



UNIVERSITÉ DE LILLE
École Graduée Biologie-Santé

THÈSE DE DOCTORAT
En vue de l'obtention du grade de
DOCTEUR DE L'UNIVERSITÉ
Discipline : Science de la vie et de la santé

**« SEQUENCAGE LONG-READ POUR L'ANALYSE DE VARIANTS DE
STRUCTURE : DE LA CARACTÉRISATION GÉNOMIQUE A
L'IMPACT FONCTIONNEL »**

Présentée et soutenue publiquement par

Jade Fauqueux

A Lille, le 15 Septembre 2025

Thèse dirigée par

le Pr Jamal Ghoumid et le Dr Thomas Smol

ULR7364 RADEME

Soutenance devant le jury composé de :

Pr Christel Depienne	University Hospital Essen	Rapporteuse
Pr Paul Kuentz	CHU de Besançon	Rapporteur
Pr Gaël Nicolas	Université de Rouen	Examineur, Président du jury
Pr Claire-Marie Dhaenens	Université de Lille	Examinatrice
Pr Jamal Ghoumid	Université de Lille	Directeur de thèse
Dr Thomas Smol	Université de Lille	Co-encadrant de thèse

Les maladies rares sont définies par leur faible prévalence et non par des caractéristiques cliniques ou pathologiques communes. Dans l'Union européenne, une maladie est dite rare lorsqu'elle touche moins d'une personne sur 2 000, soit environ 5 millions d'individus. Parmi les maladies rares, les troubles du neurodéveloppement représentent un groupe complexe, associé à une grande hétérogénéité génétique et de leurs répercussions profondes sur la qualité de vie. Les variants de structure (SVs) sont responsables de 10 à 20 % des cas. Pourtant, leur caractérisation complète reste limitée par les outils actuels, et leur interprétation demeure souvent incertaine.

C'est notamment le cas des duplications, dont l'interprétation dépend de leur organisation sous-jacente. Un gain de copies peut correspondre à des structures très différentes : duplication en tandem (directe ou inversée), duplication complexe (DUP-TRP/INV-DUP), triplication, insertion inter- ou intragénique... Cette diversité structurelle peut induire des mécanismes variés de dérégulation : triplosensibilité, haploinsuffisance, gain ou perte de fonction, rendant la caractérisation fine indispensable à une interprétation complète. Les technologies de diagnostic de routine peinent à capturer cette complexité. Le séquençage *long-read*, en particulier par la technologie Oxford Nanopore (ONT), apparaît comme une solution prometteuse. Il est jusqu'à cinq fois plus sensible pour détecter les SVs, et permet une analyse directe de leur structure.

Dans ce contexte, ce projet de thèse visait à évaluer la pertinence du séquençage ONT pour l'étude de gains de copies identifiés en diagnostic chez des patients atteints de trouble neurodéveloppementaux. L'objectif était double : affiner leur caractérisation structurelle et explorer leur impact fonctionnel quand cela était possible, afin d'en améliorer l'interprétation clinique.

Nous avons séquencé 15 patients par ONT. Tous les SVs ciblés ont pu être visualisés, et 93 % ont été appelés par un logiciel de détection de variants de structure. Cette approche a permis de révéler l'organisation précise des réarrangements et d'inférer les mécanismes en jeu (NAHR, NHEJ, MMBIR...). Dans plusieurs cas, une analyse fonctionnelle a été menée, conduisant à une reclassification en variant pathogène dans 42,7 % des situations analysées.

Ces résultats démontrent que le séquençage *long-read* offre une couverture suffisante à l'échelle du génome, y compris dans des régions complexes. Il permet de détecter des réarran-

gements ignorés par les méthodes conventionnelles, et d'en fournir une description fine, qu'il s'agisse de duplications simples ou complexes. Ainsi, cette technologie constitue un atout majeur pour l'étude des SVs dans les troubles du neurodéveloppement, en complément des approches classiques. Elle ouvre la voie à une meilleure compréhension des mécanismes impliqués, à une interprétation plus précise des variants détectés et à une amélioration du rendement diagnostique en génétique clinique.

ABSTRACT

Rare diseases are defined by their low prevalence rather than by shared clinical or pathological characteristics. In the European Union, a disease is considered rare when it affects fewer than one in 2,000 people, which corresponds to approximately 5 million individuals. Among rare diseases, neurodevelopmental disorders represent a complex group, associated with high genetic heterogeneity and significant impacts on quality of life. Structural variants (SVs) account for 10 to 20% of cases. However, their comprehensive characterization remains limited by current tools, and their interpretation is often uncertain.

This is particularly true for duplications, whose interpretation depends on their underlying structure. A copy number gain can correspond to very different architectures: tandem duplications (direct or inverted), complex duplications (DUP-TRP/INV-DUP), triplications, intergenic or intragenic insertions, etc. This structural diversity can lead to various mechanisms of deregulation, triplosensitivity, haploinsufficiency, gain or loss of function, making detailed characterization essential for accurate interpretation. Routine diagnostic technologies struggle to capture this complexity. Long-read sequencing, particularly with Oxford Nanopore Technology (ONT), emerges as a promising solution. It is up to five times more sensitive in detecting SVs and allows direct analysis of their structure.

In this context, this PhD project aimed to assess the relevance of ONT sequencing for studying copy number gains identified in diagnostics in patients with neurodevelopmental disorders. The goal was twofold: to refine the structural characterization of these variants and to explore their functional impact when possible, in order to improve their clinical interpretation.

We sequenced 15 patients using ONT. All targeted SVs could be visualized, and 93% were detected by structural variant calling software. This approach enabled us to uncover the precise organization of the rearrangements and infer the mechanisms involved (NAHR, NHEJ, MMBIR, etc.). In several cases, functional analyses were performed, leading to reclassification as pathogenic variants in 50% of the analyzed cases.

These results demonstrate that long-read sequencing provides sufficient genome-wide coverage, including in complex regions. It allows detection of rearrangements missed by conventional methods and offers fine-grained structural resolution, whether for simple or complex duplications. As such, this technology is a major asset for studying SVs in neurodevelopmental

tal disorders, complementing classical approaches. It paves the way for a better understanding of the underlying mechanisms, more accurate interpretation of detected variants, and improved diagnostic yield in clinical genetics.

ULR7364 RADEME

Maladies rares du développement et du métabolisme

Dirigée par le Pr Jamal Ghoumid, PU-PH

et

le Dr Thomas Smol, MCU-PH

Faculté de Médecine, Pôle Recherche

1, Place de Verdun

F59000 LILLE

REMERCIEMENTS

Je dédie cette thèse à toutes les personnes qui m’ont encouragée et soutenue tout au long de mon parcours universitaire. Leur présence, leurs mots et leur confiance ont été essentiels à chaque étape de cette aventure.

Merci aux membres du jury, **Christel Depienne** et **Paul Kuentz** d’avoir accepté d’être rapporteurs de mon travail ainsi que **Claire-Marie Dhaenens** et **Gaël Nicolas** d’avoir accepté d’en être examinateurs et de m’avoir accompagnée tout au long de ces années dans le cadre du jury de CSI. Je leur suis très reconnaissante pour le temps qu’ils ont consacré à la lecture de ma thèse et pour l’attention portée à mon travail.

A mes directeurs de thèse, **Thomas Smol** et **Jamal Ghoumid**.

Thomas, ce travail n’aurait jamais vu le jour sans toi. Ta confiance et ton soutien inébranlables, même face à mes innombrables “tu penses que ça va aller ?”, m’ont été d’une valeur inestimable. Merci pour tes conseils, pour tout ce que tu m’as transmis, et pour les nombreuses opportunités que tu m’as offertes. Cette aventure touche à sa fin, mais j’espère que d’autres beaux réarrangements complexes viendront encore, à l’avenir, nous donner du fil à retordre.

Jamal, un immense merci de m’avoir accueillie au sein du laboratoire, d’avoir accepté d’encadrer cette thèse et de m’avoir soutenue dans chacun de mes projets. Je te suis profondément reconnaissante pour ta confiance et ton accompagnement.

Je remercie sincèrement mes collègues de l’équipe ULR7364 RADEME :

Florence, pour ton expertise et tes conseils tout au long de ces années, ainsi que pour les opportunités de communication orale.

Frédéric, pour ton aide précieuse et tes conseils lorsque je suis arrivée au laboratoire durant mon stage de M2.

Clémence et **Fiona**, pour l’année passée ensemble à RADEME et pour vos propositions d’études de caractérisation basées sur vos patients.

Simon, pour ta disponibilité et ton aide dans la récupération des données patients.

Luc, pour ton travail acharné sur les données d’Hi-C, qui t’ont sûrement valu quelques cauchemars !

Jérôme, pour ces deux années partagées à RADEME, entre science (ou pas), culture cellulaire et entraide au quotidien.

Lauryne, pour tes coups de mains avec la culture cellulaire et les moments passés au labo.

Perrine et **Anne-Sophie**, pour votre disponibilité et votre réactivité concernant la potentielle mise en place du RNA-seq pour certains patients.

Je remercie les membres du CBP:

Caroline, tout particulièrement pour ton implication tout au long de ma thèse. Des premiers bêta-tests de séquençage long-read, au design de primers, en passant par toutes les fois où tu m'as sauvée pour des commandes : tu as toujours été là. Une grande partie des résultats n'aurait pas été aussi facile à obtenir sans ton aide précieuse !

Alexis, Heidi et **Delphine**, merci pour votre disponibilité, votre réactivité et votre aide sur toutes les techniques. Votre soutien a vraiment compté.

Martin, Jean-Pascal et **Emilie**, pour votre implication dans l'analyse bioinformatique des données de séquençage nanopore.

Seb et **Vantu**, sans qui l'accès au CBP n'aurait pas été aussi simple.

Céline, pour ta disponibilité lors des premiers essais de préparation de librairies.

Et toute l'**équipe du plateau de culture cellulaire de biochimie**, sans qui les cellules immortalisées n'auraient jamais vu le jour.

Je remercie les membres de la clinique de génétique « Guy Fontaine » :

Roseline, pour ton implication dans le projet CAR-SV, sans lequel une grande partie de nos données n'aurait jamais vu le jour.

Luisa, pour ta disponibilité durant les réunions RADEME.

Je remercie les membres de l'Institut de Génétique Médicale :

Élise, pour ton implication et ta disponibilité dans la caractérisation de notre fameux réarrangement complexe.

Marine, pour ton intérêt pour le projet CAR-SV.

Je remercie également tous les **co-auteurs des articles** de cette thèse, ainsi que **les patients** et **leurs familles**, qui ont accepté de participer à la recherche.

Merci à **l'Université de Lille** et à **l'EGBSL** pour m'avoir donné l'opportunité de réaliser cette thèse, et au **GIRCI Nord-Ouest** pour le financement des génomes long-read.

Je remercie ma famille :

Margaux et **Giovambattista**, pour votre soutien indéfectible malgré les épreuves, et pour toute l'aide que vous m'avez apportée pendant ces trois années, et bien au-delà. Merci, du fond du cœur.

Papa et **Candice**, pour m'avoir donné les moyens d'arriver là où j'en suis aujourd'hui, et pour avoir été à l'écoute de toutes mes péripéties et résultats. C'est grâce à vous.

Matilda, merci pour ton soutien. C'est à ton tour maintenant de tracer ta propre voie et de trouver celle dans laquelle tu t'épanouiras pleinement !

Enfin, un immense merci à **tous mes amis** pour leur soutien au quotidien. J'espère que vous n'en avez pas encore marre de m'entendre parler de mes recherches... parce que ce n'est pas fini.

AVANT-PROPOS

Plusieurs articles originaux publiés ou en cours de publication en premier auteur dans des revues scientifiques internationales à comité de lecture structurent ce travail de thèse :

→ **Fauqueux J, et al. Long-Read Sequencing of Recurrent FGF12 Duplications in Epilepsy: Insights into Structural Mechanisms and Aberrant Isoforms. 2025 August ; Epilepsia (Article Annexe 1)**

→ **Fauqueux J, et al. Combining long-read DNA and RNA sequencing to enhance molecular understanding of structural variations leading to copy gains. 2025 April ; Computational and Structural Biotechnology Journal ;Journal 27 (2025) 1732–1740 (Article Annexe 2)**

→ **Fauqueux J, et al. Deciphering Structural Mechanisms of Copy Gain Variants With Long-Read Sequencing. Under Submission.**

Ces articles ont également fait l'objet de présentations lors de communications orales et/ou de posters en congrès scientifiques nationaux ou internationaux :

→ **Structural characterization of recurrent duplications involving FGF12 by long-read Nanopore Sequencing. ESHG 2024, Berlin – Allemagne, 2024 (Poster)**

→ **Deciphering a complex short size RERE rearrangement: Importance of combined approaches for duplication interpretation. ESHG 2024, Berlin – Allemagne, 2024 (Poster)**

→ **Impact des duplications : comment s'en sortir ? Journée Achropuce, Paris – France, 2024 (Communication Orale)**

→ **Séquençage Long-Reads pour la détection et l'analyse des variants de structure liés aux gains de copie : Approches ADN et ARN intégrées. 2^e symposium FHU-G4 Post-Génomique, Rouen – France, 2025 (Communication orale)**

→ **Long-Read Sequencing for structural variant detection and analysis: mechanism characterization. ESHG 2025, Milan – Italie, 2025 (Poster)**

La thématique associée au syndrome MED13L a été abordée dans un article et lors de communications dont les résultats ne sont pas présentés dans ce travail.

