	89693116	11-06624	DEL	NC_000010.10:g.89684804_89693116del	8.312	1	99	PTEN	1
10	89726729	89733691	11-06624	DEL	NC_000010.10:g.89726729_89733691del	6.962	1	99	PTEN	1
10	89685150	89693116	13-06009	DEL	NC_000010.10:g.89685150_89693116del	7.966	1	99	PTEN	0
10	89726729	89732364	13-06009	DEL	NC_000010.10:g.89726729_89732364del	5.635	1	99	PTEN	0
10	89621147	89652755	15-06158	DEL	NC_000010.10:g.89621147_89652755del	31.608	1	99	PTEN	1
10	89621147	89624589	18-00909	DUP	NC_000010.10:g.89621147_89624589dup	3.442	3	27	PTEN	0
10	89685220	89690951	19600-001	DEL	NC_000010.10:g.89685220_89690951del	5.731	1	99	PTEN	0
10	89727637	89732364	19600-001	DEL	NC_000010.10:g.89727637_89732364del	4.727	1	99	PTEN	0
10	88502067	88699574	23140-001	DEL	NC_000010.10:g.88502067_88699574del	197.507	1	99	BMPR1A	1
10	89653424	89653732	26342-001	DEL	NC_000010.10:g.89653424_89653732del	0.308	1	99	PTEN	0
10	88678073	89742644	3483-001	DEL	NC_000010.10:g.88678073_89742644del	1064.571	1	99	PTEN,BMPR1A	1
16	9856965	9862986	22569-001	DEL	NC_000016.9:g.9856965_9862986del	6.021	1	99	GRIN2A	1
17	7573397	7582278	07-04795	DEL	NC_000017.10:g.7573397_7582278del	8.881	1	99	TP53	1
17	7580599	7583471	09-06646	DUP	NC_000017.10:g.7580599_7583471dup	2.872	3	87	TP53	1
17	7585364	7596077	11-06921	DEL	NC_000017.10:g.7585364_7596077del	10.713	1	99	WRAP53,TP53	1
17	7577798	7584538	13-05835	DEL	NC_000017.10:g.7577798_7584538del	6.74	1	99	TP53	1
17	7569492	7592875	17-05375	DEL	NC_000017.10:g.7569492_7592875del	23.383	1	99	TP53	1
17	7568467	7568848	17-10989	DUP	NC_000017.10:g.7568467_7568848dup	0.381	3	34	TP53	0
17	7564623	7566106	18-10734	DEL	NC_000017.10:g.7564623_7566106del	1.483	1	73	TP53	0
17	7579644	7583513	18889-001	DUP	NC_000017.10:g.7579644_7583513dup	3.869	3	99	TP53	1
17	7576579	7581267	28645-001	DUP	NC_000017.10:g.7576579_7581267dup	4.688	3	99	TP53	1
17	7572557	7574135	31014-006	DEL	NC_000017.10:g.7572557_7574135del	1.578	1	99	TP53	1
17	7572176	7593265	32539-001	DEL	NC_000017.10:g.7572176_7593265del	21.089	1	99	TP53	1
17	7570171	7604301	5415-007	DEL	NC_000017.10:g.7570171_7604301del	34.13	1	99	TP53	1
18	48555601	48582805	18-02468	DUP	NC_000018.9:g.48555601_48582805dup	27.204	3	99	SMAD4	1
18	48591693	48593761	32643-001	DUP	NC_000018.9:g.48591693_48593761dup	2.068	3	36	SMAD4	1
19	1203112	1208307	14693-002	DEL	NC_000019.9:g.1203112_1208307del	5.195	1	99	STK11	1
19	1203112	1208307	17-01927	DEL	NC_000019.9:g.1203112_1208307del	5.195	1	99	STK11	1
19	1203822	1210586	17-11160	DEL	NC_000019.9:g.1203822_1210586del	6.764	1	99	STK11	1
19	1203822	1230469	18-02613	DEL	NC_000019.9:g.1203822_1230469del	26.647	1	99	STK11	1
19	1192457	1243248	22373-001	DEL	NC_000019.9:g.1192457_1243248del	50.791	1	99	STK11	1
19	1222884	1225265	29131-001	DEL	NC_000019.9:g.1222884_1225265del	2.381	1	99	STK11	1
19	1213712	1219536	31452-001	DEL	NC_000019.9:g.1213712_1219536del	5.824	1	99	STK11	1
19	1217466	1243248	31668-001	DEL	NC_000019.9:g.1217466_1243248del	25.782	1	99	STK11	1
22	41542649	41544015	33893-001	DEL	NC_000022.10:g.41542649_41544015del	1.366	1	93	EP300	1
X	29301014	29301374	23915-001	DUP	NC_000023.10:g.29301014_29301374dup	0.36	GAIN	22	IL1RAPL1	1
X	53421612	53421892	27463-001	DUP	NC_000023.10:g.53421612_53421892dup	0.28	GAIN	22	SMC1A	0
X	153127953	153129238	30721-002	DEL	NC_000023.10:g.153127953_153129238d	1.285	LOSS	99	L1CAM	1
X	153127953	153141522	5922-001	DEL	NC_000023.10:g.153127953_153141522d	13.569	LOSS	99	L1CAM	1
X	153296359	153298052	6929-001	DEL	NC_000023.10:g.153296359_153298052d	1.693	LOSS	99	MECP2	1

Supplementary table 3 : Summary of candidate CNVs called by our CANOES-centered workflow from Gene Panel Sequencing Data.
The first three columns define the CNVs as called by CANOES, corresponding to the start of the first regions detected as deviated to the end of the last region. The last column indicates if the CNVs was confirmed or not (1=true CNV)

2.2. Recherche de CNV parmi des données massives d'exomes : application à une étude cas-témoins portant sur la maladie d'Alzheimer

Une fois notre approche initiale établie et notre pipeline validé sur plusieurs jeux de données différents, nous avons pu mettre en place l'analyse d'un jeu de données massif de cas et témoins construit dans le cadre du consortium européen ADES complété par les données du consortium américain ADSP. Ce jeu de données avait déjà été utilisé pour identifier de nouveaux facteurs de risques impliqués dans la maladie d'Alzheimer en se basant sur l'analyse des variations ponctuelles (SNV/indel), travail auquel j'ai activement participé (Hostege et al.). C'est dans le cadre de cette analyse que deux nouveaux facteurs de risque ont été identifiés : les variants rares faux sens et/ou perte de fonction des gènes *ABCA1* et *ATP8B4, en* plus des trois gènes majeurs déjà connus et répliqués : *SORL1* (identifié initialement par notre laboratoire (Nicolas et al. 2016; C. Pottier et al. 2012)*, TREM2* et *ABCA7*. Nous avons cherché à identifier les variations du nombre de copies dans ce même jeu de données, afin d'identifier de potentiels nouveaux facteurs de risque génétique.

Plusieurs problèmes ont été rencontrés lors de ces travaux. Le premier est l'origine variée des données et leur très grande hétérogénéité. En effet, les exomes ont été produits dans plusieurs laboratoires différents, sur des plateformes de séquençage différentes, et avec des kits de capture différents. Étant donné que le logiciel CANOES requière les coordonnées du kit de capture pour faire les calculs de profondeur, il nous était impossible de faire un calling global de tous les individus en une seule fois. Nous aurions pu envisager ne considérer chaque exon comme une région pour effectuer le comptage de read mais cela aurait induit plusieurs biais : tout d'abord tous les kits de capture ne capturent pas tous les gènes du génome, et pour un même gène, des différences existent dans les exons capturés (exons alternatif, régions UTR, etc). Ceci induirait des régions avec peu ou pas de reads que le logiciel pourrait potentiellement faussement interpréter comme des délétions et donc du bruit dans notre signal. Ensuite, la capture induit normalement un biais dans la profondeur moyenne principalement dû au taux de GC de la région. CANOES prend en compte ce biais en appliquant une correction en fonction du pourcentage de GC : si l'on prend les coordonnées exactes des exons, corriger sur le taux de GC ne serait plus pertinent et il faudrait alors passer par d'autres outils. Au final, et afin de réduire au maximum la variabilité technique, nous avons fait le choix d'analyser les données produites pour chaque laboratoire et pour chaque kit séparément, et ce même si un même kit était utilisé dans plusieurs laboratoires. Avec la même idée de réduire au maximum la variabilité technique et lorsque

cela était possible, nous faisions l'analyse par run de séquençage. Lorsque cela n'était pas possible, nous regroupions alors tous les individus du même laboratoire séquencé sur le même kit.

Une fois cette étape terminée, il nous a été possible d'appliquer notre pipeline et d'obtenir les résultats de détection des variations du nombre de copies. La suite du travail a alors consisté à trouver une solution, pour faire correspondre les résultats issus des différents jeux de données. Il faut pour cela remettre dans le contexte l'objectif de notre étude. Lorsque l'on travaille sur des études cas/témoins afin d'identifier de nouveaux facteurs de risques, on cherche à calculer la fréquence des différents CNV dans les différents groupes puis identifier un enrichissement par un test statistique. Lorsque l'on travaille avec des variations ponctuelles, il est facile de comparer une même variation identifiée chez plusieurs individus. Du fait de la rareté des variations d'intérêt, les tests portant sur les variations ponctuelles ont souvent recours à une agrégation à l'échelle du gène, le plus souvent de type burden test comme dans le travail du consortium (Holstege et al. 2022). Pour les variations du nombre de copies, une analyse à l'échelle du CNV nécessite d'être sûr que les coordonnées soient les mêmes pour pouvoir considérer qu'il s'agit bien du même événement, et une telle analyse aurait comme conséquence de ne s'intéresser qu'aux CNV récurrents. Quoi qu'il en soit, le fait d'avoir des kits de capture différents implique que les coordonnées ne sont jamais tout à fait les mêmes et ce même si, biologiquement, il s'agit bien du même événement. Nous avons cherché et testé plusieurs méthodes d'agrégation des CNVs avant d'arriver à la méthodologie présentée dans ces travaux. Nous avons fait le choix de décomposer les différents CNVs au niveau des transcrits affectés, de façon à travailler au niveau du plus petit dénominateur commun possible.

Enfin, une fois les fréquences établies des différents événements dans les différentes populations de cas et de témoins, nous avons choisi d'effectuer une analyse en dosage des différents transcrits : l'impact des délétions et des duplications est supposé être une diminution ou une augmentation de la production de la protéine codée par le transcrit en question, respectivement. On s'attend à observer un effet opposé entre les cas et les témoins : si la diminution de production d'une protéine est un facteur de risque de développer la maladie alors une augmentation de cette même protéine induirait un effet protecteur et serait donc surreprésenté chez les témoins. Dans le cas des délétions, il est facile de faire l'hypothèse qu'une perte partielle ou totale du gène induit le plus souvent une perte de fonction, mais dans le cas des duplications, l'interprétation est plus compliquée. En effet, lorsque l'on utilise une approche basée sur la profondeur de lecture, il est possible de détecter les duplications mais il est impossible de savoir la localisation de ces duplications. Ceci rend plus complexe l'interprétation des duplications partielles, c'est-à-dire n'affectant qu'une partie des exons d'un gène.

S'il y a duplication partielle et à distance, c'est finalement la zone d'insertion de cette duplication qui va avoir potentiellement un impact plus que l'élément dupliqué en lui-même. Si la duplication est en tandem, le risque est alors de produire un transcrit aberrant potentiellement avec perte de phase de la protéine. Devant cette difficulté à interpréter l'événement, nous avons fait le choix ne considérer que les duplications complètes des transcrits.

De plus, et parce que les bases de données de CNV sont majoritairement issues de puces ou de génomes, mais assez peu d'exomes, la filtration sur les CNV fréquents avec un certain pourcentage de chevauchement peut poser problème. Pour réduire le risque que des gènes fréquemment délétés ou dupliqués restent impactés par des CNV, nous avons travaillé sur les niveaux de filtration et finalisé l'analyse sur des jeux de transcrits haploinsuffisants ou triplosensibles.

Un des enseignements tirés de notre étude en partie 1 a été la mise en évidence d'une plus faible performance du pipeline pour les CNV à une seule cible (souvent des CNV monoexoniques), la valeur prédictive positive de notre approche progressant de 70% pour les CNVs monoexoniques à plus de 90% à partir de deux cibles. Pour réduire le risque de faux positifs, nous nous sommes donc focalisés sur les CNV emportant au moins deux cibles.

Enfin, des approches complémentaires ont été proposées, en plus du critère principal de l'étude, qui portait sur l'analyse en dosage des CNV dans les deux jeux de transcrits en comparant les malades jeunes (EOAD) aux témoins : des analyses ont impliqué également les malades à début tardif, avec un focus uniquement sur les délétions (complètes et partielles) ou sur les duplications (complètes), des analyses ciblées sur les gènes de GWAS dans la MA, et enfin des analyses combinées des délétions avec les variations ponctuelles perte de fonction.

Au total, et grâce à ces différentes adaptations, nous avons pu effectuer l'analyse de 22 319 individus et identifier de potentiels nouveaux facteur de risque pour les formes précoces de la maladie d'Alzheimer.

Les résultats principaux de notre étude sont l'identification de plusieurs loci candidats comme facteurs de risque potentiels de formes précoces de la maladie d'Alzheimer. Parmi ces différents résultats, on notera la présence de délétions dans la région 22q11.21 de manière plus importante par des patients jeunes par rapport à la population de témoins qui portent préférentiellement des duplications en miroir. Ces délétions, dites centrales, ont déjà été identifiées comme facteurs de risque de maladies du développement, et des délétions plus larges sont responsables du syndrome de DiGeorge

(McDonald-McGinn et Sullivan 2011; Morrison et al. 2020). Nous avons aussi pu identifier des délétions affectant les gènes *ABCA1* et *ABCA7*, ces gènes étant des facteurs de risque déjà identifiés dans le cadre de la maladie d'Alzheimer à travers des mécanismes perte de fonction. De plus, nous priorisons, dans une approche combinée de perte de fonction (CNV + SNV et indels), les variations de *CSTB*, en plus d'*ABCA1* et d'A*BCA7*. Comme pour toutes études d'association, il est important de répliquer ces résultats dans des études indépendantes. Différentes demandes ont été effectués et sont actuellement en cours de réalisation. La première demande a été adressée au consortium EADB, auquel nous participons aussi, qui travaille sur des données de puces à SNP et a produit plusieurs études de GWAS. En plus de la réplication, l'appel à ce consortium a un second intérêt : une partie des échantillons de notre étude est aussi incluses dans les études de GWAS d'EADB : les individus concernés ne pourront bien sûr pas être intégrés dans la réplication, mais pourront servir de validation indépendante de l'existence de certains CNV d'intérêt, d'une taille suffisante pour être détectables. La seconde cohorte interrogée est la UKBiobank.

Les résultats de notre étude décrivant l'analyse de la cohorte et les différents résultats obtenus font l'objet de ce second chapitre de résultats. L'article, présenté ci-dessous, sera prochainement soumis sur le site de PeerReview MedRxiv en attendant d'obtenir les résultats de réplications dans des cohortes indépendantes. J'ai également été amené à présenter les résultats oralement lors du congrès ADPD (Alzheimer Disease/Parkinson Disease) en mars 2021 à Barcelone, et lors des JNRB (Journées Normandes de Recherche Biomédicale) à Caen en juin 2023, ainsi que sous forme de poster lors des assises de génétique de Rennes en 2022 et dans les différentes réunions de travail du consortium ADES.

Extremely rare CNVs contributing to Alzheimer disease risk: a case-control association analysis of exome sequencing data from 22,319 individuals

Olivier Quenez[1]*, Catherine Schramm[1]*, Kévin Cassinari[1], Marc Hulsman[2,3,4,5], Anne-Claire Richard[1], Stéphane Rousseau[1], Anne Rovelet-Lecrux[1], Shahzad Ahmad[6,7], Najaf Amin[6,8], Philippe Amouyel[9], Olivia Belbin[10,11], Céline Bellenguez[9], Claudine Berr[12], Anne Boland[13], Paola Bossù[14], Femke Bouwman[4,5], Jose Bras[15,16], Jordi Clarimon[10,11], Antonio Daniele[17], Jean-François Dartigues[18], Stéphanie Debette[18,19], Jean-François Deleuze[13], Nicola Denning[20], Oriol Dols-Icardo[10,11], Cornelia M. van Duijn[6,8], Wiesje M. van der Flier[3,4], Nick C. Fox[21], Daniela Galimberti[22,23], Emmanuelle Genin[24], Johan J. P. Gille[25], Benjamin Grenier-Boley[9], Detelina Grozeva[26], Rita Guerreiro[15,16], John Hardy[27], Steffi G. Riedel-Heller[28], Clive Holmes[29], Holger Hummerich[30], M. Arfan Ikram[6], M. Kamran Ikram[6], Iris E. Jansen[3,4,31], Amit Kawalia[32], Robert Kraaij[33], Marc Lathrop[34], Sven J. van der Lee[2,3,4,5], Morgane Lacour[35], Afina W. Lemstra[3,4], Alberto Lleó[10,11], Lauren Luckcuck[26], Marcel M. A. M. Mannens[36], Rachel Marshall[26], Carlo Masullo[37], Simon Mead[30], Patrizia Mecocci[38], Alun Meggy[20], Merel O. Mol[39], Kevin Morgan[40], Alexandre Morin[35], Benedetta Nacmias[41,42], Penny J. Norsworthy[30], Florence Pasquier[43], Pau Pastor[44,45], Alfredo Ramirez[32,46,47,48,49], Rachel Raybould[20], Richard Redon[50], Marcel J. T. Reinders[5], Fernando Rivadeneira[33], Jeroen G. J. van Rooij[33,39], Natalie S. Ryan[21], Salha Saad[26], Pascual Sanchez-Juan[11,51], Philip Scheltens[3,4], Jonathan M. Schott[21], Davide Seripa[52], Daoud Sie[25], Rebecca Sims[26], Erik A. Sistermans[25], Sandro Sorbi[41,42], Resie van Spaendonk[25], Gianfranco Spalletta, John C. van Swieten[39], Niccolo' Tesi[2,3,4,5], Petty Tijms[3], André G. Uitterlinden[33], Pieter Jelle Visser[3], Michael Wagner[47,48], David Wallon[35], Julie Williams[26], Aline Zarea[35], Jean-Charles Lambert[9], Henne Holstege[2,3,4,5], Camille Charbonnier[1], Gaël Nicolas[1]

1. Univ Rouen Normandie, Normandie Univ, Inserm U1245 and CHU Rouen, Department of Genetics and CNRMAJ, F-76000 Rouen, France. 2. Genomics of Neurodegenerative Diseases and Aging, Human Genetics, Vrije Universiteit Amsterdam, Amsterdam UMC location VUmc, Amsterdam, theNetherlands. 3. Alzheimer Center Amsterdam, Neurology, Vrije Universiteit Amsterdam, Amsterdam UMC location VUmc, Amsterdam, the Netherlands. 4. Amsterdam Neuroscience, Neurodegeneration, Amsterdam, the Netherlands. 5. Delft Bioinformatics Lab, Delft University of Technology, Delft, the Netherlands. 6. Department of Epidemiology, Erasmus Medical Centre, Rotterdam, the Netherlands. 7. Leiden Academic Centre for Drug Research, Leiden, the Netherlands.

8. Nuffield Department of Population Health Oxford University, Oxford, UK. 9. Université Lille, INSERM, Centre Hospitalier Universitaire Lille, Institut Pasteur de Lille, U1167-RID-AGE facteurs de risque et déterminants moléculaires des maladies liées au vieillissement, Lille, France. 10. Department of Neurology, II B Sant Pau, Hospital de la Santa Creu i Sant Pau, Universitat Autònoma de Barcelona, Barcelona, Spain. 11. Biomedical Research Networking Center on Neurodegenerative Diseases, National Institute of Health Carlos III, Madrid, Spain. 12. Université Montpellier, INSERM, Institute for Neurosciences of Montpellier, Montpellier, France. 13. Université Paris-Saclay, Commissariat à l'énergie Atomique et aux énergies Alternatives, Centre National de Recherche en Génomique Humaine Evry, Gif-sur-Yvette, France. 14. Experimental Neuro-psychobiology Laboratory, Department of Clinical and Behavioral Neurology, Istituto di Ricovero e Cura a Carattere Scientifico Santa Lucia Foundation, Rome, Italy. 15. Department of Neurodegenerative Science, Van Andel Institute, GrandRapids, MI, USA. 16. Division of Psychiatry and Behavioral Medicine, Michigan State University College of Human Medicine, Grand Rapids, MI, USA. 17. Department of Neuroscience, Catholic University of Sacred Heart, Fondazione Policlinico Universitario A. Gemelli Istituto di Ricovero e Cura a Carattere Scientifico, Rome, Italy. 18. Université Bordeaux, INSERM, Bordeaux Population Health Research Center, Bordeaux, France. 19. Department of Neurology, Bordeaux University Hospital, Bordeaux, France. 20. UKDRI Cardiff, School of Medicine, Cardiff University, Cardiff, UK. 21. Dementia Research Centre, University College London Queen Square Institute of Neurology, London, UK. 22. Fondazione Istituto di Ricovero e Cura a Carattere Scientifico C' Granda, Ospedale Policlinico, Milan, Italy. 23. University of Milan, Milan, Italy. 24. Université Brest, INSERM, Etablissement Français du Sang, Centre Hospitalier Universitaire Brest, Unité Mixte de Recherche 1078, GGB, Brest, France. 25. Genome Diagnostics, Department of Human Genetics, VU University, AmsterdamUMC (locationVUmc), Amsterdam, the Netherlands. 26. Medical Research Council Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neuroscience, School of Medicine, Cardiff University, Cardiff, UK. 27. Reta Lila Weston Research Laboratories, Department of Molecular Neuroscience, University College London Institute of Neurology, London, UK. 28. Institute of Social Medicine, Occupational Health and Public Health, University of Leipzig, Leipzig, Germany. 29. Clinical and Experimental Science, Faculty of Medicine, University of Southampton, Southampton, UK. 30. Medical Research Council Prion Unit at University College London, University College London Institute of Prion Diseases, London, UK. 31. Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, Vrije University, Amsterdam, the Netherlands. 32. Division ofNeurogenetics and Molecular Psychiatry,

Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany. 33. Department of Internal Medicine, Erasmus Medical Centre, Rotterdam, the Netherlands. 34. McGill University and Genome Quebec Innovation Centre, Montreal, Quebec, Canada. 35. Univ Rouen Normandie, Normandie Univ, Inserm U1245 and CHU Rouen, Department of Neurology and CNRMAJ, F-76000 Rouen, France 36. Department of Human Genetics, Amsterdam UMC, University of Amsterdam, Amsterdam Reproduction and Development Research Institute, Amsterdam, the Netherlands. 37. Institute of Neurology, Catholic University of the Sacred Heart, Rome, Italy. 38. Institute of Gerontology and Geriatrics, Department of Medicine and Surgery, University of Perugia, Perugia, Italy. 39. Department of Neurology, Erasmus Medical Centre, Rotterdam, the Netherlands. 40. Human Genetics, School of Life Sciences, University of Nottingham, Nottingham, UK. 41. Department of Neuroscience, Psychology, Drug Research and Child Health University of Florence, Florence, Italy. 42. IRCCS Fondazione Don Carlo Gnocchi, Florence, Italy. 43. Université Lille, INSERM, Centre Hospitalier Universitaire Lille, UMR1172, Resources and Research Memory Center (MRRC) of Distalz, Licend, Lille, France. 44. Fundació Docència i Recerca MútuaTerrassa and Movement Disorders Unit, Department of Neurology, University Hospital MÃºtuaTerrassa, Barcelona, Spain. 45. Memory Disorders Unit, Department of Neurology, Hospital Universitari Mutua de Terrassa, Barcelona, Spain. 46. Department of Psychiatry and Glenn Biggs Institute for Alzheimer's and Neurodegenerative Diseases, San Antonio, TX, USA. 47. Department of Neurodegenerative Diseases and Geriatric Psychiatry, University Hospital Bonn, Medical Faculty, Bonn, Germany. 48. German Center for Neurodegenerative Diseases, Bonn, Germany. 49. Cluster of Excellence Cellular Stress Responses in Aging-Associated Diseases, University of Cologne, Cologne, Germany. 50. Université de Nantes, Centre Hospitalier Universitaire Nantes, Centre National de la Recherche Scientifique, INSERM, l'institut du Thorax, Nantes, France. 51. Neurology Service, Marqués de Valdecilla University Hospital (University of Cantabria and IDIVAL), Santander, Spain. 52. Laboratory for Advanced Hematological Diagnostics, Department of Hematology and Stem Cell Transplant, Lecce, Italy. 53. Laboratory of Neuropsychiatry, Department of Clinical and Behavioral Neurology, Istituto di Ricovero e Cura a Carattere Scientifico Santa Lucia Foundation, Rome, Italy.

Abstract

Rare coding single nucleotide variants (SNV) and short insertions or deletions (indels) significantly contribute to Alzheimer disease (AD) genetic risk, from pathogenic variants in autosomal dominant genes to deleterious variants with a moderate to strong effect in a handful of risk-factor genes. In contrast, copy number variants (CNV) have been scarcely studied, with the exception of a few autosomal dominant examples, such as APP duplications. We took advantage from a large case-control dataset of 22,319 exomes (4,150 early-onset AD (EOAD, onset ≤65 years), 8,519 late onset AD (LOAD) and 9,650 controls) to detect CNVs. After the identification of 17 causative CNVs in autosomal dominant genes (8 novel, 6 *APP* duplications and 2 *MAPT* duplications), we performed two analyses: (i) a protein-coding genome-wide analysis at the transcript level using a dosage strategy (EOAD versus controls) and (ii) an integrated loss-of-function (LOF) analysis gathering short truncating variants with CNV-deletions in genes prioritized in (i) and in a list of known AD risk genes from Genome-Wide Association Studies (GWAS, common variants) and in burden tests of rare SNVs and indels. We identified 21 genes on 5 different loci, with a false discovery rate (FDR) below 10%, including the so-called central region on chromosome 22q11.2 (FDR=0.0386), a region in linkage disequilibrium with the *APOE* locus on chromosome 19 (FDR=0.0271), and three single-gene loci, namely *MBL2* (FDR=0.0271), *FADS6* (FDR=0.0271), and *ADI1* (FDR=0.0916). Loss-of-function analysis helped narrowing the region of interest to the *SCARF2-KLHL22-MED15* region within the 22q11.2 region, and highlighted rare deletions in *ABCA1* and *ABCA7* as contributing to AD risk, deletions representing 10% (3/30) and 8.6% (10/115) of loss-of-function alleles of these genes, respectively, and *CTSB* loss-of-function as a candidate rare AD genetic determinant (EOAD-control OR=5.03, 95%CI=[1.5-20.7], p=0.0089). CNVs represent a minority of the deleterious alleles at known loci, as compared to SNVs and indels. Gathering CNVs with SNVs and indels may help increasing power to detect new signals.

Introduction

The etiology of Alzheimer disease (AD) is heterogeneous. Some families exhibit autosomal dominant early onset AD (EOAD, onset ≤65 years), explained by rare variants in the *PSEN1*, *PSEN2* or *APP* genes, but carriers of such highly penetrant variants represent less than 0.5% of all AD cases[1]. In non-Mendelian cases, including EOAD, AD is considered as a complex disorder with a high genetic component[2].

In complex AD, a large diversity of risk factors have been reported to date. They are usually classified according to their respective frequencies and effect sizes. Among common variants (frequency >1%),

the ε4 allele of the APOE gene is the main risk factor with odd ratios (OR) of 3-4 and 11-14 respectively for heterozygous and homozygous carriers[3]. A recent large genome-wide association study (GWAS) reported 75 loci, most of them being associated through common non-coding single nucleotide polymorphisms (SNPs) with a modest effect on AD risk[4]. On the other hand, a burden of rare (frequency <1%) to ultra-rare variants in *SORL1*, *TREM2*, *ABCA7*, *ABCA1* and *ATP8B4* as well as two recurrent rare single variants in *PLCG2* and *ABI3* demonstrated a larger diversity of effects, ranging from a modest effect for the least rare variants (e.g., *ABI3*, *TREM2* R62H) to a very strong effect for some ultra-rare variants (e.g., loss-of-function SORL1 variants)[4]–[10]. Importantly, except for a few rare recurrent variants showing nominal association with AD, evidence for rare variant association with AD risk is generally obtained following burden tests gathering truncating variants with missense, predicted deleterious variants, at the gene level. Burden tests performed on the category of truncating variants alone either showed an exome-wide level of association for *SORL1* and *ABCA7* or a suggestive signal for *ABCA1* and *TREM2*. For the latter genes, truncating variants remain extremely rare, thus likely explaining insufficient power. Importantly, genetic results are in line with known mechanisms, indeed suggesting a deleterious effect of haploinsufficiency or loss of function of *SORL1*, *TREM2*, *ABCA7* and *ABCA1*.

While the most studied category of genetic determinants in AD is represented by short variants, i.e., single nucleotide variants (SNV) or short insertions and deletions (indel), Copy Number Variants (CNV) have also been involved in AD. Duplications of the *APP* locus[11] and in-frame deletions of exon 9 or exon 9 and 10 of *PSEN1*[12], [13] cause autosomal-dominant AD. Some large common CNVs were identified by SNP arrays and associated with a modest risk of AD [14], but none of them was genome-wide significant. On the other hand, a handful of studies focusing on rare CNVs were performed using arrays, on a limited number of EOAD patients[15], [16]. Individual CNVs affecting genes of potential interest were prioritized, but segregation data in families was missing and such CNVs remained too rare to be assessed in case-control studies[17]. Large datasets should help detecting rare CNVs associated with AD risk.

Up to now, CNV studies in AD have mainly focused on chips technologies, as SNP arrays at a rather large scale or array CGH at a smaller scale. Such chips can detect CNVs with intermediate resolution, generally with a CNV size of 50 to 100 kb or more, depending on the array density. Thus, they classically miss smaller CNVs. Sequencing technologies can also be used for CNV detection and can detect CNVs of any size. Bioinformatics tools used to detect CNVs from sequencing data are based on four main approaches [18], [19]: (i) relative distribution of reads along the sequence or read depth approach, (ii)

relative positions of paired reads from each other, (iii) multiple and partial alignments of reads and (iv) de novo assembly. When working from sequencing data obtained following capture, as for exome sequencing, only the read depth approach can be applied with a high level of accuracy at the exome level[20], [21], while data obtained from whole genome sequencing require combinations of tools, also including read depth approaches although with distinct procedures and tools. We previously assessed the performances of a CNV-calling workflow based on the CANOES tool[22] which showed good sensitivity (87.25 %) and positive predictive value (85.2%) from exome sequencing data[23]. One of the largest sequencing datasets available worldwide to perform a case-control study in AD is probably the combination of the European and American consortia into the ADES-ADSP dataset, which recently unveiled burdens of rare variants in *ATP8B4* and *ABCA1* in addition to the known genes and which also has the advantage of being enriched in EOAD cases (35.7 %), among which we can expect a higher contribution of rare CNVs, as for SNVs and indels [5]. This dataset is made of a majority of exomes (68.6% after QC).

Here, we took advantage of the large number of exome sequencing data of EOAD cases and controls available in the ADES-ADSP dataset. We performed uniform CNV calling on 22,319 samples using a validated workflow and built up a specific QC pipeline to perform harmonized transcript-based analyses. Beyond novel CNVs in autosomal dominant genes, we highlight the role of extremely rare deletions in known risk factor genes as well novel candidates requiring replication in independent datasets.

METHODS

Exome sequencing dataset

We considered the megasample of exomes from ADES and ADSP datasets as described in Holstege et al[5] in a two-stage analysis. Only samples passing the quality control (QC) described in detail in Holstege et al[5] were selected for our analysis. Briefly, for the individual QC based on short sequence variants, all sequencing (Fastq) files were processed using the same BWA-GATK-based pipeline on the Cartesius supercomputer embedded in the Dutch national e-infrastructure. Samples with either a high level of variant missingness, a high suspicion of DNA contamination, a discordant genetic sex

annotation, a non-European ancestry, a high number of novel variants (compared to dbSNP v150), a deviation from standard heterozygous/homozygous or transition/transversion ratios, and relatives up to the 3rd degree were excluded. We considered as cases all individuals with a diagnosis of definite or probable AD (using the NIAA or NINCDS-ADRA criteria depending on the date of diagnosis[24], [25]), based on clinical examination and paraclinical information including CSF AD biomarkers, when available. Individuals with unclear diagnosis were excluded (e.g., Braak stage I-II in cases, or Braak stages of V-VI in controls). In addition, pathogenic SNV and indel variant carriers in a list of Mendelian dementia genes were also excluded from stage-1 data in the Holstege et al. dataset. We applied the same analysis leading to the exclusion of 22 additional pathogenic variant carriers among cases and controls from stage-2 so that the megasample fulfils the same criteria (Table S14). Detailed methods for individual QC are available as a supplementary information file in Holstege et al.[5].

Because the ADES-ADSP dataset is built from multiple studies, and to ensure a good accuracy of CANOES, the CNV caller used here, we focused on samples as homogeneous as possible in batches of at least 50 individuals (Table S1). When library preparation and sequencing batches information were not available, we grouped into CNV-calling batches, samples belonging to a same study, prepared with a same capture kit, and sequenced in a same sequencing center. Overall, our starting dataset contained 12,669 exomes from cases (including 4,150 EOAD) and 9,650 exomes from controls (see Sup Table S1).

All participants provided informed written consent as described in Holstege et al. This study, based on existing data, was approved by the CERDE ethics committee from Rouen University Hospital (CERNI notifications 2017-015 and 2019-055).

CNV Calling

CNV calling was processed following a workflow centered on CANOES[22], a tool based on the distribution of the depth of coverage information across samples. It includes a correction based on GC content of each target to reduce the background variability often observed in NGS data[26].

The workflow was applied from BAM files, as previously described[23]. BAM files were retrieved either from the pipeline described in Holstege et al. or directly downloaded from the dbGAP website (ADSP-stage 2, 1554 samples). All samples were processed on either the Cartesius supercomputer or the CEREBRO cluster from the sequencing facility of the University of Rouen Normandie. Of note, ADSP-stage 2 BAM files were aligned to the GRCh38 version of the human genome, whereas other BAM files

were aligned on the GRCh37 version. As our analysis is based on transcripts affected by CNVs and because reference genome was specific to cases or to controls, we expect that the different reference genomes do not affect our analysis. For each capture kit, targets covering the same exon and separated by less than 30 bp were merged. Then, for each sample, the number of reads covering each merged target was determined using BEDTools[27]. For each CNV-calling batch, we removed non-informative regions, namely regions where >90% of the samples showed less than 10 reads on the target. Finally, CANOES was run on each CNV-calling batch to generate CNV calls.

CNV and samples quality check

After CNV calling, we excluded all individuals with ≥50 calls, indicating an excess of variability in the sample's read distribution compared to other samples from the same CNV-calling batch and following CANOES user instructions [22]. Then, to account for potential biases due to clonal hematopoiesis associated with large mosaic, blood-specific age-related CNVs which[28], [29], we excluded carriers of CNVs that are large enough to be unlikely germline. As calling is performed from capture-based exomes and because of the presence of low complexity regions decreasing the calling accuracy, such large CNVs can be detected as multiple, smaller CNVs on a same chromosome. Thus, we computed the cumulative size of detected CNVs of the same type (deletion/duplication) for each sample and each chromosome, as well as the total chromosomal region covered from the first to last CNV per chromosome. For each cumulative size per chromosome greater than 2.5 Mb and encompassing more than 10 Mb for a given chromosome, we proceeded to a manual visualization (Figure S3) on the UCSC genome browser. Carriers of candidate large likely somatic CNVs were excluded from the analysis (Table S2).

Annotation of CNVs

For CNV annotation, we first confronted the frequency of CNV calls to two public databases: the Database of Genomic Variants (DGV), gold Standard section[30] and the non-neuro non-Finnish European section of the gnomAD database v2.1[31], [32], considering a mutual overlap of 70%. In addition, we removed CNV overlapping > 50% with segmental duplication regions, as extracted from the UCSC Table Browser (https://genome-euro.ucsc.edu/cgi-bin/hgTables). Second, in order to classify the potential effect of CNVs and to overcome the differences between capture kits that could impact

the proper classification of some CNVs and thus the results of statistical analyses, we harmonized the dataset by working at the transcript level (see supplementary information, Figure S2). Indeed, classic CNV annotation tools may not be applied to heterogeneous datasets, as the definition of a complete versus partial gene duplication or deletion usually rely on the actual definition of a given gene, based on genome coordinates. As some capture kits vary on the regions actually targeted for a given gene (e.g. capture of untranslated regions or not, capture of all coding exons from all known transcripts or not), this may lead to heterogeneous definitions of partial or complete deletions/duplications (see for a theoretical example Figure S2). Thus, we built our own annotation pipeline that relied on each individual's capture kit to define whether a given CNV partially or completely affects a given transcript. Thus, we redefined the transcripts positions according to the capture kit used (see supplementary information, Figure S2).

For all comparisons between capture kits, CNV calls and public databases, we used a combination of the BEDTools suite[27] and homemade scripts. All scripts and formats used are available at: https://github.com/U1245/ExtremelyRareCNVContributingToADRisk.

Identification of Mendelian pathogenic CNVs

Based on current literature, we considered, as pathogenic CNVs, partial deletions of *PSEN1* involving exclusively exon 9 (NM_000021.4) or both exons 9 and 10[12], [13], complete duplications of *APP*[11], *MAPT*[33], complete duplications or triplications of the SNCA gene[34], [35] as well as any coding deletion of the *GRN* gene[36]. Individuals carrying such a pathogenic CNV were then excluded from case-control analyses.

Case-control analyses

CNV Filtering

This study focused on rare CNVs for two main reasons. First, rare CNVs, similar to rare SNVs and indels, have higher chances to be associated with a moderate to high AD risk[5]. Second, CNV-calling from exome sequencing data based on read-depth comparison has been essentially validated for rare CNVs[23]. Indeed, CANOES (and other read-depth comparison tools adapted to sequencing data obtained following capture) is based on the comparison of read depth on a specific target to a matrix

computed from samples from a same batch. Thus, a very common variant (several dozens of percent in frequency) would likely be missed because of natural variability in each batch. Even if CANOES can detect low frequency variants, its performances have not been assessed precisely on such variants.

To focus on rare CNVs, we filtered out all CNVs showing at least a 70% overlap with a common CNV (frequency >1% in the above-mentioned public databases). Similarly, we filtered out CNVs with a ≥50% overlap with segmental duplication regions, as these regions are considered as highly variable across individuals and hardly callable. Because of the complexity of X-linked models in a case-control analysis, we focused here on autosomes.

Finally, we removed CNVs overlapping only one target to decrease the risk of false positive CNV calls. Indeed, in our validation study[23], the rate of false positive calls among CNVs overlapping one target reached 30% against 9.8% in CNVs called by ≥ 2 targets.

Transcripts filtering

To avoid the case-control analysis of rare CNVs affecting transcripts commonly deleted or duplicated, we built two sets of transcript sets, based on Refseq protein-coding transcripts (assessed, 26/10/2022): (i) a set of transcripts with cumulative frequency of deletions (partial or complete) in less than 1% individuals in public databases (set A, corresponding to haploinsufficient transcripts) and (ii) a set of transcripts with cumulative frequency of complete duplications in less than 1% in public databases (set B, corresponding to triplosensitive transcripts), based on the DGV gold Standard section[30] and the non-neuro non-Finnish European section of the gnomAD database v2.1 [31], [32].

Building a copy number matrix

For the case-control analysis, we built a copy number matrix per autosomal protein-coding transcript, in order to (i) harmonize heterogeneous data at the transcript level and (ii) determine the missingness and hence which transcripts were eligible for case-control analysis. For each sample and each transcript remaining after filtration, five possible states were determined: (i) missing information, i.e. the corresponding capture kit does not target the transcript or the region was not covered enough (at least 10 reads) for 90% of the samples and no CNV encompassing this transcript has been detected from targets on neighboring genes or (ii) no CNV overlaps the transcript [copy number = 2] or (iii) a

complete or (iv) a partial duplication overlaps this transcript or (v) a deletion overlaps this transcript, whether or not it is a complete or a partial duplication. In our five-state categorization, we worked under the assumption that complete deletions likely result in haploinsufficiency. To explore the dosage effect, states (ii), (iii) and (iv) were merged into a transcript copy number information (see supplementary methods for more details).

Statistical analyses

Given the previous knowledge obtained in gene-based rare variant analyses showing a larger effect among EOAD cases and given the expected extreme rarity of CNVs limiting power, we decided to (i) set the EOAD cases versus controls as the primary analysis, (ii) to perform a dosage analysis, (iii) at the transcript level. To perform an EOAD cases vs controls transcript-based association study on protein-coding genes using a dosage strategy as the primary endpoint, we worked on the fusion of set A and set B and used rare partial and complete deletions and complete duplications. In parallel, we show the results obtained among all AD vs controls. In addition, to better understand the signal identified in dosage analysis and to further assess the hypothesis that AD-associated genes may be affected by rare CNVs, we performed (i) transcript-based analyses in a deletions-only (full and partial deletions, on set A) and complete duplications-only (on set B) and (ii) a gene-based analysis of loss-of-function variants (deletion-CNVs plus truncating SNVs/indels, see supplementary methods) in a list of AD-associated genes in a GWAS study[4] and in our latest rare variants gene-based case-control study[5] and among the genes prioritized by the dosage analysis.

Dosage analyses were restricted to transcripts overlapping at least four CNVs (including at least one deletion and one complete duplication) in our dataset. For deletions (respectively duplications) analyses at transcript level, regression was performed for each transcript from set A (resp. B) overlapping at least four deletions (resp. complete duplications) in our dataset. In a sensitivity analysis, we adjusted for APOE4 or ancestry. See supplementary methods for more details.

For each statistical analysis, we performed firth logistic regression with status (EOAD vs controls or all AD vs CTRL) as dependent variable. The independent variable was either the dosage information (transcript copy number by individual), the presence/absence of a deletion (or duplication) as binary information or the presence/absence of a LOF variant in the gene (SNVs/indel or deletion-CNV), as a binary variable. More information about models is available in supplemental methods. At the

transcript and gene levels, all analyses were performed on a specific subset of individuals with available information relative to the transcript or the gene, i.e. individuals without missing information in the copy number matrix.

As many tests were performed twice or more because of the similarity of most of the transcripts-affecting CNVs, we gathered all transcripts-based tests of a given gene into one test if they all showed the same number of CNVs carriers leading thus to the same p-value, and counted separately the tests for transcripts with different numbers of CNVs carriers. We set a threshold at FDR=10% in the dosage analysis (EOAD-controls, primary outcome) to consider the signals as suggestive.

Finally, we performed a cumulate CNV analysis of all transcripts based on list of genes of interest. For that purpose, we compared carriers and non-carriers of CNVs encompassing genes associated with AD in GWAS[4], as well as genes having function in Aβ network[37]. To overcome the variability in sequencing techniques, we excluded from this analysis all transcripts with a significant differential missingness between disease status ($p<10^{-5}$ in a khi2 test).