→ **Fauqueux.J, 2024, Splice site variants in the canonical donor site of MED13L exon 7 lead to intron retention in patients with MED13L syndrome - Journal of Medical Genetics**

→ **Identification of a cryptic splicing site in MED13L intron 7 leading of truncated protein in patients with MED13L syndrome. EuroNDD 2024, Lisbonne – Portugal, 2024**

TABLE DES MATIÈRES

Table des matières

Index des figures.....	19
Index des Tables.....	22
Liste des Abréviations.....	23
Lexique des gènes - Nomenclature HUGO :.....	24
I/ Introduction.....	25
1. De la génétique humaine aux troubles neurodéveloppementaux.....	26
1.1. La génétique humaine et sa diversité.....	26
1.2. Maladies génétiques rares.....	27
1.3. Troubles neurodéveloppementaux.....	28
2. Variants génétiques, mécanismes et impact.....	30
2.1. Variants génétiques.....	31
2.1.1. Variant Nucléotidique.....	31
2.1.2. INDEL :.....	32
2.1.3. Les variants cytogénétiques.....	32
2.1.3.1. Variants intrachromosomiques simples.....	34
2.1.3.2. Variants interchromosomiques simples.....	34
2.1.3.3. Variants chromosomiques complexes.....	35

2.2. Mécanismes d'apparitions des SV.....	39
2.2.1. Variant récurrent.....	40
2.2.2. Variant non récurrent.....	43
2.2.2.1. Non-Homologous End-Joining.....	44
2.2.2.2. Mécanismes liés à la réplication.....	46
2.3. Conséquences des variants de structures.....	49
2.3.1. Sensibilité au dosage.....	49
2.3.2. Impact sur la séquence codante.....	51
2.3.3. Effet de position.....	52
2.3.3.1. Impact sur l'élément régulateur.....	53
2.3.3.2. Impact sur l'environnement 3D.....	54
3. Detection des SV.....	58
3.1. Caryotype et FISH.....	58
3.2. Analyse chromosomique par puce ADN (CGH-Array/ACPA) :.....	62
3.3. Séquençage de nouvelles générations.....	63
3.3.1. Méthode.....	63
3.3.2. Détection des variants.....	66
3.3.3. Limites.....	68
3.4. Cartographie Optique.....	70
3.5. Séquençage de troisième génération.....	71

3.5.1. Séquençage PacBio.....	71
3.5.2. Oxford Nanopore Technologie.....	72
3.5.2.1. Origine du nanopore.....	72
3.5.2.2. Type de nanopore.....	74
3.5.2.3. Principe.....	76
3.5.2.4. Spécificités.....	77
3.6. De la cytogénétique à la cytogénomique.....	79
4. Utilisation du séquençage <i>long-read</i> pour l'étude des variants de structures.....	81
5. Interprétation des variants.....	82
II/ Objectifs.....	85
III/ Matériel et Méthodes.....	87
1. Matériel.....	88
1.1. Matériel biologique.....	88
1.2. Matériels associés au séquençage Nanopore.....	88
1.2.1. Séquenceurs.....	88
1.2.2. Informatique et logiciels.....	88
2. Méthodes.....	90
2.1. CGH-Array.....	91
2.2. Librairie de séquençage ADN <i>Long-Read</i> Nanopore.....	91
2.3. Déroulement d'un <i>run</i> Nanopore sur ADN.....	91

2.3. Analyse qPCR ciblée.....	93
2.4. Culture cellulaire.....	93
IV/ Résultats.....	94
1. Approche de séquençage et panorama des données.....	95
1.1. Metrics associé au séquençage.....	95
1.2. Spécificités du séquençage par <i>adaptive sampling</i>	96
1.3. Impact des facteurs pré-analytiques sur les performances de séquençage.....	96
2. Étude de variants récurrents, soumis au mécanisme de NAHR.....	99
2.1. Article 1 – Publié dans <i>Epilepsia: Long-Read Sequencing of Recurrent FGF12 Duplications in Epilepsy: Insights into Structural Mechanisms and Aberrant Isoforms</i>	99
2.1.1. Introduction.....	99
2.1.2. Méthodes spécifiques.....	100
2.1.2.1. Déclaration éthique.....	100
2.1.2.2. Séquençage des amplicons d'ADNc.....	100
2.1.2.3. Analyse des points de cassures.....	101
2.1.3. Résultats.....	101
2.1.3.1. Description clinique et caractéristiques des EEG de UPN1.....	101
2.1.3.2. Description clinique et caractéristiques des EEG de UPN2.....	105
2.1.3.3. Caractérisation des réarrangements structuraux par séquençage LRS...107	
2.1.3.4. Séquençage LRS des isoformes majeures A et B de FGF12.....	109

2.1.4. Discussion.....	110
2.2. Etude d'un variant récurrent lié à l'X: IDS.....	117
2.2.1. Méthode complémentaire.....	117
2.2.2. Résultats.....	117
2.2.2.1. Présentation clinique et familiale du patient UPN3.....	117
2.2.2.2. Diagnostic moléculaire d'une mucopolysaccharidose de type II par sé- quençage <i>Long-Read</i>	118
2.2.3. Conclusion.....	120
3. Étude de variants non récurrents.....	121
3.1. Article 2: Publié dans CSBJ : <i>Combining long-read DNA and RNA sequencing to enhance molecular understanding of structural variations leading to copy gains</i> (149)	121
3.1.1. Introduction.....	121
3.1.2. Matériels et méthodes supplémentaires.....	123
3.1.2.1. Descriptions cliniques.....	123
3.1.2.2. Cartographie optique du génome (OGM).....	123
3.1.2.3. Librairie de séquençage ADNc <i>Long-Read</i> Nanopore.....	124
3.1.2.4. Analyse cytogénétique conventionnelle.....	124
3.1.2.5. Séquençage Hi-C.....	124
3.1.3. Résultats.....	125
3.1.3.1. Le séquençage <i>long-read</i> permet une caractérisation à l'échelle nucléoti- dique d'une triplication impliquant le gène ZMYM2.....	125

3.1.3.2. Le séquençage <i>long-read</i> révèle un variant structurel hautement complexe impliquant le gène <i>RERE</i>	128
3.2. Article 3 – En soumission: Deciphering Structural Mechanisms of Copy Gain Variants With Long-Read Sequencing.....	137
3.2.1. Introduction.....	137
3.2.2. Méthodes spécifiques.....	138
3.2.3. Résultats.....	138
3.2.3.1 Le séquençage <i>long-reads</i> (LRS) identifie des duplications en tandem associées à des mécanismes de réparation par microhomologie.....	138
3.2.3.2. Le séquençage <i>long-reads</i> révèle un mécanisme de réparation marqué par des cicatrices d'insertion aux points de rupture.....	141
3.2.3.3. Le séquençage <i>long-reads</i> permet une analyse approfondie des variants de structure complexes et des mécanismes de réparation.....	143
3.2.4. Discussion.....	145
3.3. Étude de variants associés à <i>MED13L</i>	149
3.3.1. Contextualisation.....	149
3.3.2. Résultats.....	150
3.3.2.1. Présentation clinique des patients UPN13 et UPN14.....	150
3.3.2.2. Étude de l'impact biologique des gains de copies associés à <i>MED13L</i>	150
3.3.3. Conclusion.....	152
3.4. Étude d'un variant associé à <i>GPSM2</i>	152
3.4.1. Méthode spécifique.....	152

3.4.2. Résultats.....	153
3.4.2.1. Présentation clinique du patient UPN15 et explorations par génome <i>Short-Read</i>	153
3.4.2.2. Diagnostic moléculaire du syndrome Chudley-McCullough par séquen- çage Long-Read.....	154
3.4.3. Conclusion.....	157
V/ Discussion.....	158
1. Caractérisations structurelle et moléculaire de variants de structure.....	159
1.1. Analyse par LRS : Vers un meilleur taux diagnostique ?.....	159
1.2. Que nous aurait révélé le séquençage short-read ?.....	161
1.3. Les gains de copies : une complexité sous-jacente ?.....	162
1.4. Étude associée aux mécanismes d'apparitions.....	163
2. Le séquençage long-read par la technologie Nanopore.....	164
2.1. Mise en œuvre du séquençage génomique complet sur <i>flowcell</i> Nanopore.....	164
2.2. Faut-il privilégier l' <i>Adaptive Sampling</i> ?.....	165
2.3. Vers une mise en place en routine diagnostique ?.....	167
2.4. Qu'en est-il de la seconde approche de séquençage Long-Read : PacBio ?.....	170
VI/ Conclusion.....	172
VII/ Bibliographie.....	174
VIII/ Annexes.....	190
Article Annexe 1.....	206

Index des figures

Index des figures

Figure 1: Gènes de DI identifiés au cours des années.	29
Figure 2: Type de variant génétique.....	31
Figure 3: Comptage médian des SVs totaux par génome.....	33
Figure 4: Réarrangements chromosomiques complexes.....	37
Figure 5: Schéma des trois processus associés à la chromoanagénèse.....	39
Figure 6: Recombinaison homologue (HR) pour la réparation de DSBs.....	41
Figure 7: Mécanisme de NAHR.....	43
Figure 8: Génération de SVs récurrents et non récurrents via le mécanisme de NAHR.....	44
Figure 9: Mécanisme de Non-Homologous End Joining.....	45
Figure 10: Mécanisme de réparation de l'ADN associé à la réplication.....	47
Figure 11: Translocation t(X;8) entraînant la formation d'un transcrit de fusion PKT2-THOC2	52
Figure 12 : Organisation tridimensionnelle de la chromatine et délimitation des TADs.....	55
Figure 13: Synthèse des différentes possibilités d'effet de position en lien avec l'organisation de la chromatine en TADs.....	56
Figure 14 : Stratégies pour l'interprétation des variants de structure.....	58
Figure 15: Chromosome Banding du chromosome 3.....	60

Figure 16: Hybridation In-Situ en Fluorescence démontrant une translocation t(1,22).....	61
Figure 17: CGH-Array: Principe de la technique.....	63
Figure 18: Principe du séquençage de Nouvelle Génération (NGS) par la technologie Illumina	65
Figure 19: Alignement des fragments d'ADN du séquençage short-read dépendant du ciblage.	66
Figure 20: Stratégie de détection des SVs par NGS.....	68
Figure 21: Principe de la Cartographie Optique du Génome.....	70
Figure 22: Principe du séquençage Long-Read par la technologie des SMRT-bells de PacBio	72
Figure 23: Carnet de David Deamer (1989).....	73
Figure 24: Structures de trois pores biologiques.....	75
Figure 25: Principe du séquençage Long-Read par la technologie ONT.....	76
Figure 26: Comparaison structurelle des chimies de la technologie ONT.....	79
Figure 27: Rendement des analyses des variants de structure à travers les techniques d'étude du génome.....	81
Figure 28: Stratégies pour la classification des variants.....	84
Figure 29: Protocole du projet de Thèse.....	90
Figure 30: Profil d'activité des pores au cours du séquençage de l'échantillon UPN4.....	97
Figure 31: EEG de UPN1 au début de l'épilepsie.....	102
Figure 32 : EEG de UPN1 à l'âge de 3 ans.....	104
Figure 33 : EEG de UPN2.....	106

Figure 34 : Étude des réarrangements d'ADN englobant le gène <i>FGF12</i>	108
Figure 35: Étude de l'impact sur les transcrits ARN de <i>FGF12</i>	110
Figure 36 : Évolution en trois phases de l'épilepsie et du développement neurodéveloppemental.....	115
Figure 37: Diagnostic moléculaire de UPN3.....	119
Figure 38: Analyse moléculaire du gain de deux copies chez UPN4.....	127
Figure 39: Analyse moléculaire du réarrangement complexe du gène <i>RERE</i> chez UPN5....	130
Figure 40: Analyse du génome 3D ciblant la réorganisation structurale de UPN5.....	131
Figure 41 : Analyse moléculaire des variants de gain de copies chez UPN6-9, mettant en évidence des points de cassure associés à des microhomologies.....	141
Figure 42 : Analyse moléculaire des variants de gain de copies chez UPN10, mettant en évidence des cicatrices d'insertion aux points de cassure.....	142
Figure 43 : Analyse moléculaire de variants de structure complexes chez UPN11 et UPN12	144
Figure 44: Analyses moléculaires des gains de copies impliquant <i>MED13L</i>	151
Figure 45: Analyse des variants affectant le gène <i>GPSM2</i> chez UPN15.....	155
Figure 46: Prédiction <i>in silico</i> de l'impact de l'insertion sur l'épissage de <i>GPSM2</i>	156

Index des tableaux

Tableau 1: Comparaison des différentes techniques de cytogénétique.....	80
Tableau 2: Caractéristiques de séquençage de quinze individus (UPN1–UPN15) présentant des variants de structure non résolus détectés lors de tests diagnostiques cliniques.....	98
Tableau 3: Tableau clinique des patients nouvellement rapportés et précédemment publiés présentant la duplication récurrente de 500 kb du gène FGF12.....	116

Liste des Abréviations

ACPA	Analyse Chromosomique sur Puce à ADN
ADN	Acide désoxyribonucléique
ADNc	ADN complémentaire
ARN	Acide Ribonucléique
BIR	Break-Induced Replication
CGH	Comparative Genomic Hybridization
CGR	Réarrangement Génomique Complexe
CNV	Copy Number Variant ; Variant du Nombre de Copie
DI	Déficience Intellectuelle
DSB	Double Strand Break ; Cassure Double Brin
DUP	Duplication
FISH	Fluorescence In Situ Hybridization
FosTeS	Fork Stalling and Template Switching
Hi-C	High-Throughput Chromosome Conformation Capture
HR	Homologous Recombination
INDEL	Insertion/Délétion
INV	Inversion
kb	Kilobase
LCR	Low-Copy Repeat
LINE	Long Interspersed Nuclear Elements ; éléments nucléaires dispersés Longs
LRS	Long-Read Sequencing
Mb	Mégabase
MMBIR	Microhomology-Mediated Breakage Induced Repair
NAHR	Non Allelic Homologous Recombination ; Recombinaison Homologue non Allélique
NGS	Next-Generation Sequencing
NHEJ	Non Homologous End Joining
NMD	Nonsense-Mediated mRNA Decay
ONT	Oxford Nanopore Technology
PATRR	Palindromic AT-Rich Repeat
pb	Paire de base
SINE	Short Interspersed Nuclear Element ; éléments nucléaires dispersés courts
SNV	Short Nucleotide Variant
SNP	Short Nucleotide Polymorphism
SRS	Short-Read Sequencing
SV	Structural Variant
T-LRS	Targeted Long-Read Sequencing ; Séquençage ciblé du génome
TAD	Topologically-Associated Domain
TRP	Triplcation
VSI	Variant de Signification Inconnu
WGS	Whole Genome Sequencing ; Séquençage du génome complet
WT	Wild Type

Lexique des gènes - Nomenclature HUGO :

Symbole	Nom	Position chromosomique	#MIM	Symbole	Nom	Position chromosomique	#MIM
<i>MECP2</i>	methyl-CpG binding protein 2	Xq28	300005	<i>SCN6A/SCN7A</i>	sodium voltage-gated channel alpha subunit 7	2q24.3	182392
<i>PLP1</i>	proteolipid protein 1	Xq22.2	300401	<i>IDS</i>	iduronate 2-sulfatase	Xq28	300823
<i>MED13L</i>	mediator complex subunit 13L	12q24.21	608771	<i>EOLA1-DT</i>	EOLA1 divergent transcript	Xq28	
<i>NSD1</i>	nuclear receptor binding SET domain protein 1	5q35.3	606681	<i>ZMYM2</i>	zinc finger MYM-type containing 2	13q12.11	602221
<i>STAG2</i>	STAG2 cohesin complex component	Xq25	300826	<i>RERE</i>	arginine-glutamic acid dipeptide repeats	1p36.23	605226
<i>CREBBP</i>	CREB binding lysine acetyltransferase	16p13.3	600140	<i>PSPC1</i>	paraspeckle component 1	13q12.11	612408
<i>LMNB1</i>	lamin B1	5q23.2	150340	<i>ZMYM5</i>	zinc finger MYM-type containing 5	13q12.11	616443
<i>COL1A1</i>	collagen type I alpha 1 chain	17q21.33	120150	<i>FHAD1</i>	forkhead associated phosphopeptide binding domain 1	1p36.21	620889
<i>COL1A2</i>	collagen type I alpha 2 chain	7q21.3	120160	<i>FAM230G</i>	family with sequence similarity 230 member G	22q11.21	
<i>BCR</i>	BCR activator of RhoGEF and GTPase	22q11.23	151410	<i>CSNK2A1</i>	casein kinase 2 alpha 1	20p13	115440
<i>ABL</i>	ABL proto-oncogene 1, non-receptor tyrosine kinase	9q34.12	189980	<i>TSPEAR</i>	thrombospondin type laminin G domain and EAR repeats	21q22.3	612920
<i>PTK2</i>	protein tyrosine kinase 2	8q24.3	600758	<i>ITGB2</i>	integrin subunit beta 2	21q22.3	600065
<i>THOC2</i>	THO complex subunit 2	Xq25	300395	<i>XIAP</i>	X-linked inhibitor of apoptosis	Xq25	300079
<i>SOX9</i>	SRY-box transcription factor 9	17q24.3	608160	<i>GRIK2</i>	glutamate ionotropic receptor kainate type subunit 2	6q16.3	138244
<i>CSMD1</i>	CUB and Sushi multiple domains 1	8p23.2	608397	<i>PURA</i>	purine rich element binding protein A	5q31.3	600473
<i>MEF2C</i>	myocyte enhancer factor 2C	5q14.3	600662	<i>VPS26C</i>	VPS26 endosomal protein sorting factor C	21q22.13	605298
<i>ARHGAP36</i>	Rho GTPase activating protein 36	Xq26.1	300937	<i>DYRK1A</i>	dual specificity tyrosine phosphorylation regulated kinase 1A	21q22.13	600855
<i>TP53</i>	tumor protein p53	17p13.1	191170	<i>KCNJ6</i>	potassium inwardly rectifying channel subfamily J member 6	21q22.13	600877
<i>FGF12</i>	fibroblast growth factor 12	3q28-q29	601513	<i>GPR173</i>	G protein-coupled receptor 173	Xp11.22	300253
<i>FGF13</i>	fibroblast growth factor 13	Xq26.3-q27.1	300070	<i>SMC1A</i>	structural maintenance of chromosomes 1A	Xp11.22	300040
<i>FGF11</i>	fibroblast growth factor 11	17p13.1	601514	<i>HUWE1</i>	HECT, UBA and WWE domain containing E3 ubiquitin protein ligase 1	Xp11.22	300697
<i>FGF14</i>	fibroblast growth factor 14	13q33.1	601515	<i>SIM1</i>	SIM bHLH transcription factor 1	6q16.3	603128
<i>SCN2A</i>	sodium voltage-gated channel alpha subunit 2	2q24.3	182390	<i>MED13</i>	mediator complex subunit 13	17q23.2	603308
<i>SCN5A</i>	sodium voltage-gated channel alpha subunit 5	3p22.2	600163	<i>GPSM2</i>	G protein signaling modulator 2	1p13.3	609245

I/ Introduction

1. De la génétique humaine aux troubles neurodéveloppementaux

1.1. La génétique humaine et sa diversité

Le noyau d'une cellule humaine est diploïde, contenant 46 chromosomes, répartis en 22 paires d'autosomes et une paire de gonosomes (chromosomes sexuels). Les chromosomes constituent le support de l'information génétique et sont formés de longues molécules d'acide désoxyribonucléique (ADN) en double hélice d'environ 2 mètres.(1,2)

Le génome humain comprend environ 3 milliards de paires de bases et est unique à chaque individu, à l'exception des jumeaux monozygotes. La diversité génétique, définie comme l'ensemble des variants dans les séquences d'ADN entre les individus ou les populations, est un moteur fondamental de l'évolution. Elle contribue à la variabilité phénotypique humaine et peut influencer la susceptibilité aux maladies, la réponse aux traitements ou encore certains traits physiologiques.

Les premières études sur la variation génétique ont montré que de nombreuses différences sont partagées entre les populations humaines. Deux individus pris au hasard dans une population caucasienne présentent en moyenne environ 3 à 4 millions de variants nucléotidiques, représentant environ 0,1 % de divergence génomique.(3)

Les variants génétiques les plus courants sont observés dans la plupart des populations humaines, mais certains variants plus rares sont spécifiques à certains sous-groupes. Le projet 1000 Genomes, lancé en 2008, a cartographié cette diversité en séquençant le génome de 2504 individus issus de 26 populations à travers le monde, mettant en évidence plus de 88 millions de variants. Il a révélé que de nombreux variant polymorphiques, aussi appelés SNPs (*Single Nucleotide Polymorphism*) varient considérablement en fréquence selon les populations, et qu'une partie de ces variants « non synonymes » influence la structure et la fonction des protéines. Ces différences suggèrent que certains variants génétiques pourraient être le résultat d'adaptations à des environnements particuliers.(4,5)

Au-delà de l'étude de la diversité génétique, les études d'association pangénomiques (GWAS) ont été menées à grande échelle sur des milliers de patients et de témoins sains. Grâce à cette

approche, des associations ont pu être établies entre certains variants génétiques et diverses maladies, permettant d'identifier des régions du génome particulièrement impliquées.

Si la majorité des variants génétiques sont neutres ou de faible impact, certains peuvent perturber des fonctions biologiques essentielles et être à l'origine de pathologies. C'est notamment le cas des maladies génétiques rares, qui résultent généralement de mutations délétères affectant un ou plusieurs gènes.

1.2. Maladies génétiques rares

Les maladies rares sont définies avant tout par leur faible prévalence, et non sur la base de critères pathologiques ou cliniques communs. Dans l'Union européenne, une maladie est considérée comme rare si elle touche moins d'une personne sur 2 000, soit environ ~260-450 millions de personnes dans l'ensemble de la population mondiale. Bien que peu fréquentes individuellement, ces maladies sont collectivement nombreuses : on estime qu'il en existe entre 6 000 et 10 000, affectant au total plusieurs centaines de millions de personnes dans le monde. (6)

En raison de cette définition fondée sur la fréquence, les maladies rares forment un ensemble hétérogène, englobant une grande diversité de tableaux cliniques et de mécanismes physiopathologiques. Cette thèse se concentrera sur un sous-ensemble de ces pathologies, à savoir les maladies génétiques rares, qui représenteraient environ 80 % de l'ensemble des maladies rares.

Début 2020, la base de données OMIM, qui recense les gènes humains ainsi que les maladies génétiques et leurs phénotypes associés, répertoriait 6 594 maladies associées à des anomalies génétiques caractérisées au niveau moléculaire, impliquant collectivement 4 225 gènes.(7) Ce nombre ne cesse d'augmenter au fil des années, au rythme des avancées en génomique, en phénotypage et en partage de données cliniques. Cette dynamique est illustrée par les données de la base *MorbidGenes Panel*, qui comptabilise, en juin 2025, 5 072 gènes associés à des anomalies génétiques caractérisées. Il est ainsi concevable qu'avec le temps, nous puissions accumuler suffisamment de cas humains pour nous rapprocher d'un objectif longtemps imaginé par les généticiens : un crible de mutagenèse de saturation à l'échelle du génome humain, permettant de perturber systématiquement chaque gène et d'en observer les conséquences phénotypiques. Bien qu'accessible dans un cadre expérimental direct chez

l'Humain via la technique de saturation génomique (SGE), l'étude des maladies génétiques rares représente une forme naturelle de mutagenèse, qui, par accumulation, nous aide à cartographier les fonctions essentielles de chaque gène dans l'espèce humaine.(8)

Parmi l'ensemble des maladies génétiques rares, les troubles neuro-développementaux occupent une place particulière en raison de leur complexité clinique, de leur forte hétérogénéité génétique et de leur impact majeur sur la qualité de vie des patient.

1.3. Troubles neurodéveloppementaux

Les troubles neuro-développementaux se définissent cliniquement comme un ensemble d'affections survenant au cours de la période de développement. Ces troubles sont caractérisés par un déficit du développement, entraînant des altérations du fonctionnement personnel, social ou académique.(9–11)

Parmi ces troubles, la déficience intellectuelle (DI) se manifeste par une réduction des capacités intellectuelles et adaptatives, affectant trois domaines principaux :

- Le domaine conceptuel, qui inclut les aspects cognitifs tels que la connaissance, la mémoire et le raisonnement ;
- Le domaine social, impliquant les interactions interpersonnelles, telles que l'amitié et l'empathie ;
- Le domaine pratique, qui concerne l'autonomie dans les activités de la vie quotidienne.

On parle généralement de déficience intellectuelle à partir de l'âge de 5 ans, lorsque les fonctions cognitives sont suffisamment développées pour être évaluées de manière fiable. Avant cet âge, on utilisera plus le terme de retard global de développement.(12)

Pour poser un diagnostic de DI, trois critères doivent être satisfaits :

1. Un déficit des fonctions intellectuelles
2. Un déficit des fonctions adaptatives

3. L'apparition de ces déficits durant la période de développement

La prévalence des troubles neuro-développementaux, y compris la DI, est estimée entre 1 et 3 %, avec des variations régionales notables. (10,11)

Les causes de la DI sont variées. Bien que des facteurs exogènes tels que des traumatismes ou des infections puissent être impliqués, un nombre important est d'origine génétique. La DI peut survenir de manière non syndromique (isolée), mais elle est souvent associée à d'autres signes cliniques, tels que des caractéristiques dysmorphiques, des anomalies congénitales ou encore des troubles neurocomportementaux, comme l'autisme (10,11)

Ces troubles sont caractérisés par une grande hétérogénéité clinique et génétique, ce qui rend leur diagnostic particulièrement complexe. Le premier diagnostic génétique de la DI a été posé en 1959, avec l'identification de la trisomie 21, responsable du syndrome de Down (MIM #190685), qui est l'anomalie la plus fréquente dans les cas de DI.(10,13)

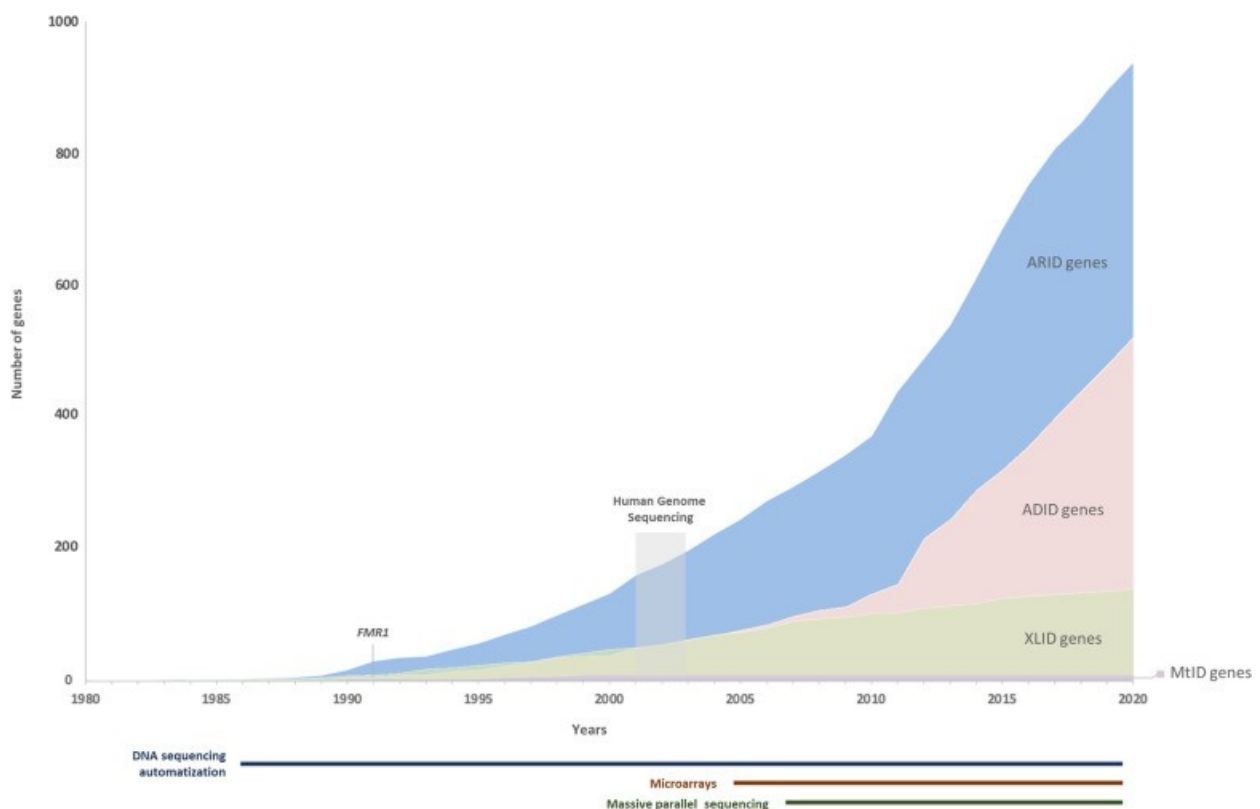


Figure 1: Gènes de DI identifiés au cours des années. ARID: Déficience Intellectuelle Autosomique Dominante; ADID: Déficience Intellectuelle Autosomique Récursive; XLID: Déficience Intellectuelle lié à l'X; MtID: Déficience Intellectuelle Mitochondriale (Figure extraite de Maia et al, 2021)

Bien que rarement traitable, un diagnostic moléculaire est essentiel pour les familles, car il permet de les accompagner dans le processus d'acceptation de la maladie et d'offrir des conseils génétiques en vue de la planification d'éventuelles grossesses futures. Les avancées technologiques récentes ont considérablement accru la capacité à établir des diagnostics moléculaires, tout en améliorant notre compréhension des causes génétiques de ces troubles.(Figure 1)(10)

À ce jour, plus de 3000 gènes associés dans le développement de la DI et candidats ont été identifiés (selon la base de donnée SysNDD) (Figure 1). Actuellement, des variants pathogènes dans environ 1 700 gènes ont été directement associés à des troubles neurologiques (OMIM, avril 2025).(7) Par ailleurs, la prévalence de la DI est plus élevée chez les hommes que chez les femmes, en raison de l'implication d'environ une centaine de gènes liés au chromosome X. Cette hétérogénéité génétique résulte non seulement de la présence de variants cytogénétiques, tels que des anomalies du nombre ou de la structure des chromosomes, mais aussi de variants génétiques ponctuels.(14)

2. Variants génétiques, mécanismes et impact

Les variants polymorphiques ou pathogènes sont très diverses, allant des variants d'un seul nucléotide (*Single Nucleotide Variant*, SNV), qui représentent des altérations ponctuelles d'une seule base de l'ADN, à des réarrangements cytogénétiques plus importants, tels que les variants du nombre de copies (*Copy Number Variants*, CNV) et les variants de structure (*Structural Variants*, SV) (Figure 2). (14,15)

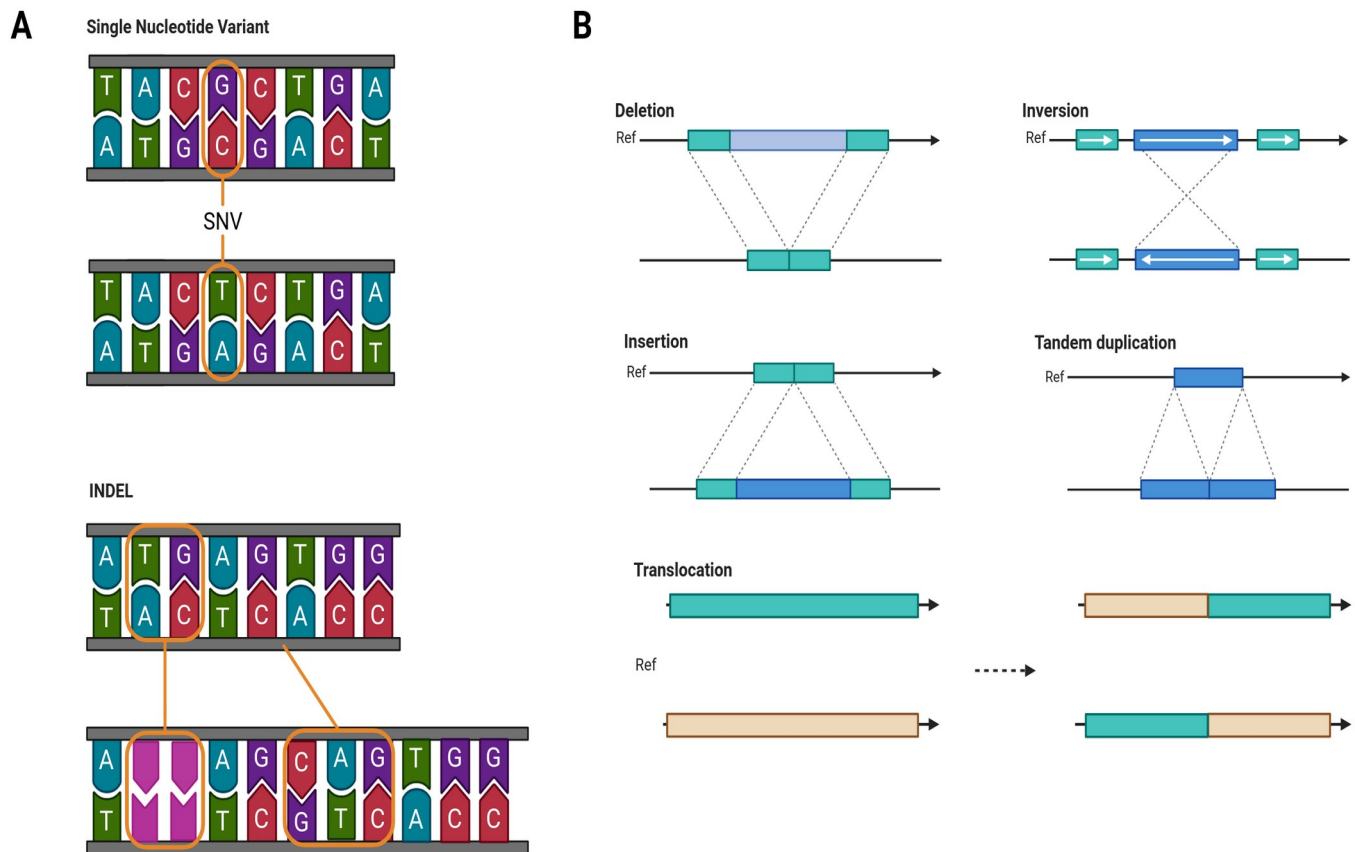


Figure 2: Type de variant génétique: (A) Exemple d'un SNV: $G > T$ et d'un INDEL (B) Exemple de SVs