Given the previous knowledge obtained in gene-based rare variant analyses showing a larger effect among EOAD cases and given the expected extreme rarity of CNVs limiting power, we decided to (i) set the EOAD cases versus controls as the primary analysis, (ii) to perform a dosage analysis, (iii) at the transcript level. To perform an EOAD cases vs controls transcript-based association study on protein-coding genes using a dosage strategy as the primary endpoint, we worked on the fusion of set A and set B and used rare partial and complete deletions and complete duplications. Dosage analyses were restricted to transcripts overlapping at least four CNVs (including at least one deletion and one complete duplication) in our dataset and to the subset of individuals with available information relative to the transcript or the gene, i.e. individuals without missing information in the copy number matrix. Because of the rarity of CNVs, association was tested via firth logistic regression with status (EOAD vs controls) as dependent variable and dosage information (transcript copy number by individual) as independent variable. Since many tests were performed twice or more because of the similarity of most of the transcripts-affecting CNVs, we gathered all transcripts-based tests of a given gene into one test if they all showed the same number of CNVs carriers leading thus to the same p-value, and counted separately the tests for transcripts with different numbers of CNVs carriers. We set a threshold at FDR=10% in this dosage EOAD analysis to consider signals as suggestive.

Several complementary analyses were implemented. For the sake of information, we show the results obtained similarly when comparing dosages between all AD and controls. Also, to better understand the signal identified in the dosage analysis and to further assess the hypothesis that AD-associated genes may be affected by rare CNVs, we performed (i) transcript-based analyses in a deletions-only (full and partial deletions, on set A) and complete duplications-only (on set B) and (ii) a gene-based analysis of loss-of-function variants (deletion-CNVs plus truncating SNVs/indels, see supplementary methods) in a list of AD-associated genes in a GWAS study[4] and in our latest rare variants gene-based case-control study[5] and among the genes prioritized by the dosage analysis. For deletions (respectively duplications) analyses at transcript level, regression was performed for each transcript from set A (resp. B) overlapping at least four deletions (resp. complete duplications) in our dataset. In a sensitivity analysis, we adjusted for APOE4 or ancestry. See supplementary methods for more details.

Finally, we performed a cumulate CNV analysis of all transcripts based on list of genes of interest. For that purpose, we compared EOAD versus controls in a model focusing on CNVs encompassing genes associated with AD in GWAS[4], as well as genes having function in Aβ network[37]. To overcome the variability in sequencing techniques, we excluded all transcripts with a significant differential missingness between disease status ($p<10^{-5}$ in a khi2 test) from this analysis. (Supplemental methods)

CNV confirmation

The main analysis provided a list of prioritized transcripts/genes. Each transcript with an FDR below 10% was carefully checked in the UCSC genome browser showing all filtered and unfiltered CNVs separately at each transcript of interest, in cases and controls, along with the study and the capture kit information (Figure S5 to S9). This allowed us to detect signals driven by genes overlapping with repeats or duplicated gene in the genome despite CNVs did not overlap the 50% threshold set for repeats in the filtration steps, or genes for which the signal dosage analysis was driven by deletions whereas the transcript was not in set A, thus not relevant. We then removed the candidate transcripts from the prioritized list as well as loci were a CNV was not orthogonally confirmed.

From each locus prioritized in the main analysis, we performed targeted validation of CNVs when DNA was available to us, using either ddPCR as previously described [30][38] or Quantitative Multiplex PCR of Short Fluorescent fragments (QMPSF)[39].

RESULTS

Detection of CNVs from exome sequencing data of 22,319 individuals

Dataset, CNV calling and QC

We initially included 22,319 exomes from previously described dataset[5] (4,150 EOAD, 8,519 LOAD and 9,650 controls). A detailed description is available in supplementary information, Table S1 and summarized in Table 2.

Following QC, we excluded 148 samples with excessive calls (≥ 50), 55 carriers of large likely somatic CNVs (26 controls, 29 LOAD cases, 1 EOAD case, see supplementary information, Table S2), and 6 samples with biased amplification due to a custom kit. There was no significant difference when comparing large likely somatic CNVs between all cases and controls. Moreover, likely somatic CNVs are linked to age at last visit (used as a proxy for age at blood sampling) and not to AD status (logistic regression; Age: OR=1.0001, SE=2.9E-5, p-value=1.14E-5; AD status: OR=1.0009, SE=7.39E-4, p-value=0.24). This suggests that such events are likely linked to clonal hematopoiesis and should thus be excluded from our germline CNV analysis study.

Among the 22,110 remaining individuals, we detected 261,190 CNVs. After excluding common CNVs and those overlapping with segmental duplication regions (see Supplementary Method), we first kept in our analyses a total of 61,053 rare CNVs including 24,262 duplications and 36,791 deletions (Table 2, Figure 1).

We provide as a supplement a catalog of genes with partial/full deletions and duplications and frequencies among EOAD cases, LOAD cases, and controls.

Detection of Mendelian pathogenic CNVs

We observed 10 carriers of a complete duplication of APP, all with EOAD (mean AAO = 49.6 years, SD = 5.4, range: 43 to 58, 3 were previously reported[40], [41]) and 6 carriers of a complete duplication of *MAPT*, thus leading to a diagnosis a primary tauopathy (mean AAO = 58.2 years, SD = 14.9, range: 45 to 87). In addition, one patient carried a partial in-frame deletion in *PSEN1* (AAO=56 years old) (already reported[13]). Of note, no carrier of a deletion of the *GRN* or any other AD differential gene was detected, with the exception of *MAPT* duplication carriers, some of them having been described previously[33], [42]. All pathogenic CNV carriers are reported in Table 1.

Case-control analyses

CNV analysis on the discovery exome sequencing dataset

After QC and exclusion of Mendelian CNV carriers, we kept 22,093 samples for case-control analyses, including 12,552 cases (4,077 EOAD, 8,457 LOAD) and 9,559 controls and a total of 45,964 rare CNVs overlapping ≥2 targets, outside of segmental duplication regions and affecting at least 1 coding transcript as described in RefSeq (23,134 deletions and 22,830 duplications).

To perform analyses based on transcript information, we focused on two sets of transcripts. Of 61,238 Refseq autosomic transcripts of 18,426 genes, set A contains 58,602 non-haplosufficient transcripts (related to 17,302 genes), of which 3305 (1043 genes) are affected by at least four deletions. Set B contains 60,220 non-triplosufficient transcripts (17,921 genes), of which 3392 (1252 genes) are affected by at least four complete duplications. The fusion of sets A and B contains 60,633 transcripts (18,141 genes), including 4997 transcripts (1717 genes) affected by at least four CNVs (including at least one deletion and one complete duplication).

CNV-dosage EOAD-control analysis

Dosage analysis prioritized 55 transcripts from 20 genes with FDR <10% (Table 3 displaying one transcript per gene when several transcripts exhibit the same p-values). Three genes were then excluded because they either overlapped with repeats or duplicated regions in the genome or the signal was driven by deletions and the transcript was not in set A (corresponding to non-haplosufficient transcripts). Of the 44 remaining transcripts from 18 genes, five independent loci were identified, as 13 genes mapped to the 22q11.2 region (FDR=0.0271 to 0.0386), and 2 other contiguous genes mapped to chromosome 19q13.32 (FDR=0.0271 to 0.0412).

The 22q11.2 region is known as a locus with rare recurrent deletions and duplications driven by non-allelic homologous recombination (NAHR) through low copy repeats (LCR) labelled LCR 22A, 22B, 22C, 22D and 22E (Figure S5). The region where all transcripts with FDR≤10% clustered is located between LCR 22B and 22D, which do not encompass the critical region for Di-George (velo-cardio-facial) 22q11.2 microdeletion syndrome (22A-22B including *TBX1*). Similar LCR 22B-22D deletions are often labelled central deletions and considered as a risk factor for developmental disorders, with incomplete penetrance and variable expression[43], [44]. We observed deletions in this region in 4 EOAD cases, and this region was also prioritized in the deletion-only analysis (Table S4 displaying the top transcripts in the deletion-only analysis with FDR<20%). Of note, while 3/4 EOAD cases exhibited a central deletion

only, the fourth EOAD case actually carried a larger LCR 22A-22D deletion, typical of DiGeorge syndrome. This case has no specific history of learning disabilities or neurodevelopmental disorders, but he was born with congenital heart disease, without any other known features of this syndrome. He presented first signs of cognitive impairment at the age of 43 and the diagnosis was confirmed with AD CSF biomarkers (Aβ42, Tau, and 181-P-Tau levels all in abnormal ranges). DNA analysis using dPCR in the patient confirmed the presence of the deletion. DNA from both unaffected parents was available and the deletion was absent, which, after parenthood confirmation using informative polymorphisms, allowed us to identify that the deletion occurred de novo in the proband (Figure S5). DNA of one of the other three EOAD cases, carrying a central deletion, was also available and we confirmed the presence of the deletion by ddPCR.

Interestingly, a smaller deletion was found in a LOAD case (LCR 22C-22D) and, strikingly, duplications mirroring the LCR 22A-22D or the 22B-22D deletions, or affecting the 22C-22D or 22C to 22E regions, were observed in 8 to 10 controls (depending on the CNV coordinates) and in 4 to 5 LOAD cases (Figure S5), although the duplication signal was not significant enough to reach the 20% FDR threshold in the duplication-only analysis (Table S5). Overall, this suggests a strong dosage effect with deletions possibly increasing AD risk and duplications possibly decreasing the risk or delaying ages at onset. Of note, DNA from 2 carriers of deletions was available and all CNVs that could be assessed were confirmed.

Overall, the region with highest level of association based on deletions in EOAD cases was the LCR 22B-22D region, as was the case for the dosage analysis, including the mirror duplications. If considering all rare CNVs from this locus, a smaller minimal region could be defined between LCR 22C and 22D but this would be based on a single LOAD case and a single control. Thus, we consider that the region of interest, to be replicated and refined, remains the LCR 22B-22D region.

Another locus was identified with two contiguous genes involved, namely *ERCC2* (p-value=1.09E-4, FDR=0.0271) and *KLC3* (p-value=4.28E-4, FDR=0.0412). Most of the duplications at this locus shared similar coordinates, also encompassing the *PPP1R13L* gene but transcripts of this gene were not considered in the analysis because they were not predicted to be entirely duplicated (Figure S6). This locus is located on chromosome 19q13.32, ~440-kb away from the APOE gene. Despite showing FDR<10% in the dosage analysis, the signal was clearly driven by an enrichment of complete duplications in cases with an EOAD>LOAD>controls pattern. We observed that the duplications were significantly more frequently carried by APOE4+ individuals, independently of the disease status (Table

S6), suggesting that rare duplications may have occurred on an APOE4-linked haplotype. Consistent with this hypothesis, the association signal did not remain significant after adjusting for APOE-ε4 dosage (1.58 [0.85 – 2.99] p=0.148) while such adjustment did not change the results on the other transcripts (Table S7). Of note, DNA from 12 carriers was available and duplications were confirmed in all of them.

In addition to these two multi-gene loci, three single-gene loci showed FDR<10%, namely *MBL2* (FDR=0.0271), *FADS6* (FDR=0.0271) and *ADI1* (FDR=0.0916). For *MBL2* and *FADS6*, the dosage signal was mainly driven by deletions. *MBL2* deletions appeared to be enriched in controls whereas *FADS6* deletions appeared to be enriched in EOAD cases (Supplemental information, Figures S7, S9). However, enrichment observed for *MBL2* did not remain after adjusting for ancestry (Supplemental results, Figure S). For ADI1, complete duplications appeared to be enriched in cases (Supplemental information, Figure S8). DNA from 12 carriers (1 *MBL2* carrier, 1 *ADI1* and 10 *FASD6*) was available and CNVs were confirmed in 11 of them.

Joint CNV and loss-of-function indels and SNVs analyses

Then, we hypothesized that the signal driven by deletions could be reinforced by LOF SNVs and indels or that such small variants could help refine genes of interest at the 22q11.2 locus. We thus performed a joint analysis of CNVs (deletions only) with LOF SNVs and indels among (i) the genes belonging to the prioritized loci in the dosage analysis and (ii) a list of AD-associated genes.

Among the candidate loci identified above, the signal seemed to be driven by duplications in cases for *ADI1* and the *ERCC2-KLC3* genes, and there was no LOF association overall in the joint analysis, which is consistent with the *ERCC2-KLC3* duplications found as enriched in APOE4+ individuals, thus not directly linked to these genes. No LOF SNV/indel was identified in the *FADS6* gene.

In the 22q11.2 central deletion locus with 13 genes involved, the joint LOF analysis trended to narrow the locus of interest to *SCARF2*, *MED15* and *KLHL22*. Although the added LOF SNV/indels did not significantly reinforce these genes, with a single splice site short delins in *SCARF2* in an EOAD case (NM_153334.4:c.1424+1_1424+2delinsTC) and 2 *MED15* LOF SNV/indel variants in 2 LOAD cases (one nonsense, one frameshift 1-bp deletion), and no control carrying LOF SNV/indels (as for deletions), but all the other genes appeared as less significant now with a few LOF variant carriers in controls and not significantly more in cases.

We then focused on known AD-associated genes. Among them, we identified a significant enrichment of LOF variants of ABCA1 (uncorrected p=0.0002) and confirmed the association of *ABCA7* LOF variants (p=0.0006), now adding deletion-CNVs to the known association of *ABCA7* LOF variants with AD[45], while the *CTSB* gene, a GWAS locus, also appeared as a candidate (p=0.0089). In *ABCA1*, we observed 3 carriers of 3 distinct deletions in *ABCA1*, impacting respectively exons 16 to 18, 32 to 34 and 47 to 50 (Figure S11). All were EOAD patients (AAO: 49, 51 and 55). The joint OR of *ABCA1* LOF variants was 5.77 [2.25-17.06], suggesting that LOF of *ABCA1* is a moderate EOAD risk factor, while burden tests gathering LOF SNV/indels with missense variants showed more modest odds ratios[5]. Overall, LOF variants in *ABCA1* remained extremely rare with only 13 EOAD cases (0.32%), 12 LOAD (0.14%) and 5 controls (0.05%) carrying such a variant and deletions represented 10% (3/30) of the LOF alleles of this gene. Deletions in *ABCA7* were observed in 4 EOAD cases (0.09%), 3 LOAD (0.04%) and 3 controls (0.03%). They represented 8.6% (10/115) of the LOF alleles of this gene, whereas a complete duplication of *ABCA7* was observed in one control (0.01%) and 2 LOAD cases (0.02%) (Figure S10). Deletions did not affect the order of magnitude of known AD or EOAD association of *ABCA7* with LOF variants in terms of OR, given the higher frequency of *ABCA7* LOF variants overall. DNA was available for all 3 *ABCA1* and 2 *ABCA7* deletion carriers and confirmed the existence of each deletion.

Interestingly, LOF variants of *MAF*, *PLEKHA1* and *EPDR1* genes (incl. one deletion for each gene) were only detected in controls, leading to the hypothesis of a protective effect of LOF of these genes, which cannot be confirmed here as such events are extremely rare (Table S8, Figure S13). *CTSB* and *APH1B* genes CNVs suggested a mirror effect with deletions of these genes observed in EOAD patients only whereas duplications were observed in controls only; although these genes did not appear in the top 10% FDR in the primary analysis. Extending the analysis of these genes to LOAD, we observed 4 LOAD patients carrying a deletion of *CTSB* as well as 1 LOAD patients (AAO=90) carrying a duplication (Figure S12). Of note, there were no carrier of CNV encompassing *APH1B* gene among LOAD.

Cumulate CNV analysis of all transcripts

Considering CNVs overall in all transcripts from sets A and B, EOAD patients were more likely to carry at least one deleted and one duplicated gene than controls (deletion: OR=1.15 [1.06;1.23], p-value=0.0003; duplication: OR=1.10[1.02;1.19], p-value=0.0125). Restricting the analysis to the GWAS list of genes, difference between EOAD and controls increased for deletions (OR=2.67 [1.51;4.74], p-value=0.0008) whereas the difference trended to be reversed for duplications (OR=0.58 [0.30;1.04], p-

value=0.0690) suggesting that a deletion in the GWAS list of genes might be a risk factor for EOAD, that full duplications of some of these genes could be protective (Table S13), and that duplications in other genes remain to be identified. Finally, analysis of the Aβ network list of genes suggested that having a deletion encompassing genes from this list might increase the risk for EOAD (OR=1.44 [1.09;1.90], p=0.0114), and more particularly genes involved in the processing and trafficking of APP and genes involved in calcium homeostasis (OR=1.64 [1.04;2.57], p=0.0336 and OR=16.42 [1.59;2208.35], p=0.0161, respectively).

DISCUSSION

In this study, we managed to detect and analyze jointly CNVs from a heterogeneous exome dataset of 22,319 individuals. We propose a quality control (QC) and analysis strategy based first on an extensive individual and sample QC from previous work[5] and, second, on CNV-specific features. Our QC allowed us to obtain a dataset harmonized at the transcript level that allowed (i) the identification of high-quality CNV calls in known AD genes (Mendelian and risk factor genes) and (ii) identification of candidate loci, despite the extreme rarity of the individual CNVs under study.

Rare CNVs are a known mechanism for Mendelian disorder since decades in the context of multiple diverse human disorders (e.g. [46]–[48]) and this includes neurodegenerative disorders. Duplications or triplications of the *SNCA* gene are responsible from autosomal dominant Parkinson disease or Dementia with Lewy Bodies[34], for example, and *APP* duplications or triplications, from Alzheimer disease with or without cerebral amyloid angiopathy [11], [49]. However, in most Mendelian neurodegenerative disorders, CNVs represent an extremely rare mechanism. For example, only one *APP* triplication has been reported so far [49] while *APP* duplications represent 9% of all solved autosomal dominant EOAD families in our series (update data from [50]). In addition to 10 *APP* duplications (7 novel), we identified 6 *MAPT* duplications in this dataset (2 novel), confirming that certain presentations of this heterogeneous primary tauopathy can mimic AD clinically and represents a differential diagnosis [51]. We did not detect any CNV in other dementia genes causing AD-like phenotypes, such as GRN, where we know that deletions also represent a low proportion of pathogenic alleles[52].

In line with Mendelian dementia genes, we show here that CNVs are also an extremely rare mechanism leading to a loss of function (LOF) of AD risk genes *ABCA1* and *ABCA7*. In *ABCA7*, where the burden of

LOF SNV/indels is known to be associated with AD, joint analysis of CNVs and LOF SNV/indels did not modify the odds ratios much, remaining in the order of magnitude of 2 to 3 [5], [53], because (i) LOF alleles are not so rare (0.37% in controls; 0.86% in EOAD cases) and (ii) CNVs represented only 8.6% of all LOF alleles in the joint analysis. In *ABCA1*, where LOF variants are much rarer, we identified three distinct deletions, adding up to 27 LOF SNVs/indels. Deletions thus represent 10% of all LOF alleles. This analysis on *ABCA1* now allows us to consider LOF of *ABCA1* as a stronger risk factor for EOAD, as compared to the known average odds ratios obtained by gathering LOF variants with missense, predicted damaging variants (OR=2.2 [1.6;2.9] for LOF SNV/indels and missense variants with REVEL score >0.75 in Holstege et al.,[5] and OR=5.77 [2.25;17.06] here in the joint LOF SNV/indels/CNV analysis). These results suggest that (i) some CNVs represent an extremely rare mechanism increasing the risk of AD but also that (ii) at the individual level, given the effect on AD risk of such LOF alleles, CNVs should not be ignored. Although risk variants are not used for genetic counseling, they may be used for the future of AD prevention in the context of precision medicine, along with other factors.

In our analysis, we identified several novel candidate loci. Of note, one of these novel suggestive associations pointed to the *APOE* locus, which was not expected from a rare CNV analysis. However, this unexpected result, suggesting that rare duplications occurred on an APOE4-associated haplotype, further strengthen the validity of our global analysis. It is also a rare example of a rare recurrent CNV in linkage disequilibrium with a common GWAS locus, highlighting that this type of event can actually happen and should be considered in genome-wide analyses.

In the analysis at the protein-coding genome level, we identified an enrichment of deletions and of duplications in cases. After focusing on the GWAS list of genes, the signal based on duplications trended to be reversed. This suggests that additional genes, not identified here, may play a role in AD determinism, and this highlights a limitation of analyzing CNVs at the protein-coding genome or gene-list level, as increasing expression or decreasing expression of genes may show opposite effects on AD pathophysiology. For example, *APP*, *ABCA7* and *ABCA1* are all AD GWAS hits, but gathering their effect into a same analysis (if *APP* duplications would not be excluded prior to case-control studies), would lead to non-interpretable results, as duplications of APP are expected to be enriched in cases but, conversely, deletions of *ABCA1* or *ABCA7* are expected to be enriched in cases. Similarly, we investigated genes related to the Aβ network. Although deletions suggestively contribute to AD, signal

is probably confused by opposite effects of an expression decrease of these genes on AD. Overall, our analysis suggest that additional genes remain to be identified despite power limitations when working with rare CNVs.

Among the loci of interest, one of the strongest signals was the 22q11.2 locus. This unexpected finding remains to be replicated. This signal was suggestive in the dosage analysis, with deletions in cases and duplications in controls, and remained suggestive in the deletions-only analysis. This suggests that some genes may influence AD risk or age at onset. Of note, carriers of the so-called central 22q11.2 deletion have a higher risk of neurodevelopmental disorders [43], [44], as do the carriers of the full 22q11.2 deletion with higher penetrance (DiGeorge syndrome). Thus, one could hypothesize that carriers may have a lower cognitive reserve, even if not reaching the threshold for a formal diagnosis of a neurodevelopmental disorder, but still sufficient to decrease the age at onset in the carriers[54]–[56]. Another hypothesis is that there would be genes influencing the pathophysiology of AD in this region. There is no gene obviously linked to AD mechanisms here, but we can still hypothesize that genes like *MED15* or *SCARF2* may be related somehow to AD pathophysiology. *SCARF2* encodes a scavenger receptor (class F scavenger receptor family) expressed in the brain and in macrophages. Although a scavenger receptor role for *SCARF2* is yet to be demonstrated, contrary to other members of this family including SCARF1, another member of this family, *MEGF10*, has been shown to be a receptor for Aβ in the brain [57], [58]. *MED15* is also expressed in the brain and belongs to a family of proteins related to other genes' expression, including inflammation and TGFβ and SMAD2/3 signal transduction, both involved in AD pathophysiology [59].

Here, we used a large existing dataset of exome sequencing data. Whole genome sequencing (WGS) with short reads may also be used to replicate such findings, although at a high computational cost. Such an analysis has been recently run on WGS of 6,646 AD cases (average age at onset: 74.6 years) and 6,938 controls from the ADSP consortium [60], [61]. A moderate but significant burden of (coding and non-coding) of CNVs was associated with AD status overall and appeared higher for rarest events (singletons). In addition, an aggregated analysis of structural variants affecting a list of known AD genes also unveiled an association with AD status. Interestingly, these structural variants included one coding partial deletion of *SORL1* and 5 coding partial deletions of *ABCA7* (all in cases), among a few other genes of interest. None of the CNVs was significantly associated with AD due to power limitations.

Interestingly, non-coding structural variants, which are not detectable by exome sequencing, were also analyzed, prioritizing for example an intronic deletion in the *ADD3* gene (11 cases and no control) and another intronic deletion in the *ITPR2* gene (33 cases, 7 controls), suggesting that increasing sample sizes with WGS data should unveil novel associations.

Novel long read sequencing technologies may also offer new opportunities. In addition to deletions and duplications, such novel technologies will enable the analysis of mobile element insertions, repeat variations, and balanced structural variants with an unprecedented accuracy [62], [63]. However, sequencing costs remain very high, limiting the expectations that we can have from such a promising technology given the rarity of the events with a putatively strong effect.

In conclusion, we conducted an exome-wide CNV screen from sequencing data and identified ultra-rare CNVs that contribute to AD risk. Beyond autosomal dominant known CNVs, we highlight rare deletions in *ABCA1* and *ABCA7* as participating to AD risk in carriers and we identify candidate loci that will require replication in larger datasets, such as *CTSB* and the 22q11.2 central region. While additional CNVs may contribute to AD risk, as suggested by the gene-list and protein-coding genome-wide analyses, CNVs remain extremely rare events. We used a transcript point of view to reduce the recurrence issue at the CNV level and to reduce the heterogeneity among datasets, but the rarity of such events remains an issue even in datasets as large as the one used here. Finally, we provide here a catalog of genes with rare CNVs in EOAD, LOAD and controls, which can be reused for meta-analyses and combination with other datasets.

ACKNOWLEDGEMENTS

CS is supported by the Rouen metropole through the “post-doctorants métropole” fundings.

This research was conducted using the funding obtained by the following study cohorts: ADES-FR, AgeCoDe-UKBonn; Barcelona SPIN; AC-EMC; ERF and Rotterdam; ADC-Amsterdam; 100-plus study; EMIF-90+; Control Brain Consortium; PERADES; UCL-DRC EOAD; ADSP. Full consortium acknowledgements and funding sources are listed in the supplementary information document

We thank the ADSP study for providing exome data from cases and controls (ADSP umbrella NG00067.v3). ADSP data were prepared, archived, and distributed by the National Institute on Aging Alzheimer's Disease Data Storage Site (NIAGADS) at the University of Pennsylvania (U24AG041689), funded by the National Institute on Aging. The ADSP umbrella contains data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at:

http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

REFERENCES

[1] C. Schramm, D. Wallon, G. Nicolas, and C. Charbonnier, 'What contribution can genetics make to predict the risk of Alzheimer's disease?', Revue Neurologique, vol. 178, no. 5, pp. 414–421, May 2022, doi: 10.1016/j.neurol.2022.03.005.

[2] M. Gatz et al., 'Role of genes and environments for explaining Alzheimer disease', Arch Gen Psychiatry, vol. 63, no. 2, Art. no. 2, Feb. 2006, doi: 10.1001/archpsyc.63.2.168.

[3] E. Genin et al., 'APOE and Alzheimer disease: a major gene with semi-dominant inheritance', Mol Psychiatry, vol. 16, no. 9, Art. no. 9, Sep. 2011, doi: 10.1038/mp.2011.52.

[4] C. Bellenguez et al., 'New insights into the genetic etiology of Alzheimer's disease and related dementias', Nat Genet, vol. 54, no. 4, Art. no. 4, Apr. 2022, doi: 10.1038/s41588-022-01024-z.

[5] H. Holstege et al., 'Exome sequencing identifies rare damaging variants in ATP8B4 and ABCA1 as risk factors for Alzheimer's disease', Nat Genet, vol. 54, no. 12, Art. no. 12, Dec. 2022, doi: 10.1038/s41588-022-01208-7.

[6] C. Schramm et al., 'Penetrance estimation of Alzheimer disease in SORL1 loss-of-function variant carriers using a family-based strategy and stratification by APOE genotypes', Genome Med, vol. 14, no. 1, Art. no. 1, Jun. 2022, doi: 10.1186/s13073-022-01070-6.

[7] R. Sims et al., 'Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease', Nat Genet, vol. 49, no. 9, Art. no. 9, Sep. 2017, doi: 10.1038/ng.3916.

[8] G. Nicolas et al., 'SORL1 rare variants: a major risk factor for familial early-onset Alzheimer's disease', Mol Psychiatry, vol. 21, no. 6, Art. no. 6, Jun. 2016, doi: 10.1038/mp.2015.121.

[9] S. Steinberg et al., 'Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease', Nat Genet, vol. 47, no. 5, Art. no. 5, May 2015, doi: 10.1038/ng.3246.

[10] T. Jonsson et al., 'Variant of TREM2 associated with the risk of Alzheimer's disease', N Engl J Med, vol. 368, no. 2, Art. no. 2, Jan. 2013, doi: 10.1056/NEJMoa1211103.

[11] A. Rovelet-Lecrux et al., 'APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy', Nat Genet, vol. 38, no. 1, Art. no. 1, Jan. 2006, doi: 10.1038/ng1718.

[12] H. Karlstrom et al., ‘Variable phenotype of Alzheimer’s disease with spastic paraparesis’, J Neurochem, vol. 104, no. 3, Art. no. 3, Feb. 2008, doi: 10.1111/j.1471-4159.2007.05038.x.

[13] K. Le Guennec et al., ‘Deletion of exons 9 and 10 of the Presenilin 1 gene in a patient with Early-onset Alzheimer Disease generates longer amyloid seeds’, Neurobiol Dis, vol. 104, pp. 97–103, Aug. 2017, doi: 10.1016/j.nbd.2017.04.020.

[14] H. Wang, L.-S. Wang, G. Schellenberg, and W.-P. Lee, ‘The role of structural variations in Alzheimer’s disease and other neurodegenerative diseases’, Front Aging Neurosci, vol. 14, p. 1073905, 2022, doi: 10.3389/fnagi.2022.1073905.

[15] B. V. Hooli et al., ‘Rare autosomal copy number variations in early-onset familial Alzheimer’s disease’, Mol Psychiatry, vol. 19, no. 6, pp. 676–681, Jun. 2014, doi: 10.1038/mp.2013.77.

[16] A. Rovelet-Lecrux et al., ‘A genome-wide study reveals rare CNVs exclusive to extreme phenotypes of Alzheimer disease’, Eur J Hum Genet, vol. 20, no. 6, pp. 613–617, Jun. 2012, doi: 10.1038/ejhg.2011.225.

[17] A. Rovelet-Lecrux et al., ‘A genome-wide study reveals rare CNVs exclusive to extreme phenotypes of Alzheimer disease’, Eur. J. Hum. Genet., vol. 20, no. 6, pp. 613–617, Jun. 2012, doi: 10.1038/ejhg.2011.225.

[18] M. Zhao, Q. Wang, Q. Wang, P. Jia, and Z. Zhao, ‘Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives’, BMC Bioinformatics, vol. 14 Suppl 11, no. Suppl 11, Art. no. Suppl 11, 2013, doi: 10.1186/1471-2105-14-S11-S1.

[19] S. M. Teo, Y. Pawitan, C. S. Ku, K. S. Chia, and A. Salim, ‘Statistical challenges associated with detecting copy number variations with next-generation sequencing’, Bioinformatics, vol. 28, no. 21, Art. no. 21, Nov. 2012, doi: 10.1093/bioinformatics/bts535.

[20] M. Gabrielaite et al., ‘A Comparison of Tools for Copy-Number Variation Detection in Germline Whole Exome and Whole Genome Sequencing Data’, Cancers (Basel), vol. 13, no. 24, Art. no. 24, Dec. 2021, doi: 10.3390/cancers13246283.

[21] J. Y. Hehir-Kwa, R. Pfundt, and J. A. Veltman, ‘Exome sequencing and whole genome sequencing for the detection of copy number variation’, Expert Rev. Mol. Diagn., vol. 15, no. 8, Art. no. 8, 2015, doi: 10.1586/14737159.2015.1053467.

[22] D. Backenroth et al., ‘CANOES: detecting rare copy number variants from whole exome sequencing data’, Nucleic Acids Res., vol. 42, no. 12, p. e97, Jul. 2014, doi: 10.1093/nar/gku345.

[23] O. Quenez et al., ‘Detection of copy-number variations from NGS data using read depth information: a diagnostic performance evaluation’, Eur J Hum Genet, vol. 29, no. 1, Art. no. 1, Jan. 2021, doi: 10.1038/s41431-020-0672-2.

[24] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, ‘Clinical diagnosis of Alzheimer’s disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer’s Disease’, Neurology, vol. 34, no. 7, Art. no. 7, Jul. 1984, doi: 10.1212/wnl.34.7.939.

[25] G. M. McKhann et al., ‘The diagnosis of dementia due to Alzheimer’s disease: recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease’, Alzheimers Dement, vol. 7, no. 3, Art. no. 3, May 2011, doi: 10.1016/j.jalz.2011.03.005.

[26] Y. Benjamini and T. P. Speed, ‘Summarizing and correcting the GC content bias in high-throughput sequencing’, Nucleic Acids Res., vol. 40, no. 10, Art. no. 10, May 2012, doi: 10.1093/nar/gks001.

[27] A. R. Quinlan and I. M. Hall, ‘BEDTools: a flexible suite of utilities for comparing genomic features’, Bioinformatics, vol. 26, no. 6, Art. no. 6, Mar. 2010, doi: 10.1093/bioinformatics/btq033.

[28] M. J. Machiela et al., ‘Mosaic chromosome 20q deletions are more frequent in the aging population’, Blood Advances, vol. 1, no. 6, Art. no. 6, Feb. 2017, doi: 10.1182/bloodadvances.2016003129.

[29] H. Bouzid et al., ‘Clonal hematopoiesis is associated with protection from Alzheimer’s disease’, Nat Med, Jun. 2023, doi: 10.1038/s41591-023-02397-2.

[30] J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer, ‘The Database of Genomic Variants: a curated collection of structural variation in the human genome’, Nucleic Acids Res, vol. 42, no. Database issue, pp. D986-992, Jan. 2014, doi: 10.1093/nar/gkt958.

[31] S. Chen et al., ‘A genome-wide mutational constraint map quantified from variation in 76,156 human genomes’, Genetics, preprint, Mar. 2022. doi: 10.1101/2022.03.20.485034.

[32] K. J. Karczewski et al., ‘Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes’, Genomics, preprint, Jan. 2019. doi: 10.1101/531210.

[33] K. Le Guennec et al., ‘17q21.31 duplication causes prominent tau-related dementia with increased MAPT expression’, Molecular Psychiatry, vol. 22, no. 8, Art. no. 8, Aug. 2017, doi: 10.1038/mp.2016.226.

[34] A. B. Singleton et al., ‘alpha-Synuclein locus triplication causes Parkinson’s disease’, Science, vol. 302, no. 5646, p. 841, Oct. 2003, doi: 10.1126/science.1090278.

[35] D. W. Miller et al., ‘Alpha-synuclein in blood and brain from familial Parkinson disease with SNCA locus triplication’, Neurology, vol. 62, no. 10, pp. 1835–1838, May 2004, doi: 10.1212/01.wnl.0000127517.33208.f4.

[36] A. Rovelet-Lecrux et al., ‘Deletion of the progranulin gene in patients with frontotemporal lobar degeneration or Parkinson disease’, Neurobiol Dis, vol. 31, no. 1, pp. 41–45, Jul. 2008, doi: 10.1016/j.nbd.2008.03.004.

[37] D. Campion, C. Pottier, G. Nicolas, K. Le Guennec, and A. Rovelet-Lecrux, ‘Alzheimer disease: modeling an Aβ-centered biological network’, Mol. Psychiatry, vol. 21, no. 7, pp. 861–871, 2016, doi: 10.1038/mp.2016.38.

[38] K. Cassinari et al., ‘A Simple, Universal, and Cost-Efficient Digital PCR Method for the Targeted Analysis of Copy Number Variations’, Clin. Chem., Jul. 2019, doi: 10.1373/clinchem.2019.304246.

[39] F. Charbonnier et al., ‘Detection of exon deletions and duplications of the mismatch repair genes in hereditary nonpolyposis colorectal cancer families using multiplex polymerase chain reaction of short fluorescent fragments’, Cancer Res., vol. 60, no. 11, pp. 2760–2763, Jun. 2000.

[40] A. Rovelet-Lecrux et al., ‘De novo deleterious genetic variations target a biological network centered on Aβ peptide in early-onset Alzheimer disease’, Mol Psychiatry, vol. 20, no. 9, pp. 1046–1056, Sep. 2015, doi: 10.1038/mp.2015.100.

[41] H.-M. Lanoiselée et al., ‘APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: A genetic screening study of familial and sporadic cases’, PLoS Med, vol. 14, no. 3, p. e1002270, Mar. 2017, doi: 10.1371/journal.pmed.1002270.

[42] D. Wallon et al., ‘Clinical and neuropathological diversity of tauopathy in MAPT duplication carriers’, Acta Neuropathol, vol. 142, no. 2, pp. 259–278, Aug. 2021, doi: 10.1007/s00401-021-02320-4.

[43] P. Rump et al., ‘Central 22q11.2 deletions’, Am J Med Genet A, vol. 164A, no. 11, pp. 2707–2723, Nov. 2014, doi: 10.1002/ajmg.a.36711.

[44] K. J. Woodward et al., ‘Atypical nested 22q11.2 duplications between LCR22B and LCR22D are associated with neurodevelopmental phenotypes including autism spectrum disorder with incomplete penetrance’, Mol Genet Genomic Med, vol. 7, no. 2, p. e00507, Feb. 2019, doi: 10.1002/mgg3.507.

[45] S. Steinberg et al., ‘Loss-of-function variants in ABCA7 confer risk of Alzheimer’s disease’, Nat Genet, vol. 47, no. 5, pp. 445–447, May 2015, doi: 10.1038/ng.3246.

[46] J. R. Lupski et al., ‘DNA duplication associated with Charcot-Marie-Tooth disease type 1A’, Cell, vol. 66, no. 2, pp. 219–232, Jul. 1991, doi: 10.1016/0092-8674(91)90613-4.

[47] J. Y. Hehir-Kwa, R. Pfundt, and J. A. Veltman, ‘Exome sequencing and whole genome sequencing for the detection of copy number variation’, Expert Rev Mol Diagn, vol. 15, no. 8, pp. 1023–1032, 2015, doi: 10.1586/14737159.2015.1053467.

[48] Y. Shen et al., ‘Clinical genetic testing for patients with autism spectrum disorders’, Pediatrics, vol. 125, no. 4, pp. e727-735, Apr. 2010, doi: 10.1542/peds.2009-1684.

[49] L. Grangeon et al., ‘Early-Onset Cerebral Amyloid Angiopathy and Alzheimer Disease Related to an APP Locus Triplication’, Neurol Genet, vol. 7, no. 5, p. e609, Oct. 2021, doi: 10.1212/NXG.0000000000000609.

[50] H.-M. Lanoiselée et al., ‘APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: A genetic screening study of familial and sporadic cases’, PLoS Med, vol. 14, no. 3, p. e1002270, Mar. 2017, doi: 10.1371/journal.pmed.1002270.

[51] D. Wallon et al., ‘Clinical and neuropathological diversity of tauopathy in MAPT duplication carriers’, Acta Neuropathol, vol. 142, no. 2, pp. 259–278, Aug. 2021, doi: 10.1007/s00401-021-02320-4.

[52] A. Rovelet-Lecrux et al., ‘Deletion of the progranulin gene in patients with frontotemporal lobar degeneration or Parkinson disease’, Neurobiol. Dis., vol. 31, no. 1, pp. 41–45, Jul. 2008, doi: 10.1016/j.nbd.2008.03.004.

[53] K. Le Guennec et al., ‘ABCA7 rare variants and Alzheimer disease risk’, Neurology, vol. 86, no. 23, pp. 2134–2137, Jun. 2016, doi: 10.1212/WNL.0000000000002627.

[54] Y. Stern, ‘Cognitive reserve in ageing and Alzheimer’s disease’, Lancet Neurol, vol. 11, no. 11, pp. 1006–1012, Nov. 2012, doi: 10.1016/S1474-4422(12)70191-6.

[55] A. Soldan et al., ‘Cognitive reserve and long-term change in cognition in aging and preclinical Alzheimer’s disease’, Neurobiol Aging, vol. 60, pp. 164–172, Dec. 2017, doi: 10.1016/j.neurobiolaging.2017.09.002.

[56] M. E. Nelson, D. J. Jester, A. J. Petkus, and R. Andel, ‘Cognitive Reserve, Alzheimer’s Neuropathology, and Risk of Dementia: A Systematic Review and Meta-Analysis’, Neuropsychol Rev, vol. 31, no. 2, pp. 233–250, Jun. 2021, doi: 10.1007/s11065-021-09478-4.

[57] K. Wilkinson and J. El Khoury, ‘Microglial scavenger receptors and their roles in the pathogenesis of Alzheimer’s disease’, Int J Alzheimers Dis, vol. 2012, p. 489456, 2012, doi: 10.1155/2012/489456.

[58] T. D. Singh et al., ‘MEGF10 functions as a receptor for the uptake of amyloid-β’, FEBS Lett, vol. 584, no. 18, pp. 3936–3942, Sep. 2010, doi: 10.1016/j.febslet.2010.08.050.

[59] R. von Bernhardi, F. Cornejo, G. E. Parada, and J. Eugenín, ‘Role of TGFβ signaling in the pathogenesis of Alzheimer’s disease’, Front Cell Neurosci, vol. 9, p. 426, 2015, doi: 10.3389/fncel.2015.00426.

[60] W.-P. Lee et al., ‘Copy Number Variation Identification on 3,800 Alzheimer’s Disease Whole Genome Sequencing Data from the Alzheimer’s Disease Sequencing Project’, Front Genet, vol. 12, p. 752390, 2021, doi: 10.3389/fgene.2021.752390.

[61] H. Wang et al., ‘Structural Variation Detection and Association Analysis of Whole-Genome-Sequence Data from 16,905 Alzheimer’s Diseases Sequencing Project Subjects’, medRxiv, p. 2023.09.13.23295505, Sep. 2023, doi: 10.1101/2023.09.13.23295505.

[62] G. A. Logsdon, M. R. Vollger, and E. E. Eichler, ‘Long-read human genome sequencing and its applications’, Nat Rev Genet, vol. 21, no. 10, pp. 597–614, Oct. 2020, doi: 10.1038/s41576-020-0236-x.

[63] D. E. Miller et al., ‘Targeted long-read sequencing identifies missing disease-causing variation’, Am J Hum Genet, vol. 108, no. 8, pp. 1436–1449, Aug. 2021, doi: 10.1016/j.ajhg.2021.06.006.