2.1. Variants génétiques

2.1.1. Variant Nucléotidique

Les SNVs sont des variants génétiques où un seul nucléotide d'une séquence d'ADN est substitué (Figure 2.A). Ce sont les formes les plus courantes de variation génétique, avec plus de 630 millions de SNVs identifiés et répertoriés dans la base de données ClinVar.(17). Ces variants peuvent se situer dans des régions codantes ou non codantes du génome et sont liées à de nombreux troubles neuro-développementaux, entraînant des effets phénotypiques divers. (17,18)

Dans le contexte d'une séquence codante, un SNV peut avoir plusieurs conséquences. Un variant est dit synonyme s'il entraîne un changement dans la séquence codante sans affecter la séquence d'acides aminés, grâce à la redondance du code génétique. À l'inverse, un variant

peut être qualifié de non-sens s'il provoque un changement dans la séquence d'acides aminés et introduit un codon stop, aboutissant ainsi à une protéine tronquée. Enfin, un variant faux-sens entraîne également une substitution d'un acide aminé pour un autre, mais sans créer de codon stop.(18)

Les SNVs sont parfois appelés polymorphismes nucléotidiques. La différence principale entre ces deux termes réside dans leur fréquence au sein de la population: un variant est qualifié de SNP lorsque sa fréquence dépasse 1 %.(18)

2.1.2. INDEL :

Les INDELs (insertions et délétions) sont des variants génétiques impliquant l'ajout ou la suppression de quelques bases nucléotidiques dans le génome, généralement entre 1 et 50 paires de base (pb) (Figure 2.A). Comme les SNVs, ces variants peuvent se situer dans des séquences codantes ou non codantes.

Lorsque des INDELs apparaissent dans une séquence codante, leur impact sur la traduction dépend de leur taille. Les INDELs qui ne sont pas des multiples de trois induisent un décalage du cadre de lecture, ce qui peut mener à la production d'une protéine tronquée ou non fonctionnelle. (18)

Une étude basée sur les données du projet 1000 génomes a révélé que les INDELs sont relativement rares, se produisant environ 16 fois moins fréquemment que les SNPs. (5,18,19)

Ces variants inférieurs à 50pb comprenant donc les INDELs ainsi que les SNPs ne font pas partis de l'étude et ne seront donc pas plus abordés.

2.1.3. Les variants cytogénétiques

Les anomalies cytogénétiques se classent en deux grandes catégories : les anomalies de nombre et les anomalies de structure. Les anomalies de nombre, également appelées aneuploïdies, sont caractérisées par un déséquilibre du nombre total de chromosomes. Parmi ces anomalies, on retrouve la trisomie, qui se traduit par la présence d'un chromosome surnuméraire alors que la monosomie, se manifeste par la perte d'un chromosome dans une paire.(20)

Ce travail ne se concentrera pas sur les anomalies de nombre de chromosomes et ne les abordera donc pas davantage.

Les anomalies de structure, qui se définissent par des modifications d'au moins 50 nucléotides, sont des altérations cytogénétiques ayant un impact majeur sur la structure globale du génome (Figure 2.B). Ces variants de structure représentent une force mutationnelle importante, modifiant l'organisation génomique et générant des effets sur les fonctions cellulaires.

Les variants de structure, loin d'être rares, sont en réalité très fréquents dans le génome humain. Une étude menée en 2020, qui a analysé des génomes complets par séquençage courts fragments sur de 14 891 échantillons, a identifié plus de 433 371 variants de structure. Cette analyse a permis d'établir un nombre médian de variants de structure dans la population générale, avec une moyenne d'environ 7 439 variants par génome (Figure 3). Ce chiffre est bien plus élevé que celui observé dans le cadre du Projet 1000 Genomes, qui avait fourni un premier aperçu de la variabilité génomique structurelle humaine.(5,21) Sur la base des estimations actuelles, la charge totale des SV par génome est environ 200 fois inférieur au nombre moyens de SNV (4×10^6) et plus de 40 fois moins que les INDELs (8×10^5). (22)

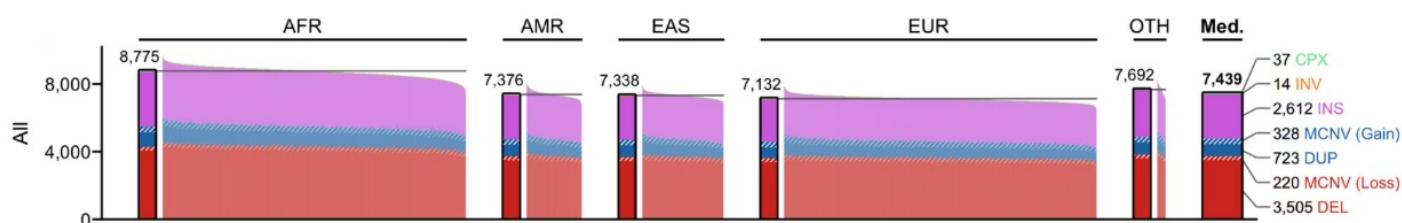


Figure 3: Comptage médian des SVs totaux par génome (Figure extraite de Collins et al, 2020)

Les variants de structure se divisent en deux catégories en fonction de leur impact sur le dosage de l'ADN : les variants déséquilibrés et les variants équilibrés. Les variants déséquilibrés, ou variants en nombre de copies (*Copy Number Variant* ; CNVs), incluent les duplications, qui entraînent un gain d'ADN, et les délétions, qui en entraînent une perte. Ces modifications perturbent le dosage de l'ADN, affectant ainsi l'équilibre génique. À l'inverse, les variants dits équilibrés, tels que les inversions ou les translocations, ne modifient pas la quantité totale d'ADN dans le génome. Toutefois, ces réarrangements structuraux peuvent avoir des conséquences pathogènes importantes, en particulier lorsque les points de cassure interrompent un gène essentiel ou déplacent un gène dans un environnement régulateur inapproprié. Ce type

de réarrangement peut ainsi altérer l'expression d'un gène, modifier les interactions entre promoteurs et *enhancers*, ou perturber l'architecture tridimensionnelle du génome, avec des effets potentiels sur la transcription.(23)

Ces réarrangements chromosomiques sont importants pour la diversité génétique et l'évolution humaine, car ils contribuent à la variabilité entre les individus. Toutefois, certains peuvent être associés à des troubles : environ 10% des patients présentant des DI ou et 5 % des patients présentant des troubles du spectre autistique présentent un variant de structure qui peut expliquer leur phénotype.(24) Cela met en évidence l'importance des variants de structure qui, bien qu'étant souvent neutres, peuvent provoquer des phénotypes cliniques lorsqu'ils affectent des gènes impliqués dans le développement ou les fonctions neuronales.(9)

Les variants de structure émergent principalement au cours de trois étapes cellulaires: pré-méiotique, méiotique et postzygotique. Cependant, bien que ces moments soient identifiés comme des périodes critiques de génération de ces variants, leur timing précis reste souvent difficile à déterminer.

2.1.3.1. Variants intrachromosomiques simples

Les variants intrachromosomiques simples englobent les délétions, les duplications et les inversions qui n'affectent qu'un seul chromosome et résultent d'un ou de deux points de cassures. Les inversions correspondent au renversement d'une portion du génome (Figure 2.B). Les délétions peuvent se localiser à l'intérieur d'un bras chromosomique (délétions interstitielles) ou tronquer l'extrémité d'un chromosome (délétions terminales) (Figure 2.B). En ce qui concerne les gains de copie, ils présentent une certaine diversité en fonction de la réinsertion de la copie surnuméraire (Figure 2.B). Celle-ci peut se faire en tandem, en miroir, intercalée (*interspersed*), ou être insérée dans une autre région de l'ADN.(23) La diversité de la réalité de l'organisation des gains de copie complexifie leur interprétation en routine diagnostique.

2.1.3.2. Variants interchromosomiques simples

Les variants interchromosomiques simples, en particulier les translocations, impliquent un échange de matériel génétique entre deux chromosomes distincts. Dans la majorité des cas, il s'agit de translocations réciproques équilibrées, où deux chromosomes échangent des seg-

ments sans perte ni gain d'information génétique. On distingue deux principaux types de translocations équilibrées : les translocations robertsoniennes, impliquant la fusion de deux chromosomes acrocentriques, et les translocations réciproques (Figure 2.B).

Bien que ces translocations soient généralement asymptomatiques, car elles ne modifient pas le nombre de copies des gènes impliqués, elles peuvent avoir des effets phénotypiques si les points de cassure impactent des gènes impliqués en pathologie. Chez les individus porteurs de translocations équilibrées, la méiose peut être perturbée, entraînant un risque d'infertilité en raison d'un arrêt méiotique. Il est aussi important de noter que bien que de nombreux porteurs de translocations équilibrées soient phénotypiquement sains, ils présentent un risque accru de transmettre ces variants à leur descendance, ce qui peut mener à des complications génétiques chez les enfants, telles qu'un déséquilibre génique ou un risque accru de fausse couche. (23,25)

À l'inverse, les translocations déséquilibrées, qui conduisent à un gain ou à une perte de matériel génétique, peuvent entraîner des anomalies cliniques telles que la trisomie ou la monosomie. Ces déséquilibres sont souvent associés à des DI et peuvent être à l'origine de l'apparition et la transmission de DI au sein des familles.(23)

2.1.3.3. Variants chromosomiques complexes

Un variant chromosomique est qualifié de complexe lorsqu'il présente au moins trois points de cassure dans le génome, une caractéristique souvent associée aux maladies onco-hématologiques mais rare en cytogénétique constitutionnelle. Ces réarrangements chromosomiques complexes présentent des organisations spécifiques, auxquelles des termes particuliers sont attribués selon leur structure.

La duplication inversée adjacente à une délétion terminale est un exemple de ce type de réarrangement.(Figure 4.A) Relativement fréquente, cette structure est présente aussi bien dans les contextes oncologiques que dans le génome constitutionnel. Un autre type courant est la duplication-normale-duplication (DUP-NML-DUP), caractérisée par des duplications entourant une région de copie normale.(Figure 4.B) Ces DUP-NML-DUP ne sont pas associées à un syndrome spécifique, car elles proviennent de divers *loci* génomiques et concernent une grande variété de gènes. Cependant, ces réarrangements peuvent dupliquer, fusionner ou per-

turber des gènes aux points de cassure, rendant cruciale leur caractérisation précise pour évaluer leur impact potentiel sur la pathologie.(23)

Les triplications se déclinent en plusieurs formes. Les triplications simples, ou de type I, apparaissent sans autre variation de copie adjacente, tandis que les triplications de type II sont accompagnées de duplications adjacentes.(Figure 4.C) Un exemple notable de réarrangement complexe impliquant une triplication est la DUP-TRP/INV-DUP, une structure associant triplication et duplications.

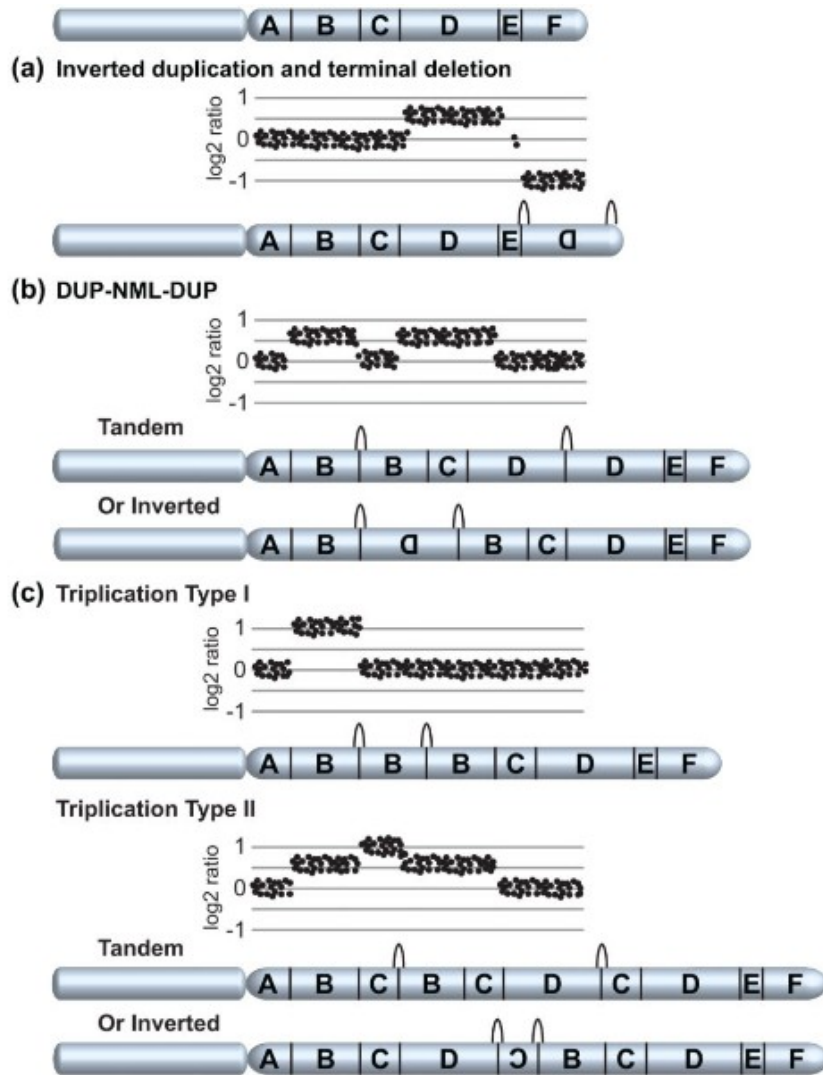


Figure 4: Réarrangements chromosomiques complexes: Caractérisation structurelle des réarrangements complexes associés aux résultats de CGH-Array (A) Duplication inversée adjacente à une délétion terminale (B) Duplication-Normal-Duplication (C) Triplication de Type I et de type II (Figure extraite de Weckselblatt, 2015)

Un exemple associé aux variants complexes sus-mentionnés sont les variants de structures associés au syndrome de duplication de *MECP2* (MIM #300260). Ce trouble neurodéveloppemental se manifeste par une hypotonie infantile, un retard global du développement, une déficience intellectuelle sévère, une absence de langage et des crises d'épilepsie. L'organisation génomique sous-jacente à ce syndrome inclut cinq catégories de variants de structure, parmi lesquelles 5 % correspondent à des DUP-NML-DUP et 20 % à des DUP-TRP/INV-DUP.(26)

Dans des cas plus complexes, on parle de chromoanagenèse, un phénomène où les chromosomes subissent des réarrangements massifs. La chromoanagenèse germinale est extrêmement rare mais a été observée chez les enfants atteints de troubles de développement. La croissance de cohortes témoins a révélé que ces événements peuvent également se produire dans la population générale à des fréquences très basses (environ 1/10000 individus).(22)

La chromoanagenèse comprend trois types spécifiques : le chromothripsis correspondant à l'éclatement des chromosomes, la chromoanasyntèse qui désigne la reconstitution des chromosomes et la chromoplexie qui est un tout-en-un désignant un remaniement massif du génome.(Figure 5).(23,27)

La chromothripsie est un réarrangement chromosomique fréquemment observé dans les cancers et originellement observé dans un cas de leucémie myéloïde chronique, où il touche environ 2 % des génomes tumoraux. Ce phénomène se distingue par un éclatement massif d'un ou plusieurs chromosomes (parfois jusqu'à cinq), suivi d'un réassemblage aléatoire des fragments chromosomiques (Figure 5). Principalement *de novo*, la chromothripsie est caractérisée par une prépondérance d'inversions et de délétions, avec des duplications relativement rares. Parmi ses principales caractéristiques, elle affecte généralement un seul des deux chromosomes concernés, présente des points de cassure groupés et entraîne une variation du nombre de copies, allant de l'hétérozygotie à une perte complète. Le réassemblage aléatoire des fragments mène à une reconstruction complexe et souvent désorganisée des chromosomes impliqués.(23,27,28)

La chromoanasyntèse est un processus de reconstitution chromosomique qui se limite à un seul chromosome ou à un *locus* spécifique (Figure 5). Contrairement à la chromothripsie, elle génère non seulement des délétions, mais aussi des duplications et triplications, impliquant entre 8 et 33 points de cassure. Les principales caractéristiques de la chromoanasyntèse incluent une combinaison complexe de réarrangements structuraux, des modifications localisées du nombre de copies avec une prédominance de duplications et triplications, ainsi que la présence de microhomologies aux points de cassure. Ces caractéristiques font de la chromoanasyntèse un réarrangement chromosomique distinct et complexe.(23,27,29)

Et enfin, la chromoplexie est un processus de remaniement massif du génome. Ce phénomène résulte de multiples translocations et de délétions inter/intra-translocations (Figure 5). Ces

phénomènes peuvent impliquer jusqu'à 8 chromosomes distincts en chaîne fermée. On va retrouver donc plusieurs chromosomes dérivés sans réelle grosse altération en nombre de copie. (27,30)

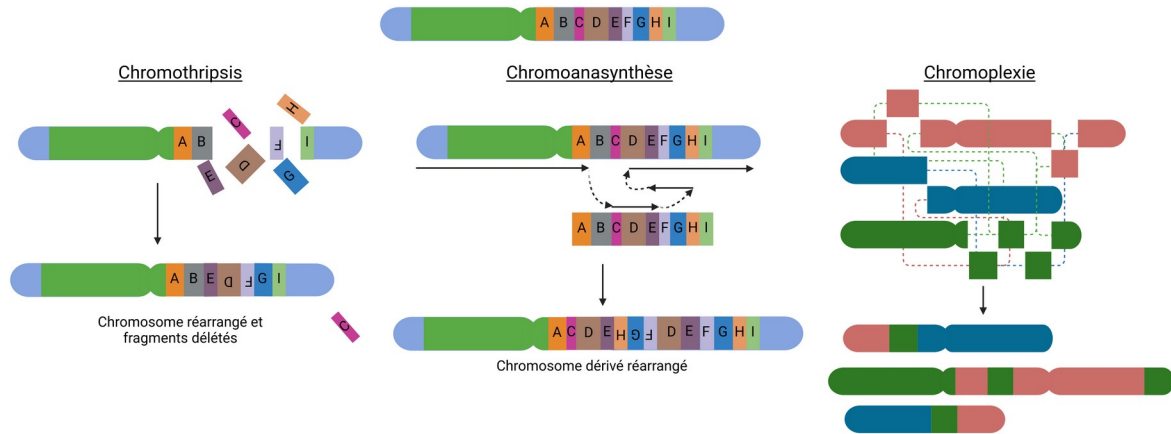


Figure 5: Schéma des trois processus associés à la chromoanagenèse: Chromothripsis, Chromosynthèse et Chromoplexie.

Tous les variants génétiques, qu'il s'agisse de SNVs, d'INDELs, de variants de structure ou de variants chromosomiques complexes, possèdent des caractéristiques distinctes qui nécessitent des approches spécifiques pour leur détection et leur interprétation. Leur impact dépend de leur nature et de leur position dans le génome, influençant l'expression et la fonction des gènes. Ces variants résultent de mécanismes mutationnels variés, tels que des erreurs de réparation, de recombinaison ou de réplication, dont l'étude est essentielle pour comprendre leur origine et leurs conséquences biologiques.

2.2. Mécanismes d'apparitions des SV

Les SVs résultent de divers mécanismes mutationnels, principalement liés à des erreurs de réparation de cassures doubles brins (*Double Strand Breaks* ; DSBs) de l'ADN, ainsi qu'à des processus de recombinaison et de réplication. Ces DSBs peuvent survenir spontanément lors de la réplication ou de la réparation de l'ADN, mais aussi être induites par des facteurs extérieurs comme les rayonnements ou certains agents chimiques. En moyenne, pendant la réplication, une DSB apparaît naturellement tous les 10^8 paires de bases, mais ce taux peut considérablement augmenter sous l'effet d'expositions. Par exemple, une mammographie peut en

induire environ 0,016 lésions/cellules, un TEP-scan environ 0,4 lésions/cellules, tandis qu'une mission spatiale de 60 jours pourrait en générer jusqu'à 2 lésions/cellules. (31,32)

Sur le plan évolutif et populationnel, plusieurs études ont tenté d'estimer le taux d'apparition spontanée de ces réarrangements au fil des générations. Ainsi, une étude basée sur le séquençage de nouvelle génération (*Next-Generation Sequencing* ; NGS) a estimé qu'en moyenne, 0,11 à 0,29 SV *de novo* apparaissent à chaque génération, une valeur probablement sous-estimée du fait des limites de détection actuelles.(22)

La détection de ces mécanismes repose sur l'analyse des "cicatrices" laissées dans le génome, qui fournissent des indices sur le type de processus ayant conduit à la formation d'un variant de structure. Ces marques génomiques permettent de retracer l'origine du mécanisme impliqué, que ce soit une erreur de recombinaison, une anomalie lors de la réplication ou une réparation inappropriée des DSBs.

2.2.1. Variant récurrent

Les variants de structures cytogénétiques sont des événements rares et spontanés dans le génome, mais certains d'entre eux apparaissent de manière récurrente dans les bases de données. Ces variants récurrents, observés chez plusieurs individus, présentent des caractéristiques communes, comme la taille, le contenu génique, et des co-localisations des points de cassure. Cette récurrence indique souvent l'existence de mécanismes biologiques sous-jacents qui facilitent leur apparition, telle que la recombinaison homologue (HR), un processus impliqué dans la réparation des DSBs.

La recombinaison homologue est essentielle pour réparer les DSBs de l'ADN, en utilisant la chromatide sœur comme modèle pour restaurer les séquences endommagées. Ce mécanisme repose sur l'alignement précis des brins d'ADN et l'homologie des séquences entre les chromatides, garantissant une copie réparée fidèle.

Le processus débute par la résection de l'ADN double-brin au niveau de la cassure, sur une distance d'environ 300 nucléotides. Cette résection génère des extrémités 3' simple brin, qui sont rapidement stabilisées par des protéines de liaison à l'ADN simple-brin. Ces protéines permettent ensuite l'appariement de la séquence endommagée avec la chromatide sœur. Cette étape aboutit à la formation d'une structure d'invasion de brin, appelée D-loop, dans laquelle

le brin lésé s'apparie avec la séquence complémentaire intacte. Une polymérase est alors recrutée pour synthétiser le brin manquant en utilisant la chromatide sœur comme matrice.

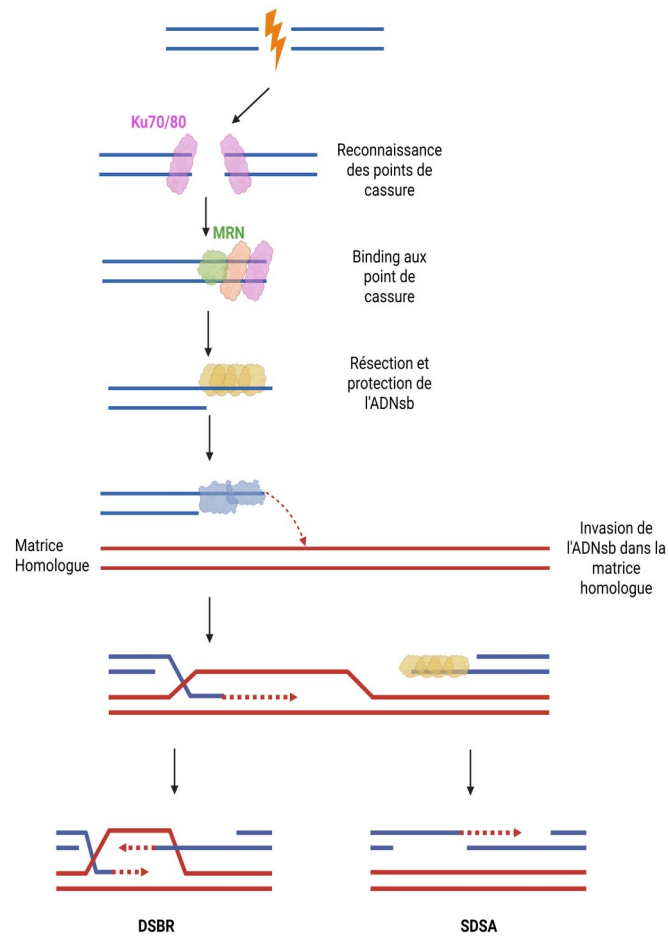


Figure 6: Recombinaison homologue (HR) pour la réparation de DSBs

À mesure que la synthèse progresse, la D-loop peut évoluer vers une jonction de Holliday. Deux voies principales de réparation sont alors possibles. Dans le modèle de réparation classique par cassure double-brin (DSBR), la seconde extrémité de la cassure participe également à la formation d'une jonction de Holliday, aboutissant à la formation d'une double jonction. Cette dernière est ensuite résolue par clivage enzymatique, ce qui peut conduire à un échange de matériel génétique entre les chromatides. À l'inverse, dans le modèle de type SDSA (*Synthesis-Dependent Strand Annealing*), le brin nouvellement synthétisé se détache de la chromatide sœur après une courte phase d'élongation. Il s'apparie alors avec l'autre extrémité de la

cassure, évitant la formation de structures intermédiaires complexes et limitant ainsi les risques de recombinaison croisée.(Figure 6)(33)

La HR, essentielle à la réparation des DSBs dans l'ADN, peut être trompée par la présence de séquences répétées, représentant 50% du contenu du génome humain.(34) Elles incluent des éléments mobiles comme les séquences *Alu*, des répétitions en tandem, et des régions à faibles nombres de copies (*Low Copy Repeats* ; LCRs), qui comprennent des duplications segmentaires (10 to 400kilobases (kb)), de longs éléments nucléaires dispersés (*Long Interspersed Nuclear Elements* ; *LINE*) (~6kb), et de courts éléments nucléaires dispersés (*Short Interspersed Nuclear Elements* ; *SINE*) (~300pb).(35) Ces séquences répétées, en particulier les grands LCRs qui représentent environ 4 à 5 % du génome, créent des zones vulnérables aux réarrangements structuraux à cause de leur forte homologie avec d'autres séquences du génome (au moins 90 % d'identité sur une longueur minimale de 1 kb).(34)

Lorsqu'une recombinaison homologue non allélique (*Non Allelic Homologous Recombination*; NAHR) se produit, elle est induite par des LCRs de compositions proches sur des chromatides sœurs, ce qui amène la cellule à engager par erreur un processus de recombinaison. Lorsque les LCRs sont alignées dans le même sens, la NAHR favorise l'apparition de délétions ou de duplications, créant des *crossing-over* inégaux.(Figure 7.A) À l'inverse, lorsque les LCRs sont orientées en sens opposé, la NAHR génère des inversions.(Figure 7.B) Ces réarrangements altèrent significativement l'organisation du génome et peuvent être observés chez différents individus avec des caractéristiques communes.(34)

La nature réciproque du mécanisme de NAHR implique qu'un réarrangement donné peut simultanément produire une délétion et une duplication d'une même région génomique. Ainsi, certains troubles génomiques présentent à la fois des formes associées à une délétion et des formes associées à une duplication de la même région.

Un exemple notable de cette dualité phénotypique est illustré par les syndromes de Smith-Magenis et de Potocki-Lupski. Le syndrome de Smith-Magenis, principalement causé par une délétion de la région 17q11.2 (MIM #182290), a son homologue dans le syndrome de Potocki-Lupski (MIM #610883), où une duplication de la région 17q11.2 est impliquée. Dans ce contexte, on retrouve trois LCRs, appelés SMS-REPs, qui jouent un rôle clé dans la survenue de réarrangements récurrents. Les SMS-REPs proximal et distal, d'environ 256 kb et 176 kb

respectivement, sont orientés dans le même sens. Cette orientation favorise la formation de délétions ou de duplications récurrentes par NAHR et explique le double syndrome. En revanche, une troisième LCR, intermédiaire, orientée en sens inverse, est impliquée dans la formation d'inversions. Les patients atteints du syndrome de Smith-Magenis tendent, par exemple, à présenter une obésité en fin d'enfance, tandis que les individus avec le syndrome de Potocki-Lupski montrent des tailles et des poids inférieurs à la moyenne. (36–38)

Bien que la plupart des variants récurrents observés soient des variants en nombre de copies, certaines translocations récurrentes existent, comme la translocation $t(11;22)(q23;q11)$. Ce type de translocation, bien que souvent asymptomatique chez les porteurs, est favorisé par des structures secondaires spécifiques de l'ADN, favorisant des DSBs et donc facilitant les réarrangements. Toutefois, les descendants des individus porteurs de cette translocation $t(11;22)$ présente un risque d'hériter d'un chromosome surnuméraire dérivé du chromosome 22, menant au syndrome d'Emanuel (MIM #609029). Ce syndrome est caractérisé par diverses anomalies congénitales, une dysmorphie cranio-faciale et un retard de développement et intellectuel important.(27–29) Ces réarrangements proviennent de la présence de séquences homologues ainsi que de *hot-spots* de DSBs situés dans des régions spécifiques du génome. Ces régions contiennent des répétitions riches en AT de type palindromique (*Palindromic AT-Rich Repeats*, PATRRs), favorisant la NAHR et conduisant ainsi à la formation de translocations. (23,25,39,40)

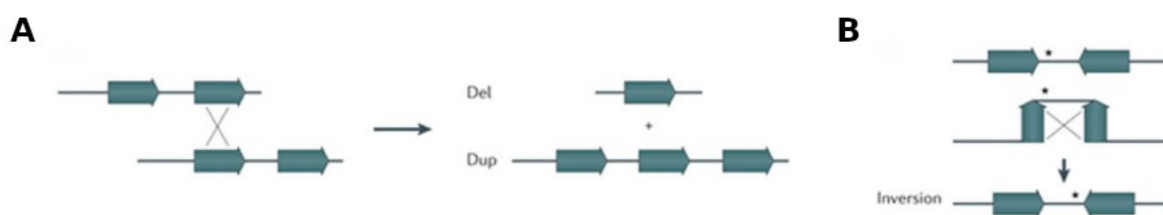


Figure 7: Mécanisme NAHR: (A) Recombinaison entre deux LCRs orientés dans le même sens générant délétions et duplications (B) Recombinaison entre deux LCRs dans une orientation inverse générant une inversion. (Figure extraite de Carvalho et Lupski, 2016)