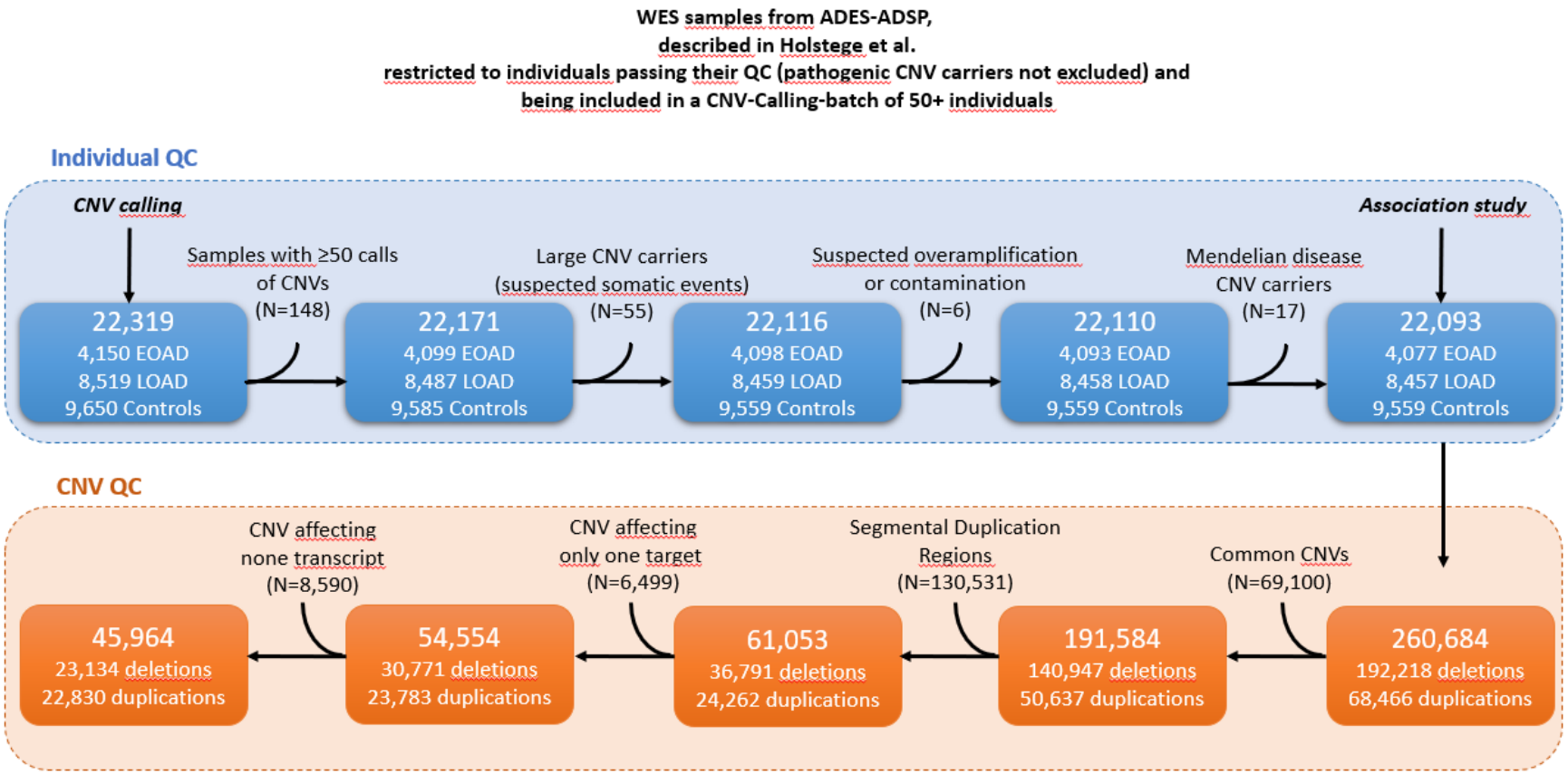


Figure 1: Quality Check (QC) pipeline

Sample	Cohort	AAO	Gender	CNV coordinate	HGVS nomenclature	Gene	Event
ALZ-0596-001[1]	ADES-FR	50	Male	chr17:43679169-44145099	NC_000017.10:g.43585921_44159772dup	MAPT	complete duplication
ALZ-0441-005[1]	ADES-FR	45	Male	chr17:43717946-44364336	NC_000017.10:g.43679455_44363695dup	MAPT	complete duplication
EXT-1114-001[1]	ADES-FR	54	Female	chr17:43679164-44364341	NC_000017.10:g.43585926_44408484dup	MAPT	complete duplication
ROU-1373-001[1]	ADES-FR	54	Male	chr17:43717946-44159945	NC_000017.10:g.43679455_44172015dup	MAPT	complete duplication
SC45A00007a	**PERADES**	**59**	**Female**	**chr17:43679204-44159910**	**NC_000017.10:g.43611648_44171922dup**	**MAPT**	**complete duplication**
C-FHS-51259	**ADSP**	**87**	**Female**	**chr17:43898720-44101538**	**NC_000017.10:g.43893949_44108840dup**	**MAPT**	**complete duplication**
EXT-0814-001[3]	ADES-FR	54	Female	chr21:22370838-28338704	NC_000021.8:g.20058403_30248689dup	APP	complete duplication
EXT-0773-001[3,4]	ADES-FR	44	Male	chr21:26946242-31066399	NC_000021.8:g.22910384_31233900dup	APP	complete duplication
EXT-0857-001[3]	ADES-FR	56	Male	chr21:26946246-28338704	NC_000021.8:g.22910379_30248689dup	APP	complete duplication
FR16640974	**Netherlands Brain Bank**	**47**	**Male**	**chr21:22370838-28338827**	**NC_000021.8:g.20230358_30248689dup**	**APP**	**complete duplication**
JW_Cardiff_TE17A00020a_DS1	**PERADES**	**45**	**Male**	**chr21:27056123-28338719**	**NC_000021.8:g.27012218_30248703dup**	**APP**	**complete duplication**
20732	**PERADES**	**53**	**Female**	**chr21:27062175-27484514**	**NC_000021.8:g.27056277_27512460dup**	**APP**	**complete duplication**
CH32A00027a	**PERADES**	**54**	**Female**	**chr21:27066065-28338719**	**NC_000021.8:g.27062287-30248703dup**	**APP**	**complete duplication**
09D04764	**ADC-Amsterdam**	**50**	**Male**	**chr21:26541838-29645932**	**NC_000021.8:g25725527_30248655dup**	**APP**	**complete duplication**
10D05060	**ADC-Amsterdam**	**43**	**Female**	**chr21:27253167-27542996**	**NC_000021.8:g.27142399_27840733dup**	**APP**	**complete duplication**
SID14617	**AgeCoDe-UKBonn**	**46**	**Female**	**chr21:26946221-27542992**	**NC_000021.8:g.25725527_27840779dup**	**APP**	**complete duplication**
EXT-0313-001[2]	ADES-FR	56	Male	chr14:73672981-73678665	NC_000014.8:g.73664857_73683810del	PSEN1	deletion exon 9-10

Table 1. Mendelien pathogenic CNVs

AAO: Age at onset, CNV: copy number variant, HGVS: Human Genome Variation Society; **in bold: novel**

[1] reported in Le Guennec et al., Mol Psychiatry, 2017

[2] reported in Le Guennec et al., Neurobiol Dis, 2017

[3] reported in Lanoiselée et al., PLoS Med, 2017

[4] reported in Rovelet-Lecrux et al., Mol Psychiatry, 2015

	EOAD	LOAD	All AD	Controls
N	4,077	8,457	12,534	9,559
Age y mean (SD)[1]	58.43 (5.09)	77.90 (7.48)	71.38 (11.42)	81.36 (12.28)
Gender N (%)				
males	1814 (44.5%)	3307 (39.1%)	5121 (40.9%)	4147 (43.4%)
females	2263 (55.5%)	5150 (60.9%)	7413 (59.1%)	5412 (56.6%)
***APOE* N (%)[2]**				
E4-non carriers	1704 (41.9%)	4640 (54.9%)	6344 (50.7%)	7535 (79.6%)
E4-heterozygous	1699 (41.7%)	3417 (40.5%)	5116 (40.9%)	1830 (19.3%)
E4-homozygous	668 (16.4%)	390 (4.6%)	1058 (8.4%)	101 (1.1%)
N. deletions mean ± SD (range)	1.18 ±1.75 (0 - 35)	1.00 ±1.38 (0 - 35)	1.06 ±1.51 (0 - 35)	1.03 ±1.41 (0 - 32)
Deletion size in bp mean (range)	36,788 (187 - 2,860,014)	37,750 (129 - 7,103,184)	37,401 (129 - 7,103,184)	35,299 (146 - 8,273,974)
N. duplications mean ± SD (range)	1.09 ±1.25 (0 - 23)	1.02 ±1.37 (0 - 36)	1.04 ±1.33 (0 - 36)	1.02 ±1.37 (0 - 33)
Duplication size in bp mean (range)	91,540 (129 - 4,847,347)	89,421 (75 - 4,809,978)	90,141 (75 - 4,847,347)	89,974 (75 - 4,840,909)

Table 2. Description of individuals included in case-control analysis

N: number, y: years, SD: standard deviation, bp: base pairs, EOAD: early onset Alzheimer disease, LOAD: late onset Alzheimer disease

[1]AAO for cases and age at last exam/inclusion for controls. Age is missing for 373 LOAD and 256 controls

[2]APOE genotype is missing for 6 EOAD, 10 LOAD and 93 controls

		N total			N carriers of CNVs (DEL/DUP)			Dosage analysis EOAD vs CTRL			Deletion analysis EOAD vs CTRL		Duplication analysis EOAD vs CTRL		
Chr	Gene	EOAD	LOAD	CTRL	EOAD	LOAD	CTRL	OR [95%CI][1]	p-value[2]	FDR	OR [95%CI]	p-value	OR [95%CI]	p-value	Transcripts
1	PRAMEF26*	2911	7302	6461	8 (8 / 0)	2 (2 / 0)	2 (1 / 1)	0,08 [0,01 ; 0,34]	2,88E-04	3,29E-02	12,61 [2,84;118,46]	3,96E-04	0,74 [0,005;13,87]	8,50E-01	NM_001306072.3
2	ADI1	4077	8457	9559	5 (0 / 5)	1 (0 / 1)	1 (1 /0)	26,58 [3,06 ; 3482,37]	1,05E-03	9,16E-02	0,78 [0,005 ; 14,65]	8,78E-01	25,82 [2,93 ; 3389,88]	1,40E-03	NM_001306077.2 ; NM_018269.4
8	MTUS1**	4077	8457	9559	54 (54/0)	38 (37/1)	57 (57/0)	0,47 [0,33;0,66]	1,95E-05	2,71E-02	related transcripts are not in set A		no duplication in EOAD nor in CTRL		NM_001001924.3; NM_001363057.2; NM_001363058.2; NM_001363059.2; NM_001363061.2
10	MBL2	4077	8457	9559	2 (2 / 0)	25 (25 / 0)	45 (44 / 1)	5,76 [2,22 ; 21,16]	5,44E-05	2,71E-02	0,13 [0,02;0,39]	2,02E-05	0,78 [0,005;14,65]	8,78E-01	NM_000242.3; NM_001378373.1; NM_001378374.1
17	FADS6	4077	8457	9559	16 (15 / 1)	11 (11 / 0)	5 (5 / 0)	0,19 [0,07 ; 0,45]	1,65E-04	2,71E-02	6,61 [2,64;19,28]	3,45E-05	7,04 [0,38 ; 1026,65]	1,89E-01	NM_178128.6
19	PLIN4**	4077	8457	9559	20 (19 / 1)	10 (10 / 0)	12 (12 / 0)	0,31 [0,15 ; 0,61]	8,38E-04	7,66E-02	related transcripts are not in set A		3,67 [1,82 ; 7,67]	2,83E-04	NM_001393888.1 ; NM_001393889.1 ; NM_001393890.1 ; NM_001393891.1; NM_001367868.2
19	KLC3	4077	8457	9559	24 (1 / 23)	33 (1 / 32)	16 (0 / 16)	3,06 [1,65 ; 5,79]	4,28E-04	4,12E-02	7,04 [0,38;1026,65]	1,89E-01	3,35 [1,8;6,41]	1,57E-04	NM_177417.3
19	ERCC2	4077	8457	9559	29 (1 / 28)	43 (2 / 41)	20 (0 / 20)	3,04 [1,74 ; 5,4]	1,09E-04	2,71E-02	7,04 [0,38;1026,65]	1,89E-01	3,28 [1,86;5,86]	4,09E-05	NM_001130867.2
19	ERCC2	4077	8457	9559	29 (1 / 28)	42 (2 / 40)	20 (0 / 20)								NM_000400.4
22	ZNF74	4077	8457	9559	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	0,04 [0,0003 ; 0,29]	2,71E-04	3,29E-02	21,12 [2,25;2799,03]	4,75E-03	0,12 [0,001;0,97]	4,58E-02	NM_001256523.2; NM_001256525.2; NM_003426.4; NM_001256524.2
22	SCARF2	4077	8457	9559	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	0,04 [0,0003 ; 0,29]	2,71E-04	3,29E-02	21,12 [2,25;2799,03]	4,75E-03	0,12 [0,001;0,97]	4,58E-02	NM_153334.7; NM_182895.5
22	KLHL22	4077	8457	9559	4 (4 / 0)	4 (0 / 4)	8 (0 / 8)	0,04 [0,0003 ; 0,30]	3,80E-04	3,86E-02	21,12 [2,25;2799,03]	4,75E-03	0,14 [0,001;1,10]	6,57E-02	NM_032775.4
22	MED15	4077	8457	9559	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	0,04 [0,0003 ; 0,29]	2,71E-04	3,29E-02	21,12 [2,25;2799,03]	4,75E-03	0,12 [0,001;0,97]	4,58E-02	NM_001003891.3; NM_001293234.2; NM_001293235.2; NM_001293236.2; NM_001293237.2; NM_015889.5
22	PI4KA	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	21,12 [2,25;2799,03]	4,75E-03	0,11 [0,001;0,87]	3,19E-02	NM_001362862.2; NM_001362863.2; NM_058004.4
22	SERPIND1	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	21,12 [2,25;2799,03]	4,75E-03	0,11 [0,001;0,87]	3,19E-02	NM_000185.4
22	SNAP29	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	21,12 [2,25;2799,03]	4,75E-03	0,11 [0,001;0,87]	3,19E-02	NM_004782.4
22	CRKL	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	21,12 [2,25;2799,03]	4,75E-03	0,11 [0,001;0,87]	3,19E-02	NM_005207.4
22	LZTR1	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	21,12 [2,25;2799,03]	4,75E-03	0,11 [0,001;0,87]	3,19E-02	NM_006767.4
22	THAP7	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	21,12 [2,25;2799,03]	4,75E-03	0,11 [0,001;0,87]	3,19E-02	NM_030573.3; NM_001008695.1
22	P2RX6	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	21,12 [2,25;2799,03]	4,75E-03	0,11 [0,001;0,87]	3,19E-02	NM_001159554.2; NM_001349874.2; NM_001349875.2; NM_001349876.2; NM_001394691.1; NM_001394692.1; NM_001394693.1; NM_001394694.1; NM_001394695.1; NM_001394696.1; NM_001394697.1; NM_005446.5
22	SLC7A4	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	21,12 [2,25;2799,03]	4,75E-03	0,11 [0,001;0,87]	3,19E-02	NM_004173.3
22	LRRC74B	3917	7799	8016	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,04 [0,0003 ; 0,29]	1,86E-04	2,71E-02	18,44 [1,97;2443,22]	7,39E-03	0,10 [0,001;0,75]	2,01E-02	NM_001291006.2

Table 3. Results from dosage analysis and subsequent analyses on deletions and duplications

Chr: chromosome; N: number, EOAD: early onset Alzheimer disease, LOAD: late onset Alzheimer disease; CTRL: controls

[1]OR should be interpreted for an increase of transcript copies by 1, i.e. a OR > 1 (resp. OR <1) means that duplications (resp. deletions) are associated with a higher disease risk

[2]non adjusted p-values, selected based on FRD < 10%

* gene overlapping with repeats or duplicated gene in the genome (despite CNVs not overlapping >50% with repeats)

** gene for which the signal in dosage analysis is drived by deletion whereas transcripts are not in set A

SUPPLEMENTAL METHODS

Sample descriptions

Cohorts

We included in our analysis a total of 22,319 samples with WES from Holstege et al [1], collected as part of the Alzheimer Disease European Sequencing consortium (ADES) comprising studies from Europe and of the Alzheimer Disease Sequencing Project (ADSP) comprising studies from the USA. Cohorts are detailed in Holsetge et al [1]. In order to homogenize EOAD/LOAD categorization across cohorts, we defined as EOAD, all cases with an age at onset (AAO) < 66 years or age at last visit < 66 if AAO is missing. Cases without numerical information for age were considered as LOAD. Cohorts are briefly described below.

All samples were selected following the quality control detailed in supplementary methods of Holstege et al[1], except samples who carry a pathogenic CNV. We included the latter in our initial dataset and excluded them later after a descriptive analysis of pathogenic CNVs found in our data set.

ADES-FR (Alzheimer Disease European Sequencing-France, France): WES of 1,731 cases (1,413 EOAD, 318 LOAD) and 1,017 controls recruited either at the French national reference center for young Alzheimer Disease at Rouen (CNRMAJ-Rouen; PHRC-GMAJ and ECASCAD study), or through the European Alzheimer's Disease Initiative (EADI) consortium [2]or the French exome project (FREX)[3].

AgeCoDe-UKBonn (Aging, Cognition and Dementia in primary care, Germany): WES of 371 cases (99 EOAD, 272 LOAD) and 1 control from the combination of the German study "Aging, Cognition, and dementia" (AgeCoDe)[4] and patients' recruitment at the interdisciplinary Memory Clinic of the department of psychiatry and department of Neurology at the University Hospital in Bonn (UKBonn). The control individual and all LOAD patients are from the AgeCoDe study and all had ≥ 75 years of age (AAO for cases, current age for control). EOAD cases were recruited through UKBonn cohort.

Barcelona-SPIN (Sant Pau Initiative on Neurodegeneration, Spain): WES of 50 EOAD patients and 9 controls recruited through the multimodal Sant Pau Initiative on Neurodegeneration (SPIN) cohort (https://santpaumemoryunit.com/our-research/spin-cohort/) [5] and for which neuropathological samples were obtained from the Neurological Tissue Bank of Biobanc-HospitalClinic-IDIBAPS.

AC-EMC (Alzheimer center - Erasmus University Medical Center, The Netherlands): WES of 108 cases (86 EOAD, 22 LOAD) from the Alzheimer center Erasmus MC cohort (AC-EMC) that are patients referred to the Department of Neurology of the Erasmus Medical center of Rotterdam.

ERF (Erasmus Rucphen Family, The Netherlands): WES of 4 cases (1 EOAD, 3 LOAD) and 319 controls from the Erasmus Rucphen Family (ERF) study, a family-based cohort study that is embedded in the Genetic Research in Isolated Populations (GRIP) program in the Southwest of the Netherlands.

Rotterdam study (The Netherlands): WES of 366 cases (3 EOAD, 363 LOAD) and 1,514 controls recruited in the prospective population-based cohort from Rotterdam study [6], focused on chronic disabling conditions of the elderly[7].

ADC-Amsterdam (Amsterdam medical center, The Netherlands): WES of 762 cases (502 EOAD, 260 LOAD) and 303 controls from patients who visit the memory clinic of the Alzheimer center at the Amsterdam University medical center [8]. Controls are individuals diagnosed with psychiatric and subjective cognitive complaints.

Netherlands Brain Bank (The Netherlands): WES of 169 cases (51 EOAD, 118 LOAD) and 52 controls obtained from DNA isolated from brain tissues.

100-plus Study (The Netherlands): WES of 64 LOAD patients (AAO ≥ 100 years) and 282 controls ≥ 100 years of age from the prospective cohort study of cognitively healthy centenarians [9].

EMIF-AD 90-plus Study (European Medical Information Framework for Alzheimer's Disease, The Netherlands): WES of 69 controls ≥ 90 years of age [10].

CBC (Control Brain consortium, UK): WES of 111 cases (33 EOAD, 78 LOAD) and 250 controls derived from brain banks in UK and USA [11].

PERADES (Defining Genetic, Polygenic and Environmental Risk for Alzheimer's Disease using multiple powerful cohorts, focused Epigenetics and Stem cell metabolomics UK): WES of 3,450 cases (1,110 EOAD, 2,340 LOAD) and 609 controls recruited across UK, Italy and Spain for the "Defining Genetic, Polygenic and Environmental Risk for AD" study.

ADSP (American Alzheimer Disease Sequencing Project, USA): WES of 5,502 cases (817 EOAD, 4,685 LOAD) and 5,228 controls from the ADSP Discovery phase (stage 1 in Holstege et al) and the ADSP Discovery extension and Augmentation phase (stage 2 in Holstege et al). Of note, controls from the

Discovery phase were all ≥ 60 years old and were selected as those having the least probability of converting to AD by age 85 [12].

BAM files sources

For most of the WES from Holstege et al. [1], BAM files were retrieved from the original processing pipeline after quality control. BAM were aligned on the GRCh37 version of the human genome. BAM from the ADSP Discovery extension and Augmentation phase were downloaded directly from the database of Genotypes and Phenotypes (dbGAP) website (N=1588), aligned on GRCh38. Thus, CNV calling and annotations were made for each sample according to the genome version. Then, data were merged to be analyzed jointly only at the transcript level.

CNV calling and quality control

Optimization of CANOES source code for large dataset

We grouped samples into calling batches as homogeneous as possible according to technical properties. Thus, depending on available information, samples were grouped primarily by sequencing batches, otherwise sequencing center else study. However, the latter option led to subsets of samples that were too large for an optimal use of CANOES as the time needed for processing an individual is proportional to the size of the input dataset.

To improve performances (computational time and resources needed) of CANOES on large datasets including ≥ 100 individuals, we slightly modified the original version of CANOES. Indeed, during the process, CANOES needs to generate an NxN matrix based on read counts similarity between samples. In the original version, CANOES may be executed sequentially or in parallel over a list of samples. In both cases, the NxN matrix was newly generated before the CNV calling for each sample, despite its consistency across all processed samples. If the time needed to completely analyze (including the matrix generation and the CNV calling) is less than 14 minutes for a dataset of 100 samples, it can go up to more than 11 hours for large datasets including 1000 samples. In the new version of CANOES, we separated the matrix generation from the calling process, such that the NxN matrix is generated only once for an input dataset. Then the calling process may be executed in parallel over all samples.

In a comparative study of the two CANOES versions in a similar setting of computational resources and parallelization (192 Gb of RAM and 32 threads), we observed that the modification helped to save ~2 minutes and ~9 hours for datasets with respectively ~100 and ~1000 samples (Figure S1).

Large mosaic CNV carriers

Due to its high sensibility, CANOES is able to call mosaic events appearing in > 30% of cells [13]. Such mosaic CNVs could be acquired early by mutation during the development process, or later in case of blood-specific, age-related events [14]. Both are associated with a lower frequency of cells carrying the mutation leading to a weak signal. Combined with the fact that they encompass low complexity regions, their calling may be less accurate and such large CNVs can be called as multiple, smaller CNVs. Thus, we computed for each sample and each chromosome, the cumulative size of detected CNVs of the same type (deletion/duplication). For each cumulative size per chromosome greater than 2.5Mb encompassing a range of 10Mb, we proceeded to a manual visualization through the UCSC genome browser. Literature review was performed to support decision making. Carriers of candidate large blood-specific, age-related CNVs were excluded from the analysis.

CNV filtering

Frequency

For each CNV, we proposed to compute two frequencies by considering independently information from two public databases: the Database of Genomic Variants (DGV) gold Standard section [15] and the non-neuro non-Finnish European section of the gnomAD database v2.1 [16]. For each CNV and each public dataset, corresponding frequency was computed as follows:

$$\text{frequency} = \frac{n_{\text{carriers}}}{n_{\text{total}}}$$

where $n_{carriers}$ corresponds to the number of carriers of a given CNV (of the same type [deletion or duplication]) overlapping mutually ≥ 70% with the CNV of interest; n_{total} is the total number of individuals included in the study revealing the candidate CNV. In case of several candidate CNVs from

several studies, all carriers of such CNVs were added together in the numerator $n_{carriers}$ whereas the denominator (n_{total}) corresponded to the highest sample size among all studies contributing to candidate CNVs in order to count individuals contributing to several candidate CNVs only once. Besides this may lead to an overestimation of the frequency, it makes our definition of "rare CNV" more stringent.

All CNVs with a frequency >1% in at least one above-mentioned public databases were excluded from the analyses.

Segmental duplications

Segmental duplication regions were extracted from the UCSC Table Browser (https://genome-euro.ucsc.edu/cgi-bin/hgTables). For each CNV, we computed its proportion overlapping segmental duplication regions. All CNV with a ≥50% overlap with segmental duplication regions were excluded from the analyses.

Information at the transcript level

To avoid potential bias due to heterogeneity of capture kits, we worked at the transcript level. This allowed us (i) to adapt the number of analyzable individuals for each transcript, (ii) to more easily highlight transcripts of interest affected by multiples CNVs with different coordinates, and (iii) to exclude transcripts affected by a recurrent small transcript-specific CNV without excluding the other transcripts of the same gene that were not affected by it. For each deletion and each duplication, we reported if it encompasses partially or completely a transcript.

To define partial and complete deletion or duplication of a transcript, we could not rely on annotation tools based on transcript coordinates defined by public databases as, for example, RefSeq. Indeed, each capture kit has its own specificity (some of them targeting UTR or specific alternative exons, while the others do not), leading to transcripts falsely reported as partially affected despite they should be considered as full events (Figure S2). Consequently, a transcript was considered as completely deleted or duplicated if all targets linked to this transcript were affected by the event.

Analyses and statistical tests

For each analysis, we compared EOAD cases versus controls at first-choice analysis. In parallel, we show the results obtained among all AD vs controls. In models detailed below, "cases" refers either to EOAD cases or to all AD cases. Each case-control comparison is performed using Firth's logistic regression. Based on a log-likelihood penalization, this method avoids infinite estimates due to separability problem that may arise in rare events analysis [17].

CNV-dosage analysis

For each transcript and each individual, the dosage information refers to the number of copies. Dosage equals 2 for individuals without any deletion nor full duplication affecting the transcript. Each duplication affecting the entire transcript increases the dosage by 1, whereas any deletion affecting partially or fully the transcript decreases the dosage by 1. Firth's logistic regression model is:

$$\text{logit}\left[\mathbb{P}(\text{case} \mid \text{dosage})\right] = \beta_0 + \beta_1 \times \text{dosage}$$

such that OR (=exp(β1)) associated with dosage information should be interpreted for an increase of transcript copies by 1, i.e. a OR > 1 (resp. OR <1) means that duplications (resp. deletions) are associated with a higher disease risk.

Deletions/duplications analyses

For each transcript and each individual, the deletion (resp. full duplication) information refers to a binary variable: absence/presence of any deletions (resp. full duplication). Firth's logistic regression model is:

$$\text{logit}\left[\mathbb{P}(\text{case} \mid \text{presence of a CNV})\right] = \beta_0 + \beta_1 \times \mathbb{1}_{(\text{presence of a CNV})}$$

such that OR (=exp(β1)) associated with CNV (any deletion or full duplication) information should be interpreted for a CNV presence versus absence, i.e. a OR > 1 (resp. OR <1) means that carrying a CNV is associated with a higher (resp. lower) disease risk.

Joint CNV and loss-of-function indels and SNVs analyses

At the gene level, we assessed the effect of all loss-of-function (LOF) variants in a list of AD risk genes and candidate genes. We included in this analysis all genes previously associated with AD in a GWAS

study [18] and in our latest gene-based rare variants exome case-control study [1] as well as genes prioritized by the dosage EOAD – controls analysis (FDR≤10%).

LOF variants were defined as following:

any deletion (partial or full) affecting at least one transcript within the gene. Only transcripts with less than 25% of missingness in the whole dataset were considered for this analysis;

SNV/indel with the highest prediction of a loss-of-function variant. To filter SNV/indels, we used the LOFTEE software[19] labeled as "High Confidence", as in Holstege et al. (i.e., nonsense, canonical splice site variants and frameshift indels not affecting the last exon or the 50 last bp of the penultimate exon with no close rescue site as predicted by MaxEntScan software [20]). In addition, we performed an additional check on splicing variants, which can have diverse consequences that are more difficult to predict than nonsense variants and frameshift indels. We filtered out splicing variants affecting in-frame coding exons of less than 33 amino-acid residues (<100bp), as the likelihood of a LOF effect is lower. We also added manually the recurrent NM_019112.4:c.5570+5G>C ABCA7 recurrent variant with a demonstrated effect on splicing and considered as a LOF variant [21]. We considered for the analysis all LoF-SNV with < 25% of missingness, without differential missingness between cases and controls (based on a Fisher's exact test and p-value > 10-30) and a Variant Batch Detector (VBD) score < 20 [1].

Individuals with missing information in at least one transcript considered for deletion or in > 80% of gene position were not included in the analysis.

For each gene and each individual, the LOF information refers to a binary variable: absence/presence of a LOF variant as defined above. Firth's logistic regression model is:

$$\text{logit}\left[\mathbb{P}(\text{case} \mid \text{presence of a LOF})\right] = \beta_0 + \beta_1 \times \mathbb{1}_{(\text{presence of a LOF})}$$

such that OR (=exp(β1)) associated with LOF information should be interpreted for a LOF presence versus absence, i.e. a OR > 1 (resp. OR <1) means that carrying a LOF is associated with a higher (resp. lower) disease risk.

Sensitivity analyses

Sensitivity analyses include either adjustment for:

- ancestry based on the principal component analysis performed on the 1000G samples and described in Holstege et al. [1]. Firth's logistic model includes the first ten principal components and is thus:

$$\text{logit}\left[\mathbb{P}(\text{case} \mid \text{X, ancestry})\right] = \beta_0 + \beta_1 \times \text{X} + \alpha_1 \times \text{PC}_1 + ... + \alpha_{10} \times \text{PC}_{10}$$

- APOE4 dosage being 0 for non-carriers of e4 allele, 1 for e4-heterozygote carriers and 2 for 4e-homozygote carriers. Firth's logistic model is thus:

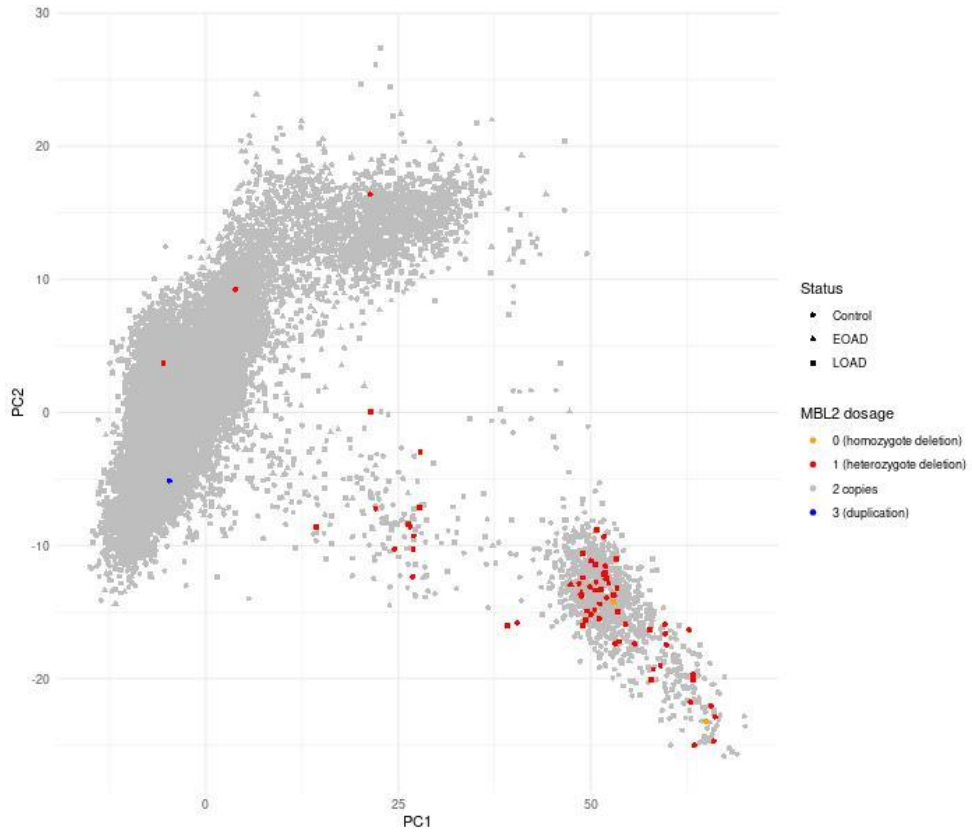
$$\text{logit}\left[\mathbb{P}(\text{case} \mid \text{X}, APOE)\right] = \beta_0 + \beta_1 \times \text{X} + \beta_2 \times APOE \text{ dosage}$$

The X variable represents either the dosage, the CNV or the LOF variant, and β1 coefficient should be interpreted as described above according to the X variable.

All statistical analyses were performed using R version 3.6 [22]. Firth's logistic regressions were performed using the "logistf" function from "logistf" package [23].

SUPPLEMENTAL RESULTS

In the sensitivity analysis, we adjusted our model for ancestry using the first ten principal components as covariates. This adjustment did not affect ORs nor p-values in the dosage analysis except for two genes. First, variability associated with dosage in analysis of FADS6 gene is higher after adjustment leading to 2-order of magnitude lower p-value. Second, in MBL2 gene analysis, OR becomes unsignificant since deletions observed is population dependent. Indeed, all carriers share a similar ancestry (see below the projection of CNV carriers in MBL2 region on the two main axes for ancestry components).



Detailed acknowledgements

SURF supercomputer facility

Part of the work in this manuscript was carried out on the Cartesius supercomputer, which is embedded in the Dutch national e-infrastructure with the support of SURF Cooperative. Computing hours were granted in 2016, 2017, 2018 and 2019 to H. Holstege by the Dutch Research Council (project name: '100plus'; project numbers 15318 and 17232).

ADES-FR

This study was funded by grants from the Clinical Research Hospital Program from the French ministry of Health (GMAJ, PHRC, 2008/067), the CNR-MAJ, the JPND PERADES, Equipe FRM DEQ20170336711, and Fondation Alzheimer (ECASCAD study). This research was supported by the Laboratory of Excellence GENMED (Medical Genomics) grant no. ANR-10-LABX-0013 managed by the National Research Agency (ANR) part of the Investment for the Future program. This work was also supported

by Foundation Alzheimer, the Institut Pasteur de Lille, Inserm, the Haut-de-France and Lille Métropole Communauté Urbaine council, and the French government's LABEX (laboratory of excellence program investment for the future) DISTALZ grant (Development of Innovative Strategies for a Transdisciplinary approach to Alzheimer's disease). The 3C Study supports are listed on the Study Website (www.three-city-study.com). This work is a collaboration between CEADRF-Jacob-CNRGH-CHU de Rouen. This work did benefit from the support of the France Génomique National infrastructure, funded as part of the "Investissements d'Avenir" program managed by the Agence Nationale pour la Recherche (contract ANR-10-INBS-09).

AgeCoDe-UKBonn

The AgeCoDe cohort was funded in part by the German Federal Ministry of Education and Research (BMBF) (grants KNDD 01GI0710, 01GI0711, 01GI0712, 01GI0713, 01GI0714, 01GI0715, 01GI0716, 01ET1006B). Sequencing of AgeCoDe sample was in part funded by the German Research Foundation (DFG) grant RA 1971/6-1 to Alfredo Ramirez.

Barcelona- SPIN

Support for Jordi Clarimon provided by Maratón RTVE (Spain). Support for Oriol Dols provided by the Association for Frontotemporal Degeneration (Clinical Research Postdoctoral Fellowship, AFTD).

AC-EMC

Exome sequencing was funded by Alzheimer Nederland.

ERF

The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4- 2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-

CT-2002- 01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043).

Rotterdam Study

The generation and management of the exome sequencing data for the Rotterdam Study was executed by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Netherlands. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for diseases in the Elderly (RIDE) (014-93-015; RIDE2), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO), the Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810, and by a Complementation Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL; www.bbmri.nl ; project number CP2010-41). We thank Mr. Pascal Arp, Ms. Mila Jhamai, Mr. and Marijn Verkerk, for their help in creating the RS-Exome Sequencing database.

ADC-Amsterdam

We thank all study participants and all personnel involved in data collection for the contributing studies. Research of Alzheimer center Amsterdam is part of the neurodegeneration research program of Amsterdam Neuroscience. Alzheimer Center Amsterdam is supported by Stichting Alzheimer Nederland and Stichting VUmc fonds. The clinical database structure was developed with funding from Stichting Dioraphte. This work was supported by Stichting Alzheimer Nederland (WE.09-2014-06, WE.05-2010-06); Stichting Dioraphte; Internationale Stichting Alzheimer Onderzoek (#11519); JPNDPERADES (ZonMw 733051022): Centralized Facility for Sequence to Phenotype analyses (ZonMW 9111025); Netherlands Consortium for Healthy Aging (NCHA 050-060-810); Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL CP2010-41); Netherlands Genomics

Initiative (NGI)/NWO. This study is further supported by ABOARD, a public-private partnership receiving funding from ZonMW (#73305095007) and Health~Holland, Topsector Life Sciences & Health (PPP-allowance; #LSHM20106). This research is performed by using data from the Parelsnoer Institute an initiative of the Dutch Federation of University Medical Centres (www.parelsnoer.org).

100-plus Study

Cohort collection and exome sequencing of the 100-plus Study cohort was supported by Stichting Alzheimer Nederland (WE.09-2014-03); HorstingStuit Foundation, VUmc Foundation, and the Dioraphte Foundation (Project 17020403), Memorabel (ZonMW project number #733050814, #733050512) and Stichting VUmcFonds. Additional support is from ABOARD, a public-private partnership receiving funding from ZonMW (#73305095007) and Health~Holland, Topsector Life Sciences & Health (PPP-allowance; #LSHM20106).

EMIF-AD 90+

The EMIF-AD 90+ Study was funded by the EU/EFPIA Innovative Medicines Initiative Joint Undertaking EMIF grant agreement no. 115372.

CBC: Control Brain Consortium

This work was supported by the UK Dementia Research Institute which receives its funding from DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK, Medical Research Council (award number MR/N026004/1). Wellcome Trust Hardy (award number 202903/Z/16/Z), Dolby Family Fund; National Institute for Health Research University College London Hospitals Biomedical Research Centre; BRCNIHR Biomedical Research Centre at University College London Hospitals NHS Foundation Trust and University College London. J. Hardy was supported by the Dolby Foundation and the JPND PERADES. J.B. and R.G were supported by the National Institute on Aging of the National Institutes of Health under Award Number R01AG067426. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

PERADES

We thank all individuals who participated in the study. We also want to express our gratitude to the MRC Centre Core Team for the laboratory support and the Advanced Research Computing at Cardiff University (ARCCA) for the computational support. Cardiff University was supported by the Medical Research Council. Cardiff University was also supported by the European Joint Programme for Neurodegenerative Disease, Alzheimer's Research UK, the Welsh Assembly Government, and a donation from the Moondance Charitable Foundation. Cardiff University acknowledges the support of the UK Dementia Research Institute, which receives its funding from UK DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK. Cambridge University acknowledges support from the MRC. The University of Southampton acknowledges support from the Alzheimer's Society. ARUK provided support to Nottingham University. Join Dementia Research (JDR) is funded by the Department of Health and delivered by the National Institute for Health Research in partnership with Alzheimer Scotland, Alzheimer's Research UK and Alzheimer's Society. IRCCS Santa Lucia Foundation acknowledges the Italian Ministry of Health for financial support (IMH_RC) of this study. The Centro de Biologia de Molecular Severo Ochoa (CSIS-UAM), CIBERNED, Instituto de Investigacion Sanitaria la Paz, University Hospital La Paz and the Universidad Autonoma de Madrid were supported by grants from the Ministerio de Educacion y Ciencia and the Ministerio de Sanidad y Consumo (Instituto de Salud Carlos III), and an institutional grant of the Fundacion Ramon Areces to the CMBSO. Thanks to I. Sastre and Dr A Martinez-Garcia for DNA preparation, and Drs P Gil and P Coria for their recruitment efforts. Department of Neurology, University Hospital Mutua de Terrassa, Terrassa, Barcelona, Spain was supported by CIBERNED, Centro de Investigacion Biomedica en Red de Enfermedades Neurodegenerativas, Instituto de Salud Carlos III, Madrid Spain and acknowledges Maria A Pastor (Department of Neurology, University of Navarra Medical School and Neuroimaging Laboratory, Center for Applied Medical Research, Pamplona, Spain), Manuel Seijo-Martinez (Department of Neurology, Hospital do Salnes, Pontevedra, Spain), Ramon Rene, Jordi Gascon and Jaume Campdelacreu (Department of Neurology, Hospital de Bellvitage, Barcelona, Spain) for providing DNA samples. Hospital de la Sant Pau, Universitat Autonoma de Spain acknowledges support from the Spanish Ministry of Economy and Competitiveness (grant number PI12/01311), and from Generalitat de Catalunya (2014SGR-235). The Santa Lucia Foundation and the Fondazione Ca' Granda IRCCS Ospedale Policlinico, Italy, acknowledge the Italian Ministry of Health (grant RC 10.11.12.13/A)

UCL-DRC EOAD

This work was supported by the Medical Research Council (UK), the Biomedical Research Centre at University College London Hospitals NHS Foundation Trust and charitable donations to the UCL Dementia Research Centre.

ADSP

The Alzheimer's Disease Sequencing Project (ADSP) is comprised of two Alzheimer's Disease (AD) genetics consortia and three National Human Genome Research Institute (NHGRI) funded Large Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer's Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National Institute of Health (NIH) institutes and other foreign 104 governmental and non-governmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs. Schellenberg, Farrer, Pericak- Vance, Mayeux, and Haines); U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate and the Discovery Extension Phase analysis is supported through U01AG052411 to Dr. Goate, U01AG052410 to Dr. Pericak-Vance and U01 AG052409 to Drs. Seshadri and Fornage, U54 AG052427 to Drs. Schellenberg and Wang, and R01 AG054060 to Dr Naj. The ADGC cohorts include: Adult Changes in Thought (ACT) (UO1 AG006781, UO1 HG004610, UO1 HG006375, U01 HG008657), the Alzheimer's Disease Centers (ADC) (P30 AG019610, P30 AG013846, P50 AG008702, P50 AG025688, P50 AG047266, P30 AG010133, P50 AG005146, P50 AG005134, P50 AG016574, P50 AG005138, P30 AG008051, P30 AG013854, P30 AG008017, P30 AG010161, P50 AG047366, P30 AG010129, P50 AG016573, P50 AG016570, P50 AG005131, P50 AG023501, P30 AG035982, P30 AG028383, P30 AG010124, P50 AG005133, P50 AG005142, P30 AG012300, P50 AG005136, P50 AG033514, P50 AG005681, and P50 AG047270), the Chicago Health and Aging Project (CHAP) (R01 AG11101, RC4 AG039085, K23 AG030944), Indianapolis Ibadan (R01 AG009956, P30 AG010133), the Memory and Aging Project (MAP) (R01 AG17917), Mayo Clinic (MAYO) (R01 AG032990, U01 AG046139, R01 NS080820, RF1 AG051504, P50 AG016574), Mayo Parkinson's Disease controls (NS039764, NS071674, 5RC2HG005605), University of Miami (R01 AG027944, R01 AG028786, R01 AG019085, IIRG09133827, A2011048), the Multi-Institutional Research in Alzheimer's Genetic Epidemiology Study (MIRAGE) (R01 AG09029, R01 AG025259), the National Cell Repository for

Alzheimer’s Disease (NCRAD) (U24 AG21886), the National Institute on Aging Late Onset Alzheimer's Disease Family Study (NIA- LOAD) (R01 AG041797), the Religious Orders Study (ROS) (P30 AG10161, R01 AG15819), the Texas Alzheimer’s Research and Care Consortium (TARCC) (funded by the Darrell K Royal Texas Alzheimer's Initiative), Vanderbilt University/Case Western Reserve University (VAN/CWRU) (R01 AG019757, R01 AG021547, R01 AG027944, R01 AG028786, P01 NS026630, and Alzheimer’s Association), the Washington Heights-Inwood Columbia Aging Project (WHICAP) (RF1 AG054023), the University of Washington Families (VA Research Merit Grant, NIA: P50AG005136, R01AG041797, NINDS: R01NS069719), the Columbia University HispanicEstudio Familiar de Influencia Genetica de Alzheimer (EFIGA) (RF1 AG015473), the University of Toronto (UT) (funded by Wellcome Trust, Medical Research Council, Canadian Institutes of Health Research), and Genetic Differences (GD) (R01 AG007584). The CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI) infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle) and the neurology working group is supported by the National Institute on Aging (NIA) R01 grant AG033193. This work was also supported by National Institute on Aging grants R01 AG048927 to Dr. Farrer, RF1 AG054080 to Dr. Beecham, U24 AG056270 to Dr. Mayeux, RF1 AG057519 to Dr. Farrer, U01 AG062602 to Dr. Farrer, R01 AG067501 to Dr. Mayeux, and U19 AG068753 to Dr. Farrer. The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry- Austria (ASPS/PRODEM-Aus), the Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmus Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project I904),the EU Joint Programme - Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project (Austria, Ministry of Science) and the Medical University of Graz and the Steiermärkische Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Promotion agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435. ARIC research is carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive data in ARIC is collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, 2U01HL096917 from the NIH (NHLBI, NINDS, NIA and NIDCD), and with previous brain MRI examinations funded by R01- HL70825 from the NHLBI. CHS research was supported by contracts

HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also supported by additional grants from the NIA (R01s AG054076, AG049607 and AG033040 and NINDS (R01 NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007- 2013)/grant agreement HEALTH-F4- 2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002- 01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies are grateful to their participants, faculty and staff. The content of these manuscripts is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the U.S. Department of Health and Human Services. The four LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), The American Genome Center at the Uniformed Services University of the Health Sciences (U01AG057659), and the Washington University Genome Institute (U54HG003079). Biological samples and associated phenotypic data used in primary data analyses were stored at Study Investigators institutions, and at the National Cell Repository for Alzheimer’s Disease (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated Phenotypic Data used in primary and secondary data analyses were provided by Study Investigators,

the NIA funded Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center (NACC, U01AG016976) and the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA This research was supported in part by the Intramural Research Program of the National Institutes of health, National Library of Medicine. Contributors to the Genetic Analysis Data included Study Investigators on projects that were individually funded by NIA, and other NIH institutes, and by private U.S. organizations, or foreign governmental or nongovernmental organizations. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] H. Holstege et al., 'Exome sequencing identifies rare damaging variants in ATP8B4 and ABCA1 as risk factors for Alzheimer's disease', Nat Genet, vol. 54, no. 12, pp. 1786–1794, Dec. 2022, doi: 10.1038/s41588-022-01208-7.

[2] J.-C. Lambert et al., ‘Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer’s disease’, Nat Genet, vol. 41, no. 10, pp. 1094–1099, Oct. 2009, doi: 10.1038/ng.439.

[3] E. Genin, ‘The French Exome (FREX) Project: A Population-based panel of exomes to help filter out common local variants.’, Genetic Epidemiology, vol. 41, no. 7, pp. 644–709, Nov. 2017, doi: 10.1002/gepi.22062.

[4] T. Luck et al., ‘Mild cognitive impairment in general practice: age-specific prevalence and correlate results from the German study on ageing, cognition and dementia in primary care patients (AgeCoDe)’, Dement Geriatr Cogn Disord, vol. 24, no. 4, pp. 307–316, 2007, doi: 10.1159/000108099.

[5] D. Alcolea et al., ‘The Sant Pau Initiative on Neurodegeneration (SPIN) cohort: A data set for biomarker discovery and validation in neurodegenerative disorders’, Alzheimers Dement (N Y), vol. 5, pp. 597–609, 2019, doi: 10.1016/j.trci.2019.09.005.

[6] M. A. Ikram et al., ‘Objectives, design and main findings until 2020 from the Rotterdam Study’, Eur J Epidemiol, vol. 35, no. 5, pp. 483–517, May 2020, doi: 10.1007/s10654-020-00640-5.

[7] A. Hofman et al., ‘The Rotterdam Study: 2014 objectives and design update’, Eur J Epidemiol, vol. 28, no. 11, pp. 889–926, Nov. 2013, doi: 10.1007/s10654-013-9866-z.

[8] W. M. van der Flier and P. Scheltens, ‘Amsterdam Dementia Cohort: Performing Research to Optimize Care’, J Alzheimers Dis, vol. 62, no. 3, pp. 1091–1111, 2018, doi: 10.3233/JAD-170850.

[9] H. Holstege et al., ‘The 100-plus Study of cognitively healthy centenarians: rationale, design and cohort description’, Eur J Epidemiol, vol. 33, no. 12, pp. 1229–1249, Dec. 2018, doi: 10.1007/s10654-018-0451-3.

[10] N. Legdeur et al., ‘Resilience to cognitive impairment in the oldest-old: design of the EMIF-AD 90+ study’, BMC Geriatr, vol. 18, no. 1, p. 289, Nov. 2018, doi: 10.1186/s12877-018-0984-z.

[11] R. Guerreiro et al., ‘TREM2 variants in Alzheimer’s disease’, N Engl J Med, vol. 368, no. 2, pp. 117–127, Jan. 2013, doi: 10.1056/NEJMoa1211851.

[12] G. W. Beecham et al., ‘The Alzheimer’s Disease Sequencing Project: Study design and sample selection’, Neurology Genetics, vol. 3, no. 5, p. e194, Oct. 2017, doi: 10.1212/NXG.0000000000000194.

[13] K. Le Guennec et al., ‘17q21.31 duplication causes prominent tau-related dementia with increased MAPT expression’, Molecular Psychiatry, vol. 22, no. 8, pp. 1119–1125, Aug. 2017, doi: 10.1038/mp.2016.226.

[14] M. J. Machiela et al., ‘Mosaic chromosome 20q deletions are more frequent in the aging population’, Blood Advances, vol. 1, no. 6, pp. 380–385, Feb. 2017, doi: 10.1182/bloodadvances.2016003129.

[15] J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer, ‘The Database of Genomic Variants: a curated collection of structural variation in the human genome’, Nucleic Acids Res, vol. 42, no. Database issue, pp. D986-992, Jan. 2014, doi: 10.1093/nar/gkt958.

[16] R. L. Collins et al., ‘An open resource of structural variation for medical and population genetics’, Genomics, preprint, Mar. 2019. doi: 10.1101/578674.

[17] G. Heinze and M. Schemper, ‘A solution to the problem of separation in logistic regression’, Statist. Med., vol. 21, no. 16, pp. 2409–2419, Aug. 2002, doi: 10.1002/sim.1047.

[18] C. Bellenguez et al., ‘New insights into the genetic etiology of Alzheimer’s disease and related dementias’, Nat Genet, vol. 54, no. 4, pp. 412–436, Apr. 2022, doi: 10.1038/s41588-022-01024-z.

[19] K. J. Karczewski et al., ‘The mutational constraint spectrum quantified from variation in 141,456 humans’, Nature, vol. 581, no. 7809, pp. 434–443, May 2020, doi: 10.1038/s41586-020-2308-7.

[20] G. Yeo and C. B. Burge, ‘Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals’, J Comput Biol, vol. 11, no. 2–3, pp. 377–394, 2004, doi: 10.1089/1066527041410418.

[21] S. Steinberg et al., ‘Loss-of-function variants in ABCA7 confer risk of Alzheimer’s disease’, Nat Genet, vol. 47, no. 5, pp. 445–447, May 2015, doi: 10.1038/ng.3246.

[22] R Core Team, ‘R: A language and environment for statistical computing.’ R Foundation for Statistical Computing, Vienna, Austria., 2019. [Online]. Available: https://www.R-project.org/.

[23] G. Heinze, M. Ploner, and L. Jiricka, ‘logistf: Firth’s Bias-Reduced Logistic Regression.’ 2022. [Online]. Available: https://CRAN.R-project.org/package=logistf

Cohort	Capture kit	Reference genome	N after QC	N for association analysis
100-plus Study	Nimblegen v3	GRCh37	247	247
100-plus Study	Agilent v6	GRCh37	90	90
EMIF-AD 90-plus Study	Agilent v6	GRCh37	68	68
AC-EMC	Agilent v6	GRCh37	39	39
AC-EMC	Nimblegen v2	GRCh37	67	67
ADC-Amsterdam	Nimblegen v3	GRCh37	296	294
ADC-Amsterdam	Agilent v6	GRCh37	758	758
ADES-FR	Agilent v5	GRCh37	1327	1319
ADES-FR	Agilent v5 + UTR	GRCh37	821	821
ADES-FR	Agilent v6 + UTR	GRCh37	460	460
ADES-FR	Agilent v4	GRCh37	106	106
ADSP	Nimblegen VCRome 2.1	GRCh37	5124	5123
ADSP	Illumina Rapid Capture Exome	GRCh37	3942	3942
ADSP	Illumina Rapid Capture Exome	GRCh38	96	96
ADSP	Agilent v5	GRCh38	200	200
ADSP	Nimblegen VCRome Custom	GRCh38	480	480
ADSP	Nimblegen v3	GRCh38	778	778
AgeCoDe-UKBonn	Nimblegen v2	GRCh37	370	369
Barcelona-SPIN	Nimblegen v3	GRCh37	59	59
Netherlands Brain Bank	Agilent v6	GRCh37	212	211
ERF	Agilent v4	GRCh37	322	322
PERADES	Nextera v1,2	GRCh37	4034	4030
Rotterdam Study	Nimblegen v2	GRCh37	1865	1865
CBC	Nimblegen v2	GRCh37	60	60
CBC	Illumina TruSeq	GRCh37	290	290
Total			**22111**	**22094**

Table S1: Sample repartition among cohort

N: number, QC: quality control, EMIF-AD: European Medical Information Framework for Alzheimer's Disease, AC-ECM: Alzheimer Center - Erasmus Medical Center, ADC: Amsterdam medical center, ADES-FR: Alzheimer Disease European Sequencing-France, ADSP: American Alzheimer Disease Sequencing Project; AgeCoDe-UKBonn: Aging, Cognition, and Dementia in primary care, SPIN: Sant Pau Initiative on Neurodegeneration, ERF: Erasmus Rucphen Family, PERADES: Defining Genetic, Polygenic and Environmental Risk for Alzheimer's Disease using multiple powerful cohorts, focussed Epigenetics and Stem cell metabolomics CBC: Control Brain Consortium

Sample	Chr	Type of event	Cumulate CNV size call	Start	End	Referent genome	Exclusion decision	Comments	Literature review	Age	Gender	Status	Cohort
A-LOAD-LD007196-CL-NCR-8008287363	1	Deletion	9559508	184477417	197146373	GRCh38	Excluded	These events are probably releted to an acquired mosaic deletion as it covers region of 12.5 Mb from 1q25 to 1q31, a locus associated with genomic syndrome of intellectual disability.	Höglund et al., 2003	90	Male	LOAD	ADSP
A-WCAP-WC003720-BL-COL-37039BL1	1	Deletion	24570931	211312652	235883583	GRCh38	Excluded	These events are probably releted to an acquired mosaic deletion as it covers region of 24.5 Mb from 1q32 to 1q42, a locus associated with genomic syndrome of intellectual disability.	Shaffer et al., 2007	80	Female	Control	ADSP
A-MIA-UM001958	1	Deletion	11557459	237655105	249212564	GRCh37	Excluded	These events are probably releted to an acquired mosaic deletion as it covers region of 11 Mb from 1q43 to 1q44 (chromosome 1 long arm telomeric extremity), a locus associated with genomic syndrome of intellectual disability.	van Bon et al., 2008	76	Male	LOAD	ADSP
10622	1	Duplication	14263793	31194249	45672024	GRCh37	Excluded	This duplication of 14 Mb from 1p34 to 1p35 is part of recurrent cytogenetic rearrangements in chronic lymphocytic leukemia.	Cavazzini et al., 2009	78	Male	LOAD	PERADES
A-NCRD-NC014940-CL-NCR-8008287827	2	Duplication	148329462	41459	242095067	GRCh38	Excluded	Trisomy 2 is a rare yet recurrent finding in myelodysplastic syndrome and occurs more frequently in acute myeloid leukaemia.	Czepulkowski et al., 2003	68	Female	LOAD	ADSP
A-ADC-AD008487 [1]	3	Duplication	48812328	107429299	197896723	GRCh37	Excluded	These events lead to a duplication of the long arm of chromosome 3. 3q21q26 rearrangements were found in 3% of t-MDS/t-AML	Block et al., 2002	67	Male	LOAD	ADSP
C-CHS-51712	4	Deletion	4477743	4861626	10586580	GRCh37	Excluded	Events from 4,861,627 to 10,586,580 are related to a deletion of 5.7Mb in 4p16 region. Such an event has been described before in leukaemia (acquired mosaic deletion) or in Wolf-Hirchhorn syndrome (constitutional abnormality).	Cavazzini et al., 2009	83	Female	Control	ADSP
A-LOAD-LD010850	4	Deletion	3048748	NA	NA	GRCh37	Not excluded	Manual vizualisation showed that this individual carries two independant deletions on chromosome 4 of respectively 742,706 bp (chr4:57164395-57907100) and 2,306,044 bp (chr4:104510838-106816881). The larger deletion appears in a polymorphic region.		82	Male	Control	ADSP

[1] This sample carries two CNVs selected as potential lareg mosaic: a duplication on chromosome 3 as well as a deletion on chromosome 13,

Sample	Chr	Type of event	Cumulate CNV size call	Start	End	Referent genome	Exclusion decision	Comments	Literature review	Age	Gender	Status	Cohort
A-MAP-MA000504	5	Duplication	3160051	NA	NA	GRCh37	Not excluded	Manual vizualisation showed that this individual carries two independant duplications on chromosome 5 of respectively 3,158,890 bp (chr5:50680374-53839263) and 11,620 bp (chr5:68849396-68861016). The larger duplication appears in a polymorphic region closed to the centromere.		88	Male	LOAD	ADSP
A-MIA-UM001958	5	Duplication	37948774	136957785	180687816	GRCh37	Excluded	These events are probably releted to an acquired mosaic deletion as it covers region of 43,7 Mb, from 5q31 to 5q25 (chromosome 5 long arm telomeric extremity), mainly detected in myelodysplastic/myeloproliferative neoplasms and acute myeloid leukemia.	Haase D, 2008.	76	Male	LOAD	ADSP
A-NCRD-NC016047-CL-NCR-8008288034 [2]	6	Deletion	70748470	203388	170584569	GRCh38	Excluded	Monosomy 6 (-6), not compatible with survival when constitutional. Somatic event rare but present in some myeloid leukemia.	Jang et al., 2015	74	Female	LOAD	ADSP
A-LOAD-LD005084	8	Duplication	100492852	363079	146279544	GRCh37	Excluded	Trisomy 8 is one of the major anomalies additional to the t(9;22), with i(17q), + der(22), before +19; found as a unique additional anomaly in 10%, with other in 25% of chronic myelogenous leukemia cases with clonal evolution.	Paulsson et al.,2007	81	Female	LOAD	ADSP
20304	8	Duplication	8229884	1808011	146279545	GRCh37	Excluded	Trisomy 8 is one of the major anomalies additional to the t(9;22), with i(17q), + der(22), before +19; found as a unique additional anomaly in 10%, with other in 25% of chronic myelogenous leukemia cases with clonal evolution.	Paulsson et al.,2007	75	Female	LOAD	PERADES
WD_061	8	Duplication	2776250	2813051	146056413	GRCh37	Excluded	Trisomy 8 is one of the major anomalies additional to the t(9;22), with i(17q), + der(22), before +19; found as a unique additional anomaly in 10%, with other in 25% of chronic myelogenous leukemia cases with clonal evolution.	Paulsson et al.,2007	80	Female	Control	ADES-FR
09D04577	9	Deletion	3745539	42366578	70490142	GRCh37	Excluded	28Mb deletion included the heterochromatic centromeric section of the chr9, a recurrent cytogenetic abnormality in acute myeloid leukaemia (AML).	Peniket et al., 2005	48	Male	EOAD	ADC-Amsterdam

[2] This sample carries two CNVs selected as potential lareg mosaic: a deletion on chromosome 6 as well as a duplication on chromosome 15,

Sample	Chr	Type of event	Cumulate CNV size call	Start	End	Referent genome	Exclusion decision	Comments	Literature review	Age	Gender	Status	Cohort
15R1099	9	Deletion	21163194	45376571	69901148	GRCh37	Excluded	24Mb deletion included the heterochromatic centromeric section of the chr9, a recurrent cytogenetic abnormality in acute myeloid leukaemia (AML).	Peniket et al., 2005	101	Female	Control	100-plus Study
20014	9	Duplication	36743260	15907	69256891	GRCh37	Excluded	This events are related to an acquired trisomy 9 is an abnormality found in 2,2% of chronic myeloproliferative disorders other than CML	Bacher et al., 2005	70	Female	LOAD	PERADES
Anglia_10503	9	Duplication	66648608	116748	141069324	GRCh37	Excluded	This events are related to an acquired trisomy 9 is an abnormality found in 2,2% of chronic myeloproliferative disorders other than CML	Bacher et al., 2005	80	Male	LOAD	PERADES
10210	10	Duplication	74182382	292666	135379064	GRCh37	Excluded	These events are realted to an acquired trisomy 10 appearing in less than 1% of acute leukemia.	Czepulkowski et al., 2002	91	Female	Control	Rotterdam Study
C-FHS-51043	11	Deletion	15684839	99690274	115375113	GRCh37	Excluded	This event is probably related to an acquired deletion in 11q region which has been associated with non-Hodgkin's lymphoma.	Cuneo et al., 2000	89	Male	LOAD	ADSP
C-RS-50633	11	Deletion	4136197	61593557	68673723	GRCh37	Excluded	The larger event is probably related to an acquired deletion in 11q region which has been associated with non-Hodgkin's lymphoma.	Cuneo et al., 2000	90	Male	Control	ADSP
A-NCRD-NC009851-CL-NCR-8008288019 [3]	11	Duplication	37678181	94129316	134387714	GRCh38	Excluded	Partial gain of the long arm of chromosome 11, containing the unrearranged mixed lineage leukaemia gene (*KMT2A*, lysine (K)-specific methyltransferase 2A) is a rare but recurrent anomaly in myeloid malignancies, and is often associated with an older age and a higly complex karyotype.	Winters et al., 2017	90	Male	LOAD	ADSP
10263	12	Deletion	10772835	11034804	21807639	GRCh37	Excluded	T-cell prolymphocytic leukemia is a rare form of mature leukemia which occurs in adults and in younger patients suffering ataxia telangiectasia. Among others, complex chromosome aberrations of chromosome 12 have been described in this disease. We searched for deletions of the 12p13 region as the result of these chromosome rearrangements.	Hetet et al., 2000	73	Male	Control	Rotterdam Study
A-NCRD-NC007070-CL-NCR-8015320875	12	Duplication	34859191	84861630	133202875	GRCh38	Excluded	Trisomy 12 is found in one third of cytogenetically abnormal chronic lymphocytic leukemia by conventional karyotype, and in about 12-54% of cases when interphase FISH is used.	Matutes et al., 1996	72	Female	LOAD	ADSP

[3] This sample carries two CNVs selected as potential large mosaic: a duplication on chromosome 11 and a deletion on chromosome 18

Sample	Chr	Type of event	Cumulate CNV size call	Start	End	Referent genome	Exclusion decision	Comments	Literature review	Age	Gender	Status	Cohort
A-GDF-GD000340	15	Duplication	8258772	31324891	102261526	GRCh37	Excluded	These events are related to an acquired trisomy 15 already observed in hematological malignancies (rare, marked male predominance, found mostly in adults with a median age of 77 years).	Smith et al., 1996	86	Male	Control	ADSP
FR16641240	15	Duplication	70464010	22016089	102417144	GRCh37	Excluded	These events are related to an acquired trisomy 15 already observed in hematological malignancies (rare, marked male predominance, found mostly in adults with a median age of 77 years).	Smith et al., 1996	102	Male	Control	100-plus Study
20679	15	Duplication	54921812	NA	NA	GRCh37	Excluded	These events are related to an acquired trisomy 15 already observed in hematological malignancies (rare, marked male predominance, found mostly in adults with a median age of 77 years).	Smith et al., 1996	77	Male	LOAD	PERADES
A-WCAP-WC003934-BL-COL-49239BL1	15	Duplication	2760315	2513895	14415147	GRCh38	Not excluded	These events are related to an acquired trisomy 15 already observed in hematological malignancies (rare, marked male predominance, found mostly in adults with a median age of 77 years).		79	Male	Control	ADSP
FR21322539	15	Duplication	10070027	NA	NA	GRCh37	Excluded	These events are related to an acquired trisomy 15 already observed in hematological malignancies (rare, marked male predominance, found mostly in adults with a median age of 77 years).	Smith et al., 1996	79	Male	LOAD	ADC-Amsterdam
C-RS-50517	15	Duplication	2889141	NA	NA	GRCh37	Excluded	trisomy 15 already observed in hematological malignancies (rare, marked male predominance, found mostly in adults with a median age of 77 years).	Smith et al., 1996	86	Male	Control	ADSP
A-ACT-AC001924	15	Duplication	3384348	NA	NA	GRCh37	Excluded	These events are related to an acquired trisomy 15 already observed in hematological malignancies (rare, marked male predominance, found mostly in adults with a median age of 77 years).	Smith et al., 1996	84	Male	LOAD	ADSP
A-ACT-AC001055	17	Deletion	21442815	NA	NA	GRCh37	Excluded	acute non lymphocytic leuemia/myelodysplastic syndromes (AML/MDS), chronic myelogenous leukemia (CML) in blast crisis	Rajan et al., 2015	83	Female	Control	ADSP
A-NCRD-NC009851-CL-NCR-8008288019 [3]	18	Deletion	10968785	NA	NA	GRCh38	Excluded	monosomy 18	Younes et al., 1994	70	Male	LOAD	ADSP
A-RAS-RA000176-CL-NCR-8011737027 [4]	18	Deletion	13646406	NA	NA	GRCh38	Excluded	Complex chromosomal rearrangement implicated the major part of the chr18	Younes et al., 1994	86	Female	LOAD	ADSP

[3] This sample carries two CNVs selected as potential large mosaic: a duplication on chromosome 11 and a deletion on chromosome 18

[4] This sample alternates deletions and duplications on chromosome 18

Sample	Chr	Type of event	Cumulate CNV size call	Start	End	Referent genome	Exclusion decision	Comments	Literature review	Age	Gender	Status	Cohort
20266	19	Duplication	5341695	NA	NA	GRCh37	Not excluded	Manual vizualisation showed that this individual carries several duplications on chromosome 19 appearing in high polymorphic region, which does not constitute a reason of exclusion	Berger et a., 1992	74	Male	LOAD	PERADES
11062	19	Duplication	3322757	NA	NA	GRCh37	Not excluded	Manual vizualisation showed that this individual carries several duplications on chromosome 19 appearing in high polymorphic region, which does not constitute a reason of exclusion	Berger et a., 1992	80	Female	Control	Rotterdam Study
12651	19	Duplication	3262352	NA	NA	GRCh37	Not excluded	Manual vizualisation showed that this individual carries several duplications on chromosome 19 appearing in high polymorphic region, which does not constitute a reason of exclusion	Berger et a., 1992	76	Male	Control	Rotterdam Study
A-WCAP-WC000258	20	Deletion	13318877	31035398	44354275	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	86	Female	Control	ADSP
C-FHS-51745	20	Deletion	8632616	33203845	43882375	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	83	Female	Control	ADSP
A-ADC-AD000856	20	Deletion	16265777	32981616	49247393	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	90	Male	Control	ADSP
A-ADC-AD009324	20	Deletion	20820057	35425904	57899516	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	90	Male	Control	ADSP
C-RS-50755	20	Deletion	13057606	36393597	49458439	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	90	Male	Control	ADSP
A-LOAD-LD002344-BL-NCR-8008288154	20	Deletion	7784399	32447226	46550585	GRCh38	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	86	Female	LOAD	ADSP
B00G7JW	20	Deletion	24198333	31015992	55214325	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	87	Female	Control	ADES-FR
10828	20	Deletion	12474100	30897614	45362502	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	95	Male	Control	Rotterdam Study

Sample	Chr	Type of event	Cumulate CNV size call	Start	End	Referent genome	Exclusion decision	Comments	Literature review	Age	Gender	Status	Cohort
15R0185	20	Deletion	6517448	32974735	44527790	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	100	Male	LOAD	100-plus Study
15R3660	20	Deletion	10714706	35796488	50217935	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	102	Female	Control	100-plus Study
FR16641041	20	Deletion	11491722	39316579	50808301	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	94	Female	Control	90-plus Study
6260	20	Deletion	13103758	30946467	44050225	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	77	Female	LOAD	PERADES
24063	20	Deletion	22413624	32161987	56228165	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	na	Male	Control	PERADES
PMAD214	20	Deletion	4578350	30818619	47858745	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	na	Female	LOAD	PERADES
A-MAP-MA001233	20	Deletion	9997242	35125107	49626876	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	90	Female	Control	ADSP
B00FNA7	20	Deletion	3933893	30917992	49185098	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	84	Female	LOAD	ADES-FR
20428	20	Deletion	9934811	32161987	50734259	GRCh37	Excluded	Acquired deletion of the long arm of chromosome 20 already described in various types of hematological disorders	Wang et al., 1998	85	Male	LOAD	PERADES
A-ADC-AD005681 [5]	20	Deletion	45082728	68349	52105741	GRCh37	Excluded	Complex chromosomal rearrangement implicated the major part of the chr20	Stevens-Kroef et al., 2000	86	Male	Control	ADSP
A-ADC-AD005681 [5]	20	Duplication	9549605	52185654	62919174	GRCh37	Excluded	Complex chromosomal rearrangement implicated the major part of the chr20	Stevens-Kroef et al., 2000	86	Male	Control	ADSP
A-WCAP-WC003965-BL-COL-41735BL1	21	Deletion	2663828	10454232	13118060	GRCh38	Excluded	Manual visualization show a large CNV of 2,663,828 bp (chr21:10454232-13118060) at the centromeric region.		80	Male	LOAD	ADSP
A-MAP-MA000096	22	Duplication	4231153	19422258	50926172	GRCh37	Excluded	Acquired duplication of chromosome 22 long arm already discribed in acute mueloid leukemia.	Vosberg et Greif, 2019	88	Female	Control	ADSP

Table S2: List of large mosaic CNV detected in the whole dataset

EOAD: Early Onset Alzheimer Disease, LOAD: Late Onset Alzheimer Disease

[5] This sample alternates deletions and duplications on chromosome 20.

		N total			N carriers of CNVs (DEL/DUP)			Dosage analysis all AD vs CTRL		Deletion analysis all AD vs CTRL		Duplication analysis all AD vs CTRL		
Chr	Gene	EOAD	LOAD	CTRL	EOAD	LOAD	CTRL	OR [95%CI]	p-value	OR [95%CI]	p-value	OR [95%CI]	p-value	Transcripts
1	PRAMEF26*	2911	7302	6461	8 (8 / 0)	2 (2 / 0)	2 (1 / 1)	0,18 [0,02 ; 0,78]	1,90E-02	4,43 [1,04;41,07]	4,29E-02	0,21 [0,001;3,95]	2,98E-01	NM_001306072.3
2	ADI1	4077	8457	9559	5 (0 / 5)	1 (0 / 1)	1 (1 /0)	12,64 [1,53 ; 1644,41]	1,32E-02	0,25 [0,002 ; 4,77]	3,62E-01	9,92 [1,73 ; 1293,57]	3,19E-02	NM_001306077.2 ; NM_018269.4
8	MTUS1**	4077	8457	9559	54 (54/0)	38 (37/1)	57 (57/0)	0,80 [0,59;1,08]	1,47E-01			2,29 [0,12;333,88]	5,91E-01	NM_001001924.3; NM_001363057.2; NM_001363058.2; NM_001363059.2; NM_001363061.2
10	MBL2	4077	8457	9559	2 (2 / 0)	25 (25 / 0)	45 (44 / 1)	1,95 [1,24 ; 3,10]	3,49E-03	0,47 [0,29;0,75]	1,56E-03	0,25 [0,002;4,77]	3,62E-01	NM_000242.3; NM_001378373.1; NM_001378374.1
17	FADS6	4077	8457	9559	16 (15 / 1)	11 (11 / 0)	5 (5 / 0)	0,31 [0,12 ; 0,7]	3,58E-03	3,68 [1,58;10,31]	1,67E-03	2,29 [0,12;333,88]	5,91E-01	NM_178128.6
19	PLIN4**	4077	8457	9559	20 (19 / 1)	10 (10 / 0)	12 (12 / 0)	0,59 [0,30 ; 1,10]	1,01E-01			1,80 [0,95 ; 3,62]	7,14E-02	NM_001393888.1 ; NM_001393889.1 ; NM_001393890.1 ; NM_001393891.1; NM_001367868.2
19	KLC3	4077	8457	9559	24 (1 / 23)	33 (1 / 32)	16 (0 / 16)	2,33 [1,4 ; 4,06]	8,40E-04	3,81 [0,31;526,00]	3,25E-01	2,57 [1,52;4,60]	2,97E-04	NM_177417.3
19	ERCC2	4077	8457	9559	29 (1 / 28)	43 (2 / 41)	20 (0 / 20)	2,31 [1,47 ; 3,76]	2,14E-04	5,34 [0,52;717,94]	1,81E-01	2,59 [1,62;4,36]	4,58E-05	NM_001130867.2
19	ERCC2	4077	8457	9559	29 (1 / 28)	42 (2 / 40)	20 (0 / 20)	2,27 [1,44 ; 3,71]	2,88E-04			2,56 [1,59;4,30]	6,34E-05	NM_000400.4
22	ZNF74	4077	8457	9559	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	0,28 [0,09 ; 0,78]	1,30E-02	6,87 [0,73;909,82]	1,01E-01	0,36 [0,11;1,06]	6,37E-02	NM_001256523.2; NM_001256525.2; NM_003426.4; NM_001256524.2
22	SCARF2	4077	8457	9559	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	0,28 [0,09 ; 0,78]	1,30E-02	6,87 [0,73;909,82]	1,01E-01	0,36 [0,11;1,06]	6,37E-02	NM_153334.7; NM_182895.5
22	KLHL22	4077	8457	9559	4 (4 / 0)	4 (0 / 4)	8 (0 / 8)	0,31 [0,09 ; 0,86]	2,33E-02	6,87 [0,73;909,82]	1,01E-01	0,40 [0,12;1,22]	1,09E-01	NM_032775.4
22	MED15	4077	8457	9559	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	0,28 [0,09 ; 0,78]	1,30E-02	6,87 [0,73;909,82]	1,01E-01	0,36 [0,11;1,06]	6,37E-02	NM_001003891.3; NM_001293234.2; NM_001293235.2; NM_001293236.2; NM_001293237.2; NM_015889.5
22	PI4KA	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,31 [0,1 ; 0,77]	1,08E-02	8,39 [0,95;1101,69]	5,67E-02	0,40 [0,13;1,08]	7,19E-02	NM_001362862.2; NM_001362863.2; NM_058004.4
22	SERPIND1	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,31 [0,1 ; 0,77]	1,08E-02	8,39 [0,95;1101,69]	5,67E-02	0,40 [0,13;1,08]	7,19E-02	NM_000185.4
22	SNAP29	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,31 [0,1 ; 0,77]	1,08E-02	8,39 [0,95;1101,69]	5,67E-02	0,40 [0,13;1,08]	7,19E-02	NM_004782.4
22	CRKL	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,31 [0,1 ; 0,77]	1,08E-02	8,39 [0,95;1101,69]	5,67E-02	0,40 [0,13;1,08]	7,19E-02	NM_005207.4
22	LZTR1	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,31 [0,1 ; 0,77]	1,08E-02	8,39 [0,95;1101,69]	5,67E-02	0,40 [0,13;1,08]	7,19E-02	NM_006767.4
22	THAP7	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,31 [0,1 ; 0,77]	1,08E-02	8,39 [0,95;1101,69]	5,67E-02	0,40 [0,13;1,08]	7,19E-02	NM_030573.3; NM_001008695.1
22	P2RX6	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,31 [0,1 ; 0,77]	1,08E-02	8,39 [0,95;1101,69]	5,67E-02	0,40 [0,13;1,08]	7,19E-02	NM_001159554.2; NM_001349874.2; NM_001349875.2; NM_001349876.2; NM_001394691.1; NM_001394692.1; NM_001394693.1; NM_001394694.1; NM_001394695.1; NM_001394696.1; NM_001394697.1; NM_005446.5
22	SLC7A4	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,31 [0,1 ; 0,77]	1,08E-02	8,39 [0,95;1101,69]	5,67E-02	0,40 [0,13;1,08]	7,19E-02	NM_004173.3
22	LRRC74B	3917	7799	8016	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,29 [0,1 ; 0,72]	7,21E-03	7,53 [0,85;988,39]	7,39E-02	0,36 [0,12;0,97]	4,38E-02	NM_001291006.2

Table 3. Results from dosage analysis and subsequent analyses on deletions and duplications

Chr: chromosome; N: number, EOAD: early onset Alzheimer disease, LOAD: late onset Alzheimer disease; CTRL: controls

[1]OR should be interpreted for an increase of transcript copies by 1, i.e. a OR > 1 (resp. OR <1) means that duplications (resp. deletions) are associated with a higher disease risk

[2]non adjusted p-values, selected based on FRD < 10%

* gene overlapping with repeats or duplicated gene in the genome (despite CNVs not overlapping >50% with repeats)

** gene for which the signal in dosage analysis is driven by deletion whereas transcripts are not in set A

		N total			N carriers of deletions			Firth regression					
								EOAD vs CTRL			all AD vs CTRL		
Chr	Gene	EOAD	LOAD	CTRL	EOAD	LOAD	CTRL	OR [95%CI]	p-value	FDR	OR [95%CI]	p-value	Transcripts
1	PRAMEF5*	2994	7878	7683	8	2	1	14,58 [3,29 ; 136,95]	1,68E-04	6,11E-02	4,95 [1,16 ; 45,88]	2,80E-02	NM_001013407.5
1	PRAMEF26*	2911	7302	6461	8	2	1	12,61 [2,84; 118,46]	3,96E-04	1,08E-01	4,43 [1,04 ; 41,07]	4,29E-02	NM_001306072.3
1	RCC1	4048	8396	9359	27	11	29	2,16 [1,28; 3,65]	4,29E-03	1,85E-01	0,98 [0,61 ; 1,60]	9,39E-01	NM_001048199.3
		4077	8457	9559	27	11	29	2,19 [1,30 ; 3,70]	3,65E-03	1,85E-01	0,99 [0,62 ; 1,62]	9,84E-01	NM_001381865.2; NM_001381866.2
1	SUCO	4077	8457	9559	4	3	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	11,44 [1,39 ; 1485,47]	1,80E-02	NM_001282750.2; NM_001282751.2; NM_014283.5; NM_016227.4
1	NPL	4077	8457	9559	15	13	12	2,91 [1,38 ; 6,26]	5,14E-03	1,94E-01	1,74 [0,91 ; 3,51]	9,24E-02	NM_001200050.2; NM_001200051.2; NM_001200052.2; NM_001200056.2; NM_030769.3
4	MMRN1	4077	8457	9559	5	4	0	25,82 [2,93 ; 3389,88]	1,41E-03	1,85E-01	14,50 [1,84 ; 1869,33]	5,71E-03	NM_001371403.1; NM_007351.3
5	WDR36	4077	8457	9559	11	7	4	6,00 [2,14 ; 20,10]	5,41E-04	1,18E-01	3,14 [1,22 ; 10,08]	1,62E-02	NM_139281.3
10	CCDC7	4077	8457	9559	4	2	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	9,92 [1,17 ; 1293,57]	3,19E-02	NM_001321115.2; NM_001395015.1
10	MBL2	4077	8457	9559	2	25	44	0,13 [0,03 ; 0,39]	2,02E-05	1,82E-02	0,47 [0,29 ; 0,75]	1,56E-03	NM_000242.3 ; NM_001378373.1; NM_001378374.1
11	NELL1	4077	8457	9559	4	1	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	8,39 [0,95 ; 1101,69]	5,67E-02	NM_001288713.1; NM_001288714.1; NM_006157.5; NM_201551.2
11	AHNAK	4077	8457	9559	4	0	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	6,87 [0,73 ; 909,82]	1,01E-01	NM_001346445.2; NM_001346446.2; NM_001620.3
16	METTL9	4077	8457	9559	12	10	7	3,92 [1,62 ; 10,19]	2,61E-03	1,85E-01	2,29 [1,05 ; 5,62]	3,73E-02	NM_001077180.3; NM_001288659.2; NM_001288660.2; NM_016025.5
16	IGSF6	4077	8457	9559	12	10	7	3,92 [1,62 ; 10,19]	2,61E-03	1,85E-01	2,29 [1,05 ; 5,62]	3,73E-02	NM_005849.4
16	OTOA	4077	8457	9559	12	10	7	3,92 [1,62 ; 10,19]	2,61E-03	1,85E-01	2,29 [1,05 ; 5,62]	3,73E-02	NM_001161683.2; NM_144672.4; NM_170664.3
17	FADS6	4077	8457	9559	15	11	5	6,63 [2,65 ; 19,34]	3,34E-05	1,82E-02	3,68 [1,58 ; 10,32]	1,65E-03	NM_178128.6
22	ZNF74	4077	8457	9559	4	0	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	6,87 [0,73 ; 909,82]	1,01E-01	NM_001256523.2; NM_001256524.2; NM_001256525.2; NM_003426.4
22	SCARF2	4077	8457	9559	4	0	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	6,87 [0,73 ; 909,82]	1,01E-01	NM_153334.7; NM_182895.5
22	KLHL22	4077	8457	9559	4	0	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	6,87 [0,73 ; 909,82]	1,01E-01	NM_032775.4
22	MED15	4077	8457	9559	4	0	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	6,87 [0,73 ; 909,82]	1,01E-01	NM_001003891.3; NM_001293234.2; NM_001293235.2; NM_001293236.2; NM_001293237.2; NM_015889.5
22	PI4KA	4077	8457	9559	4	1	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	8,39 [0,95 ; 1101,69]	5,67E-02	NM_001362862.2; NM_001362863.2; NM_058004.4
22	SERPIND1	4077	8457	9559	4	1	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	8,39 [0,95 ; 1101,69]	5,67E-02	NM_000185.4
22	SNAP29	4077	8457	9559	4	1	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	8,39 [0,95 ; 1101,69]	5,67E-02	NM_004782.4
22	CRKL	4077	8457	9559	4	1	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	8,39 [0,95 ; 1101,69]	5,67E-02	NM_005207.4
22	AIFM3	4077	8457	9559	4	1	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	8,39 [0,95 ; 1101,69]	5,67E-02	NM_001018060.3; NM_001146288.2; NM_001386814.1; NM_144704.3
22	LZTR1	4077	8457	9559	4	1	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	8,39 [0,95 ; 1101,69]	5,67E-02	NM_006767.4
22	THAP7	4077	8457	9559	4	1	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	8,39 [0,95 ; 1101,69]	5,67E-02	NM_001008695.1; NM_030573.3
22	P2RX6	4077	8457	9559	4	1	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	8,39 [0,95 ; 1101,69]	5,67E-02	NM_001159554.2; NM_001349874.2; NM_001349875.2; NM_001349876.2; NM_001394691.1; NM_001394692.1; NM_001394693.1; NM_001394694.1; NM_001394695.1; NM_001394696.1; NM_001394697.1; NM_005446.5
22	SLC7A4	4077	8457	9559	4	1	0	21,12 [2,25 ; 2799,03]	4,75E-03	1,85E-01	8,39 [0,95 ; 1101,69]	5,67E-02	NM_004173.3

Table S4: Results for deletion analysis (without adjustment)

[1]OR should should be interpreted for a deletion presence versus absence, i.e. a OR > 1 (resp. OR <1) means that carrying a deletion is associated with a higher (resp. lower) disease risk.