2.2.2. Variant non récurrent

Les variants de structure non récurrents sont des réarrangements caractérisés par une taille différente à un *locus* donné chez des individus non apparentés. Bien que distincts entre eux, ces

réarrangements peuvent malgré tout générer des phénotypes similaires chez les individus qui en sont porteurs. Ce phénomène s'explique lorsque la région de chevauchement minimale de ces réarrangements affecte un gène sensible au dosage. (34)

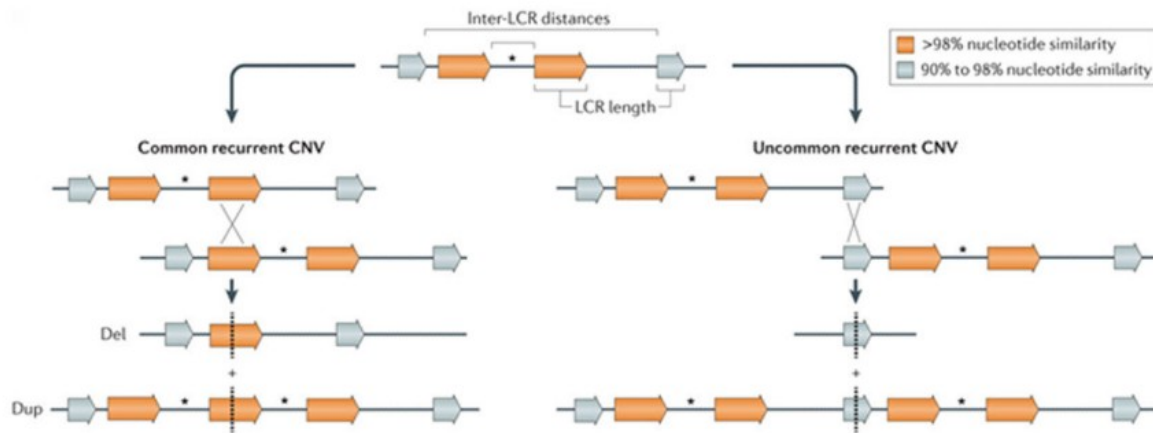


Figure 8: Génération de SVs récurrents et non récurrents via le mécanisme de NAHR. (Figure extraite de Carvalho et Lupski, 2016)

Certains variants complexes peuvent toutefois être influencés par la présence de LCRs avec une homologie de séquence moins importante que lors de SVs récurrents. En particulier, des structures comme les DUP-TRP/INV-DUP sont fréquemment médiées par ces LCRs. Les LCRs jouent ici un rôle en facilitant la NAHR, qui peut conduire à des réarrangements spécifiques. (Figure 8). Cependant, la majorité des réarrangements non récurrents se forment en l'absence de LCR et résultent plutôt d'autres mécanismes d'instabilité génomique. (34)

Deux principaux mécanismes vont permettre d'aboutir à ces réarrangements non récurrents : le *Non-Homologous End-Joining Repair* (NHEJ) et les mécanismes associés à la réplication.

2.2.2.1. Non-Homologous End-Joining

L'un des principaux mécanismes de réparation des cassures double brin est le *Non-Homologous End-Joining*, qui se divise en deux variantes : le NHEJ classique (c-NHEJ) et le NHEJ alternatif (a-NHEJ ou MMEJ). À la différence de la recombinaison homologue, qui utilise des séquences répétées pour réparer l'ADN de manière précise, le c-NHEJ est un processus de réparation non fidèle. Cela signifie qu'il ne nécessite pas de longues séquences d'homologie pour fonctionner. Toutefois, dans certains cas, il peut faire appel à de courtes régions de microhomologie, généralement de 2 à 3 bases, pour faciliter la réparation.

Dans le c-NHEJ, la réparation débute par la reconnaissance des extrémités d'ADN cassées et une stabilisation de ces extrémités et une protection contre la dégradation. Une nucléase, intervient pour ajuster les extrémités d'ADN en éliminant les bases endommagées, créant ainsi des extrémités franches. Ces modifications sont nécessaires pour permettre l'action de la ligase. Enfin, les extrémités sont reliées pour rétablir la continuité de l'ADN.(34,37)

Comme mentionné, le c-NHEJ est non conservatif, c'est-à-dire qu'il ne garantit pas une copie exacte de la séquence initiale, laissant souvent des cicatrices dans l'ADN que cela soit des petites insertions ou délétions. Cette réparation imparfaite peut également donner naissance à divers types de SVs, qui ne sont pas uniquement limités aux réarrangements en nombre de copies. En effet, toute DSB réparée par c-NHEJ peut potentiellement aboutir à des variants de structure équilibrés, tels que des translocations ou des inversions, ainsi qu'à des variants déséquilibrés, tels que les duplications et délétions.

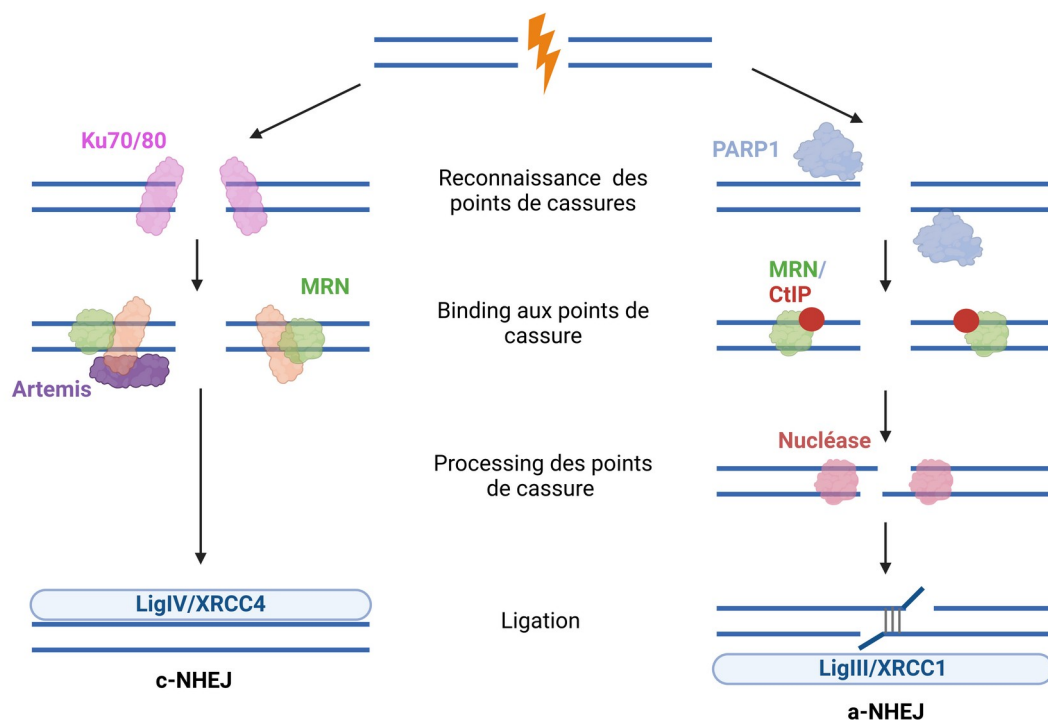


Figure 9: Mécanisme de Non-Homologous End Joining: c-NHEJ (panel gauche); a-NHEJ (panel droit).

Le a-NHEJ, ou MMEJ, devient actif lorsque le c-NHEJ n'est pas viable. Contrairement au c-NHEJ, qui dépend de protéines spécifiques pour stabiliser et réparer les cassures d'ADN, le a-NHEJ s'appuie sur la micro-homologie. Ce mécanisme est fréquemment observé lorsque les

extrémités d'ADN sont plus dégradées ou difficiles à réparer par c-NHEJ, permettant ainsi de réparer les cassures dans des conditions moins favorables.(Figure 9)(34,37)

La réparation des cassures double-brin débute par la reconnaissance des extrémités d'ADN lésées. Une fois que la protéine est recrutée et fixée sur les extrémités, elle induit une ouverture de la chromatine environnante, notamment via des modifications post-traductionnelles des histones, afin de faciliter l'accès aux autres facteurs de réparation. Un complexe multi-protéique est ensuite mobilisé pour traiter les extrémités endommagées, initiant leur résection. Cette étape génère des extrémités 3' propices à l'appariement. La résection est poursuivie par d'autres cofacteurs, permettant l'exposition de courtes régions de micro-homologie. Ces séquences servent de points d'ancrage pour aligner les deux brins d'ADN. Enfin, l'étape de ligation vient souder les extrémités, assurant ainsi la restauration de la continuité du génome.(Figure 9)(37,41)

Cependant, cette réparation est moins fidèle que celle du c-NHEJ. Le processus peut entraîner des délétions, car l'ADN cassé doit parfois être raccourci pour permettre l'alignement des extrémités via des régions de micro-homologie. Cela rend le a-NHEJ plus susceptible d'introduire des variants génétiques, notamment des délétions, tout en permettant la réparation des cassures lorsque le c-NHEJ est défaillant. Chez les mammifères, le c-NHEJ reste le mécanisme principal, tandis qu'environ 10 % des réparations par jonction des extrémités sont effectuées par le a-NHEJ/MMEJ.(42,43)

L'étude des points de cassure est cruciale pour identifier le mécanisme à l'origine des variants de structure et évaluer le rôle du NHEJ. Ce mécanisme génère principalement des points de cassure simples, parfois associés à une courte homologie de 2 à 3 bases, ainsi qu'à de petites délétions ou insertions. En revanche, lorsque les cassures présentent des cicatrices avec une micro-homologie plus étendue (> 3 bases) et l'insertion d'éléments courts, le mécanisme impliqué semble davantage lié aux processus de réplication de l'ADN, ce qui suggère une origine distincte des réarrangements structuraux.

2.2.2.2. Mécanismes liés à la réplication

Les mécanismes de réparation des cassures double brin pendant la réplication sont essentiels pour préserver l'intégrité du génome. En effet, lors de la réplication de l'ADN, des bloquages peuvent survenir, perturbant ainsi la continuité du processus et entraînant potentiellement des

SVs. Trois mécanismes principaux sont impliqués dans la réparation de ces cassures au cours de la réplication : le *Break-Induced Replication* (BIR), le *Microhomology-Mediated Break-Induced Replication* (MMBIR) et le *Fork Stalling and Template Switching* (FoSTeS). (Figure 10)

Dans le mécanisme BIR, seule une extrémité est impliquée dans la réparation, et la réplication reprend à partir de cette extrémité, envoyant le surplomb 3' envahir une séquence homologue (70-100pb). Cela forme une D-loop, où le brin d'ADN cassé s'aligne et se répare en utilisant la séquence homologue comme modèle pour la synthèse. Cette réparation est assurée par l'ADN polymérase et peut s'étendre sur de longues distances, parfois jusqu'au télomère. Bien que ce mécanisme de réparation soit précis, il peut mener à des réarrangements tels que des délétions ou des duplications dans les régions voisines de la cassure.(Figure 10)(23,34)

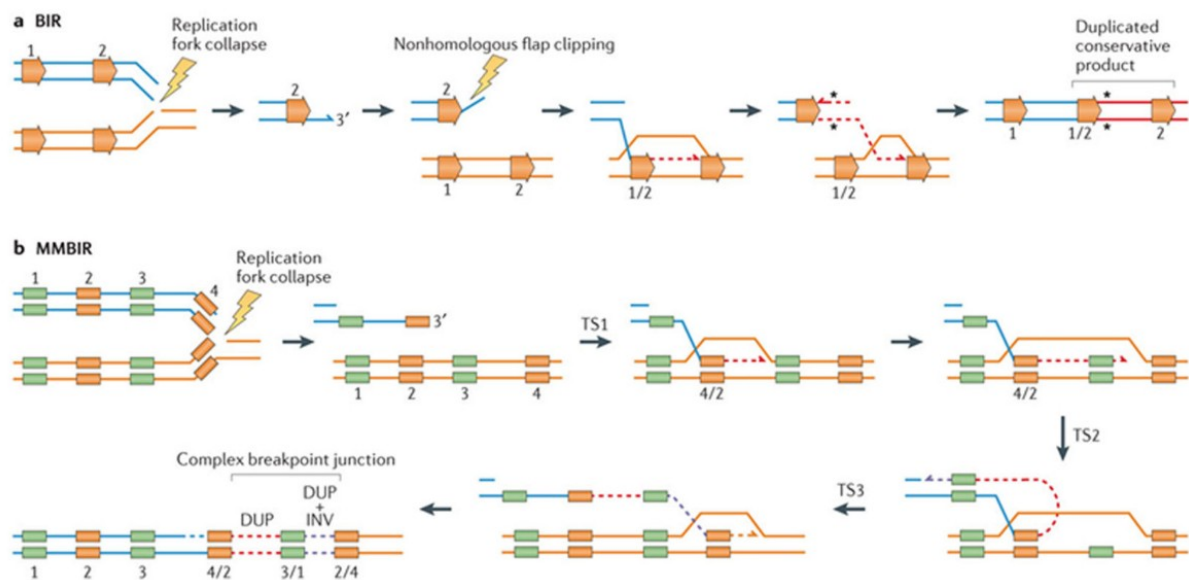


Figure 10: Mécanismes de réparation de l'ADN associé à la réplication.(A) Le BIR est initié lors de l'effondrement d'une fourche de réplication en présence de longues régions homologues.(B) Le MMBIR est initié dans des conditions similaires, mais repose sur de courtes régions homologues. (Figure extraite de Carvalho et Lupski, 2016)

Le MMBIR est une variante du BIR. Contrairement au BIR classique, qui nécessite une longue séquence d'homologie pour l'invasion du brin cassé, le MMBIR repose sur de courtes

régions de micro-homologie (2-9pb). Ce mécanisme intervient principalement lorsque la fourche de réplication est bloquée ou effondrée.

Dans ce processus, l'invasion du brin est facilitée par ces micro-homologies, et une polymérase à faible processivité est utilisée pour prolonger la synthèse d'ADN. En raison de la faible fidélité de cette réparation, le MMBIR est particulièrement impliqué dans la formation de variants de structure non récurrents, tels que des délétions, des duplications ou des inversions, rendant ainsi le génome plus vulnérable aux réarrangements.(Figure 10)

Le mécanisme FoSTeS se produit lorsqu'une fourche de réplication est bloquée ou rencontre un obstacle, ce qui entraîne un arrêt temporaire de la réplication. Ce phénomène a été observé dans des réarrangements génomiques complexes, comme ceux associés au syndrome de Pelizeus-Merzbacher (PMD, MIM #312080), une leucodystrophie hypomyélinale rare, principalement causée par des mutations dans le gène *PLP1* (protéolipide 1).(44) Lors de ce processus, la fourche de réplication peut basculer entre différentes régions du génome, générant des réarrangements sans points de cassure récurrents. Ce mécanisme utilise des micro-homologies (2-9pb), qui servent de points d'ancrage pour la reprise de la réplication sur un nouveau brin d'ADN. Par conséquent, de la même façon que le MMBIR, le FoSTeS mène à la formation de variants de structure uniques et imprévisibles.