* gene overlapping with repeats or duplicated gene in the genome (despite CNVs not overlapping >50% with repeats)

		N total			N carriers of complete duplications			Firth regression					
								EOAD vs CTRL			all AD vs CTRL		
Chr	Gene	EOAD	LOAD	CTRL	EOAD	LOAD	CTRL	OR [95%CI]	p-value	FDR	OR [95%CI]	p-value	Transcripts
7	ZNF862	4077	8457	9559	6	8	0	30,52 [3,61 ; 3980,95]	4,17E-04	1,83E-01	22,14 [2,97 ; 2829,46]	3,29E-04	NM_001099220.3
19	KLC3	4077	8457	9559	23	32	16	3,35 [1,80 ; 6,40]	1,57E-04	1,03E-01	2,57 [1,52 ; 4,60]	2,97E-04	NM_177417.3
19	ERCC2	4077	8457	9559	28	41	20	3,27 [1,86 ; 5,86]	4,09E-05	5,37E-02	2,59 [1,61 ; 4,36]	4,58E-05	NM_001130867.2
		4077	8457	9559	28	40	20				2,56 [1,59 ; 4,29]	6,34E-05	NM_000400.4

Table S5: Results for duplication analysis

[1]OR should should be interpreted for a duplication presence versus absence, i.e. a OR > 1 (resp. OR <1) means that carrying a duplication is associated with a higher (resp. lower) disease risk.

		e4 non carrier	e4 heterozygous	e4 homozygous	p-value[1]
CTRL	N duplication non carriers	7528	1819	100	9,60E-05
	N duplication carriers (%)	7 (0,0009%)	11 (0,006%)	1 (0,0099%)	
LOAD	duplication non carriers	4629	3394	384	2,67E-04
	N duplication carriers (%)	11 (0,0024%)	23 (0,0067%)	6 (0,0154%)	
EOAD	duplication non carriers	1700	1683	660	4,21E-03
	N duplication carriers (%)	4 (0,0023%)	16 (0,0094%)	8 (0,012%)	

Table S6. Repartition of APOE e4 status among carriers and non carriers of the duplication at KLC3-ERCC2 locus on chromosome 19 (based on NM_001130867.2 transcript) for each disease status group

[1] p-values were obtained using a Fisher exact test

Chr	Gene	N total			N carriers of CNVs (DEL/DUP)			Dosage analysis				Transcripts
								EOAD vs CTRL		all AD vs CTRL		
		EOAD	LOAD	CTRL	EOAD	LOAD	CTRL	OR [95%CI]	p-value[4]	OR [95%CI]	p-value	
1	PRAMEF26*	2906	7298	6448	8 (8 / 0)	2 (2 / 0)	2 (1 / 1)	0,10 [0,01 ; 0,57]	7,70E-03	0,23 [0,02 ; 1,04]	5,74E-02	NM_001306072.3
2	ADI1	4071	8447	9466	5 (0 / 5)	1 (0 / 1)	1 (1 /0)	41,64 [4,49 ; 5511,47]	3,55E-04	13,17 [1,53 ; 1723,44]	1,38E-02	NM_001306077.2 ; NM_018269.4
8	MTUS1**	4071	8447	9466	54 (54/0)	38 (37/1)	57 (57/0)	0,41 [0,28;0,60]	4,03E-06	0,80 [0,58;1,08]	1,46E-01	NM_001001924.3; NM_001363057.2; NM_001363058.2; NM_001363059.2; NM_001363061.2
10	MBL2	4071	8447	9466	2 (2 / 0)	25 (25 / 0)	45 (44 / 1)	5,66 [2,01 ; 22,08]	3,45E-04	1,81 [1,14 ; 2,93]	1,20E-02	NM_000242.3; NM_001378373.1; NM_001378374.1
17	FADS6	4071	8447	9466	16 (15 / 1)	11 (11 / 0)	5 (5 / 0)	0,13 [0,05 ; 0,49]	1,69E-05	0,25 [0,09 ; 0,57]	6,72E-04	NM_178128.6
19	PLIN4**	4071	8447	9466	20 (19 / 1)	10 (10 / 0)	12 (12 / 0)	0,30 [0,14 ; 0,63]	1,79E-03	0,62 [0,31 ; 1,20]	1,59E-01	NM_001393888.1 ; NM_001393889.1 ; NM_001393890.1 ; NM_001393891.1; NM_001367868.2
19	KLC3	4071	8447	9466	24 (1 / 23)	32 (1 / 31)	16 (0 / 16)	1,37 [0,69 ; 2,75]	3,65E-01	1,43 [0,83 ; 2,57]	1,98E-01	NM_177417.3
19	ERCC2	4071	8447	9466	29 (1 / 28)	42 (2 / 40)	19 (0 / 19)	1,58 [0,85 ; 2,99]	1,48E-01	1,47 [0,90 ; 2,49]	1,25E-01	NM_001130867.2
19	ERCC2	4071	8447	9466	29 (1 / 28)	41 (2 / 39)	19 (0 / 19)			1,47 [0,90 ; 2,48]	1,30E-01	NM_000400.4
22	ZNF74	4071	8447	9466	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	0,04 [0,0003 ; 0,37]	1,25E-03	0,30 [0,09 ; 0,87]	2,51E-02	NM_001256523.2; NM_001256525.2; NM_003426.4; NM_001256524.2
22	SCARF2	4071	8447	9466	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	0,04 [0,0003 ; 0,37]	1,25E-03	0,30 [0,09 ; 0,87]	2,51E-02	NM_153334.7; NM_182895.5
22	KLHL22	4071	8447	9466	4 (4 / 0)	4 (0 / 4)	8 (0 / 8)	0,05 [0,0003 ; 0,42]	2,54E-03	0,36 [0,10 ; 1,03]	5,77E-02	NM_032775.4
22	MED15	4071	8447	9466	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	0,04 [0,0003 ; 0,37]	1,25E-03	0,30 [0,09 ; 0,87]	2,51E-02	NM_001003891.3; NM_001293234.2; NM_001293235.2; NM_001293236.2; NM_001293237.2; NM_015889.5
22	PI4KA	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,04 [0,0003 ; 0,39]	1,75E-03	0,38 [0,12 ; 0,98]	4,61E-02	NM_001362862.2; NM_001362863.2; NM_058004.4
22	SERPIND1	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,04 [0,0003 ; 0,39]	1,75E-03	0,38 [0,12 ; 0,98]	4,61E-02	NM_000185.4
22	SNAP29	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,04 [0,0003 ; 0,39]	1,75E-03	0,38 [0,12 ; 0,98]	4,61E-02	NM_004782.4
22	CRKL	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,04 [0,0003 ; 0,39]	1,75E-03	0,38 [0,12 ; 0,98]	4,61E-02	NM_005207.4
22	LZTR1	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,04 [0,0003 ; 0,39]	1,75E-03	0,38 [0,12 ; 0,98]	4,61E-02	NM_006767.4
22	THAP7	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,04 [0,0003 ; 0,39]	1,75E-03	0,38 [0,12 ; 0,98]	4,61E-02	NM_030573.3; NM_001008695.1
22	P2RX6	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,04 [0,0003 ; 0,39]	1,75E-03	0,38 [0,12 ; 0,98]	4,61E-02	NM_001159554.2; NM_001349874.2; NM_001349875.2; NM_001349876.2; NM_001394691.1; NM_001394692.1; NM_001394693.1; NM_001394694.1; NM_001394695.1; NM_001394696.1; NM_001394697.1; NM_005446.5
22	SLC7A4	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,04 [0,0003 ; 0,39]	1,75E-03	0,38 [0,12 ; 0,98]	4,61E-02	NM_004173.3
22	LRRC74B	3911	7795	7975	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,05 [0,0004 ; 0,41]	1,87E-03	0,35 [0,11 ; 0,93]	3,45E-02	NM_001291006.2

Table S7.a APOE4-adjusted dosage analysis

* gene overlapping with repeats or duplicated gene in the genome (despite CNVs not overlapping >50% with repeats)

** gene for which the signal in dosage analysis is driven by deletion whereas transcripts are not in set A

		N total			N carriers of CNVs (DEL/DUP)			Deletion analysis				Duplication analysis				
								EOAD vs CTRL		all AD vs CTRL		EOAD vs CTRL		all AD vs CTRL		
Chr	Gene	EOAD	LOAD	CTRL	EOAD	LOAD	CTRL	OR [95%CI]	p-value	OR [95%CI]	p-value	OR [95%CI]	p-value	OR [95%CI]	p-value	Transcripts
1	PRAMEF26*	2906	7298	6448	8 (8 / 0)	2 (2 / 0)	2 (1 / 1)	6,54 [1,70;97,89]	9,04E-03	3,73 [0,79;35,96]	1,03E-01	1,50 [0,01;28,08]	8,12E-01	0,34 [0,002;6,37]	4,78E-01	NM_001306072.3
2	ADI1	4071	8447	9466	5 (0 / 5)	1 (0 / 1)	1 (1 /0)	1,51 [0,01 ; 28,36]	8,07E-01	0,40 [0,003 ; 7,44]	5,46E-01	41,29 [4,43;5468,10]	3,99E-04	11,53 [1,28 ; 1519,57]	2,54E-02	NM_001306077.2 ; NM_018269.4
8	MTUS1**	4071	8447	9466	54 (54/0)	38 (37/1)	57 (57/0)	related transcripts are not in set A				duplication in EOAD nor in CT		2,29 [0,12;333,88]	5,91E-01	NM_001001924.3; NM_001363057.2; NM_001363058.2; NM_001363059.2; NM_001363061.2
10	MBL2	4071	8447	9466	2 (2 / 0)	25 (25 / 0)	45 (44 / 1)	0,14 [0,03;0,44]	1,75E-04	0,51 [0,31;0,83]	6,94E-03	1,51 [0,01;28,36]	8,07E-01	0,40 [0,003;7,44]	5,46E-01	NM_000242.3; NM_001378373.1; NM_001378374.1
17	FADS6	4071	8447	9466	16 (15 / 1)	11 (11 / 0)	5 (5 / 0)	8,68 [3,31;26,05]	7,33E-06	4,31 [1,82;12,24]	5,16E-04	0,65 [0,03 ; 95,17]	8,01E-01	0,31 [0,02;45,87]	5,24E-01	NM_178128.6
19	PLIN4**	4071	8447	9466	20 (19 / 1)	10 (10 / 0)	12 (12 / 0)	related transcripts are not in set A				3,89 [1,81 ; 8,53]	5,16E-04	1,74 [0,89 ; 3,58]	1,07E-01	NM_001393888.1 ; NM_001393889.1 ; NM_001393890.1 ; NM_001393891.1; NM_001367868.2
19	KLC3	4071	8447	9466	24 (1 / 23)	32 (1 / 31)	16 (0 / 16)	2,98 [0,16;434,75]	4,73E-01	3,95 [0,27;565,83]	3,37E-01	1,47 [0,74;3,00]	2,73E-01	1,56 [0,89;2,86]	1,24E-01	NM_177417.3
19	ERCC2	4071	8447	9466	29 (1 / 28)	42 (2 / 40)	19 (0 / 19)	2,98 [0,16;434,75]	4,73E-01	4,53 [0,38;631,08]	2,62E-01	1,69 [0,90;3,23]	1,04E-01	1,62 [0,97;2,81]	6,57E-02	NM_001130867.2
19	ERCC2	4071	8447	9466	29 (1 / 28)	41 (2 / 39)	19 (0 / 19)							1,61 [0,96;2,80]	6,86E-02	NM_000400.4
22	ZNF74	4071	8447	9466	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	17,49 [1,56;2428,63]	1,67E-02	5,16 [0,49;705,22]	1,98E-01	0,13 [0,001;1,13]	6,80E-02	0,37 [0,10;1,13]	8,17E-02	NM_001256523.2; NM_001256525.2; NM_003426.4; NM_001256524.2
22	SCARF2	4071	8447	9466	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	17,49 [1,56;2428,63]	1,67E-02	5,16 [0,49;705,22]	1,98E-01	0,13 [0,001;1,13]	6,80E-02	0,37 [0,10;1,13]	8,17E-02	NM_153334.7; NM_182895.5
22	KLHL22	4071	8447	9466	4 (4 / 0)	4 (0 / 4)	8 (0 / 8)	17,49 [1,56;2428,63]	1,67E-02	5,16 [0,49;705,22]	1,98E-01	0,18 [0,001;1,59]	1,47E-01	0,45 [0,12;1,42]	1,75E-01	NM_032775.4
22	MED15	4071	8447	9466	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	17,49 [1,56;2428,63]	1,67E-02	5,16 [0,49;705,22]	1,98E-01	0,13 [0,001;1,13]	6,80E-02	0,37 [0,10;1,13]	8,17E-02	NM_001003891.3; NM_001293234.2; NM_001293235.2; NM_001293236.2; NM_001293237.2; NM_015889.5
22	PI4KA	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	17,49 [1,56;2428,63]	1,67E-02	5,83 [0,60;783,93]	1,48E-01	0,16 [0,001;1,31]	9,90E-02	0,48 [0,15;1,33]	1,61E-01	NM_001362862.2; NM_001362863.2; NM_058004.4
22	SERPIND1	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	17,49 [1,56;2428,63]	1,67E-02	5,83 [0,60;783,93]	1,48E-01	0,16 [0,001;1,31]	9,90E-02	0,48 [0,15;1,33]	1,61E-01	NM_000185.4
22	SNAP29	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	17,49 [1,56;2428,63]	1,67E-02	5,83 [0,60;783,93]	1,48E-01	0,16 [0,001;1,31]	9,90E-02	0,48 [0,15;1,33]	1,61E-01	NM_004782.4
22	CRKL	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	17,49 [1,56;2428,63]	1,67E-02	5,83 [0,60;783,93]	1,48E-01	0,16 [0,001;1,31]	9,90E-02	0,48 [0,15;1,33]	1,61E-01	NM_005207.4
22	LZTR1	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	17,49 [1,56;2428,63]	1,67E-02	5,83 [0,60;783,93]	1,48E-01	0,16 [0,001;1,31]	9,90E-02	0,48 [0,15;1,33]	1,61E-01	NM_006767.4
22	THAP7	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	17,49 [1,56;2428,63]	1,67E-02	5,83 [0,60;783,93]	1,48E-01	0,16 [0,001;1,31]	9,90E-02	0,48 [0,15;1,33]	1,61E-01	NM_030573.3; NM_001008695.1
22	P2RX6	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	17,49 [1,56;2428,63]	1,67E-02	5,83 [0,60;783,93]	1,48E-01	0,16 [0,001;1,31]	9,90E-02	0,48 [0,15;1,33]	1,61E-01	NM_001159554.2; NM_001349874.2; NM_001349875.2; NM_001349876.2; NM_001394691.1; NM_001394692.1; NM_001394693.1; NM_001394694.1; NM_001394695.1; NM_001394696.1; NM_001394697.1; NM_005446.5
22	SLC7A4	4071	8447	9466	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	17,49 [1,56;2428,63]	1,67E-02	5,83 [0,60;783,93]	1,48E-01	0,16 [0,001;1,31]	9,90E-02	0,48 [0,15;1,33]	1,61E-01	NM_004173.3
22	LRRC74B	3911	7795	7975	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	15,16 [1,33;2111,62]	2,53E-02	5,15 [0,53;695,32]	1,87E-01	0,14 [0,001;1,14]	7,06E-02	0,43 [0,13;1,21]	1,11E-01	NM_001291006.2

Table S7.b APOE4-adjusted dosage analysis

* gene overlapping with repeats or duplicated gene in the genome (despite CNVs not overlapping >50% with repeats)

** gene for which the signal in dosage analysis is driven by deletion whereas transcripts are not in set A

	Position in hg19			Number of individuals included in LoF analysis			Number of carriers of LoF variants (including CNV-deletion)			Firth regression				Reason of inclusion	Transcripts used for deletion definition
										EOAD versus CTRL		all AD versus CTRL			
Gene	Chr	Start	End	EOAD	LOAD	CTRL	EOAD	LOAD	CTRL	OR [95%CI]	p-value	OR [95%CI]	p-value		
ADI1	2	3,501,690	3,523,350	4077	8457	9559	1 (0)	2 (0)	5 (1)	0.64 [0.07 ; 3.20]	0.6130	0.49 [0.11 ; 1.82]	0.2824	Prioritized genes from dosage analysis(FDR < 10%)	NM_001306077.2; NM_018269.4
ADAM17	2	9,628,614	9,695,959	4077	8457	9559	2 (0)	1 (1)	4 (0)	1.30 [0.23 ; 5.87]	0.7411	0.59 [0.13 ; 2.43]	0.4612	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_001382777.1; NM_001382778.1; NM_003183.6
WDR12	2	203,738,983	203,776,000	4077	8457	9558	0 (0)	3 (3)	2 (0)	0.47 [0.003 ; 5.76]	0.5964	1.07 [0.21 ; 6.41]	0.9370	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_001371664.1; NM_018256.4
MME	3	154,741,990	154,901,518	7077	8457	9559	10 (0)	24 (3)	11 (0)	2.14 [0.91 ; 4.99]	0.0795	2.29 [1.21 ; 4.67]	0.0099	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_001354644.1; NM_000902.5; NM_001354642.2; NM_001354643.1; NM_007287.4; NM_007288.3; NM_007289.4
IDUA	4	980,784	998,352	4068	8423	9510	7 (1)	17 (2)	14 (0)	1.21 [0.47 ; 2.84]	0.6760	1.29 [0.68 ; 2.52]	0.4425	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_000203.5; NM_001363576.1
CLNK	4	10,488,018	10,686,390	4077	8457	9559	2 (0)	5 (0)	1 (1)	3.91 [0.52 ; 42.72]	0.1785	3.81 [0.83 ; 36.17]	0.0882	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_052964.4
RASGEF1C	5	179,527,794	179,636,000	4077	8457	9559	1 (1)	2 (0)	3 (1)	1.00 [0.10 ; 6.11]	0.9962	0.76 [0.16 ; 3.59]	0.7203	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_175062.4
HLA-DQA1	6	32,605,182	32,611,460	4053	8427	9508	14 (13)	15 (14)	25 (25)	1.33 [0.68 ; 2.51]	0.3875	0.88 [0.52 ; 1.51]	0.6405	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_002122.5
EPDR1	7	37,960,241	37,991,530	4077	8453	9558	0 (0)	0 (0)	3 (1)	0.33 [0.003 ; 3.45]	0.4095	0.11 [0.001 ; 1.12]	0.0645	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_001242946.2; NM_017549.5; NM_001242948.2
CTSB	8	11,700,032	11,725,590	4077	8457	9559	7 (1)	10 (4)	3 (0)	5.03 [1.50 ; 20.71]	0.0089	3.82 [1.36 ; 14.51]	0.0092	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_001317237.2; NM_001384714.1; NM_001384723.1; NM_001384724.1; NM_001384725.1; NM_001384726.1; NM_001384727.1; NM_001384728.1; NM_001908.5; NM_147780.4; NM_147781.4; NM_147782.4; NM_147783.4
ABCA1	9	107,543,286	107,690,000	4077	8457	9559	13 (3)	12 (0)	5 (0)	5.77 [2.25 ; 17.06]	0.0002	3.54 [1.52 ; 9.95]	0.0025	Genes associated with AD in LoF analysis (Holstege et al)	NM_005502.4
MBL2	10	54,525,139	54,532,540	3604	7273	9199	3 (2)	19 (19)	40 (40)	0.22 [0.06 ; 0.57]	0.0008	0.47 [0.28 ; 0.78]	0.0032	Prioritized genes from dosage analysis(FDR < 10%)	NM_000242.3; NM_001378373.1; NM_001378374.1

	Position in hg19			Number of individuals included in LoF analysis			Number of carriers of LoF variants (including CNV-deletion)			Firth regression					
										EOAD versus CTRL		all AD versus CTRL		Reason of inclusion	Transcripts used for deletion definition
Gene	Chr	Start	End	EOAD	LOAD	CTRL	EOAD	LOAD	CTRL	OR [95%CI]	p-value	OR [95%CI]	p-value		
PLEKHA1	10	124,134,094	124,191,871	4076	8457	9559	0 (0)	0 (0)	2 (1)	0.47 [0.003 ; 5.76]	0.5966	0.15 [0.001 ; 1.87]	0.1520	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_001001974.4; NM_001195608.2; NM_001330178.2; NM_001377230.1; NM_001377231.1; NM_001377232.1; NM_001377234.1; NM_001377235.1; NM_001377237.1; NM_001377238.1; NM_001377240.1; NM_001377241.1; NM_001377242.1; NM_001377243.1; NM_001377244.1; NM_001377245.1; NM_001377246.1; NM_001377247.1; NM_001377248.1; NM_001377249.1; NM_001377250.1; NM_001377251.1; NM_001377252.1; NM_001377253.1; NM_001377254.1; NM_001377255.1; NM_001377256.1; NM_001377257.1; NM_001377258.1; NM_021622.5
MS4A4A	11	60,048,138	60,076,440	4077	8457	9559	1 (1)	3 (0)	1 (0)	2.34 [0.19 ; 28.88]	0.4671	2.29 [0.42 ; 22.82]	0.3519	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_001243266.2; NM_024021.4; NM_148975.3
SLC24A4	14	92,788,924	92,967,820	4077	8457	9559	1 (0)	2 (1)	1 (0)	2.34 [0.19 ; 28.88]	0.4671	1.78 [0.29 ; 18.36]	0.5417	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_001378620.1; NM_153646.4; NM_153647.4; NM_153648.4
APH1B	15	63,569,803	63,601,320	4077	8457	9559	3 (1)	6 (0)	3 (0)	2.34 [0.50 ; 11.06]	0.2666	2.07 [0.66 ; 8.29]	0.2223	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_001145646.2; NM_031301.4
DOC2A	16	30,016,834	30,022,340	3935	8101	8581	2 (2)	2 (1)	3 (2)	1.56 [0.26 ; 8.00]	0.5985	0.92 [0.22 ; 4.10]	0.9029	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_001282062.1; NM_001282063.2; NM_001282068.2; NM_003586.3
MAF	16	79,627,734	79,634,630	4077	8457	9558	0 (0)	0 (0)	2 (1)	0.47 [0.003 ; 5.76]	0.5964	0.15 [0.001 ; 1.87]	0.1519	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_001031804.3; NM_005360.5
PLCG2	16	81,812,895	81,996,290	4077	8457	9559	0 (0)	9 (3)	5 (2)	0.21 [0.002 ; 1.88]	0.1957	1.32 [0.47 ; 4.06]	0.6031	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_002661.5
PRDM7	16	90,122,973	90,143,730	4077	8457	9559	3 (0)	17 (0)	17 (1)	0.47 [0.12 ; 1.32]	0.1610	0.89 [0.47 ; 1.71]	0.7291	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_001098173.2
MYO15A	17	18,012,069	18,083,110	4077	8457	9559	16 (1)	27 (1)	40 (0)	0.95 [0.52 ; 1.66]	0.8743	0.82 [0.53 ; 1.26]	0.3608	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_016239.4
WNT3	17	44,839,871	44,896,050	4077	8454	9553	0 (0)	2 (1)	2 (0)	0.47 [0.003 ; 5.76]	0.5962	0.76 [0.12 ; 4.93]	0.7619	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_030753.5

	Position in hg19			Number of individuals included in LoF analysis			Number of carriers of LoF variants (including CNV-deletion)			Firth regression				Reason of inclusion	Transcripts used for deletion definition
										EOAD versus CTRL		all AD versus CTRL			
Gene	Chr	Start	End	EOAD	LOAD	CTRL	EOAD	LOAD	CTRL	OR [95%CI]	p-value	OR [95%CI]	p-value		
FADS6	17	72,873,473	72,889,705	4077	8457	9559	15 (15)	11 (11)	5 (5)	6.63 [2.65 ; 19.34]	3.34E-05	3.68 [1.58 ; 10.32]	0.0017	Prioritized genes from dosage analysis(FDR < 10%)	NM_178128.6
ABCA7	19	1,040,105	1,065,571	4077	8457	9559	35 (4)	44 (3)	36 (3)	2.29 [1.44 ; 3.65]	0.0006	1.66 [1.13 ; 2.49]	0.0090	Genes associated with AD in LoF analysis (Holstege et al)	NM_019112.4
KLC3	19	45,843,998	45,854,778	4077	8457	9554	11 (1)	20 (1)	30 (0)	0.88 [0.43 ; 1.69]	0.7174	0.79 [0.48 ; 1.30]	0.3464	Prioritized genes from dosage analysis(FDR < 10%)	NM_177417.3
ERCC2	19	45,854,649	45,873,845	4077	8457	9559	4 (1)	12 (2)	11 (0)	0.92 [0.27 ; 2.57]	0.8758	1.09 [0.52 ; 2.39]	0.8142	Prioritized genes from dosage analysis(FDR < 10%)	NM_000400.4; NM_001130867.2
SIGLEC11	19	50,452,241	50,464,420	4077	8455	9555	8 (7)	12 (8)	11 (6)	1.73 [0.69 ; 4.19]	0.2333	1.36 [0.67 ; 2.89]	0.3987	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_052884.3; NM_001135163.1
LILRB2	19	54,777,675	54,785,033	4077	8457	9559	1 (1)	3 (2)	4 (3)	0.78 [0.08 ; 4.22]	0.7896	0.76 [0.20 ; 2.95]	0.6848	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_001080978.4; NM_001278403.3; NM_001278404.3; NM_001278405.2; NM_005874.5; NM_001278406.2
RBCK1	20	388,942	412,783	4075	8431	9534	1 (0)	2 (2)	2 (0)	1.40 [0.13 ; 10.56]	0.7467	1.07 [0.21 ; 6.41]	0.9372	Genes associated with AD in GWAS analysis (Bellenguez et al)	NM_001323960.2; NM_001323956.2; NM_001323958.2; NM_006462.6; NM_031229.4
ZNF74	22	20,748,440	20,762,740	4077	8458	9559	4 (4)	1 (0)	4 (0)	2.35 [0.61 ; 9.09]	0.2077	0.93 [0.26 ; 3.47]	0.9120	Prioritized genes from dosage analysis(FDR < 10%)	NM_001256523.2; NM_001256524.2; NM_001256525.2; NM_003426.4
SCARF2	22	20,778,873	20,792,110	4034	8281	9042	5 (4)	0 (0)	0 (0)	24.68 [2.80 ; 3240.77]	0.0017	8.08 [0.92 ; 1060.65]	0.0623	Prioritized genes from dosage analysis(FDR < 10%)	NM_153334.7; NM_182895.5
KLHL22	22	20,795,806	20,850,170	4077	8457	9559	4 (4)	0 (0)	0 (0)	21.12 [2.25 ; 2799.03]	0.0047	6.87 [0.73 ; 909.82]	0,1011	Prioritized genes from dosage analysis(FDR < 10%)	NM_032775.4
MED15	22	20,861,896	20,941,900	4077	8457	9559	4 (4)	2 (0)	0 (0)	21.12 [2.25 ; 2799.03]	0.0047	9.92 [1.17 ; 1293.57]	0.0319	Prioritized genes from dosage analysis(FDR < 10%)	NM_001003891.3; NM_001293234.2; NM_001293235.2; NM_001293236.2; NM_015889.5; NM_001293237.2
PI4KA	22	21,061,979	21,213,100	4077	8457	9559	7 (4)	7 (1)	5 (0)	3.20 [1.07 ; 10.22]	0.0379	2.01 [0.79 ; 5.90]	0.1446	Prioritized genes from dosage analysis(FDR < 10%)	NM_001362862.2; NM_001362863.2; NM_058004.4
SERPIND1	22	21,128,383	21,142,008	4077	8457	9559	5 (4)	4 (1)	3 (0)	3.69 [0.98 ; 15.91]	0.0531	2.07 [0.66 ; 8.29]	0.2223	Prioritized genes from dosage analysis(FDR < 10%)	NM_000185.4
SNAP29	22	21,213,292	21,245,501	4077	8457	9559	4 (4)	2 (1)	2 (0)	4.22 [0.94 ; 24.23]	0.0607	1.98 [0.51 ; 10.79]	0.3368	Prioritized genes from dosage analysis(FDR < 10%)	NM_004782.4

	Position in hg19			Number of individuals included in LoF analysis			Number of carriers of LoF variants (including CNV-deletion)			Firth regression					
										EOAD versus CTRL		all AD versus CTRL		Reason of inclusion	Transcripts used for deletion definition
Gene	Chr	Start	End	EOAD	LOAD	CTRL	EOAD	LOAD	CTRL	OR [95%CI]	p-value	OR [95%CI]	p-value		
CRKL	22	21,271,714	21,308,037	4051	8364	9471	4 (4)	2 (1)	1 (0)	7.02 [1.30 ; 70.01]	0.0230	3.31 [0.70 ; 31.73]	0.1414	Prioritized genes from dosage analysis(FDR < 10%)	NM_005207.4
LZTR1	22	21,336,558	21,353,326	4077	8457	9559	10 (4)	12 (1)	14 (0)	1.70 [0.75 ; 3.74]	0.1998	1.18 [0.62 ; 2.34]	0.6149	Prioritized genes from dosage analysis(FDR < 10%)	NM_006767.4
THAP7	22	21,354,061	21,356,404	4074	8444	9541	4 (4)	1 (1)	1 (0)	7.03 [1.30 ; 70.13]	0.0229	2.79 [0.56 ; 27.25]	0.2249	Prioritized genes from dosage analysis(FDR < 10%)	NM_001008695.1; NM_030573.3
P2RX6	22	21,369,464	21,382,302	4077	8457	9559	13 (4)	13 (1)	16 (0)	1.92 [0.92 ; 3.94]	0.0813	1.23 [0.67 ; 2.31]	0.5145	Prioritized genes from dosage analysis(FDR < 10%)	NM_001159554.2; NM_001349874.2; NM_001349875.2; NM_001349876.2; NM_001394691.1; NM_001394692.1; NM_001394693.1; NM_001394694.1; NM_001394695.1; NM_001394696.1; NM_001394697.1; NM_005446.5
SLC7A4	22	21,383,007	21,386,847	4077	8455	9553	5 (4)	1 (1)	7 (0)	1.72 [0.54 ; 5.15]	0.3442	0.66 [0.22 ; 1.91]	0.4384	Prioritized genes from dosage analysis(FDR < 10%)	NM_004173.3
LRRC74B	22	21,400,249	21,418,457	3917	7799	8016	6 (4)	5 (1)	6 (0)	2.05 [0.67 ; 6.24]	0.2012	1.21 [0.47 ; 3.37]	0.6946	Prioritized genes from dosage analysis(FDR < 10%)	NM_001291006.2

Table S8: LoF variants analysis

Only genes with at least one CNV are displayed in this table

Sample	Cohort	AAO	Gender	Status	Type	CNV coordinate	HGVS nomenclature	Fully encompassed Genes	partially encompassed Genes
EFA-429-001	ADES-FR	43	male	EOAD	Deletion	chr22:18873820-21846407	NC_000022.10:g.18871041_21921807del	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
FR16641215	Netherlands Brain Bank	64	female	EOAD	Deletion	chr22:20705732-21414861	NC_000022.10:g.20705196_21480084del	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
FR16642636	ADC-Amsterdam	53	male	EOAD	Deletion	chr22:20708601-21537450	NC_000022.10:g.20706580_21539553del	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
ROU-0165-001	ADES-FR	51	male	EOAD	Deletion	chr22:20717833-21414845	NC_000022.10:g.20398711_21576182del	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
JW_Anglia_10158	PERADES	86	female	LOAD	Deletion	chr22:21062310-21414817	NC_000022.10:g.21045693_21480534del	CRKL,LRRC74B,LZTR1,P2RX6,PI4KA,SERPIND1,SLC7A4,SNAP 29,THAP7	
700003	PERADES	56	female	EOAD	Duplication	chr22:20891401-21088843	NC_000022.10:g.20873283_21096512dup		MED15,PI4KA
EXT-PUC-001	ADES-FR	51	male	EOAD	Duplication	chr22:21322064-21340318	NC_000022.10:g.21341672_21341672dup		LZTR1
B00E9CV	ADES-FR	78	female	LOAD	Duplication	chr22:18893812-21411601	NC_000022.10:g.18655928_21799164dup	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
A-ACT-AC002268	ADSP	81	female	LOAD	Duplication	chr22:18893887-21568164	NC_000022.10:g.18835841_21570212dup	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
A-ADC-AD003697	ADSP	74	male	LOAD	Duplication	chr22:20705059-21414817	NC_000022.10:g.20653814_21480535dup	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
A-TARC-TC000847	ADSP	67	female	LOAD	Duplication	chr22:20718475-21570368	NC_000022.10:g.20717966_21570758dup	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
A-ROS-RS000801	ADSP	87	female	LOAD	Duplication	chr22:21044318-22453623	NC_000022.10:g.20940993_22516590dup	CRKL,LRRC74B,LZTR1,P2RX6,PI4KA,SERPIND1,SLC7A4,SNAP 29,THAP7	
1865116	ERF	55	female	Control	Duplication	chr22:18893812-21411601	NC_000022.10:g.18655928_21799164dup	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
FR21322518	ADC-Amsterdam	50	male	Control	Duplication	chr22:18667400-21639009	NC_000022.10:g.18656711_21661880dup	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
A-WCAP-WC003992-BL-COL-57689BL1	ADSP	78	female	Control	Duplication	chr22:18773782-21193485	NC_000022.10:g.19139966_21617839dup	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
A-WCAP-WC003999-BL-COL-40463BL1	ADSP	75	female	Control	Duplication	chr22:18855517-21109723	NC_000022.10:g.19230001_21524610dup	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
C-RS-50711	ADSP	90	female	Control	Duplication	chr22:20705059-21414817	NC_000022.10:g.20653814_21480535dup	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
FR16641043	EMIF-AD 90-plus Study	91	female	Control	Duplication	chr22:20705076-21579718	NC_000022.10:g.20657675_21637213dup	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
WD_478	ADES-FR	85	male	Control	Duplication	chr22:20717833-20825852	NC_000022.10:g.20398713_20843263dup	SCARF2,ZNF74	KLHL22
A-ADC-AD000379	ADSP	90	male	Control	Duplication	chr22:20720856-21414817	NC_000022.10:g.20718608_21480535dup	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
WD_139	ADES-FR	81	male	Control	Duplication	chr22:20731392-21414845	NC_000022.10:g.20729961_21576182dup	CRKL,KLHL22,LRRC74B,LZTR1,MED15,P2RX6,PI4KA,SCARF2, SERPIND1,SLC7A4,SNAP29,THAP7,ZNF74	
C-RS-50996	ADSP	90	male	Control	Duplication	chr22:20812093-20940993	NC_000022.10:g.20800968_21044317dup	MED15	KLHL22
FR16641086	EMIF-AD 90-plus Study	88	female	Control	Duplication	chr22:21044363-21680972	NC_000022.10:g.21035063_21724331dup	CRKL,LRRC74B,LZTR1,P2RX6,PI4KA,SERPIND1,SLC7A4,SNAP 29,THAP7	
WD_131	ADES-FR	84	female	Control	Duplication	chr22:21056494-21414845	NC_000022.10:g.21045672_21576182dup	CRKL,LRRC74B,LZTR1,P2RX6,PI4KA,SERPIND1,SLC7A4,SNAP 29,THAP7	
A-ACT-AC002023	ADSP	86	male	Control	Duplication	chr22:21322232-21340187	NC_000022.10:g.21304135_21341791dup		LZTR1

Table S9. CNV encompassing the 22q11.21 locus

Only genes within the locus are reported

		N total			N carriers of CNVs (DEL/DUP)			main Dosage analysis EOAD vs CTRL			Dosage analysis with PC adjustment EOAD vs CTRL			
Chr	Gene	EOAD	LOAD	CTRL	EOAD	LOAD	CTRL	OR [95%CI][1]	p-value[2]	FDR	OR [95%CI]	p-value[1]	FDR	Transcripts
1	PRAMEF26*	2911	7302	6461	8 (8 / 0)	2 (2 / 0)	2 (1 / 1)	0,08 [0,01 ; 0,34]	2,88E-04	3,29E-02	0,07 [0,01 ; 0,34]	3,65E-04	8,07E-02	NM_001306072.3
1	HJV	4077	5457	9559	8 (0 / 8)	8 (1 / 7)	8 (4 / 4)	4,91 [1,71 ; 16,69]	2,82E-03	1,99E-01	7,34 [2,31 ; 27,48]	5,97E-04	8,33E-02	NM_001316767.2; NM_001379352.1; NM_213653.4
2	ADI1	4077	8457	9559	5 (0 / 5)	1 (0 / 1)	1 (1 /0)	26,58 [3,06 ; 3482,37]	1,05E-03	9,16E-02	26,50 [2,88 ; 3518,08]	1,59E-03	1,45E-01	NM_001306077.2 ; NM_018269.4
8	MTUS1**	4077	8457	9559	54 (54/0)	38 (37/1)	57 (57/0)	0,47 [0,33;0,66]	1,95E-05	2,71E-02	0,58 [0,40 ; 0,84]	3,75E-03	1,91E-01	NM_001001924.3; NM_001363057.2; NM_001363058.2; NM_001363059.2; NM_001363061.2
10	MBL2	4077	8457	9559	2 (2 / 0)	25 (25 / 0)	45 (44 / 1)	5,76 [2,22 ; 21,16]	5,44E-05	2,71E-02	1,74 [0,55 ; 7,60]	3,78E-01	9,91E-01	NM_000242.3; NM_001378373.1; NM_001378374.1
17	FADS6	4077	8457	9559	16 (15 / 1)	11 (11 / 0)	5 (5 / 0)	0,19 [0,07 ; 0,45]	1,65E-04	2,71E-02	0,32 [0,12 ; 0,80]	1,42E-02	5,40E-01	NM_178128.6
19	PLIN4**	4077	8457	9559	20 (19 / 1)	10 (10 / 0)	12 (12 / 0)	0,31 [0,15 ; 0,61]	8,38E-04	7,66E-02	0,34 [0,16 ; 0,70]	3,59E-03	8,07E-02	NM_001393888.1 ; NM_001393889.1 ; NM_001393890.1 ; NM_001393891.1; NM_001367868.2
19	KLC3	4077	8457	9559	24 (1 / 23)	33 (1 / 32)	16 (0 / 16)	3,06 [1,65 ; 5,79]	4,28E-04	4,12E-02	2,79 [1,48 ; 5,36]	1,50E-03	1,45E-01	NM_177417.3
19	ERCC2	4077	8457	9559	29 (1 / 28)	43 (2 / 41)	20 (0 / 20)	3,04 [1,74 ; 5,4]	1,09E-04	2,71E-02	2,87 [1,62 ; 5,16]	3,32E-04	8,07E-02	NM_001130867.2
19	ERCC2	4077	8457	9559	29 (1 / 28)	42 (2 / 40)	20 (0 / 20)							NM_000400.4
22	ZNF74	4077	8457	9559	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	0,04 [0,0003 ; 0,29]	2,71E-04	3,29E-02	0,03 [0,0002 ; 0,30]	6,37E-04	8,33E-02	NM_001256523.2; NM_001256525.2; NM_003426.4; NM_001256524.2
22	SCARF2	4077	8457	9559	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	0,04 [0,0003 ; 0,29]	2,71E-04	3,29E-02	0,03 [0,0002 ; 0,30]	6,37E-04	8,33E-02	NM_153334.7; NM_182895.5
22	KLHL22	4077	8457	9559	4 (4 / 0)	4 (0 / 4)	8 (0 / 8)	0,04 [0,0003 ; 0,30]	3,80E-04	3,86E-02	0,03 [0,0003 ; 0,33]	1,14E-03	1,26E-01	NM_032775.4
22	MED15	4077	8457	9559	4 (4 / 0)	4 (0 / 4)	9 (0 / 9)	0,04 [0,0003 ; 0,29]	2,71E-04	3,29E-02	0,03 [0,0002 ; 0,30]	6,37E-04	8,33E-02	NM_001003891.3; NM_001293234.2; NM_001293235.2; NM_001293236.2; NM_001293237.2; NM_015889.5
22	PI4KA	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	0,03 [0,0002 ; 0,29]	4,41E-04	8,07E-02	NM_001362862.2; NM_001362863.2; NM_058004.4
22	SERPIND1	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	0,03 [0,0002 ; 0,29]	4,41E-04	8,07E-02	NM_000185.4
22	SNAP29	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	0,03 [0,0002 ; 0,29]	4,41E-04	8,07E-02	NM_004782.4
22	CRKL	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	0,03 [0,0002 ; 0,29]	4,41E-04	8,07E-02	NM_005207.4
22	LZTR1	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	0,03 [0,0002 ; 0,29]	4,41E-04	8,07E-02	NM_006767.4
22	THAP7	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	0,03 [0,0002 ; 0,29]	4,41E-04	8,07E-02	NM_030573.3; NM_001008695.1
22	P2RX6	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	0,03 [0,0002 ; 0,29]	4,41E-04	8,07E-02	NM_001159554.2; NM_001349874.2; NM_001349875.2; NM_001349876.2; NM_001394691.1; NM_001394692.1; NM_001394693.1; NM_001394694.1; NM_001394695.1; NM_001394696.1; NM_001394697.1; NM_005446.5
22	SLC7A4	4077	8457	9559	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,03 [0,0003 ; 0,28]	1,92E-04	2,71E-02	0,03 [0,0002 ; 0,29]	4,41E-04	8,07E-02	NM_004173.3
22	LRRC74B	3917	7799	8016	4 (4 / 0)	6 (1 / 5)	10 (0 / 10)	0,04 [0,0003 ; 0,29]	1,86E-04	2,71E-02	0,03 [0,0002 ; 0,29]	4,41E-04	8,28E-02	NM_001291006.2

Table S12. Results from dosage analysis after adjustment on ancestry

Chr: chromosome; N: number, EOAD: early onset Alzheimer disease, LOAD: late onset Alzheimer disease; CTRL: controls

[1]OR should be interpreted for an increase of transcript copies by 1, i.e. a OR > 1 (resp. OR <1) means that duplications (resp. deletions) are associated with a higher disease risk

[2]non adjusted p-values, selected based on FRD < 10%

* gene overlapping with repeats or duplicated gene in the genome (despite CNVs not overlapping >50% with repeats)

** gene for which the signal in dosage analysis is driven by deletion whereas transcripts are not in set A

		N total			EOAD vs CTRL (firth regression)		All AD vs CTRL (firth regression)		N carriers		
List of genes		EOAD	LOAD	CTRL	OR [95%CI][1]	p-value	OR [95%CI]	p-value	EOAD	LOAD	CTRL
All genes	deletion	4077	8457	9559	1,15 [1,06 ; 1,23]	3,41E-04	1,04 [0,98 ; 1,09]	1,94E-01	1756	338	3802
	duplication	4077	8457	9559	1,10 [1,02 ; 1,19]	1,25E-02	1,06 [1,00 ; 1,12]	3,31E-02	1408	2816	3091
Genes in GWAS list	deletion	4077	8457	9559	2,67 [1,51 ; 4,74]	7,98E-04	2,02 [1,26 ; 3,35]	2,98E-03	25	34	22
	duplication	4077	8457	9559	0,58 [0,30 ; 1,04]	6,90E-02	0,67 [0,45 ; 1,00]	5,25E-02	12	32	50
Genes involved in Abeta network	deletion	4077	8457	9559	1,44 [1,09 ; 1,90]	1,14E-02	1,26 [1,01 ; 1,57]	3,62E-02	81	138	133
Genes involved in Processing or traffficking of Abeta	deletion	4077	8457	9559	1,64 [1,04 ; 2,57]	3,36E-02	0,96 [0,65 ; 1,42]	8,33E-01	32	26	46
Genes involved in mediator of Abeta toxicity / calcium homeostasis	deletion	4077	8457	9559	16,42 [1,59 ; 2208,35]	1,61E-02	11,45 [1,40 ; 1485,47]	1,80E-02	3	4	0

Table S13. Association Results aggregated by lists of genes

Sample	Cohort	AAO / Age at last « visit »	Gender	APOE	Status	Gene	BaseChange	AAChange
FR16618481	ADC-Amsterdam	51	2	34	EOAD	PSEN2	c.715A>G	p.Met239Val
FR16640975	Netherlands Brain Bank	34	2	33	EOAD	PSEN1	c.617G>A	p.Gly206Asp
FR16642626	ADC-Amsterdam	56	2	33	EOAD	PSEN1	c.786G>C	p.Leu262Phe
FR16640980	Netherlands Brain Bank	57	1	33	EOAD	PSEN1	c.791C>T	p.Pro264Leu
FR16642720	Netherlands Brain Bank	59	1	33	EOAD	PSEN1	c.791C>T	p.Pro264Leu
FR16640953	Netherlands Brain Bank	42	2	33	EOAD	PSEN1	c.1151G>C	p.Gly384Ala
FR16618475	Netherlands Brain Bank	32	1	23	EOAD	PSEN1	c.1254G>C	p.Leu418Phe
FR16640939	Netherlands Brain Bank	74	2	33	LOAD	APP	c.2149G>A	p.Val717Ile
G-KGAD-KA000745-UNK-WU-8002617069	ADSP	54	2	24	EOAD	APP	c.2146A>G	p.Ile716Val
G-KGAD-KA000751-BL-WU-8002616270	ADSP	53	1	33	EOAD	PSEN1	c.677T>G	p.Leu226Arg
FR16640931	Netherlands Brain Bank	47	1	34	EOAD	PSEN1	c.779C>T	p.Ala260Val
FR16642644	ADC-Amsterdam	56	2	34	EOAD	PSEN1	c.1130G>T	p.Arg377Met
FR21322134	ADC-Amsterdam	42	2	33	Control	TARDBP	c.1144G>A	p.Ala382Thr
G-KGAD-KA000743-BL-WU-8015455061	ADSP	41	1	22	Control	MAPT	c.1216C>T	p.Arg406Trp
G-KGAD-KA000168-BL-WU-8298765	ADSP	61	1	23	EOAD	GRN	c.709-2A>G	SpliceVariant
G-KGAD-KA000243-UNK-WU-8005874997	ADSP	54	2	33	Control	GRN	c.911G>A	p.Trp304*
G-KGAD-KA000738-UNK-WU-8002616127	ADSP	60	2	33	EOAD	GRN	c.1145delC	p.Thr382fs
EXT-1733-001	ADES-FR	58	1	33	EOAD	GRN	c.1403_1404insGTGAGTGCCTCCCTGCCTGCCCCTGCCTGGGTGGGAGCTGGGCCTGCTGCCAGTTGCCCCAT	p.Leu469fs
A-WCAP-WC004049-BL-COL-40538BL1	ADSP	84	1	34	LOAD	NOTCH3	c.2095T>C	p.Cys699Arg
G-KGAD-KA000377-UNK-WU-24094552	ADSP	69	2	44	LOAD	NOTCH3	c.3691C>T	p.Arg1231Cys
FR16617223	ADC-Amsterdam	72	2	33	LOAD	NOTCH3	c.3691C>T	p.Arg1231Cys
FR16642490	ADC-Amsterdam	54	2	34	EOAD	NOTCH3	c.3568C>T	p.Arg1190Cys

Table S14. Pathogenic variant carriers excluded from stage-2 ADES/ADSP dataset

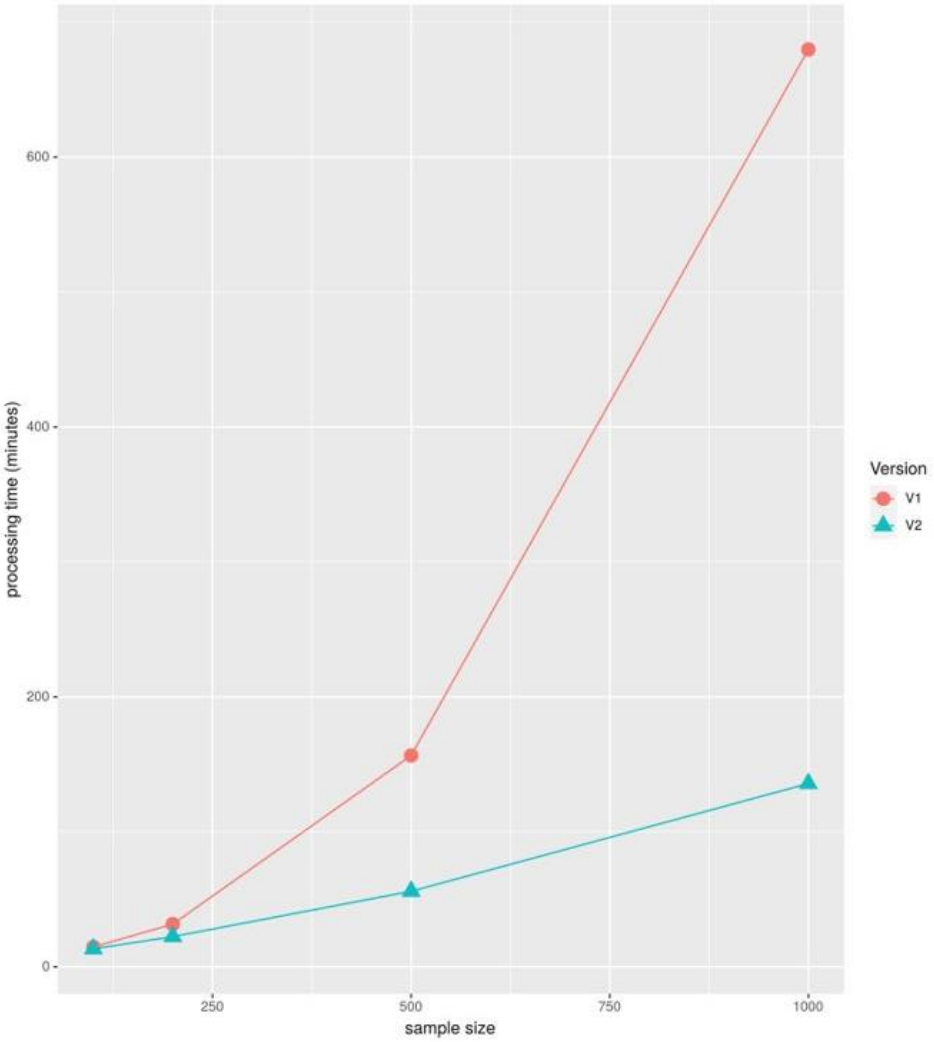


Figure S1. Time required for the calling step of our pipeline with two versions of CANOES
The two versions of CANOES differ only on the parallelization step, there is no differences on CNVs calling.
The same data have been processed using the same computational server with 32 threads and 192Gb of RAM. The parallelization used all threads, with the exception of the 1000 samples datasets, limited to 16 threads due to RAM limitation due to the matrix size.

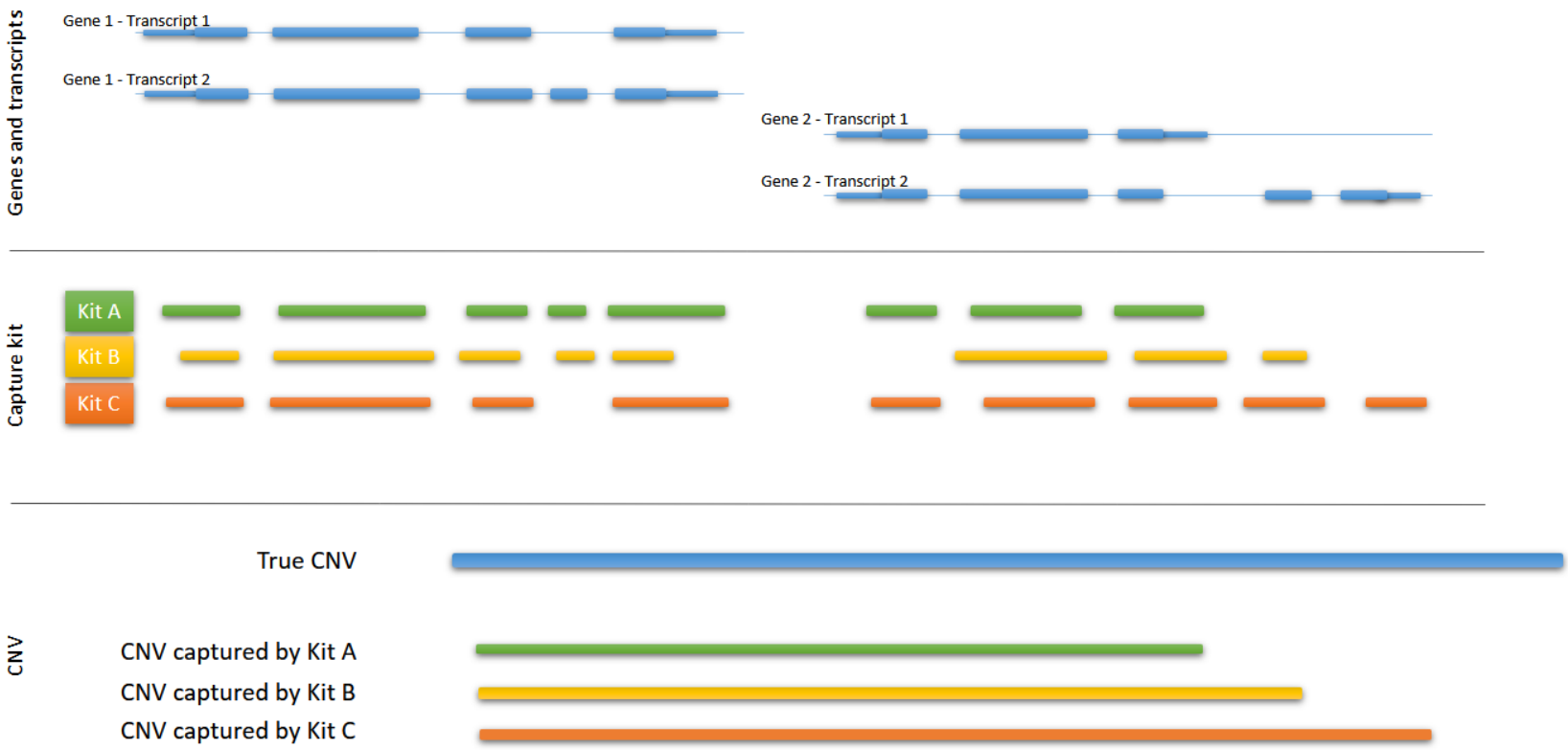


Figure S2. Theoretical example of definition of CNV coordinates according to transcript and capture kit coordinates.
On the top, a fictious example of two genes with two transcripts each. Larger blue rectangles represent exons whereas thinner rectangles represent non coding regions.
On the middle, an example of a difference of captured zone between different kits.
On the bottom, the fictious true CNV, in blue, and the coordinates attributed by each capture kit to this CNV.
In this example, the true CNV could be considered as encompassing fully Gene 2 when called from data obtained following capture by kit C, whereas the same CNV would be annotated as partially affecting Gene A regarding kits B and C if using reference coordinates to define a transcript. Using the proposed definition here (all targets are encompassed defines a fully encompassed transcript), this CNV is considered as fully affecting gene A in all three examples.

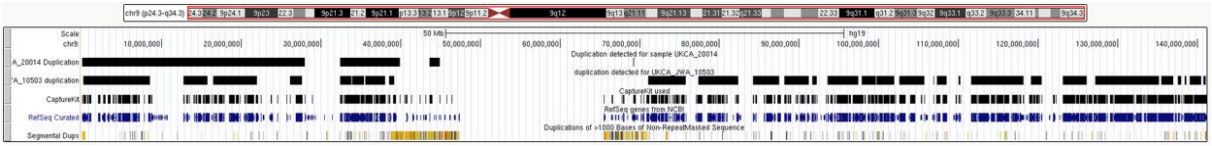


Figure S3. Example of a large CNV on chromosome 9 suggesting a blood-specific event (probable clonal hematopoiesis)
This figure was obtained from the UCSC genome browser and represents large mosaic duplications detected in two samples from the PERADES dataset on chromosome 9 (GRCh37/hg19). The first two rows display duplications detected across chromosome 9. Black rectangles indicate duplications as defined by CANOES. Blue rectangles indicate how multiple calls should be merged as one unique CNV. The third line indicates the position of capture kit targets (both samples have the same capture kit). The fourth line indicates genes position according to RefSeq. Finally, the last line indicates positions of segmental duplication regions.

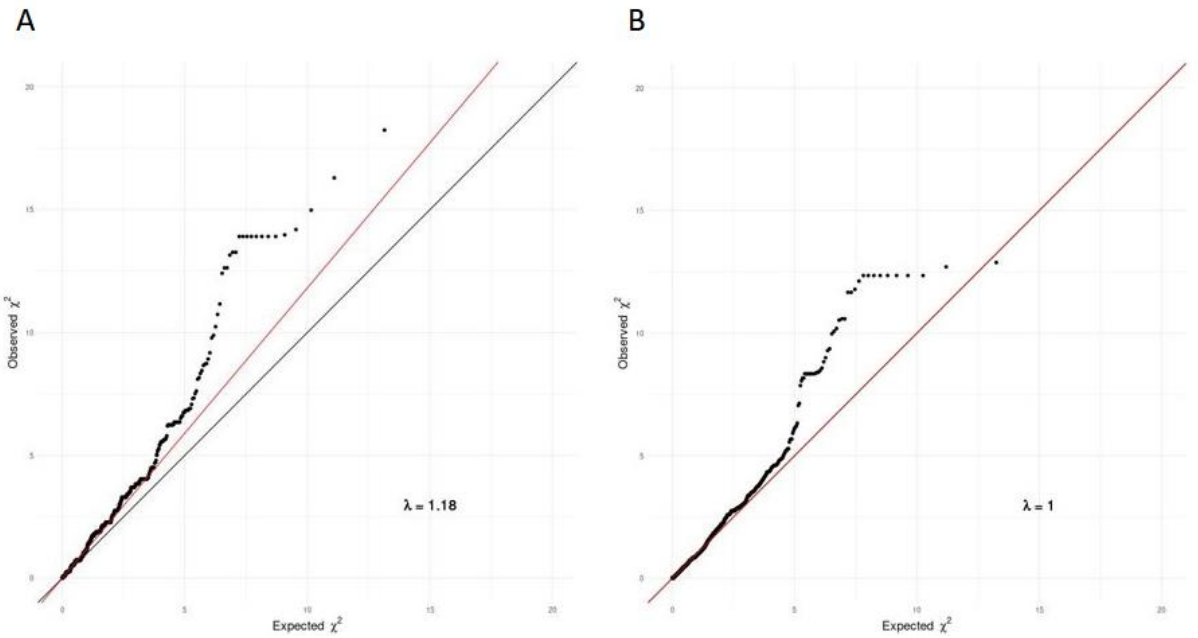


Figure S4. Comparison of results before and after adjustment for ancestry.
1. P-value qq-plot for main dosage analysis.
2. B. P-value qq-plot for ancestry adjusted dosage analysis.

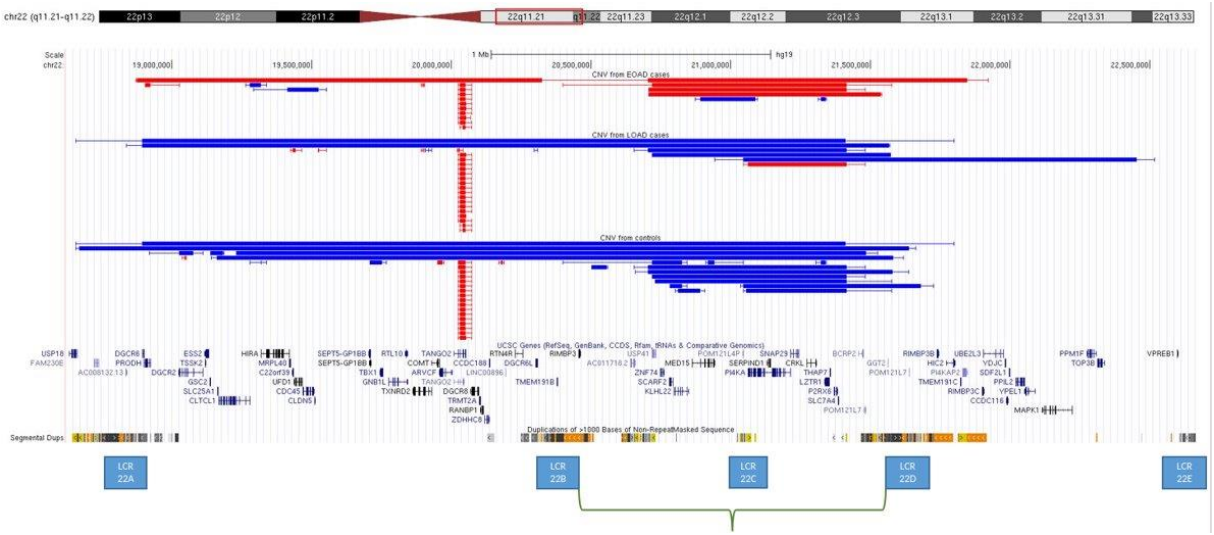


Figure S5. 22q11.2 locus as displayed on the UCSC genome browser
In red: deletions
In blue: duplications
Large areas represent coordinates of CNVs as detected by CANOES. Thinner areas show breakpoints uncertainties, i.e. regions between two targets of the sample-specific capture kit.
Genes are indicated below, and the regions of segmental duplications appear at the bottom part.
Blue rectangles show regions of low copy repeats underlying recurrent rearrangements by non-allelic homologous reparation (NAHR).
The green area indicates the locus identified in our study

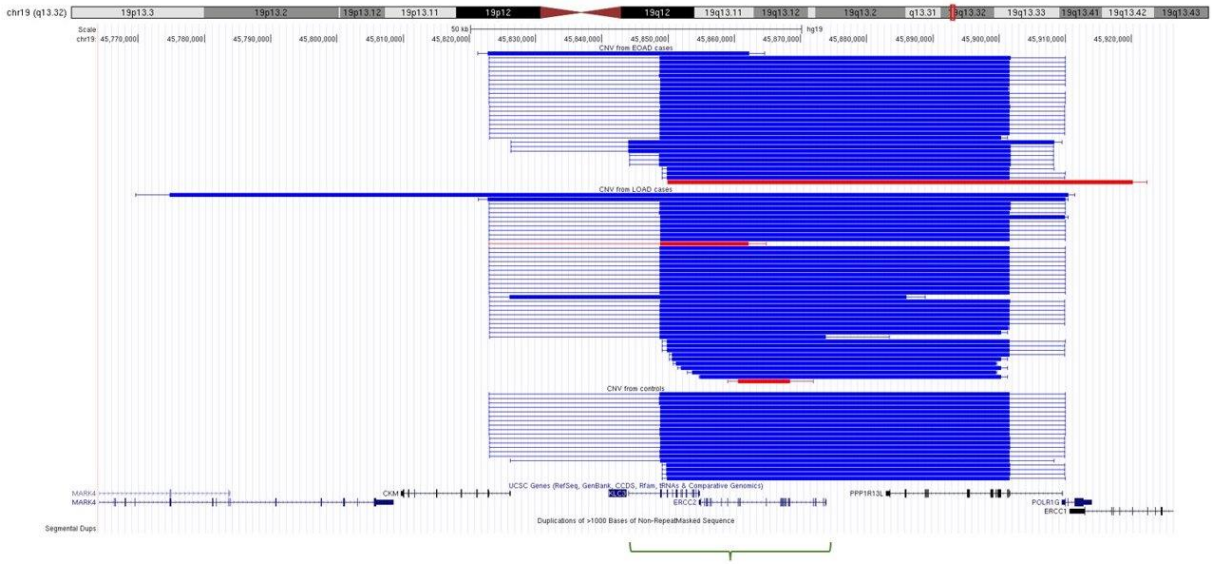


Figure S6. *ERCC2-KLC3* locus on chr19 as displayed on the UCSC genome browser
In red: deletions
In blue: duplications
Large areas represent coordinates of CNVs as detected by CANOES. Thinner areas show breakpoints uncertainties, i.e. regions between two targets of the sample-specific capture kit.
Genes are indicated below, and the regions of segmental duplications appear at the bottom part.
The green area indicates the locus identified in our study

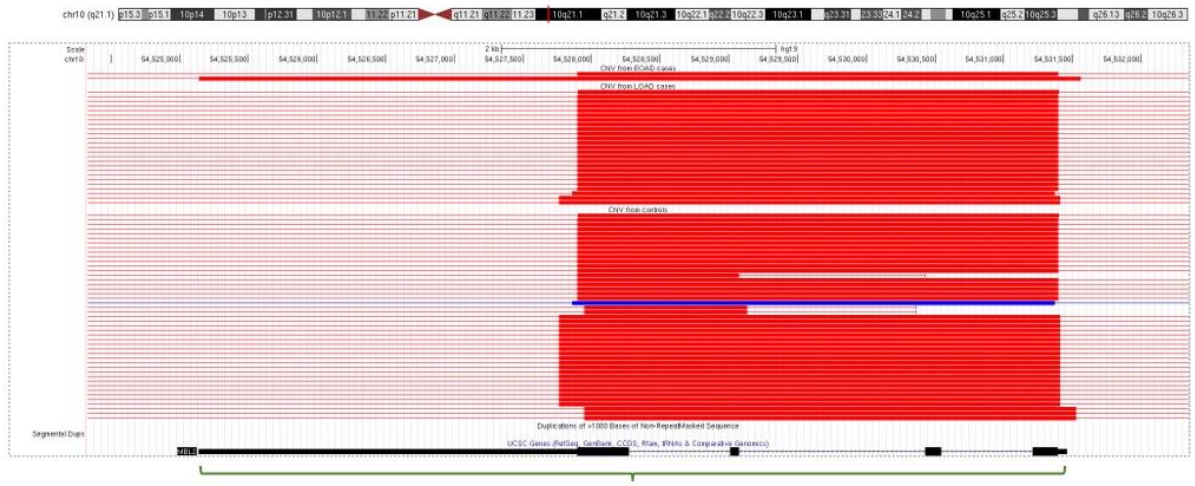


Figure S7. *MBL2* locus as displayed on the UCSC genome browser
In red: deletions
In blue: duplications
Large areas represent coordinates of CNVs as detected by CANOES. Thinner areas show breakpoints uncertainties, i.e. regions between two targets of the sample-specific capture kit.
Genes are indicated below, and the regions of segmental duplications appear at the bottom part.
The green area indicates the locus identified in our study

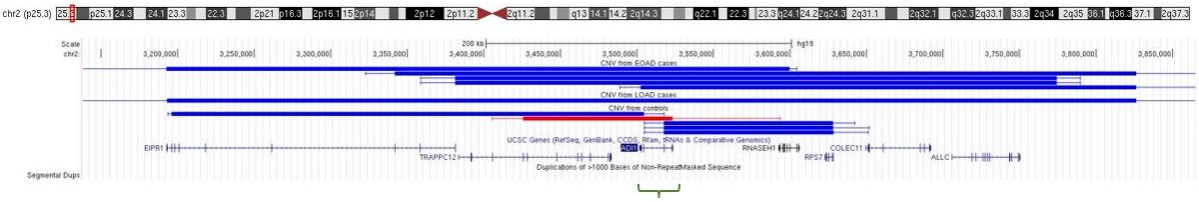


Figure S8. *ADI1* locus as displayed on the UCSC genome browser
In red: deletions
In blue: duplications
Large areas represent coordinates of CNVs as detected by CANOES. Thinner areas show breakpoints uncertainties, i.e. regions between two targets of the sample-specific capture kit.
Genes are indicated below, and the regions of segmental duplications appear at the bottom part.
The green area indicates the locus identified in our study

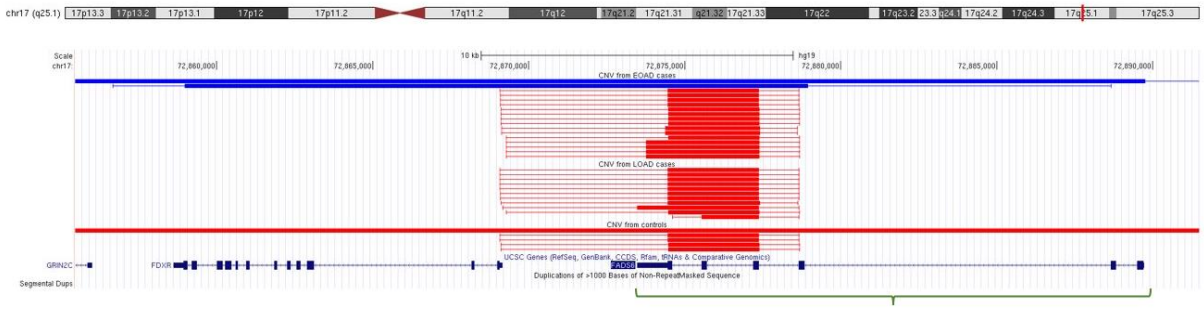


Figure S9. *FADS6* locus as displayed on the UCSC genome browser
In red: deletions
In blue: duplications
Large areas represent coordinates of CNVs as detected by CANOES. Thinner areas show breakpoints uncertainties, i.e. regions between two targets of the sample-specific capture kit.
Genes are indicated below, and the regions of segmental duplications appear at the bottom part.

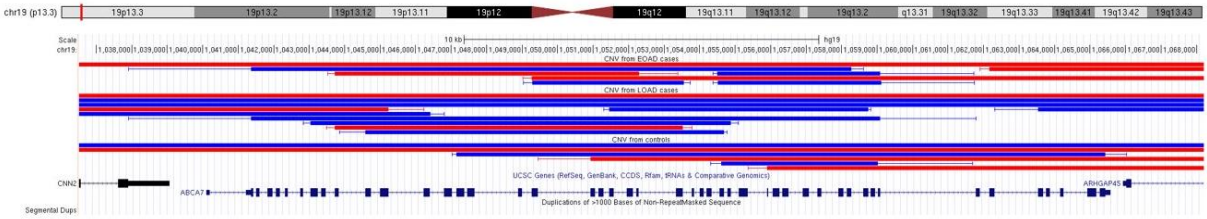


Figure S10. *ABCA7* locus as displayed on the UCSC genome browser
In red: deletions
In blue: duplications
Large areas represent coordinates of CNVs as detected by CANOES. Thinner areas show breakpoints uncertainties, i.e. regions between two targets of the sample-specific capture kit.
Genes are indicated below, and the regions of segmental duplications appear at the bottom part.

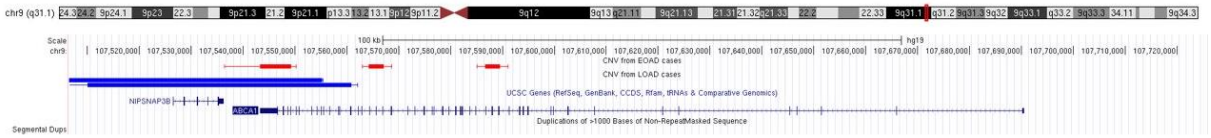


Figure S11. *ABCA1* locus as displayed on the UCSC genome browser
In red: deletions
In blue: duplications
Large areas represent coordinates of CNVs as detected by CANOES. Thinner areas show breakpoints uncertainties, i.e. regions between two targets of the sample-specific capture kit.
Genes are indicated below, and the regions of segmental duplications appear at the bottom part.

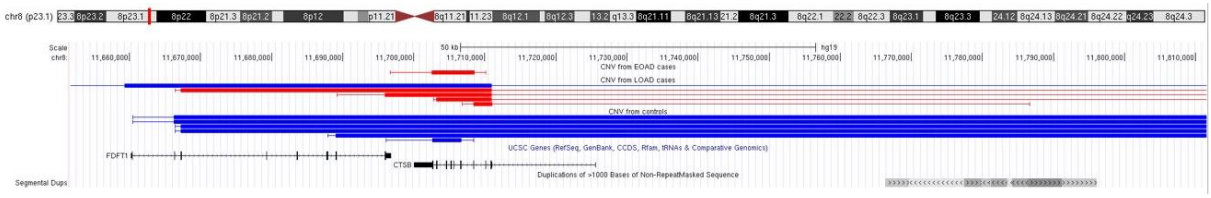


Figure S12. *CTSB* locus as displayed on the UCSC genome browser
In red: deletions
In blue: duplications
Large areas represent coordinates of CNVs as detected by CANOES. Thinner areas show breakpoints uncertainties, i.e. regions between two targets of the sample-specific capture kit.
Genes are indicated below, and the regions of segmental duplications appear at the bottom part.

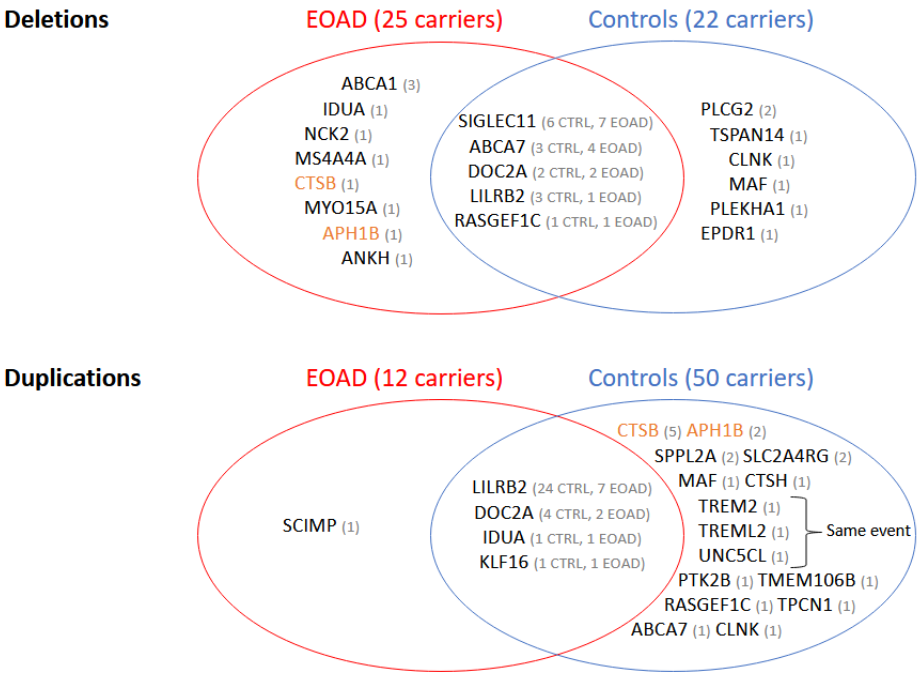


Figure S13. Genes from GWAS list being deleted or duplicated in our dataset.
Gray numbers corresponds to number of carriers of CNV encompassing the gene. Orange colored genes are those suggesting a mirror effect.

3. Discussion

3.1. L'évolution de la détection des variations de structure

3.1.1. Utilisation actuelle des approches de séquençage

L'étude des variations de structure reste l'un des défis majeurs dans l'étude de nombreuses pathologies et dans la structure globale de notre génome. Il est facile de faire un parallèle entre l'étude des variations ponctuelles (SNVs/indels) et celle des variations de structures. La première difficulté est la détection des événements de manière précise, puis vient la question de l'interprétation. Dans le cas des variations ponctuelles, la qualité des séquences produites permet d'obtenir, en dehors de régions complexes répétées, une confiance de plus de 99% dans le calling. Cette qualité de détection est loin d'être atteinte dans le cas des variations de structure, surtout lorsque l'on considère les technologies les plus répandues dans le cadre de la recherche et du diagnostic.

A l'heure actuelle, les technologies de séquençage majoritairement mises en place dans le cadre de la recherche et du diagnostic restent celles de seconde génération, et plus particulièrement par capture (panel de gène ou exome). La raison principale reste principalement le coût : un séquençage de génome reste un peu plus élevé plus que celui d'un exome, lui-même étant plus élevé que celui d'un panel de gènes, mais surtout le coût bioinformatique reste une différence majeure. La question est donc de savoir ce qui est le plus avantageux entre nombre d'individus séquencés et les régions séquencées. Dans le cadre de la recherche de variants rares comme nous l'avons effectué dans le cadre de ces travaux, la faible fréquence des variations du nombre de copies et le nombre de loci potentiellement associés à un risque de développer la maladie conduit à la nécessité d'avoir des grands effectifs.

3.1.2. Séquençage ShortReads : de l'exome vers le génome

Le passage de l'exome vers le génome a été induit par plusieurs facteurs. Tout d'abord l'évolution des séquenceurs a permis de rentre plus accessible ce type de séquençage, que cela soit en termes de coût ou de temps de séquençage. Cette évolution des séquenceurs s'est aussi accompagnée d'évolutions bioinformatiques et informatiques : le développement de solutions plus optimisées et l'augmentation de la puissance des machines de calcul ont rendu possible l'analyse rapide des données. Ensuite, dans le cadre de l'exploitation des variations ponctuelles, l'évolution des connaissances des régions non

codantes a ouvert le plein potentiel des génomes. Grâce à cela, nous sommes passés d'un génome qui était exploité plus comme un exome avec de meilleures couverture dans certaines régions (Meienberg et al. 2016) à une exploitation complète du génome, à la fois des régions codantes et non codantes. Enfin, l'intérêt de plus en plus important envers les variations de structure conduit à passer des techniques ACPA (SNP array / CGH array) vers les techniques de séquençage.

D'un point de vue bioinformatique et comme nous l'avons vu au cours de ces travaux, l'utilisation du séquençage d'exome nous limite principalement à utiliser les approches basées sur la profondeur de lecture. Il est techniquement possible d'utiliser les autres approches mais les résultats sont généralement très décevants car peu sensibles et bruités, les points de cassure ciblés par ces approches étant rarement localisés dans les régions codantes ou à proximité. Dans le cadre du génome, il est possible d'utiliser les autres approches (orientation des paires, splitreads, assemblage de novo) et ainsi combiner différents outils. L'apport est alors important : non seulement il est possible d'identifier tous les types de variations équilibrées (inversion, translocation, insertion, ...) en plus des variations du nombre de copies, mais l'utilisation de différents outils basés sur des signaux différents permet de croiser les différents résultats afin de venir améliorer la détection et la filtration des variants : si plusieurs outils convergent pour un événement, la probabilité que ce dernier soit vrai augmente, bien que certains événements restent impossibles à détecter par certaines techniques, par essence, comme les points de cassure de remaniements médiés par des séquences répétées par exemple.

De plus, la combinaison de plusieurs outils permet aussi de comprendre certaines structures, comme dans l'exemple de la Figure 32. Notre pipeline utilisant CANOES (basé sur la profondeur de lecture) a identifié la duplication complète du gène *SMAD4*, alors que le logiciel GRIDSS (Cameron et al. 2017) qui utilise les 3 autres méthodes, a identifié de multiples délétions à l'intérieur de la région. En observant plus précisément les fichiers d'alignements, les délétions identifiées correspondent aux coordonnées des introns. Nous avons conclu à la présence d'une rétrocopie chez le patient, cette rétrocopie étant fréquente dans la population générale (Chatron et al. 2019).

Le passage au séquençage en génome complet est donc une étape importante qui permet d'obtenir une augmentation du nombre d'événements détectés et de manière plus précise (Hehir-Kwa, Tops, et Kemmeren 2018), et par la même occasion d'augmenter le nombre de diagnostics réalisés (Hehir-Kwa, Pfundt, et Veltman 2015). Ce passage au séquençage de génome complet est déjà mis en place sur les laboratoires AURAGEN et SeqOIA mis en place dans le cadre du Plan France Médecine Génomique pour le diagnostic des maladies rares et des cancers, lesquels rendant à la fois les SNV et les CNVs.

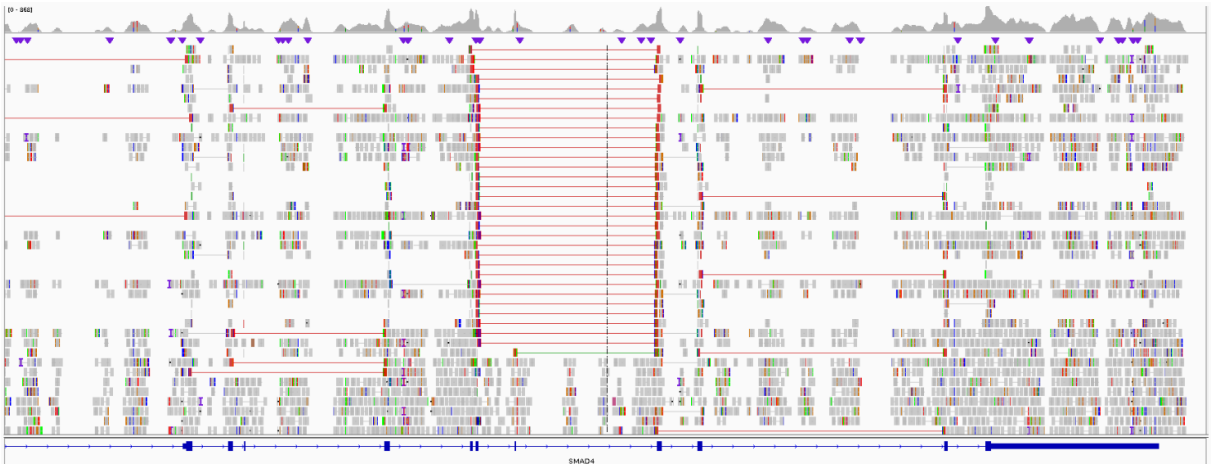


Figure 32 : Visualisation d'une rétrocopie du gène SMAD4.
On observe dans la partie supérieure des pics de profondeur correspondant à certains exons (en gris). Dans la partie inférieure, des paires de reads (en rouge) présente des tailles d'insert anormalement élevés, révélant des délétions coïncidant avec les exons (en bleue)

3.1.3. Le génome de référence : un génome ou des génomes ?

L'utilisation de plus en plus importante de séquençage de génome permet d'identifier et de caractériser de nombreux haplotypes présents dans la population générale. Ces haplotypes étant par définition fréquents dans la population, ils sont détectés de manière récurrente dans les analyses. Il est donc nécessaire de les prendre en compte et de les filtrer de la même manière que pour les variations ponctuelles fréquentes. Contrairement à ces dernières, les variations de structure ont un impact important dans l'alignement et afin de pouvoir identifier correctement des variations de structure rares, il faut pouvoir identifier correctement les haplotypes de l'individu étudié.

Il est donc nécessaire de faire évoluer le génome de référence afin de prendre en compte ces différents haplotypes. C'est le cas de la version hg38 du génome de référence qui inclut un certain nombre de ces haplotypes. Malheureusement, leur utilisation reste encore rare, voire anecdotique. Ceci est dû à plusieurs points bloquants. Tout d'abord pendant l'étape d'alignement des reads sur le génome de référence, les logiciels vont considérer que les reads peuvent à la fois correspondre au chromosome de référence "classique" et en même temps à l'haplotype, conduisant à un alignement multiple qui était jusqu'à maintenant traité comme un signe de mauvaise qualité. De plus, la majorité des bases de données ne tiennent pas compte de ces haplotypes et ne conserve l'information que pour les chromosomes de référence. Cela signifie que même si l'on aligne et que l'on exploite les données

présentes sur les haplotypes, Il ne sera pas ou très peu possible de récupérer des données de fréquence ou même des connaissances issues de bases de données spécialisées.

L'évolution du génome de référence prend actuellement deux directions qui se révèleront sûrement complémentaire. Il y a tout d'abord la voie établie par le consortium T2T (Telomere-to-telomere) qui, comme nous l'avons rapidement vu, vise à compléter entièrement la séquence connue du génome humain, et plus particulièrement les zones de faible complexité de ce dernier (Nurk et al. 2022). Afin de réaliser la complétion du génome, il a été nécessaire de combiner de très nombreuses approches, avec à la fois du séquençage de troisième génération (technologie Oxford Nanopore et PacBio) associé à de la cartographie optique et du séquençage de seconde génération. Au final, ils ont obtenu une séquence complète, mais pour une seule molécule d'ADN entièrement homozygote issue d'une môle hydatiforme, c'est à dire une anomalie rare de la grossesse qui se manifeste par la croissance d'une masse cellulaire dont tous les chromosomes paternels sont dupliqués. Ceci permet d'être sûr d'avoir une seule molécule à séquencer et non deux. Le problème de cette approche est que l'on retire toute l'information de diversité qui a été ajoutée au fur et à mesure de la construction du génome de référence actuel.

L'autre voie d'évolution est la construction d'un pangénome qui prendra en compte le besoin de représentativité de toute la diversité du génome humain (Miga et Wang 2021; Sherman et Salzberg 2020). Le principe consiste non pas à avoir un génome linéaire mais un ensemble de génome assemblé sous forme de graphes (Eizenga et al. 2020). De cette manière, les haplotypes ne sont plus une séquence alternative d'une référence mais un embranchement possible dans une structure complète. Séquencer un individu consiste alors à définir le parcours à effectuer à travers le graphe des possibilités pour obtenir sa séquence personnelle. Au niveau international, le consortium du HPRC (Human Pangenome Reference Consortium) travaille à l'établissement de ce dernier (T. Wang et al. 2022). Ce passage d'un génome de référence linéaire à un pangénome organisé sous forme de graphe nécessite tout de même de repenser et de remettre en place tout un pipeline d'analyse. Actuellement le pipeline standard bio-informatique établi par le Broad Institute (McKenna et al. 2010) et constitué de la suite BWA puis GATK est le plus utilisé dans les laboratoires, mais celui-ci n'est pas le plus adapté pour le travail sur du graphe. Cela signifie donc qu'il faudra mettre en place de nouveaux aligneurs et de nouveaux logiciels de détection de variants adaptés à ces structures en graphe. À la fin du processus bioinformatique, la manière dans sont stockés les variants est aussi à redéfinir : il faut passer d'un système de coordonnées chromosome/position à un système basé sur les haplotypes, obligeant à revoir la façon d'appréhender la "position" d'un variant. Ceci conduit aussi à revoir la manière dont

sont construites les bases de données actuelles, qui devront à la fois permettre l'inclusion des nouveaux variants tout en reportant toutes les variations connues sur ce nouveau système de coordonnées. Tous ces changements devront être pris en compte avant de passer sur une référence pangénomique.

3.1.4. Evolution technologique : séquençage de 3ème génération et cartographie optique

Toutes ces évolutions ont été possibles principalement grâce au développement des techniques de séquençage de 3ème génération et, en parallèle de ces dernières, les techniques de cartographie optique. Ces technologies de séquençage permettent de résoudre les problèmes posés par la présence des longues séquences répétées et qui diminuent la qualité des alignements des courts fragments dans ces régions (Midha, Wu, et Chiu 2019). Grâce à cela, la détection des variations de structure est grandement améliorée (Figure 33), permettant d'obtenir une résolution jusqu'à maintenant difficilement atteignable avec les technologies shortReads (Fujimoto et al. 2021; Midha, Wu, et Chiu 2019). Les technologies longRead ne présentent pas non plus de biais dans les régions séquencées du fait de l'absence d'étapes d'amplification avant séquençage, la couverture du génome est alors homogène sur l'ensemble de la séquence (Pollard et al. 2018). Enfin, l'utilisation de fragments longs permet le phasage des variants, phasage qui peut avoir son importance dans le cadre de l'étude de pathologies avec une transmission récessive.