Ces deux mécanismes associés sont toujours nommés ensemble. Ce sont des modèles proposés et aucune donnée ne permet de nommer un mécanisme préférentiellement. La base principale de leur réparation se base sur la micro-homologie et dans la plupart des variants observés, une micro-homologie sera observée à la jonction du point de cassure. Cependant dans certains cas, la micro-homologie peut se retrouver à distance du point de cassure (entre 5 et 100pb).(45)

Ces mécanismes sont essentiels pour maintenir l'intégrité génomique, mais leur potentiel à générer des erreurs pendant la réparation peut mener à des réarrangements génomiques variés, certains récurrents et d'autres non récurrents. En effet, les mécanismes de MMBIR et de FoSTeS sont très souvent nommés pour être la cause d'apparition de variants de structures non récurrents, que cela soit des variants de gains de copies, des inversions, des translocations mais aussi de variants beaucoup plus complexes. (34,37)

Les variants de structure complexes mentionnés précédemment, tels que les DUP-NML-DUP et les DUP-TRP/INV-DUP, sont bien documentés et étroitement liés aux mécanismes de réparation impliqués dans leur apparition. La faible processivité des polymérases joue un rôle dans ces réarrangements, en favorisant des « sauts » entre différentes régions du génome, ce qui explique la formation de ces structures complexes.(34,44,46)

Ces 4 mécanismes décrits ici ne sont cependant pas les seuls mécanismes pouvant aboutir à génération de variants de structures. Les mécanismes responsables de variants complexes sont encore plus divers et peuvent comprendre des cascades mutationnelles en plusieurs étapes à la mauvaise ségrégation des chromatides et peuvent aussi être actuellement inconnus.

2.3. Conséquences des variants de structures

L'étude de l'impact des variants génétiques constitue un défi majeur pour le diagnostic et le suivi des patients. Une analyse menée sur 465 familles présentant des troubles du neurodéveloppement a révélé que, parmi l'ensemble des variants candidats détectés, 203 ont été classés comme pathogènes ou probablement pathogènes. Parmi ceux-ci, 65 % (132/203) étaient des SNV, 22 % (45/203) des INDELs et 12 % des variants de structure.(47)

Les SNV et INDELs, largement étudiés en diagnostic moléculaire, sont majoritairement situés dans les régions codantes, où ils altèrent la séquence protéique en modifiant les acides aminés, ou dans les introns, où ils perturbent l'épissage. Ces mécanismes, bien documentés, facilitent généralement leur interprétation, à l'exception des variants faux-sens, dont l'impact fonctionnel peut être plus difficile à prédire.

En revanche, les variants de structure présentent une grande diversité et des conséquences plus complexes à interpréter. Bien qu'il a été démontré que plus de 49 % des SV détectés sont petits et inférieurs à 500pb, des très grands variants de structures subsistent et sont ceux qui vont poser le plus de difficulté à l'interprétation.(22)

2.3.1. Sensibilité au dosage

Les variants déséquilibrés, comme mentionné précédemment, correspondent aux variants du nombre de copies, incluant les duplications et les délétions qui entraînent respectivement un

gain ou une perte de matériel génétique. Dans certains cas, des CNVs de grandes tailles peuvent englober un ou plusieurs gènes entiers, induisant ainsi des conséquences fonctionnelles significatives. Ces altérations peuvent entraîner soit une haploinsuffisance, soit une triplosensibilité. La plupart des gènes impliqués dans ces mécanismes sont associés à des maladies autosomiques dominantes.

L'haploinsuffisance se produit lorsque l'expression d'un seul allèle ne suffit pas à assurer le bon fonctionnement du gène dans la cellule, on parle d'intolérance à une perte de copie. Un exemple bien documenté est le gène *MED13L*, dont la perte d'une copie est associée au syndrome MED13L (MIM #616789). Ce syndrome se manifeste par une plagiocéphalie, un strabisme, une absence de langage, un retard du développement et une DI. (48) C'est aussi le cas du syndrome de Sotos (MIM #117550) par la délétion du gène *NSD1*.

À l'inverse, la triplosensibilité désigne une intolérance à l'augmentation du dosage d'un gène, c'est-à-dire une sensibilité au gain de copie. Un exemple caractéristique est le gène *STAG2*. Des études menées sur plusieurs familles présentant une duplication en Xq25 ont permis d'identifier une région minimale critique de 173 kb contenant uniquement *STAG2*, qui code pour une sous-unité du complexe cohésine. Les individus porteurs de cette duplication présentent un retard des acquisitions, des troubles du langage, une hypotonie, une DI, des troubles du comportement ainsi qu'une dysmorphie faciale caractéristique.(49) D'autres gènes sont connus tels que *CREBBP* (MIM #600140), *LMNB1* (#150340), *MECP2* (MIM #300005) ou encore *PLP1* (MIM #300401).

Ainsi, les variants déséquilibrés, en modifiant le nombre de copies de gènes essentiels, peuvent avoir des conséquences fonctionnelles majeures. L'haploinsuffisance et la triplosensibilité illustrent l'importance du dosage génique, comme le démontrent les exemples de *MED13L* et *STAG2*. Ces mécanismes soulignent l'impact potentiel des CNVs sur le développement, justifiant leur analyse approfondie dans un contexte clinique et génétique.

La taille du déséquilibre génomique joue également un rôle déterminant dans la probabilité d'un effet pathogène. En effet, un variant de grande taille a plus de chances d'inclure des gènes connus pour leur implication dans des pathologies, augmentant ainsi la probabilité d'une conséquence clinique significative. Toutefois, bien que de larges délétions impliquant plusieurs gènes pathogènes puissent expliquer le phénotype par un mécanisme d'haploinsuffi-

sance, la pathogénicité de nombreuses duplications ne peut être attribuée uniquement à la présence d'une copie supplémentaire mais peuvent impacter le génome de manière différente. (50)

2.3.2. Impact sur la séquence codante

Certains variants de structure, qu'ils soient équilibrés ou déséquilibrés, peuvent présenter des points de cassure situés à l'intérieur d'un gène. L'impact fonctionnel d'un tel remaniement dépend du type de variant et de la conservation ou non d'une copie intacte du gène concerné.

Dans certains cas, ces variants n'ont pas de conséquences majeures, comme c'est le cas lors de duplications en tandem où le gène partiellement dupliqué retrouve ses deux copies fonctionnelles. Cependant, lorsque la séquence d'un allèle est altérée, l'impact peut être plus important. En effet, certains variants de structure peuvent entraîner une interruption du gène du gène, ce qui conduit à la production de transcrits tronqués. Ces derniers peuvent alors être dégradés par le mécanisme de *nonsense-mediated mRNA decay* (NMD), entraînant ainsi une haploinsuffisance.

D'autres réarrangements intragéniques peuvent modifier profondément l'architecture du gène concerné. Par exemple, une duplication en tandem totalement intragénique peut allonger le gène et introduire des exons supplémentaires, modifiant ainsi le cadre de lecture et la structure protéique. De même, une inversion intragénique peut perturber l'ordre des exons et engendrer la production d'une protéine non fonctionnelle ou au contraire acquérir une nouvelle fonction. Selon la nature du remaniement et le rôle du gène impliqué, ces altérations peuvent conduire à une perte de fonction ou, dans certains cas, à un gain de fonction.

Dans des cas plus complexes, la protéine altérée peut induire un effet dominant-négatif, une situation qui se situe à la frontière entre les mutations perte de fonction et gain de fonction. La protéine altérée perd sa fonction initiale, tout en interférant avec la fonction normale des protéines produites par l'allèle sain. Ce mécanisme peut rendre certaines pathologies plus délétères à l'état hétérozygote qu'à l'état homozygote, comme c'est le cas des gènes *COL1A1* et *COL1A2*, impliqués dans les ostéoporoses imparfaites (MIM #166200). (50,51)

Un cas aussi retrouvé en constitutionnel mais de façon plus importante en oncologie sont les fusions de gènes. En effet, des translocations peuvent survenir lorsque les points de cassure se

situent chacun dans un gène. Avec la translocation de ceux-ci, on peut retrouver 2 gènes différents fusionnés. Le cas le plus connu est le chromosome de Philadelphie qui est une translocation $t(9;22)(q34.1;q11.2)$ qui crée le gène de fusion *BCR-ABL* en lien avec la leucémie myéloïde chronique (CML, MIM #608232). (62,63) Un seul cas de fusion de gènes lié à un remaniement constitutionnel a été décrit. Il s'agit de la fusion des gènes *PKT2* et *THOC2* (MIM #300395) suite à une translocation $t(X;8)(q25;q24.3)$ chez une patiente présentant une ataxie congénitale, une déficience cognitive et une hypoplasie congénitale cérébelleuse (Figure 11). (60)

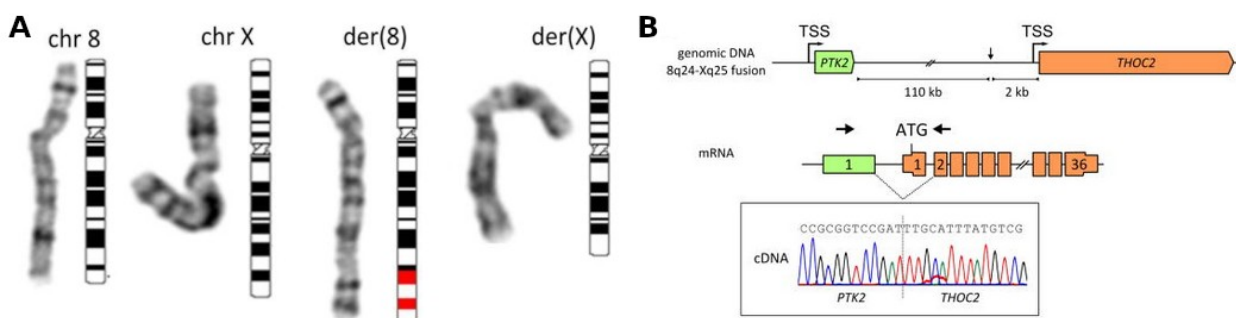


Figure 11: Translocation $t(X;8)$ entraînant la formation d'un transcrit de fusion *PKT2-THOC2*.(A) Caryotype montrant un chromosome X et un chromosome 8 WT, ainsi que les chromosomes X et 8 dérivés (*der(8)* et *der(X)*) résultant de la translocation.(B) Schéma illustrant la formation du transcrit de fusion au niveau de l'ADN et de l'ARNm, accompagné de la validation par séquençage Sanger de la séquence ARNm fusionnée. (Figure extraite de Di Gregorio et al., 2013)

2.3.3. Effet de position

Dans certains cas, les variants de structure n'affectent ni des gènes sensibles au dosage, et leurs points de cassure ne se situent pas dans des régions codantes. Toutefois, ils peuvent induire un effet de position, modifiant l'environnement génique et perturbant ainsi l'expression d'un gène.

Le premier mécanisme repose sur la *variegation* par effet de position (*Position Effect Variegation*, PEV), où un gène est déplacé d'un environnement euchromatinien (chromatine décondensée) permissif vers un environnement hétérochromatinien plus condensé, défavorable à sa transcription. Ce phénomène, largement étudié chez *Drosophila melanogaster*, n'a pas encore

été clairement démontré chez l'Homme. (52) Toutefois, un mécanisme s'en rapprochant a été proposé dans certains cas de translocations entre un autosome et le chromosome X chez des individus de sexe féminin (XX). La translocation d'une séquence autosomale sur le chromosome X peut entraîner sa co-inactivation en raison du processus d'inactivation physiologique de l'un des deux chromosomes X, conduisant ainsi à une hyperméthylation et à une possible répression transcriptionnelle de cette région autosomale.(53)

Les autres mécanismes, mieux caractérisés chez l'Homme, sont associés aux SVs perturbant l'interaction entre un gène et ses éléments régulateurs, compromettant ainsi son expression. La régulation transcriptionnelle repose sur trois composants clés:

- Le promoteur, site de recrutement de la machinerie transcriptionnelle.
- Les *enhancers* et *silencers*, régions d'ADN non codantes, modulant l'expression d'un gène en interagissant avec son promoteur.
- L'environnement chromatinien local, influençant ces interactions en fonction de l'organisation tridimensionnelle du génome.(54)

Si un variant de structure perturbe l'un de ces éléments, l'expression du gène peut être altérée.

2.3.3.1. Impact sur l'élément régulateur

Les *enhancers*, des régions spécifiques non codantes de l'ADN, jouent un rôle clé dans la régulation de la transcription et l'expression des gènes. Découverts il y a environ 40 ans par Schaffer et ses collaborateurs, leur importance a été révélée par une séquence de 72 pb du virus simien 40 (SV40) capable de stimuler l'activité de transcription du gène de la β -globine de lapin, indépendamment de l'orientation ou de la distance par rapport au gène cible.(55)

Les *enhancers* sont donc assez difficiles à détecter car ils sont dispersés dans les 98% du génome humain qui n'est pas codant, ce qui démontre d'un grand espace de recherche. De plus, ils jouent un rôle sur la transcription en agissant en *cis* sur les promoteurs des gènes, situés en amont du site d'initiation de la transcription. L'emplacement des *enhancers* est variable, ils peuvent se retrouver en amont ou en aval des gènes, avoir une action sur des gènes assez éloignés sur le plan linéaire et peuvent aussi réguler différents gènes. Pour que l'activité transcriptionnelle se fasse, les *enhancers* ainsi que les promoteurs disposent de site de liaison à l'ADN

qui permettent la fixation de facteurs de transcription et aussi à l'ARN polymérase en vue de la transcription.(55,56)

Les SVs peuvent directement affecter un élément régulateur, soit en altérant sa séquence, soit en modifiant l'environnement associé. D'une part, une modification de la séquence d'un *enhancer* peut perturber ses sites de fixation à l'ADN, compromettant ainsi son interaction avec les facteurs de transcription. D'autre part, une duplication ou une délétion survenant au sein d'un même environnement chromatinien peut entraîner une variation du nombre de copies des éléments régulateurs interagissant avec un gène donné, impactant potentiellement son expression.(Figure 13)

Un exemple bien documenté concerne le *locus SOX9* : la duplication de ses éléments régulateurs entraîne une réversion sexuelle chez les individus 46,XX (MIM #278850), tandis que leur délétion provoque une réversion sexuelle chez les individus 46,XY (MIM #616425). En revanche, aucune conséquence pathogène n'a été rapportée pour les inversions de régions contenant des enhancers.(57)

Dans un contexte plus directement lié à la DI, des variants ciblant un enhancer régulant *CSMD1* ont été identifiées, entraînant des perturbations de la neurogénèse.(58)

2.3.3.2. Impact sur l'environnement 3D

2.3.3.2.1. Domaines topologiquement associés

Le génome humain, bien que principalement décrit de façon linéaire est de plus en plus représenté en structure 3D. En effet, le génome est compartimenté en territoires bien distincts qui va permettre la mise en place d'une compartimentalisation des gènes pour réguler leur expression transcriptionnelle.