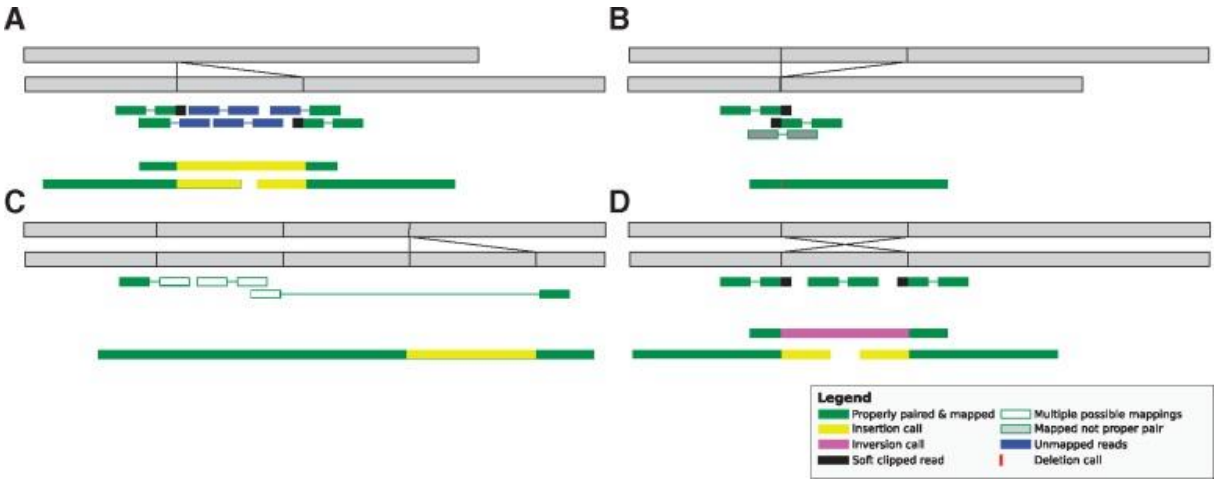


Figure 33 : Apport du long read pour la détection des variations de structure.
Pour chaque réarrangement sont représentés à la fois des short reads (partie supérieure) et des long reads (partie inférieure). A : Insertion de grande taille. B : Délétion de grande taille. C : Duplication dans une région multi allélique. D : Inversion d'un fragment. Image adaptée de Pollard et al., 2018

Les deux technologies actuelles de séquençage de 3ème génération présentent chacune des avantages et inconvénients. La technologie SMRT de la société PacBio, basée sur une circularisation des fragments d'ADN, permet d'obtenir des fragments allant jusqu'à 100 kb au maximum, et une taille moyenne à environ 20 kb. L'intérêt de circulariser la molécule est de pouvoir la lire de multiples fois et ainsi de corriger les éventuelles erreurs lors du séquençage, produisant ainsi des reads HiFi (High Fidelity) en faisant un niveau de qualité similaire voire supérieur aux séquençages de 2ème génération. La technologie d'Oxford Nanopore, basée sur l'utilisation de canaux ioniques et la mesure de modification du champ électrique lors du passage des molécules, est capable de produire des molécules très longues, certains projets ayant réussi à obtenir des fragments de plus de 800 Kb. Le point faible de ces technologies est la qualité de séquençage qui présente un taux d'erreur relativement élevé, bien que les dernières évolutions de la chimie aient réduit ces inconvénients.

En fonction de la longueur des fragments souhaités, on choisira de préférence la technologie SMRT si l'on recherche à la fois des variations ponctuelles et de grande taille en un seul séquençage, ou bien la technologie ONT si on recherche avant tout à produire de très grands fragments, que l'on combinera par exemple avec des données de séquençage short reads pour l'analyse des variations ponctuelles si nécessaire.

Seule la question du séquençage d'ADN a été abordé, mais ces technologies permettent bien sûr l'analyse transcriptomique, mais aussi la détection de la méthylation de l'ADN pour des analyses épigénétiques.

Le facteur limitant pour l'emploi de ces technologies reste actuellement le coût qui est supérieur à celui d'un génome ou d'un exome séquencé en shortReads. Au fur et à mesure que les coûts de production des séquences vont diminuer, de nombreux projets vont basculer depuis les séquençages short reads vers le séquençage long reads, de la même manière que de nombreux projets sont passés du séquençage d'exome vers le génome avec la réduction du coût de séquençage à la base. Au-delà du coût du séquençage en lui-même, il faut prendre en compte la partie informatique qui ne cesse elle aussi d'augmenter. Les besoins de stockage et de calcul pour le passage de l'exome au génome sont importants, et la transition vers le séquençage de 3ème génération le sera tout autant. Les différents outils utilisés pour la détection des SV à partir de ces données (tel que pbsv ou sniffles) utilisent des approches combinatoires (Smolka et al. 2022; Töpfer 2022) qui se révèlent plus consommatrices en ressources que les outils basés sur des approches uniques.

En complément des technologies de séquençage, les technologies de cartographie optique, telle que la plateforme SAPHYR de la société Bionano, génèrent des cartes génétiques complètes et sont capable de détecter des variations de structure avec une résolution de 500 pb. Le principal défaut de cette technique et l'absence totale de lecture base à base, qui ne permet pas en un seul passage d'avoir et la séquence et la structure du génome séquencé. Cette technique relativement coûteuse (environ 1300€ par échantillon) sera donc utilisée dans des maladies induites par des grands réarrangements chromosomique, ou bien en complément d'autres approches de séquençage.

3.2. Impact biologique des variations de structure

3.2.1. L'interprétation des variations de structures

La première grande difficulté avec les variations de structure est, comme nous avons pu le voir jusqu'ici, leurs détections. La question suivante est celle de l'interprétation qui reste encore dans certains cas relativement complexe. Afin d'interpréter les variations de structure, il est important de considérer à la fois le contenu mais aussi les éventuels points d'insertion de la variation.

En fonction du type d'événement, les conséquences ne seront pas les mêmes. Dans le cas où l'on observe une délétion, il faut se poser la question du contenu de cette dernière : affecte-t-elle un gène codant ? Si oui, est ce que celui-ci est haploinsuffisant ou non ? Si la délétion est intronique, affecte-t-elle l'épissage ou un élément de régulation de l'expression génique ? Et enfin si elle est intergénique, emporte-t-elle des éléments de régulation impactant des gènes proches ? Dans le cas des duplications, il faut définir si ces dernières chevauchent entièrement un gène ainsi que les éléments de régulation associés. Il faut aussi prendre en compte la localisation du point d'insertion de la duplication si celle-ci n'est pas en tandem, et aussi définir l'orientation de la duplication. Des questions similaires vont se poser concernant les inversions et les insertions d'éléments mobiles. Au final, c'est principalement le contenu génique des événements et les points d'insertion ou de cassure qui sont les éléments clés pour pouvoir interpréter l'effet d'une variation.

Certains logiciels, tel qu'AnnotSV (Geoffroy et al. 2018), annotent les variations et fournissent une partie des informations nécessaires à l'interprétation, tel que les éléments affectés ou bordant l'événement (gène, élément de régulation, etc) ainsi que les fréquences dans les bases de données publiques tel que la DGV ou gnomAD-SV. Néanmoins, ce genre d'outil présente un certain nombre de limites telles que la non prise en compte des points d'insertion des duplications ou leur orientation.

De plus, comme nous l'avons rencontré dans notre étude, le caractère "complet" d'un gène affecté par une variation peut être biaisé par la méthode de détection : si les cibles de capture ne couvrent pas exactement l'ensemble du gène, l'événement sera rapporté comme partiel. Dans le cas d'un séquençage avec peu d'individus à la fois (comme dans le cadre du diagnostic par exemple), il est possible de parcourir le fichier et de vérifier ces différentes discordances qui peuvent être introduites. Mais dans le cas d'une étude contenant plusieurs milliers d'individus, cette curation manuelle est difficilement applicable, et il faut alors trouver une méthode pour résoudre ce problème.

Dans le cas de notre étude, nous avons adapté notre stratégie aux données auxquelles nous avions accès. Notre jeu de données étant constitué de séquençages en exome, nous avons dû travailler avec les différents kits de capture qui variaient entre les études : la recherche de récurrence au niveau des CNVs étant difficile à réaliser sans perdre de l'information, nous avons fait le choix de travailler au niveau des gènes/transcrits. Les différents kits de capture ne ciblant pas les mêmes exons et ne couvrant pas les transcrits de manière égale, nous avons redéfini les coordonnées de chaque transcrit pour chaque kit de capture. Enfin, nous avons assimilé les délétions de transcrits, qu'ils soient affectés complètement ou partiellement, comme une perte de fonction en nous focalisant sur les gènes potentiellement haploinsuffisants. Pour les duplications, nous avons considéré uniquement les transcrits entièrement affectés, considérant que leur duplication conduirait à une augmentation de la production de la protéine.

En conclusion, l'interprétation fine de chaque événement nécessite une curation manuelle et une expertise, avec un retour aux données brutes afin de bien interpréter dans un premier temps, la structure réelle de l'événement puis son impact potentiel sur la pathologie d'intérêt.

3.2.2. L'apport des variations de structures dans la recherche et le diagnostic

L'implication des SV dans un certain nombre de pathologies humaines est bien identifiée, depuis les maladies chromosomiques jusqu'aux délétions ou duplications responsables de maladies mendéliennes, tel que nous l'avons vu en introduction de ce document. La part des diagnostics imputables aux variations de structure varie en fonction de la pathologie, avec par exemple environ 10 % de diagnostics expliqués par des SV dans les troubles du spectre autistique (Shen et al. 2010; Tammimies et al. 2015) et entre 10 et 20% dans des cas de troubles neurodéveloppementaux (Hehir-Kwa, Pfundt, et Veltman 2015; Pfundt et al. 2017).

La validation de l'implication d'un SV dans une pathologie peut se faire par différentes approches : soit une recherche d'enrichissement dans des cohortes cas témoins, soit par l'identification de variations rares, voire ultra rares dans des familles ou des cohortes très bien caractérisées de patients. Dans le cas des événements rares, voire singletons, on s'appuiera généralement sur des preuves biologiques pour affirmer l'effet sur la pathologie. Cette approche est efficace quand le risque associé est fort, signe d'un effet biologique important. Pour étudier l'effet des facteurs de risque ultra rares, la difficulté principale est la récurrence trop faible pour permettre une validation CNV par CNV. L'équipe du Pr Jacquemont a proposé de quantifier l'impact d'un CNV en calculant un score basé sur différentes caractéristiques du contenu génique de l'événement (Huguet et al. 2018). On peut par exemple se baser sur le nombre de gènes, la fréquence de ces derniers dans la population, leur score de pression de sélection tel que la pLI, et bien sûr les connaissances établies liées au phénotype étudié. On considère un score élevé si l'événement porte de nombreux gènes fortement conservés et ayant un lien avec notre pathologie. Néanmoins, il reste difficile d'utiliser ce type d'approche dans le cadre du conseil génétique pour les CNV, bien que les arguments d'interprétation des sociétés savantes recoupent en partie ce type d'argument (taille, contenu génique, tolérance à la perte de fonction, etc).

Les limites de détection (selon la technologie employée) et d'interprétation des CNV conduisent à une sous-évaluation de la part des SV dans l'origine génétique des pathologies. Il est probable que la proportion de SV va augmenter dans les années qui viennent, entre l'étude de manière plus systématique de ces dernières, que cela soit par du séquençage de 2ème ou de 3ème génération, mais aussi avec une résolution de plus en plus élevée, révélant des délétions ou des insertions de l'ordre de l'exon, jusqu'alors très peu étudiées dans des études d'enrichissement de cas et de témoins.

3.2.3. Impact des CNVs dans la maladie d'Alzheimer

Les origines génétiques de la maladie d'Alzheimer sont complexes, allant de formes autosomiques dominantes provenant de variations dans les gènes *APP*, *PSEN1*, *PSEN2* à des formes complexes largement déterminées par l'accumulation de facteurs de risque. Parmi ces facteurs de risque, il existe une grande variabilité à la fois dans la fréquence et dans le risque associé à ces variations. Jusqu'à maintenant, seuls des SNV ont été associés de manière significative avec la maladie d'Alzheimer, ces facteurs ayant été identifiés soit par des études de GWAS pour les variants fréquents, soit par des études de séquençage pour les variants rares. Les SV, et plus particulièrement les CNVs, n'avaient été

identifiés que dans les formes autosomiques dominantes (duplication complète du gène *APP*, délétion partielle du gène *PSEN1*). Tous les autres CNVs potentiellement impliqués dans la pathologie sont des événements fréquents identifiés dans des études de puces de génotypage mais sans réplication dans des études indépendantes, ou bien des variants ultra rares, voire des singletons, et pour lesquels un lien biologique potentiel a été identifié. Il est donc nécessaire de rassembler de très grandes cohortes de cas et de témoins afin d'avoir la puissance statistique nécessaire pour détecter un signal.

Dans notre étude, nous avons réussi à identifier 18 gènes affectés de manière différente entre les cas et les témoins répartis sur 5 loci, avec des p-valeurs associées variant entre 1,05x10-3 et 5,44x10-5. L'un des résultats les plus intéressants concerne la région centrale de la délétion 22.q11.2, un facteur de risque (aussi connu sous le nom de PIEV - Pénétrance Incomplète et/ou Expressivité Variable) de maladie du neurodéveloppement (Rump et al. 2014; Woodward et al. 2019). Le lien entre le réarrangement et la pathologie reste encore à préciser. Une explication pourrait être un effet aspécifique de la délétion, celle-ci induisant potentiellement une diminution de la réserve cognitive du patient (Nelson et al. 2021; Soldan et al. 2017; Stern 2012) : si un individu possède une réserve cognitive plus faible, la perte neuronale induite par la maladie aura des conséquences cognitives détectables potentiellement plus précoces. Si on considère deux individus avec le même fond génétique compatible avec le développement d'une maladie d'Alzheimer, celui présentant la délétion 22q11.2 en subirait alors les effets plus précocement. Si cet effet aspécifique est confirmé, on peut alors se poser la question de l'impact des autres PIEV identifiés dans les troubles neurodéveloppementaux. Une perspective de prolongement de nos travaux pourrait donc être d'aller regarder spécifiquement dans ces régions d'intérêt et de voir si un enrichissement est détecté chez les cas ou chez les témoins. Il faudra pour cela revenir aux données brutes d'origine, les nombreuses filtrations appliquées pouvant exclure les CNVs dans ces régions d'intérêt.

En plus de ces 5 loci à répliquer, nous avons pu identifier des délétions dans des gènes préalablement identifiés comme facteur de risque pour la maladie, avec 10 délétions du gène *ABCA7* (4 chez des EOAD, 3 chez des LOAD et 3 chez des témoins) et 3 délétions partielles du gène *ABCA1*, toutes identifiées chez des EOAD. Ces délétions renforcent le signal initialement détecté à partir des SNV perte de fonction. Ceci confirme le mécanisme perte de fonction comme facteur de risque pour ces gènes, les délétions représentant 9% des variations perte de fonction pour *ABCA7* et 10% pour *ABCA1*. En plus de ces deux gènes identifiés à partir de variations rares, Nous avons identifié des délétions et variations perte de fonction du gène *CTSB*, gène qui avait été identifié à partir des études de GWAS,

indiquant que c'est un mécanisme perte de fonction qui pourrait médier le risque pour la maladie, sans pour autant connaitre l'effet biologique associé au locus commun de GWAS.

De manière intéressante, nous avons identifié lors des différents contrôles qualité des grands réarrangements emportant soit un bras chromosomique, soit un chromosome complet. Dans la majorité des cas, ces grands événements sont incompatibles avec la vie ou un développement normal (Berger et al. 1992; Matutes et al. 1996; Paulsson et Johansson 2007) qui permettrait l'inclusion du porteur, soit comme un cas de maladie d'Alzheimer, soit comme un témoin. Ayant accès à l'ADN de certains des porteurs (K Le Guennec et al. 2017), nous avons pu confirmer l'événement et observer que celui-ci était présent en mosaïque. Nous en avons donc conclu que ces CNVs sont potentiellement somatiques, probablement spécifiques du tissu étudié, le sang, car ils ne sauraient être constitutionnels. L'hypothèse principale est qu'il s'agisse de clones hématopoïétiques, c'est à dire des populations de cellules sanguines provenant d'une même cellule souche hématopoïétique. Avec l'âge et la multiplication des mitoses que subissent les tissus en constant renouvèlement, des mutations somatiques peuvent apparaitre lors de la mitose. Certaines de ces mutations confèrent un avantage sélectif à la cellule qui le porte, par exemple des mutations perte de fonction dans le gène *TP53* ou des mutations récurrentes dans le gène *DNMT3A* (Jakubek, Reiner, et Honigberg 2023). Ainsi, un clone hématopoïétique peut se constituer et représenter une proportion significative de toutes les cellules hématopoïétiques souches et leurs cellules filles différenciées. Avec le temps, ces clones peuvent eux-mêmes acquérir d'autres mutations somatiques qui les caractérisent, et parmi elles, ces sous populations peuvent présenter des réarrangements chromosomiques qui ne sont pas éliminés du fait du caractère immortel ou quasiment, du clone initial ayant introduit dans son génome une mutation dite driver. (D. G. Wang et al. 1998). Plusieurs études ont été menées sur ces clones hématopoïétiques (où CHIP – Clonal Hematopoiesis of Indeterminate Potential). Ces CHIP peuvent conduire à des troubles hématologiques tel que des leucémies avec un risque multiplié 10 d'hémopathies malignes (Chung et al. 2000; Lu et al. 2009) mais elles peuvent d'également rester sans conséquence clinique pendant un très long moment. Elles ont également été associées au risque cardiovasculaire avec un risque multiplié par deux d'infarctus du myocarde ou cérébral (Evans, Sano, et Walsh 2020). Le facteur le plus associé au risque de CHIP est l'âge, avec des preuves fortes d'une accumulation avec l'âge du porteur de CHIP et de mutations dites driver (Acuna-Hidalgo et al. 2017). Une étude a tenté d'établir le lien entre ces CHIPs et la maladie d'Alzheimer (Bouzid et al. 2023), identifiant un potentiel effet protecteur de ces derniers (OR=0.64 [0.52 - 0.79], p-valeur = 3.8×10^{-5}). L'hypothèse physiopathologique serait un lien entre les cellules de l'immunité innée (monocytes) et un passage de ces cellules dans le cerveau

pour y subir une différenciation microgliale et lutter contre la neurodégénérescence, faisant également le lien entre certains gènes associés à la maladie d'Alzheimer et qui sont d'expression et de fonction microgliale, comme *TREM2*. Bien que ces résultats étaient corrigés sur l'âge, nous n'avons dans notre étude pas reproduit ces résultats (OR= 1.0009, SE=7.39x10^{-4}, *P*=0.24) en analyse multivariée, mais avons relié la présence des CHIP uniquement à l'âge des porteurs (OR=1.0001, SE=2.9x10^{-5}, *P*=1.14x10^{-5}). Il existe des différences méthodologiques, comme le fait que nous ne détections pas les CHIP de la même manière : elles sont identifiées à travers les variations ponctuelles driver, à l'origine des CHIP, dans la littérature, alors que nous identifions des grands CNV qui sont des potentiels marqueurs de CHIPs. Nous allons poursuivre cette étude en nous intéressant également aux variations driver. Par ailleurs, nous avons travaillé à partir d'études cas-témoins quand d'autres études ont travaillé en partie à partir de cohortes qui sont probablement moins biaisées, mais également à partir des données d'exomes d'ADSP, ce qui est plus surprenant car nous avons également étudié ces données. Cette analyse des CHIP n'étant qu'une partie de notre QC, nous n'avons pour le moment pas poussé les analyses. Il faudra néanmoins aller plus loin sur cet aspect, et faire une analyse séparée d'ADSP et d'ADES, une analyse des CNV mais aussi des variations driver, restreindre l'analyse aux échantillons avec du sang comme prélèvement (certains patients ont eu un séquençage d'ADN extrait de cerveau), et restreindre aux participants avec APOE 3-3 du fait d'un biais de sélection dans ADSP et comme fait dans Bouzid et al., 2023.

Afin de prolonger le projet, plusieurs possibilités sont envisageables. La première possibilité est de continuer l'exploration du jeu de données actuelles en relâchant certains filtres de façon à aller explorer d'autres événements. Nous pourrions explorer les CNV monoexoniques, exclus de l'analyse actuelle à cause de la valeur prédictive positive qui est seulement de 70% pour les CNVs mono exoniques contre plus de 90% pour les CNVs emportant au moins deux exons. Un travail important de curation sera nécessaire, mais ces petits événements ne sont généralement pas exploités dans les grandes études et peuvent potentiellement contenir des variations non détectées jusqu'à maintenant. Ensuite, et afin de gagner en puissance, nous pourrions augmenter l'effectif de l'étude en séquençant de nouveaux patients et de nouveaux témoins, toujours en utilisant de la capture d'exome. Pour être le plus efficace possible il faudrait que tous ces individus soient séquencés avec le même kit de capture de façon à limiter la variabilité. Une autre possibilité envisagée est de passer au séquençage de génome avec des courtes lectures ou de 3ème génération sur la technologie SMRT de PacBio de façon à pouvoir analyser à la fois les régions non codantes et les variations de structure ainsi que les répétitions.

En conclusion, les CNVs sont une partie minoritaire, en proportion, des allèles responsables ou associés à des maladies humaines. Dans le contexte global de la maladie, ils sont difficiles à identifier car rares dans la population mais au niveau individuel, ils peuvent expliquer à eux seuls l'origine de la pathologie, d'où un intérêt très important et un enjeu de détection et d'interprétation pour certains. En cela, les variations de structure sont à considérer de manière similaire aux SNVs rares : difficiles à identifier car peu présent et nécessité de les agréger pour identifier un signal significatif et pourtant ayant un rôle très important au niveau individuel, avec une difficulté supplémentaire provenant de la détection des CNVs. Les CNVs sont aussi à l'origine d'avancées dans la compréhension des mécanismes impliqués dans les pathologies en apportant des arguments physiopathologiques, tel que les duplications du gène *APP* dans la maladie d'Alzheimer ou celle du gène *SNCA* dans la maladie de Parkinson.

D'un point de vue méthodologique, l'utilisation de plusieurs milliers d'individus avec une variabilité technique importante marque le passage au "big data", nécessitant l'utilisation de ressources de calcul toujours plus importantes et la mise en place de méthodes spécifiques à cette volumétrie de données. Néanmoins, plusieurs grands défis restent devant nous, avec l'intégration de données toujours plus variées. Cette variabilité provenant de l'origine des données, issues de séquençage d'exome ou de génome, et produit par des technologies de short ou de long read, et provenant de populations diverses. La variabilité proviendra aussi du type des variations étudiées et la nécessité dans le futur d'intégrer tous les types de variations de façon à obtenir une vision globale des facteurs génétiques des pathologies. Ceci passera aussi par l'interprétation du non codant qui reste dans la majorité des études peu exploité et surtout très peu interprété. La diversité des variations non codantes étant très importante, elle nécessitera un retour à l'analyse individuelle, nécessitant une étude au cas par cas avant de pouvoir généraliser et revenir à des analyses globales.

Références bibliographiques

1000 Genomes Project Consortium, Gonçalo R. Abecasis, David Altshuler, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Richard A. Gibbs, Matt E. Hurles, et Gil A. McVean. 2010. « A Map of Human Genome Variation from Population-Scale Sequencing ». *Nature* 467 (7319): 1061-73. https://doi.org/10.1038/nature09534.

1000 Genomes Project Consortium, Goncalo R. Abecasis, Adam Auton, Lisa D. Brooks, Mark A. DePristo, Richard M. Durbin, Robert E. Handsaker, Hyun Min Kang, Gabor T. Marth, et Gil A. McVean. 2012. « An Integrated Map of Genetic Variation from 1,092 Human Genomes ». *Nature* 491 (7422): 56-65. https://doi.org/10.1038/nature11632.

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. « A Global Reference for Human Genetic Variation ». *Nature* 526 (7571): 68-74. https://doi.org/10.1038/nature15393.

Abyzov, Alexej, Alexander E. Urban, Michael Snyder, et Mark Gerstein. 2011. « CNVnator: An Approach to Discover, Genotype, and Characterize Typical and Atypical CNVs from Family and Population Genome Sequencing ». *Genome Research* 21 (6): 974-84. https://doi.org/10.1101/gr.114876.110.

Acuna-Hidalgo, Rocio, Hilal Sengul, Marloes Steehouwer, Maartje van de Vorst, Sita H. Vermeulen, Lambertus A. L. M. Kiemeney, Joris A. Veltman, Christian Gilissen, et Alexander Hoischen. 2017. « Ultra-Sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life ». *American Journal of Human Genetics* 101 (1): 50-64. https://doi.org/10.1016/j.ajhg.2017.05.013.

Aloui, Chaker, Stéphanie Guey, Eva Pipiras, Manoelle Kossorotoff, Sophie Guéden, Michaelle Corpechot, Pierre Bessou, et al. 2020. « Xq28 Copy Number Gain Causing Moyamoya Disease and a Novel Moyamoya Syndrome ». *Journal of Medical Genetics* 57 (5): 339-46. https://doi.org/10.1136/jmedgenet-2019-106525.

Backenroth, Daniel, Jason Homsy, Laura R. Murillo, Joe Glessner, Edwin Lin, Martina Brueckner, Richard Lifton, Elizabeth Goldmuntz, Wendy K. Chung, et Yufeng Shen. 2014. « CANOES: Detecting Rare Copy Number Variants from Whole Exome Sequencing Data ». *Nucleic Acids Research* 42 (12): e97. https://doi.org/10.1093/nar/gku345.

Baker, Emily, Rebecca Sims, Ganna Leonenko, Aura Frizzati, Janet C. Harwood, Detelina Grozeva, GERAD/PERADES, et al. 2019. « Gene-Based Analysis in HRC Imputed Genome Wide Association Data Identifies Three Novel Genes for Alzheimer's Disease ». *PloS One* 14 (7): e0218111. https://doi.org/10.1371/journal.pone.0218111.

Batzer, Mark A., et Prescott L. Deininger. 2002. « Alu Repeats and Human Genomic Diversity ». *Nature Reviews Genetics* 3 (5): 370-79. https://doi.org/10.1038/nrg798.

Bellenguez, Céline, Camille Charbonnier, Benjamin Grenier-Boley, Olivier Quenez, Kilan Le Guennec, Gaël Nicolas, Ganesh Chauhan, et al. 2017. « Contribution to Alzheimer's Disease Risk of Rare Variants in TREM2, SORL1, and ABCA7 in 1779 Cases and 1273 Controls ». *Neurobiology of Aging* 59 (novembre): 220.e1-220.e9. https://doi.org/10.1016/j.neurobiolaging.2017.07.001.

Bellenguez, Céline, Fahri Küçükali, Iris E. Jansen, Luca Kleineidam, Sonia Moreno-Grau, Najaf Amin, Adam C. Naj, et al. 2022. « New Insights into the Genetic Etiology of Alzheimer's Disease and Related Dementias ». *Nature Genetics* 54 (4): 412-36. https://doi.org/10.1038/s41588-022-01024-z.

Benjamini, Yuval, et Terence P. Speed. 2012. « Summarizing and Correcting the GC Content Bias in High-Throughput Sequencing ». *Nucleic Acids Research* 40 (10): e72. https://doi.org/10.1093/nar/gks001.

Bennett, Simon T., Colin Barnes, Anthony Cox, Lisa Davies, et Clive Brown. 2005. « Toward the 1,000 Dollars Human Genome ». *Pharmacogenomics* 6 (4): 373-82. https://doi.org/10.1517/14622416.6.4.373.

Berger, R., M. Le Coniat, J. Derré, M. A. Flexor, et J. Hillion. 1992. « Abnormalities of Chromosome 18 in Myelodysplastic Syndromes and Secondary Leukemia ». *Cancer Genetics and Cytogenetics* 63 (2): 97-99. https://doi.org/10.1016/0165-4608(92)90387-n.

Bignell, Graham R., Jing Huang, Joel Greshock, Stephen Watt, Adam Butler, Sofie West, Mira Grigorova, et al. 2004. « High-Resolution Analysis of DNA Copy Number Using Oligonucleotide Microarrays ». *Genome Research* 14 (2): 287-95. https://doi.org/10.1101/gr.2012304.

Bis, Joshua C., Xueqiu Jian, Brian W. Kunkle, Yuning Chen, Kara L. Hamilton-Nelson, William S. Bush, William J. Salerno, et al. 2020. « Whole Exome Sequencing Study Identifies Novel Rare and Common Alzheimer's-Associated Variants Involved in Immune Response and Transcriptional Regulation ». *Molecular Psychiatry* 25 (8): 1859-75. https://doi.org/10.1038/s41380-018-0112-7.

Bouzid, Hind, Julia A. Belk, Max Jan, Yanyan Qi, Chloé Sarnowski, Sara Wirth, Lisa Ma, et al. 2023. « Clonal Hematopoiesis Is Associated with Protection from Alzheimer's Disease ». *Nature Medicine*, juin. https://doi.org/10.1038/s41591-023-02397-2.

Bowles, Kathryn R., Derian A. Pugh, Yiyuan Liu, Tulsi Patel, Alan E. Renton, Sara Bandres-Ciga, Ziv Gan-Or, et al. 2022. « 17q21.31 Sub-Haplotypes Underlying H1-Associated Risk for Parkinson's Disease Are Associated with LRRC37A/2 Expression in Astrocytes ». *Molecular Neurodegeneration* 17 (1): 48. https://doi.org/10.1186/s13024-022-00551-x.

Bradfield, Jonathan P., Hui-Qi Qu, Kai Wang, Haitao Zhang, Patrick M. Sleiman, Cecilia E. Kim, Frank D. Mentch, et al. 2011. « A Genome-Wide Meta-Analysis of Six Type 1 Diabetes Cohorts Identifies Multiple Associated Loci ». Édité par Mark I. McCarthy. *PLoS Genetics* 7 (9): e1002293. https://doi.org/10.1371/journal.pgen.1002293.

Brooks, William S., John B. J. Kwok, Jillian J. Kril, G. Anthony Broe, Peter C. Blumbergs, Anthony E. Tannenberg, Phillipa J. Lamont, Philippa Hedges, et Peter R. Schofield. 2003. « Alzheimer's Disease with Spastic Paraparesis and "cotton Wool" Plaques: Two Pedigrees with PS-1 Exon 9 Deletions ». *Brain: A Journal of Neurology* 126 (Pt 4): 783-91. https://doi.org/10.1093/brain/awg084.

Brouwers, N., C. Van Cauwenberghe, S. Engelborghs, J.-C. Lambert, K. Bettens, N. Le Bastard, F. Pasquier, et al. 2012. « Alzheimer Risk Associated with a Copy Number Variation in the Complement Receptor 1 Increasing C3b/C4b Binding Sites ». *Molecular Psychiatry* 17 (2): 223-33. https://doi.org/10.1038/mp.2011.24.

Callaway, Ewen. 2015. « How Elephants Avoid Cancer ». *Nature*, octobre, nature.2015.18534. https://doi.org/10.1038/nature.2015.18534.

Cameron, Daniel L., Jan Schröder, Jocelyn Sietsma Penington, Hongdo Do, Ramyar Molania, Alexander Dobrovic, Terence P. Speed, et Anthony T. Papenfuss. 2017. « GRIDSS: Sensitive and Specific Genomic Rearrangement Detection Using Positional de Bruijn Graph Assembly ». *Genome Research* 27 (12): 2050-60. https://doi.org/10.1101/gr.222109.117.

Campion, D., C. Dumanchin, D. Hannequin, B. Dubois, S. Belliard, M. Puel, C. Thomas-Anterion, et al. 1999. « Early-Onset Autosomal Dominant Alzheimer Disease: Prevalence, Genetic Heterogeneity, and Mutation Spectrum ». *American Journal of Human Genetics* 65 (3): 664-70. https://doi.org/10.1086/302553.

Campion, D., C. Pottier, G. Nicolas, K. Le Guennec, et A. Rovelet-Lecrux. 2016. « Alzheimer Disease: Modeling an Aβ-Centered Biological Network ». *Molecular Psychiatry* 21 (7): 861-71. https://doi.org/10.1038/mp.2016.38.

Carvalho, Claudia M. B., et James R. Lupski. 2016. « Mechanisms Underlying Structural Variant Formation in Genomic Disorders ». *Nature Reviews. Genetics* 17 (4): 224-38. https://doi.org/10.1038/nrg.2015.25.

Cassinari, Kévin, Olivier Quenez, Géraldine Joly-Hélas, Ludivine Beaussire, Nathalie Le Meur, Mathieu Castelain, Alice Goldenberg, et al. 2019. « A Simple, Universal, and Cost-Efficient Digital PCR Method for the Targeted Analysis of Copy Number Variations ». *Clinical Chemistry*, juillet. https://doi.org/10.1373/clinchem.2019.304246.

Caulin, Aleah F., et Carlo C. Maley. 2011. « Peto's Paradox: Evolution's Prescription for Cancer Prevention ». *Trends in Ecology & Evolution* 26 (4): 175-82. https://doi.org/10.1016/j.tree.2011.01.002.

Charbonnier, F., G. Raux, Q. Wang, N. Drouot, F. Cordier, J. M. Limacher, J. C. Saurin, A. Puisieux, S. Olschwang, et T. Frebourg. 2000. « Detection of Exon Deletions and Duplications of the Mismatch Repair Genes in Hereditary Nonpolyposis Colorectal Cancer Families Using Multiplex Polymerase Chain Reaction of Short Fluorescent Fragments ». *Cancer Research* 60 (11): 2760-63.

Chatron, Nicolas, Kevin Cassinari, Olivier Quenez, Stéphanie Baert-Desurmont, Claire Bardel, Marie-Pierre Buisine, Eduardo Calpena, et al. 2019. « Identification of Mobile Retrocopies during Genetic Testing: Consequences for Routine Diagnosis ». *Human Mutation* 40 (11): 1993-2000. https://doi.org/10.1002/humu.23845.

Chen, Ken, John W Wallis, Michael D McLellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D McGrath, et al. 2009. « BreakDancer: An Algorithm for High-Resolution Mapping of Genomic Structural Variation ». *Nature Methods* 6 (9): 677-81. https://doi.org/10.1038/nmeth.1363.

Chen, Siwei, Laurent C. Francioli, Julia K. Goodrich, Ryan L. Collins, Masahiro Kanai, Qingbo Wang, Jessica Alföldi, et al. 2022. « A Genome-Wide Mutational Constraint Map Quantified from Variation in 76,156 Human Genomes ». Preprint. Genetics. https://doi.org/10.1101/2022.03.20.485034.

Chen, Xiaoyu, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, et Christopher T. Saunders. 2016. « Manta: Rapid Detection of Structural Variants and Indels for Germline and Cancer Sequencing Applications ». *Bioinformatics* 32 (8): 1220-22. https://doi.org/10.1093/bioinformatics/btv710.

Chen, Zhongbo, Jason A. Chen, Aleksey Shatunov, Ashley R. Jones, Stephanie N. Kravitz, Alden Y. Huang, Lauren Lawrence, et al. 2019. « Genome-Wide Survey of Copy Number Variants Finds MAPT Duplications in Progressive Supranuclear Palsy ». *Movement Disorders: Official Journal of the Movement Disorder Society* 34 (7): 1049-59. https://doi.org/10.1002/mds.27702.

Chrestian, Nicolas. 1993. « Hereditary Neuropathy with Liability to Pressure Palsies ». In *GeneReviews®*, édité par Margaret P. Adam, Jerry Feldman, Ghayda M. Mirzaa, Roberta A. Pagon, Stephanie E. Wallace, Lora JH Bean, Karen W. Gripp, et Anne Amemiya. Seattle (WA): University of Washington, Seattle. http://www.ncbi.nlm.nih.gov/books/NBK1392/.

Chung, C. Y., H. Kantarjian, M. Haidar, P. Starostik, T. Manshouri, C. Gidel, E. Freireich, M. Keating, et M. Albitar. 2000. « Deletions in the 13q14 Locus in Adult Lymphoblastic Leukemia: Rate of Incidence and Relevance ». *Cancer* 88 (6): 1359-64.

Collins, Rory. 2012. « What Makes UK Biobank Special? » *The Lancet* 379 (9822): 1173-74. https://doi.org/10.1016/S0140-6736(12)60404-8.

Collins, Ryan L., Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C. Francioli, Amit V. Khera, et al. 2020. « A Structural Variation Reference for Medical and Population Genetics ». *Nature* 581 (7809): 444-51. https://doi.org/10.1038/s41586-020-2287-8.

Collins, Ryan L., Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Amit V. Khera, Laurent C. Francioli, et al. 2019. « An Open Resource of Structural Variation for Medical and Population Genetics ». Preprint. Genomics. https://doi.org/10.1101/578674.

Crane, Paul K., Tatiana Foroud, Thomas J. Montine, et Eric B. Larson. 2017. « Alzheimer's Disease Sequencing Project Discovery and Replication Criteria for Cases and Controls: Data from a Community-Based Prospective Cohort Study with Autopsy Follow-Up ». *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 13 (12): 1410-13. https://doi.org/10.1016/j.jalz.2017.09.010.

David, Stéphanie, Joana Ferreira, Olivier Quenez, Anne Rovelet-Lecrux, Anne-Claire Richard, Marc Vérin, Snejana Jurici, et al. 2016. « Identification of Partial SLC20A2 Deletions in Primary Brain Calcification Using Whole-Exome Sequencing ». *European Journal of Human Genetics: EJHG* 24 (11): 1630-34. https://doi.org/10.1038/ejhg.2016.50.

Dennis, Megan Y., et Evan E. Eichler. 2016. « Human Adaptation and Evolution by Segmental Duplication ». *Current Opinion in Genetics & Development* 41 (décembre): 44-52. https://doi.org/10.1016/j.gde.2016.08.001.

Dennis, Megan Y., Xander Nuttle, Peter H. Sudmant, Francesca Antonacci, Tina A. Graves, Mikhail Nefedov, Jill A. Rosenfeld, et al. 2012. « Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication ». *Cell* 149 (4): 912-22. https://doi.org/10.1016/j.cell.2012.03.033.

Doğan, Mustafa, Recep Eröz, Mehmet Tecellioğlu, Alper Gezdirici, Betül Çevik, et İbrahim Barış. 2022. « Clinical and Molecular Findings in a Turkish Family Who Had a (c.869- 1G>A) Splicing Variant in PSEN1 Gene with A Rare Condition: The Variant Alzheimer's Disease with Spastic Paraparesis ». *Current Alzheimer Research* 19 (3): 223-35. https://doi.org/10.2174/1567205019666220414101251.

Dupont, Jean-Michel, Matthieu Egloff, Paul Kuentz, Valérie Malan, Chantal Missirian, Céline Pebrel-Richard, Serge Romana, Caroline Rooryck-Thambo, Anne-Claude Tabet, et Detlef Trost. 2022. « Guide d'interprétation des CNVs ». https://acpa-achropuce.com/wp-content/uploads/2022/09/20220930-Aide-a-l-interpretation-CNV-revise-2022-VF.pdf.

Eizenga, Jordan M., Adam M. Novak, Jonas A. Sibbesen, Simon Heumos, Ali Ghaffaari, Glenn Hickey, Xian Chang, et al. 2020. « Pangenome Graphs ». *Annual Review of Genomics and Human Genetics* 21 (août): 139-62. https://doi.org/10.1146/annurev-genom-120219-080406.

English, Adam C, William J Salerno, et Jeffrey G Reid. 2014. « PBHoney: Identifying Genomic Variants via Long-Read Discordance and Interrupted Mapping ». *BMC Bioinformatics* 15 (1): 180. https://doi.org/10.1186/1471-2105-15-180.

Evans, Megan A., Soichi Sano, et Kenneth Walsh. 2020. « Cardiovascular Disease, Aging, and Clonal Hematopoiesis ». *Annual Review of Pathology* 15 (janvier): 419-38. https://doi.org/10.1146/annurev-pathmechdis-012419-032544.

Exome Aggregation Consortium, Monkol Lek, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, et al. 2016. « Analysis of Protein-Coding Genetic Variation in 60,706 Humans ». *Nature* 536 (7616): 285-91. https://doi.org/10.1038/nature19057.

Firth, Helen V., Shola M. Richards, A. Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M. Pettett, et Nigel P. Carter. 2009. « DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources ». *The American Journal of Human Genetics* 84 (4): 524-33. https://doi.org/10.1016/j.ajhg.2009.03.010.

Fromer, Menachem, Jennifer L. Moran, Kimberly Chambert, Eric Banks, Sarah E. Bergen, Douglas M. Ruderfer, Robert E. Handsaker, et al. 2012. « Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth ». *American Journal of Human Genetics* 91 (4): 597-607. https://doi.org/10.1016/j.ajhg.2012.08.005.

Fu, Wenqing, Timothy D. O'Connor, Goo Jun, Hyun Min Kang, Goncalo Abecasis, Suzanne M. Leal, Stacey Gabriel, et al. 2013. « Analysis of 6,515 Exomes Reveals the Recent Origin of Most Human Protein-Coding Variants ». *Nature* 493 (7431): 216-20. https://doi.org/10.1038/nature11690.

Fujimoto, Akihiro, Jing Hao Wong, Yukiko Yoshii, Shintaro Akiyama, Azusa Tanaka, Hitomi Yagi, Daichi Shigemizu, Hidewaki Nakagawa, Masashi Mizokami, et Mihoko Shimada. 2021. « Whole-Genome Sequencing with Long Reads Reveals Complex Structure and Origin of Structural Variation in Human Genetic Variations and Somatic Mutations in Cancer ». *Genome Medicine* 13 (1): 65. https://doi.org/10.1186/s13073-021-00883-1.

Genin, E. 2017. « The French Exome (FREX) Project: A Population-Based Panel of Exomes to Help Filter out Common Local Variants. » *Genetic Epidemiology* 41 (7): 644-709. https://doi.org/10.1002/gepi.22062.

Genin, E., D. Hannequin, D. Wallon, K. Sleegers, M. Hiltunen, O. Combarros, M. J. Bullido, et al. 2011. « APOE and Alzheimer Disease: A Major Gene with Semi-Dominant Inheritance ». *Molecular Psychiatry* 16 (9): 903-7. https://doi.org/10.1038/mp.2011.52.

GenomeAsia100K Consortium, Jeffrey D. Wall, Eric W. Stawiski, Aakrosh Ratan, Hie Lim Kim, Changhoon Kim, Ravi Gupta, et al. 2019. « The GenomeAsia 100K Project Enables Genetic Discoveries across Asia ». *Nature* 576 (7785): 106-11. https://doi.org/10.1038/s41586-019-1793-z.

Geoffroy, Véronique, Yvan Herenger, Arnaud Kress, Corinne Stoetzel, Amélie Piton, Hélène Dollfus, et Jean Muller. 2018. « AnnotSV: An Integrated Tool for Structural Variations Annotation ». *Bioinformatics (Oxford, England)* 34 (20): 3572-74. https://doi.org/10.1093/bioinformatics/bty304.

Gianfrancesco, Olympia, Bethany Geary, Abigail L. Savage, Kimberley J. Billingsley, Vivien J. Bubb, et John P. Quinn. 2019. « The Role of SINE-VNTR-Alu (SVA) Retrotransposons in Shaping the Human Genome ». *International Journal of Molecular Sciences* 20 (23): 5977. https://doi.org/10.3390/ijms20235977.

Gong, Liang, Chee-Hong Wong, Wei-Chung Cheng, Harianto Tjong, Francesca Menghi, Chew Yee Ngan, Edison T. Liu, et Chia-Lin Wei. 2018. « Picky Comprehensively Detects High-Resolution Structural Variants in Nanopore Long Reads ». *Nature Methods* 15 (6): 455-60. https://doi.org/10.1038/s41592-018-0002-6.

Grangeon, Lou, Kévin Cassinari, Stéphane Rousseau, Bernard Croisile, Maïté Formaglio, Olivier Moreaud, Jean Boutonnat, et al. 2021. « Early-Onset Cerebral Amyloid Angiopathy and Alzheimer Disease Related to an APP Locus Triplication ». *Neurology. Genetics* 7 (5): e609. https://doi.org/10.1212/NXG.0000000000000609.

Gu, Wenli, Feng Zhang, et James R Lupski. 2008. « Mechanisms for human genomic rearrangements ». *PathoGenetics* 1 (novembre): 4. https://doi.org/10.1186/1755-8417-1-4.

Gudmundsson, Sanna, Moriel Singer-Berk, Nicholas A. Watts, William Phu, Julia K. Goodrich, Matthew Solomonson, Genome Aggregation Database Consortium, Heidi L. Rehm, Daniel G. MacArthur, et Anne O'Donnell-Luria. 2022. « Variant Interpretation Using Population Databases: Lessons from GnomAD ». *Human Mutation* 43 (8): 1012-30. https://doi.org/10.1002/humu.24309.

Gurdasani, Deepti, Tommy Carstensen, Fasil Tekola-Ayele, Luca Pagani, Ioanna Tachmazidou, Konstantinos Hatzikotoulas, Savita Karthikeyan, et al. 2015. « The African Genome Variation Project Shapes Medical Genetics in Africa ». *Nature* 517 (7534): 327-32. https://doi.org/10.1038/nature13997.

Halldorsson, Bjarni V., Hannes P. Eggertsson, Kristjan H. S. Moore, Hannes Hauswedell, Ogmundur Eiriksson, Magnus O. Ulfarsson, Gunnar Palsson, et al. 2022. « The Sequences of 150,119

Genomes in the UK Biobank ». *Nature* 607 (7920): 732-40. https://doi.org/10.1038/s41586-022-04965-x.

Handsaker, Robert E, Joshua M Korn, James Nemesh, et Steven A McCarroll. 2011. « Discovery and Genotyping of Genome Structural Polymorphism by Sequencing on a Population Scale ». *Nature Genetics* 43 (3): 269-76. https://doi.org/10.1038/ng.768.

Hardy, John, et Dennis J. Selkoe. 2002. « The Amyloid Hypothesis of Alzheimer's Disease: Progress and Problems on the Road to Therapeutics ». *Science (New York, N.Y.)* 297 (5580): 353-56. https://doi.org/10.1126/science.1072994.

Hehir-Kwa, Jayne Y., Rolph Pfundt, et Joris A. Veltman. 2015. « Exome Sequencing and Whole Genome Sequencing for the Detection of Copy Number Variation ». *Expert Review of Molecular Diagnostics* 15 (8): 1023-32. https://doi.org/10.1586/14737159.2015.1053467.

Hehir-Kwa, Jayne Y., Bastiaan B. J. Tops, et Patrick Kemmeren. 2018. « The Clinical Implementation of Copy Number Detection in the Age of Next-Generation Sequencing ». *Expert Review of Molecular Diagnostics* 18 (10): 907-15. https://doi.org/10.1080/14737159.2018.1523723.

Heinzen, Erin L., Anna C. Need, Kathleen M. Hayden, Ornit Chiba-Falek, Allen D. Roses, Warren J. Strittmatter, James R. Burke, Christine M. Hulette, Kathleen A. Welsh-Bohmer, et David B. Goldstein. 2010. « Genome-Wide Scan of Copy Number Variation in Late-Onset Alzheimer's Disease ». *Journal of Alzheimer's Disease: JAD* 19 (1): 69-77. https://doi.org/10.3233/JAD-2010-1212.

Heller, David, et Martin Vingron. 2019. « SVIM: Structural Variant Identification Using Mapped Long Reads ». Édité par Inanc Birol. *Bioinformatics* 35 (17): 2907-15. https://doi.org/10.1093/bioinformatics/btz041.

Hert, Daniel G., Christopher P. Fredlake, et Annelise E. Barron. 2008. « Advantages and Limitations of Next-Generation Sequencing Technologies: A Comparison of Electrophoresis and Non-Electrophoresis Methods ». *Electrophoresis* 29 (23): 4618-26. https://doi.org/10.1002/elps.200800456.

Holstege, Henne, Marc Hulsman, Camille Charbonnier, Benjamin Grenier-Boley, Olivier Quenez, Detelina Grozeva, Jeroen G. J. van Rooij, et al. 2022. « Exome Sequencing Identifies Rare Damaging Variants in ATP8B4 and ABCA1 as Risk Factors for Alzheimer's Disease ». *Nature Genetics* 54 (12): 1786-94. https://doi.org/10.1038/s41588-022-01208-7.

Hooli, B. V., Z. M. Kovacs-Vajna, K. Mullin, M. A. Blumenthal, M. Mattheisen, C. Zhang, C. Lange, G. Mohapatra, L. Bertram, et R. E. Tanzi. 2014. « Rare Autosomal Copy Number Variations in Early-Onset Familial Alzheimer's Disease ». *Molecular Psychiatry* 19 (6): 676-81. https://doi.org/10.1038/mp.2013.77.

Huguet, Guillaume, Catherine Schramm, Elise Douard, Lai Jiang, Aurélie Labbe, Frédérique Tihy, Géraldine Mathonnet, et al. 2018. « Measuring and Estimating the Effect Sizes of Copy Number Variants on General Intelligence in Community-Based Samples ». *JAMA Psychiatry* 75 (5): 447-57. https://doi.org/10.1001/jamapsychiatry.2018.0039.

Hussain, Suleman S., Rahul Majumdar, Grace M. Moore, Himanshi Narang, Erika S. Buechelmaier, Maximilian J. Bazil, Pavithran T. Ravindran, et al. 2021. « Measuring Nonhomologous End-Joining, Homologous Recombination and Alternative End-Joining Simultaneously at an Endogenous Locus in Any Transfectable Human Cell ». *Nucleic Acids Research* 49 (13): e74. https://doi.org/10.1093/nar/gkab262.

International Human Genome Sequencing Consortium. 2004. « Finishing the Euchromatic Sequence of the Human Genome ». *Nature* 431 (7011): 931-45. https://doi.org/10.1038/nature03001.

Iqbal, Zamin, Mario Caccamo, Isaac Turner, Paul Flicek, et Gil McVean. 2012. « De Novo Assembly and Genotyping of Variants Using Colored de Bruijn Graphs ». *Nature Genetics* 44 (2): 226-32. https://doi.org/10.1038/ng.1028.

Istrail, Sorin, Granger G. Sutton, Liliana Florea, Aaron L. Halpern, Clark M. Mobarry, Ross Lippert, Brian Walenz, et al. 2004. « Whole-Genome Shotgun Assembly and Comparison of Human Genome Assemblies ». *Proceedings of the National Academy of Sciences of the United States of America* 101 (7): 1916-21. https://doi.org/10.1073/pnas.0307971100.

Jakubek, Yasminka A., Alexander P. Reiner, et Michael C. Honigberg. 2023. « Risk Factors for Clonal Hematopoiesis of Indeterminate Potential and Mosaic Chromosomal Alterations ». *Translational Research: The Journal of Laboratory and Clinical Medicine* 255 (mai): 171-80. https://doi.org/10.1016/j.trsl.2022.11.009.

Kallioniemi, A., O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, et D. Pinkel. 1992. « Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors ». *Science (New York, N.Y.)* 258 (5083): 818-21. https://doi.org/10.1126/science.1359641.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. « The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans ». *Nature* 581 (7809): 434-43. https://doi.org/10.1038/s41586-020-2308-7.

Klambauer, Günter, Karin Schwarzbauer, Andreas Mayr, Djork-Arné Clevert, Andreas Mitterecker, Ulrich Bodenhofer, et Sepp Hochreiter. 2012. « Cn.MOPS: Mixture of Poissons for Discovering Copy Number Variations in next-Generation Sequencing Data with a Low False Discovery Rate ». *Nucleic Acids Research* 40 (9): e69. https://doi.org/10.1093/nar/gks003.

Koning, A. P. Jason de, Wanjun Gu, Todd A. Castoe, Mark A. Batzer, et David D. Pollock. 2011. « Repetitive Elements May Comprise over Two-Thirds of the Human Genome ». *PLoS Genetics* 7 (12): e1002384. https://doi.org/10.1371/journal.pgen.1002384.

Korbel, Jan O., Alexej Abyzov, Xinmeng Jasmine Mu, Nicholas Carriero, Philip Cayting, Zhengdong Zhang, Michael Snyder, et Mark B. Gerstein. 2009. « PEMer: A Computational Framework with Simulation-Based Error Models for Inferring Genomic Structural Variants from Massive Paired-End Sequencing Data ». *Genome Biology* 10 (2): R23. https://doi.org/10.1186/gb-2009-10-2-r23.

Krumm, Niklas, Peter H. Sudmant, Arthur Ko, Brian J. O'Roak, Maika Malig, Bradley P. Coe, NHLBI Exome Sequencing Project, Aaron R. Quinlan, Deborah A. Nickerson, et Evan E. Eichler. 2012. « Copy Number Variation Detection and Genotyping from Exome Sequence Data ». *Genome Research* 22 (8): 1525-32. https://doi.org/10.1101/gr.138115.112.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. « Initial Sequencing and Analysis of the Human Genome ». *Nature* 409 (6822): 860-921. https://doi.org/10.1038/35057062.

Lanoiselée, Hélène-Marie, Gaël Nicolas, David Wallon, Anne Rovelet-Lecrux, Morgane Lacour, Stéphane Rousseau, Anne-Claire Richard, et al. 2017. « APP, PSEN1, and PSEN2 Mutations in Early-Onset Alzheimer Disease: A Genetic Screening Study of Familial and Sporadic Cases ». *PLoS Medicine* 14 (3): e1002270. https://doi.org/10.1371/journal.pmed.1002270.

Layer, Ryan M., Colby Chiang, Aaron R. Quinlan, et Ira M. Hall. 2014. « LUMPY: A Probabilistic Framework for Structural Variant Discovery ». *Genome Biology* 15 (6): R84. https://doi.org/10.1186/gb-2014-15-6-r84.

Le Guennec, K, O Quenez, G Nicolas, D Wallon, S Rousseau, A-C Richard, J Alexander, et al. 2017. « 17q21.31 duplication causes prominent tau-related dementia with increased MAPT expression ». *Molecular Psychiatry* 22 (8): 1119-25. https://doi.org/10.1038/mp.2016.226.

Le Guennec, Kilan, Sarah Veugelen, Olivier Quenez, Maria Szaruga, Stéphane Rousseau, Gaël Nicolas, David Wallon, et al. 2017. « Deletion of Exons 9 and 10 of the Presenilin 1 Gene in a Patient with Early-Onset Alzheimer Disease Generates Longer Amyloid Seeds ». *Neurobiology of Disease* 104 (août): 97-103. https://doi.org/10.1016/j.nbd.2017.04.020.

Lee, Jennifer A., Claudia M. B. Carvalho, et James R. Lupski. 2007. « A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders ». *Cell* 131 (7): 1235-47. https://doi.org/10.1016/j.cell.2007.11.037.

Lee, Wan-Ping, Albert A. Tucci, Mitchell Conery, Yuk Yee Leung, Amanda B. Kuzma, Otto Valladares, Yi-Fan Chou, et al. 2021. « Copy Number Variation Identification on 3,800 Alzheimer's Disease Whole Genome Sequencing Data from the Alzheimer's Disease Sequencing Project ». *Frontiers in Genetics* 12: 752390. https://doi.org/10.3389/fgene.2021.752390.

Levy, Dan, Michael Ronemus, Boris Yamrom, Yoon-ha Lee, Anthony Leotta, Jude Kendall, Steven Marks, et al. 2011. « Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders ». *Neuron* 70 (5): 886-97. https://doi.org/10.1016/j.neuron.2011.05.015.

Li, Zhichao, Xiaosen Jiang, Mingyan Fang, Yong Bai, Siyang Liu, Shujia Huang, et Xin Jin. 2023. « CMDB: The Comprehensive Population Genome Variation Database of China ». *Nucleic Acids Research* 51 (D1): D890-95. https://doi.org/10.1093/nar/gkac638.

Liang, Zhengyu, et Xiang-Dong Fu. 2021. « 3D Genome Encoded by LINE and SINE Repeats ». *Cell Research* 31 (6): 603-4. https://doi.org/10.1038/s41422-021-00485-x.

Lieber, Michael R. 2008. « The Mechanism of Human Nonhomologous DNA End Joining ». *The Journal of Biological Chemistry* 283 (1): 1-5. https://doi.org/10.1074/jbc.R700039200.

Littlejohns, Thomas J., Jo Holliday, Lorna M. Gibson, Steve Garratt, Niels Oesingmann, Fidel Alfaro-Almagro, Jimmy D. Bell, et al. 2020. « The UK Biobank Imaging Enhancement of 100,000 Participants: Rationale, Data Collection, Management and Future Directions ». *Nature Communications* 11 (1): 2624. https://doi.org/10.1038/s41467-020-15948-9.

Lu, Gary, C. Cameron Yin, L. Jeffrey Medeiros, et Lynne V. Abruzzo. 2009. « Deletion 15q as the Sole Abnormality in Acute Myeloid Leukemia: Report of Three Cases and Review of the Literature ». *Cancer Genetics and Cytogenetics* 188 (2): 118-23. https://doi.org/10.1016/j.cancergencyto.2008.09.006.

Lupski, J. R. 1999. « Charcot-Marie-Tooth Polyneuropathy: Duplication, Gene Dosage, and Genetic Heterogeneity ». *Pediatric Research* 45 (2): 159-65. https://doi.org/10.1203/00006450-199902000-00001.

Lupski, J. R., et C. A. Garcia. 1992. « Molecular Genetics and Neuropathology of Charcot-Marie-Tooth Disease Type 1A ». *Brain Pathology (Zurich, Switzerland)* 2 (4): 337-49. https://doi.org/10.1111/j.1750-3639.1992.tb00710.x.

Lupski, J. R., R. M. de Oca-Luna, S. Slaugenhaupt, L. Pentao, V. Guzzetta, B. J. Trask, O. Saucedo-Cardenas, et al. 1991. « DNA Duplication Associated with Charcot-Marie-Tooth Disease Type 1A ». *Cell* 66 (2): 219-32. https://doi.org/10.1016/0092-8674(91)90613-4.

Lynch, M., et J. S. Conery. 2000. « The Evolutionary Fate and Consequences of Duplicate Genes ». *Science (New York, N.Y.)* 290 (5494): 1151-55. https://doi.org/10.1126/science.290.5494.1151.

MacDonald, Jeffrey R., Robert Ziman, Ryan K. C. Yuen, Lars Feuk, et Stephen W. Scherer. 2014. « The Database of Genomic Variants: A Curated Collection of Structural Variation in the Human Genome ». *Nucleic Acids Research* 42 (Database issue): D986-992. https://doi.org/10.1093/nar/gkt958.

Matutes, E., D. Oscier, J. Garcia-Marco, J. Ellis, A. Copplestone, R. Gillingham, T. Hamblin, D. Lens, G. J. Swansbury, et D. Catovsky. 1996. « Trisomy 12 Defines a Group of CLL with Atypical Morphology: Correlation between Cytogenetic, Clinical and Laboratory Features in 544 Patients ». *British Journal of Haematology* 92 (2): 382-88. https://doi.org/10.1046/j.1365-2141.1996.d01-1478.x.

McDonald-McGinn, Donna M., et Kathleen E. Sullivan. 2011. « Chromosome 22q11.2 Deletion Syndrome (DiGeorge Syndrome/Velocardiofacial Syndrome) ». *Medicine* 90 (1): 1-18. https://doi.org/10.1097/MD.0b013e3182060469.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. « The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data ». *Genome Research* 20 (9): 1297-1303. https://doi.org/10.1101/gr.107524.110.

Meienberg, Janine, Rémy Bruggmann, Konrad Oexle, et Gabor Matyas. 2016. « Clinical Sequencing: Is WGS the Better WES? » *Human Genetics* 135 (3): 359-62. https://doi.org/10.1007/s00439-015-1631-9.

Midha, Mohit K., Mengchu Wu, et Kuo-Ping Chiu. 2019. « Long-Read Sequencing in Deciphering Human Genetics to a Greater Depth ». *Human Genetics* 138 (11-12): 1201-15. https://doi.org/10.1007/s00439-019-02064-y.

Miga, Karen H., et Ting Wang. 2021. « The Need for a Human Pangenome Reference Sequence ». *Annual Review of Genomics and Human Genetics* 22 (août): 81-102. https://doi.org/10.1146/annurev-genom-120120-081921.

Miguel, Laetitia, Anne Rovelet-Lecrux, Pascal Chambon, Géraldine Joly-Helas, Stéphane Rousseau, David Wallon, Stéphane Epelbaum, et al. 2022. « Generation of 17q21.31 Duplication IPSC-Derived Neurons as a Model for Primary Tauopathies ». *Stem Cell Research* 61 (mai): 102762. https://doi.org/10.1016/j.scr.2022.102762.

Morrison, Sinead, Samuel J. R. A. Chawner, Therese A. M. J. van Amelsvoort, Ann Swillen, Claudia Vingerhoets, Elfi Vergaelen, David E. J. Linden, Stefanie Linden, Michael J. Owen, et Marianne B. M. van den Bree. 2020. « Cognitive Deficits in Childhood, Adolescence and Adulthood in 22q11.2 Deletion Syndrome and Association with Psychopathology ». *Translational Psychiatry* 10 (1): 53. https://doi.org/10.1038/s41398-020-0736-7.

Nelson, Monica E., Dylan J. Jester, Andrew J. Petkus, et Ross Andel. 2021. « Cognitive Reserve, Alzheimer's Neuropathology, and Risk of Dementia: A Systematic Review and Meta-Analysis ». *Neuropsychology Review* 31 (2): 233-50. https://doi.org/10.1007/s11065-021-09478-4.

Nicolas, G., C. Charbonnier, D. Wallon, O. Quenez, C. Bellenguez, B. Grenier-Boley, S. Rousseau, et al. 2016. « SORL1 Rare Variants: A Major Risk Factor for Familial Early-Onset Alzheimer's Disease ». *Molecular Psychiatry* 21 (6): 831-36. https://doi.org/10.1038/mp.2015.121.

Nijkamp, Jurgen F., Marcel A. Van Den Broek, Jan-Maarten A. Geertman, Marcel J. T. Reinders, Jean-Marc G. Daran, et Dick De Ridder. 2012. « *De Novo* Detection of Copy Number Variation by Co-Assembly ». *Bioinformatics* 28 (24): 3195-3202. https://doi.org/10.1093/bioinformatics/bts601.

Nobile, Carlo, Luisa Toffolatti, Francesca Rizzi, Barbara Simionati, Vincenzo Nigro, Barbara Cardazzo, Tomaso Patarnello, Giorgio Valle, et Gian Antonio Danieli. 2002. « Analysis of 22 Deletion Breakpoints in Dystrophin Intron 49 ». *Human Genetics* 110 (5): 418-21. https://doi.org/10.1007/s00439-002-0721-7.

Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, et al. 2022. « The Complete Sequence of a Human Genome ». *Science (New York, N.Y.)* 376 (6588): 44-53. https://doi.org/10.1126/science.abj6987.

O'Roak, Brian J., Pelagia Deriziotis, Choli Lee, Laura Vives, Jerrod J. Schwartz, Santhosh Girirajan, Emre Karakoc, et al. 2011. « Exome Sequencing in Sporadic Autism Spectrum Disorders Identifies Severe de Novo Mutations ». *Nature Genetics* 43 (6): 585-89. https://doi.org/10.1038/ng.835.

Osborne, L. R. 1999. « Williams-Beuren Syndrome: Unraveling the Mysteries of a Microdeletion Disorder ». *Molecular Genetics and Metabolism* 67 (1): 1-10. https://doi.org/10.1006/mgme.1999.2844.

Outtaleb, Fatima Zahra, Rachida Errahli, Nora Imelloul, Ghizlane Jabrane, Nadia Serbati, et Hind Dehbi. 2020. « [Trisomy 18 or postnatal Edward´s syndrome: descriptive study conducted at the University Hospital Center of Casablanca and literature review] ». *The Pan African Medical Journal* 37: 309. https://doi.org/10.11604/pamj.2020.37.309.26205.

Paulsson, K., et B. Johansson. 2007. « Trisomy 8 as the Sole Chromosomal Aberration in Acute Myeloid Leukemia and Myelodysplastic Syndromes ». *Pathologie-Biologie* 55 (1): 37-48. https://doi.org/10.1016/j.patbio.2006.04.007.

Perry, George H., Nathaniel J. Dominy, Katrina G. Claw, Arthur S. Lee, Heike Fiegler, Richard Redon, John Werner, et al. 2007. « Diet and the Evolution of Human Amylase Gene Copy Number Variation ». *Nature Genetics* 39 (10): 1256-60. https://doi.org/10.1038/ng2123.

Peto, R, F J Roe, P N Lee, L Levy, et J Clack. 1975. « Cancer and Ageing in Mice and Men ». *British Journal of Cancer* 32 (4): 411-26. https://doi.org/10.1038/bjc.1975.242.

Pfundt, Rolph, Marisol del Rosario, Lisenka E.L.M. Vissers, Michael P. Kwint, Irene M. Janssen, Nicole de Leeuw, Helger G. Yntema, et al. 2017. « Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders ». *Genetics in Medicine* 19 (6): 667-75. https://doi.org/10.1038/gim.2016.163.

Plagnol, Vincent, James Curtis, Michael Epstein, Kin Y. Mok, Emma Stebbings, Sofia Grigoriadou, Nicholas W. Wood, et al. 2012. « A Robust Model for Read Count Data in Exome Sequencing Experiments and Implications for Copy Number Variant Calling ». *Bioinformatics (Oxford, England)* 28 (21): 2747-54. https://doi.org/10.1093/bioinformatics/bts526.

Pollard, Martin O., Deepti Gurdasani, Alexander J. Mentzer, Tarryn Porter, et Manjinder S. Sandhu. 2018. « Long Reads: Their Purpose and Place ». *Human Molecular Genetics* 27 (R2): R234-41. https://doi.org/10.1093/hmg/ddy177.

Pottier, C., D. Hannequin, S. Coutant, A. Rovelet-Lecrux, D. Wallon, S. Rousseau, S. Legallic, et al. 2012. « High Frequency of Potentially Pathogenic SORL1 Mutations in Autosomal Dominant Early-Onset Alzheimer Disease ». *Molecular Psychiatry* 17 (9): 875-79. https://doi.org/10.1038/mp.2012.15.

Pottier, Cyril, David Wallon, Anne Rovelet Lecrux, David Maltete, Stephanie Bombois, Snejana Jurici, Thierry Frebourg, Didier Hannequin, et Dominique Campion. 2012. « Amyloid-β Protein Precursor Gene Expression in Alzheimer's Disease and Other Conditions ». *Journal of Alzheimer's Disease: JAD* 28 (3): 561-66. https://doi.org/10.3233/JAD-2011-111148.

Prihar, G., A. Verkkoniem, J. Perez-Tur, R. Crook, S. Lincoln, H. Houlden, M. Somer, et al. 1999. « Alzheimer Disease PS-1 Exon 9 Deletion Defined ». *Nature Medicine* 5 (10): 1090. https://doi.org/10.1038/13383.

Quinlan, Aaron R., Royden A. Clark, Svetlana Sokolova, Mitchell L. Leibowitz, Yujun Zhang, Matthew E. Hurles, Joshua C. Mell, et Ira M. Hall. 2010. « Genome-Wide Mapping and Assembly of Structural Variant Breakpoints in the Mouse Genome ». *Genome Research* 20 (5): 623-35. https://doi.org/10.1101/gr.102970.109.

Rafi, Syed K., et Merlin G. Butler. 2020. « The 15q11.2 BP1-BP2 Microdeletion (Burnside-Butler) Syndrome: In Silico Analyses of the Four Coding Genes Reveal Functional Associations with Neurodevelopmental Phenotypes ». *International Journal of Molecular Sciences* 21 (9): 3296. https://doi.org/10.3390/ijms21093296.

Raghavan, Neha S., Adam M. Brickman, Howard Andrews, Jennifer J. Manly, Nicole Schupf, Rafael Lantigua, Charles J. Wolock, et al. 2018. « Whole-Exome Sequencing in 20,197 Persons for Rare Variants in Alzheimer's Disease ». *Annals of Clinical and Translational Neurology* 5 (7): 832-42. https://doi.org/10.1002/acn3.582.

Rausch, Tobias, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes, et Jan O. Korbel. 2012. « DELLY: Structural Variant Discovery by Integrated Paired-End and Split-Read Analysis ». *Bioinformatics* 28 (18): i333-39. https://doi.org/10.1093/bioinformatics/bts378.

Reiter, L. T., P. J. Hastings, E. Nelis, P. De Jonghe, C. Van Broeckhoven, et J. R. Lupski. 1998. « Human Meiotic Recombination Products Revealed by Sequencing a Hotspot for Homologous Strand Exchange in Multiple HNPP Deletion Patients ». *American Journal of Human Genetics* 62 (5): 1023-33. https://doi.org/10.1086/301827.

Richardson, Sandra R., Aurélien J. Doucet, Huira C. Kopera, John B. Moldovan, José Luis Garcia-Perez, et John V. Moran. 2015. « The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes ». *Microbiology Spectrum* 3 (2): MDNA3-0061-2014. https://doi.org/10.1128/microbiolspec.MDNA3-0061-2014.

Riggs, Er, Dm Church, K Hanson, Vl Horner, Eb Kaminsky, Rm Kuhn, Ke Wain, et al. 2012. « Towards an Evidence-Based Process for the Clinical Interpretation of Copy Number Variation ». *Clinical Genetics* 81 (5): 403-12. https://doi.org/10.1111/j.1399-0004.2011.01818.x.

Rizzo, Jason M., et Michael J. Buck. 2012. « Key Principles and Clinical Applications of "next-Generation" DNA Sequencing ». *Cancer Prevention Research (Philadelphia, Pa.)* 5 (7): 887-900. https://doi.org/10.1158/1940-6207.CAPR-11-0432.

Roller, Eric, Sergii Ivakhno, Steve Lee, Thomas Royce, et Stephen Tanner. 2016. « Canvas: Versatile and Scalable Detection of Copy Number Variants ». *Bioinformatics* 32 (15): 2375-77. https://doi.org/10.1093/bioinformatics/btw163.

Rovelet-Lecrux, A., C. Charbonnier, D. Wallon, G. Nicolas, M. N. J. Seaman, C. Pottier, S. Y. Breusegem, et al. 2015. « De Novo Deleterious Genetic Variations Target a Biological Network Centered on Aβ Peptide in Early-Onset Alzheimer Disease ». *Molecular Psychiatry* 20 (9): 1046-56. https://doi.org/10.1038/mp.2015.100.

Rovelet-Lecrux, Anne, Didier Hannequin, Gregory Raux, Nathalie Le Meur, Annie Laquerrière, Anne Vital, Cécile Dumanchin, et al. 2006. « APP Locus Duplication Causes Autosomal Dominant Early-Onset Alzheimer Disease with Cerebral Amyloid Angiopathy ». *Nature Genetics* 38 (1): 24-26. https://doi.org/10.1038/ng1718.

Rovelet-Lecrux, Anne, Solenn Legallic, David Wallon, Jean-Michel Flaman, Olivier Martinaud, Stéphanie Bombois, Adeline Rollin-Sillaire, et al. 2012. « A Genome-Wide Study Reveals Rare CNVs Exclusive to Extreme Phenotypes of Alzheimer Disease ». *European Journal of Human Genetics: EJHG* 20 (6): 613-17. https://doi.org/10.1038/ejhg.2011.225.

Rubnitz, J., et S. Subramani. 1984. « The Minimum Amount of Homology Required for Homologous Recombination in Mammalian Cells ». *Molecular and Cellular Biology* 4 (11): 2253-58. https://doi.org/10.1128/mcb.4.11.2253-2258.1984.

Rump, Patrick, Nicole de Leeuw, Anthonie J. van Essen, Corien C. Verschuuren-Bemelmans, Hermine E. Veenstra-Knol, Mariëlle E. M. Swinkels, Wilma Oostdijk, et al. 2014. « Central 22q11.2 Deletions ». *American Journal of Medical Genetics. Part A* 164A (11): 2707-23. https://doi.org/10.1002/ajmg.a.36711.

Saitsu, Hirotomo, Hitoshi Osaka, Shirou Sugiyama, Kenji Kurosawa, Takeshi Mizuguchi, Kiyomi Nishiyama, Akira Nishimura, et al. 2012. « Early Infantile Epileptic Encephalopathy Associated with the Disrupted Gene Encoding Slit-Robo Rho GTPase Activating Protein 2 (SRGAP2) ». *American Journal of Medical Genetics. Part A* 158A (1): 199-205. https://doi.org/10.1002/ajmg.a.34363.

Sánchez-Juan, Pascual, Sonia Moreno, Itziar de Rojas, Isabel Hernández, Sergi Valero, Montse Alegret, Laura Montrreal, et al. 2019. « The MAPT H1 Haplotype Is a Risk Factor for Alzheimer's Disease in APOE E4 Non-Carriers ». *Frontiers in Aging Neuroscience* 11: 327. https://doi.org/10.3389/fnagi.2019.00327.

Schatz, David G., et Patrick C. Swanson. 2011. « V(D)J Recombination: Mechanisms of Initiation ». *Annual Review of Genetics* 45: 167-202. https://doi.org/10.1146/annurev-genet-110410-132552.

Schellenberg, G. D., S. S. Deeb, M. Boehnke, E. M. Bryant, G. M. Martin, T. H. Lampe, et T. D. Bird. 1987. « Association of an Apolipoprotein CII Allele with Familial Dementia of the Alzheimer Type ». *Journal of Neurogenetics* 4 (2-3): 97-108.

Schouten, Jan P., Cathal J. McElgunn, Raymond Waaijer, Danny Zwijnenburg, Filip Diepvens, et Gerard Pals. 2002. « Relative Quantification of 40 Nucleic Acid Sequences by Multiplex Ligation-Dependent Probe Amplification ». *Nucleic Acids Research* 30 (12): e57. https://doi.org/10.1093/nar/gnf056.

Schramm, C., D. Wallon, G. Nicolas, et C. Charbonnier. 2022. « What Contribution Can Genetics Make to Predict the Risk of Alzheimer's Disease? » *Revue Neurologique* 178 (5): 414-21. https://doi.org/10.1016/j.neurol.2022.03.005.

Sharp, Andrew J., Devin P. Locke, Sean D. McGrath, Ze Cheng, Jeffrey A. Bailey, Rhea U. Vallente, Lisa M. Pertz, et al. 2005. « Segmental Duplications and Copy-Number Variation in the Human Genome ». *American Journal of Human Genetics* 77 (1): 78-88. https://doi.org/10.1086/431652.

Shen, Yiping, Kira A. Dies, Ingrid A. Holm, Carolyn Bridgemohan, Magdi M. Sobeih, Elizabeth B. Caronna, Karen J. Miller, et al. 2010. « Clinical Genetic Testing for Patients with Autism Spectrum Disorders ». *Pediatrics* 125 (4): e727-735. https://doi.org/10.1542/peds.2009-1684.

Sherman, Rachel M., et Steven L. Salzberg. 2020. « Pan-Genomics in the Human Genome Era ». *Nature Reviews. Genetics* 21 (4): 243-54. https://doi.org/10.1038/s41576-020-0210-7.

Sims, Rebecca, Sven J. van der Lee, Adam C. Naj, Céline Bellenguez, Nandini Badarinarayan, Johanna Jakobsdottir, Brian W. Kunkle, et al. 2017. « Rare Coding Variants in PLCG2, ABI3, and TREM2 Implicate Microglial-Mediated Innate Immunity in Alzheimer's Disease ». *Nature Genetics* 49 (9): 1373-84. https://doi.org/10.1038/ng.3916.

Sirkis, Daniel W., Luke W. Bonham, Taylor P. Johnson, Renaud La Joie, et Jennifer S. Yokoyama. 2022. « Dissecting the Clinical Heterogeneity of Early-Onset Alzheimer's Disease ». *Molecular Psychiatry* 27 (6): 2674-88. https://doi.org/10.1038/s41380-022-01531-9.

Sleegers, Kristel, Nathalie Brouwers, Ilse Gijselinck, Jessie Theuns, Dirk Goossens, Jan Wauters, Jurgen Del-Favero, Marc Cruts, Cornelia M. van Duijn, et Christine Van Broeckhoven. 2006. « APP Duplication Is Sufficient to Cause Early Onset Alzheimer's Dementia with Cerebral Amyloid Angiopathy ». *Brain: A Journal of Neurology* 129 (Pt 11): 2977-83. https://doi.org/10.1093/brain/awl203.

Smith, J. Gustav, et Christopher Newton-Cheh. 2015. « Genome-Wide Association Studies of Late-Onset Cardiovascular Disease ». *Journal of Molecular and Cellular Cardiology* 83 (juin): 131-41. https://doi.org/10.1016/j.yjmcc.2015.04.004.

Smolka, Moritz, Luis F. Paulin, Christopher M. Grochowski, Dominic W. Horner, Medhat Mahmoud, Sairam Behera, Ester Kalef-Ezra, et al. 2022. « Comprehensive Structural Variant Detection: From Mosaic to Population-Level ». Preprint. Bioinformatics. https://doi.org/10.1101/2022.04.04.487055.

Soldan, Anja, Corinne Pettigrew, Qing Cai, Jiangxia Wang, Mei-Cheng Wang, Abhay Moghekar, Michael I. Miller, Marilyn Albert, et BIOCARD Research Team. 2017. « Cognitive Reserve and Long-Term Change in Cognition in Aging and Preclinical Alzheimer's Disease ». *Neurobiology of Aging* 60 (décembre): 164-72. https://doi.org/10.1016/j.neurobiolaging.2017.09.002.

Stankiewicz, Paweł, et James R. Lupski. 2002. « Genome Architecture, Rearrangements and Genomic Disorders ». *Trends in Genetics: TIG* 18 (2): 74-82. https://doi.org/10.1016/s0168-9525(02)02592-1.

Stern, Yaakov. 2012. « Cognitive Reserve in Ageing and Alzheimer's Disease ». *The Lancet. Neurology* 11 (11): 1006-12. https://doi.org/10.1016/S1474-4422(12)70191-6.

Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. « An Integrated Map of Structural Variation in 2,504 Human Genomes ». *Nature* 526 (7571): 75-81. https://doi.org/10.1038/nature15394.

Sulak, Michael, Lindsey Fong, Katelyn Mika, Sravanthi Chigurupati, Lisa Yon, Nigel P Mongan, Richard D Emes, et Vincent J Lynch. 2016. « TP53 Copy Number Expansion Is Associated with the Evolution of Increased Body Size and an Enhanced DNA Damage Response in Elephants ». *ELife* 5 (septembre): e11994. https://doi.org/10.7554/eLife.11994.

Svanvik, N., A. Ståhlberg, U. Sehlstedt, R. Sjöback, et M. Kubista. 2000. « Detection of PCR Products in Real Time Using Light-up Probes ». *Analytical Biochemistry* 287 (1): 179-82. https://doi.org/10.1006/abio.2000.4824.

Sykes, P. J., S. H. Neoh, M. J. Brisco, E. Hughes, J. Condon, et A. A. Morley. 1992. « Quantitation of Targets for PCR by Use of Limiting Dilution ». *BioTechniques* 13 (3): 444-49.

Szigeti, Kinga, Deepika Lal, Yanchun Li, Rachelle S. Doody, Kirk Wilhelmsen, Li Yan, Song Liu, Changxing Ma, et Texas Alzheimer Research and Care Consortium. 2013. « Genome-Wide Scan for Copy Number Variation Association with Age at Onset of Alzheimer's Disease ». *Journal of Alzheimer's Disease: JAD* 33 (2): 517-23. https://doi.org/10.3233/JAD-2012-121285.

Taliun, Daniel, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, et al. 2021. « Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program ». *Nature* 590 (7845): 290-99. https://doi.org/10.1038/s41586-021-03205-y.

Tammimies, Kristiina, Christian R. Marshall, Susan Walker, Gaganjot Kaur, Bhooma Thiruvahindrapuram, Anath C. Lionel, Ryan K. C. Yuen, et al. 2015. « Molecular Diagnostic Yield of Chromosomal Microarray Analysis and Whole-Exome Sequencing in Children With Autism Spectrum Disorder ». *JAMA* 314 (9): 895. https://doi.org/10.1001/jama.2015.10078.

Tattini, Lorenzo, Romina D'Aurizio, et Alberto Magi. 2015. « Detection of Genomic Structural Variants from Next-Generation Sequencing Data ». *Frontiers in Bioengineering and Biotechnology* 3: 92. https://doi.org/10.3389/fbioe.2015.00092.

Tjio, J. H. 1978. « The Chromosome Number of Man ». *American Journal of Obstetrics and Gynecology* 130 (6): 723-24. https://doi.org/10.1016/0002-9378(78)90337-x.

Toffolatti, Luisa, Barbara Cardazzo, Carlo Nobile, Gian Antonio Danieli, Francesca Gualandi, Francesco Muntoni, Steve Abbs, et al. 2002. « Investigating the Mechanism of Chromosomal Deletion: Characterization of 39 Deletion Breakpoints in Introns 47 and 48 of the Human Dystrophin Gene ». *Genomics* 80 (5): 523-30.

Töpfer, Armin. 2022. « pbsv ». https://github.com/PacificBiosciences/pbsv.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, et al. 2001. « The Sequence of the Human Genome ». *Science (New York, N.Y.)* 291 (5507): 1304-51. https://doi.org/10.1126/science.1058040.

Vidal, R., J. Ghiso, T. Wisniewski, et B. Frangione. 1996. « Alzheimer's Presenilin 1 Gene Expression in Platelets and Megakaryocytes. Identification of a Novel Splice Variant ». *FEBS Letters* 393 (1): 19-23. https://doi.org/10.1016/0014-5793(96)00845-9.

Vogelstein, B., et K. W. Kinzler. 1999. « Digital PCR ». *Proceedings of the National Academy of Sciences of the United States of America* 96 (16): 9236-41. https://doi.org/10.1073/pnas.96.16.9236.

Voigt, S., P. C. de Kruijff, E. A. Koemans, I. Rasing, E. S. van Etten, G. M. Terwindt, Mjp van Osch, M. A. van Buchem, Maa van Walderveen, et Mjh Wermer. 2022. « Cerebellar Hemorrhages in Patients with Dutch-Type Hereditary Cerebral Amyloid Angiopathy ». *International Journal of*

Stroke: Official Journal of the International Stroke Society 17 (6): 637-44. https://doi.org/10.1177/17474930211043663.

Wallon, David, Susana Boluda, Anne Rovelet-Lecrux, Manon Thierry, Julien Lagarde, Laetitia Miguel, Magalie Lecourtois, et al. 2021. « Clinical and Neuropathological Diversity of Tauopathy in MAPT Duplication Carriers ». *Acta Neuropathologica* 142 (2): 259-78. https://doi.org/10.1007/s00401-021-02320-4.

Wang, David G., Jian-Bing Fan, Chia-Jen Siao, Anthony Berno, Peter Young, Ron Sapolsky, Ghassan Ghandour, et al. 1998. « Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome ». *Science* 280 (5366): 1077-82. https://doi.org/10.1126/science.280.5366.1077.

Wang, Hui, Beth A. Dombroski, Po-Liang Cheng, Albert Tucci, Ya-Qin Si, John J. Farrell, Jung-Ying Tzeng, et al. 2023. « Structural Variation Detection and Association Analysis of Whole-Genome-Sequence Data from 16,905 Alzheimer's Diseases Sequencing Project Subjects ». *MedRxiv: The Preprint Server for Health Sciences*, septembre, 2023.09.13.23295505. https://doi.org/10.1101/2023.09.13.23295505.

Wang, Hui, Jinchuan Xing, Deepak Grover, Dale J. Hedges, Kyudong Han, Jerilyn A. Walker, et Mark A. Batzer. 2005. « SVA Elements: A Hominid-Specific Retroposon Family ». *Journal of Molecular Biology* 354 (4): 994-1007. https://doi.org/10.1016/j.jmb.2005.09.085.

Wang, Jianxin, Lei Song, Deepak Grover, Sami Azrak, Mark A. Batzer, et Ping Liang. 2006. « DbRIP: A Highly Integrated Database of Retrotransposon Insertion Polymorphisms in Humans ». *Human Mutation* 27 (4): 323-29. https://doi.org/10.1002/humu.20307.

Wang, Kai, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F.A. Grant, Hakon Hakonarson, et Maja Bucan. 2007. « PennCNV: An Integrated Hidden Markov Model Designed for High-Resolution Copy Number Variation Detection in Whole-Genome SNP Genotyping Data ». *Genome Research* 17 (11): 1665-74. https://doi.org/10.1101/gr.6861907.

Wang, Ting, Lucinda Antonacci-Fulton, Kerstin Howe, Heather A. Lawson, Julian K. Lucas, Adam M. Phillippy, Alice B. Popejoy, et al. 2022. « The Human Pangenome Project: A Global Resource to Map Genomic Diversity ». *Nature* 604 (7906): 437-46. https://doi.org/10.1038/s41586-022-04601-8.

Woodward, Karen J., Julie Stampalia, Hannah Vanyai, Hashika Rijhumal, Kim Potts, Fiona Taylor, Joanne Peverall, et al. 2019. « Atypical Nested 22q11.2 Duplications between LCR22B and LCR22D Are Associated with Neurodevelopmental Phenotypes Including Autism Spectrum Disorder with Incomplete Penetrance ». *Molecular Genetics & Genomic Medicine* 7 (2): e00507. https://doi.org/10.1002/mgg3.507.

Ye, Kai, Marcel H. Schulz, Quan Long, Rolf Apweiler, et Zemin Ning. 2009. « Pindel: A Pattern Growth Approach to Detect Break Points of Large Deletions and Medium Sized Insertions from Paired-End Short Reads ». *Bioinformatics (Oxford, England)* 25 (21): 2865-71. https://doi.org/10.1093/bioinformatics/btp394.

Zayed, Hatem. 2016. « The Qatar Genome Project: Translation of Whole-Genome Sequencing into Clinical Practice ». *International Journal of Clinical Practice* 70 (10): 832-34. https://doi.org/10.1111/ijcp.12871.

Zeitouni, Bruno, Valentina Boeva, Isabelle Janoueix-Lerosey, Sophie Loeillet, Patricia Legoix-né, Alain Nicolas, Olivier Delattre, et Emmanuel Barillot. 2010. « SVDetect: A Tool to Identify Genomic Structural Variations from Paired-End and Mate-Pair Sequencing Data ». *Bioinformatics (Oxford, England)* 26 (15): 1895-96. https://doi.org/10.1093/bioinformatics/btq293.

Zhao, Min, Qingguo Wang, Quan Wang, Peilin Jia, et Zhongming Zhao. 2013. « Computational Tools for Copy Number Variation (CNV) Detection Using next-Generation Sequencing Data: Features

and Perspectives ». *BMC Bioinformatics* 14 Suppl 11 (Suppl 11): S1. https://doi.org/10.1186/1471-2105-14-S11-S1.