Les domaines topologiquement associés (*Topologically Associated Domains* ; TADs) ont été initialement identifiés grâce à la technique 5C, mais ont été entièrement cartographiés via le séquençage de la conformation de la chromatine, aussi appelé Hi-C (High-Throughput Chromosome Conformation Capture) . Les TADs sont des domaines chromatinien où les interactions entre régions génomiques proches (intra-TAD) sont favorisées, contrairement aux interactions entre domaines voisins (inter-TAD) (Figure 12.A). Ces domaines, généralement de quelques centaines de kb à 1-2 Mégabases (Mb) (de moyenne d'environ 800 kb), sont organi-

sés en boucles, délimitées par des éléments dits "isolants" qui jouent le rôle de frontières. Cette organisation contribue à favoriser les interactions de proximité et à limiter les contacts entre TADs adjacents.(59,60)

De plus, la formation de TADs favorise les interactions enhancer-promoter, facilitant donc l'activité transcriptionnelle en maintenant ces paires à l'intérieur des mêmes boucles et en restreignant les interactions avec d'autres éléments régulateurs extérieurs.(60)

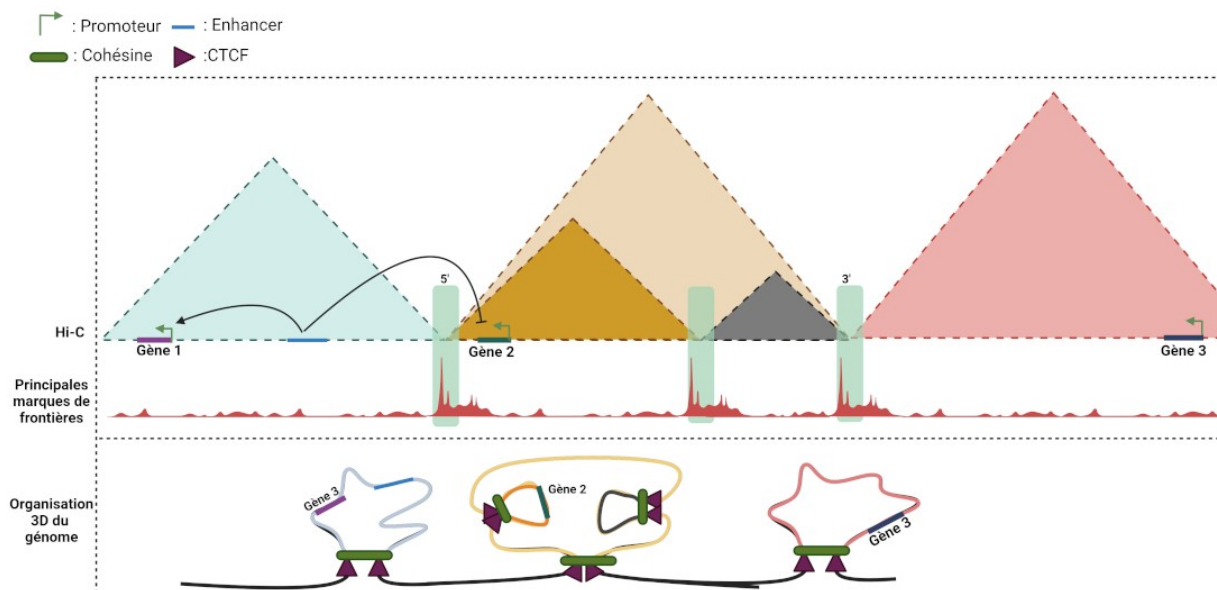


Figure 12 : Organisation tridimensionnelle de la chromatine et délimitation des TADs. Représentation des TADs identifiés par Hi-C (panel supérieur), son organisation tridimensionnelle (panel inférieur) et localisations des marques spécifiques insulinatrices (panel central).

2.3.3.2.2. Impact

Les variants de structure peuvent également perturber l'architecture des TADs, modifiant ainsi l'environnement transcriptionnel des gènes.(61)

Lorsqu'une duplication chevauche deux TADs distincts, elle peut conduire à la formation d'un néo-TAD, résultant de la fusion partielle des deux domaines initiaux.(Figure 13) Cette réorganisation peut entraîner l'établissement de nouvelles interactions entre des *enhancers* et des promoteurs appartenant initialement à des TADs distincts. Deux scénarios sont envisageables :

- **Adoption d'un enhancer** : si la duplication inclut un gène issu du premier TAD et un enhancer provenant du second, ce dernier peut être recruté de manière aberrante par le promoteur du gène du premier TAD. (62) On parle d'hijacking d'enhancer.
- **Absence d'impact transcriptionnel** : si la duplication ne contient pas l'ensemble des éléments fonctionnels nécessaires (par exemple, un gène sans un enhancer ou inversement), aucune modification de l'expression génique n'est observée.

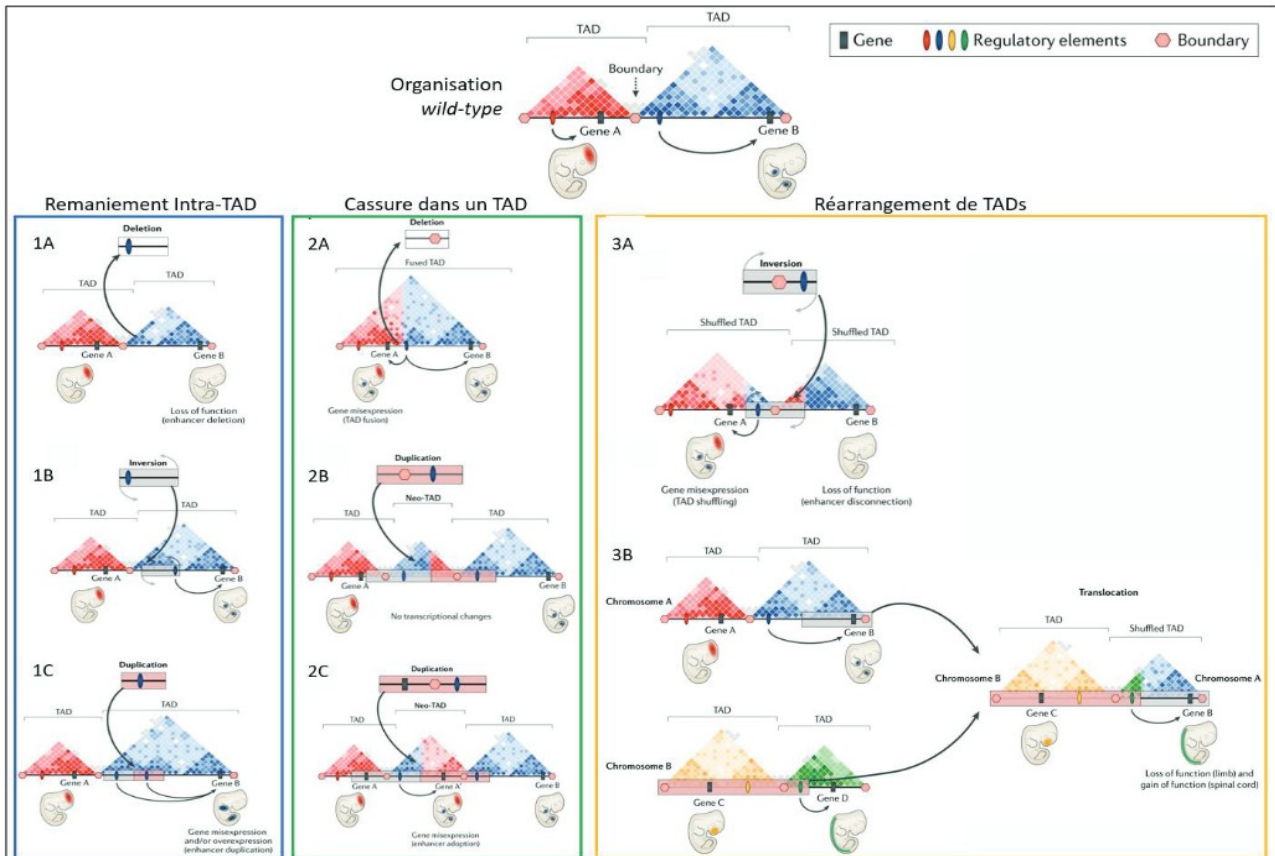


Figure 13: Synthèse des différentes possibilités d'effet de position en lien avec l'organisation de la chromatine en TADs (Figure adaptée de Spielmann et al., 2018 et extraite de N.Chatron, 2020)

De plus, la délétion d'une frontière de TAD peut fusionner deux domaines adjacents, favorisant des contacts anormaux entre un enhancer et un promoteur appartenant à des TADs distincts. (Figure 13) Ce mécanisme a été mis en évidence avec la délétion de la frontière située en 5' du TAD contenant *LMNB1*, entraînant une leucodystrophie démyélinisante de l'adulte (MIM #621061). (63) Enfin, un remodelage des TADs peut également survenir dans le cas de variants de structure équilibrés, tels que les translocations et les inversions ou de variants plus complexes. Ce phénomène, appelé "TAD shuffling", correspond à une réorganisation des

frontières de TADs, entraînant la fusion de domaines.(Figure 13) Cette réorganisation peut conduire à une adoption aberrante *d'enhancers* par des promoteurs inappropriés, modifiant ainsi l'expression des gènes.

Un cas de translocation entraînant une partie du TAD du gène *MEF2C* a été observé, causant une perte de contact avec ses *enhancers* naturels et par conséquent une perte de fonction causant une hypoplasie du corps calleux.(61) Un second cas documenté impliquant une duplication insertionnelle modifiant l'architecture chromatinienne tridimensionnelle au *locus ARH-GAP36* a été publié, impactant l'expression des voies de signalisation osseuse et entraînant une calcification ectopique progressive, où les tissus mous du patient se transforment en os. (81)

Les variants de structure, ont 2 voies d'étude vis à vis de leur impact : l'impact sur les séquences codantes et les gènes et l'impact sur l'environnement et les régions non codantes. Un schéma d'interprétation dans l'étude de l'impact des variants de structure peut donc être établie afin d'étudier les différentes possibilités associés. (Figure 14)

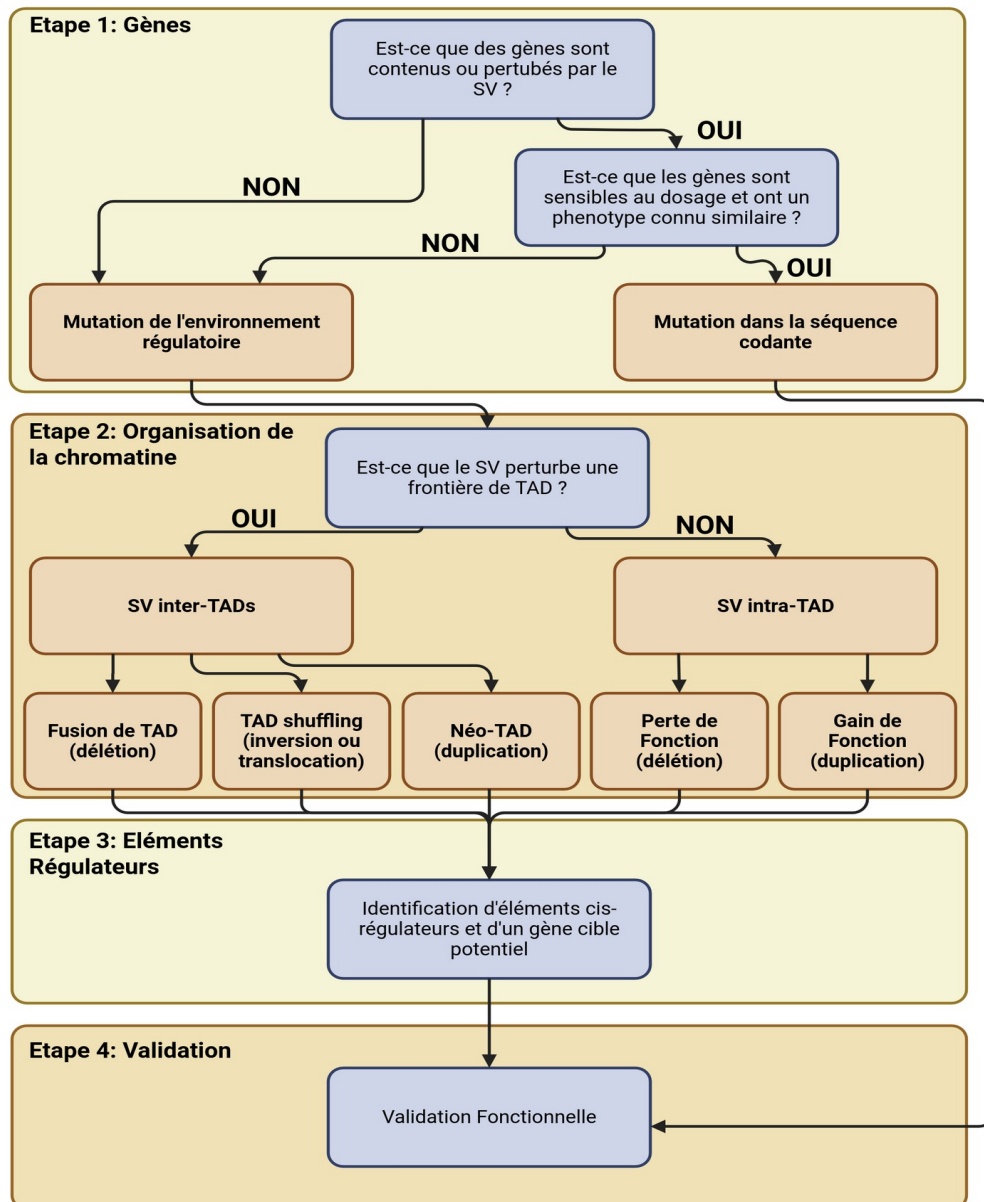


Figure 14 : Stratégies pour l'interprétation des variants de structure.(Figure adaptée de Spielmann et al., 2018)

3. Détection des SV

De nombreuses pathologies sont causées par des variants, et leur détection est cruciale pour établir un diagnostic précis. La découverte de ces variants et le diagnostic des patients reposent sur un processus diagnostique structuré, qui inclut plusieurs techniques.

3.1. Caryotype et FISH

Le caryotype humain a été décrit pour la première fois en 1956 par Tjio et Levan, marquant une avancée majeure dans le domaine de la cytogénétique.(65) En effet, depuis les 33 années

précédentes, au moins 40 études avaient confirmés la présence de 48 chromosomes alors que le nombre correct s'est révélé être 46. Cette correction a pu être obtenue par l'utilisation de la colchicine permettant l'accumulation de mitoses.(66)

Les techniques d'identification des chromosomes ont ensuite connu un essor considérable dans les années 1970, grâce au développement de méthodes de coloration différentielles permettant une analyse fine des structures chromosomiques. Ces méthodes reposent donc sur l'arrêt par la colchicine du cycle cellulaire en métaphase, étape où les chromosomes sont les plus condensés, suivi d'un traitement des cellules avec une solution hypotonique permettant la propagation des chromosomes puis d'une coloration sélective mettant en évidence l'euchromatine et l'hétérochromatine.(67)

La première méthode de banding, introduite en 1970 par Torbjörn Caspersson, utilise un colorant fluorescent, la quinacrine, pour révéler des motifs spécifiques sur les bandes cytogénétiques, une technique connue sous le nom de Q-banding.(68,69) Peu après, le G-banding a été mis au point par Drets et Shaw en 1971, reposant sur un traitement protéolytique par trypsine, dégradant certaines protéines chromosomiques, suivi d'une coloration au Giemsa.(Figure 15) (70) Cette méthode met en évidence des bandes sombres dans les régions riches en bases adénine-thymine (AT), tandis que les régions euchromatiques riches en guanine-cytosine (GC) restent peu colorées.

En parallèle, le R-banding a été développé par Verma et Lubs, également basé sur une coloration au Giemsa, mais avec un prétraitement thermique des chromosomes.(71) Ce chauffage déstabilise préférentiellement les hélices riches en AT, permettant ainsi de colorer spécifiquement les régions riches en GC.(Figure 15)(72) Par conséquent, les techniques de G-banding et de R-banding produisent des patrons de coloration inversés, offrant une complémentarité précieuse dans l'analyse des structures chromosomiques.(67,73,74)

Ces techniques de coloration permettent d'analyser avec précision les bandes cytogénétiques et, par conséquent, d'identifier d'éventuels remaniements chromosomiques. Le caryotype offre ainsi la possibilité de visualiser divers réarrangements structuraux, tels que les translocations interchromosomiques, les réarrangements intrachromosomiques (très grands variants en nombre de copies) ainsi que des anomalies complexes comme la chromothripsie et la chro-

moanasynthèse. Cette technique est considérée comme un examen cytogénétique de référence et même de première intention jusqu'à dans les années 2000.

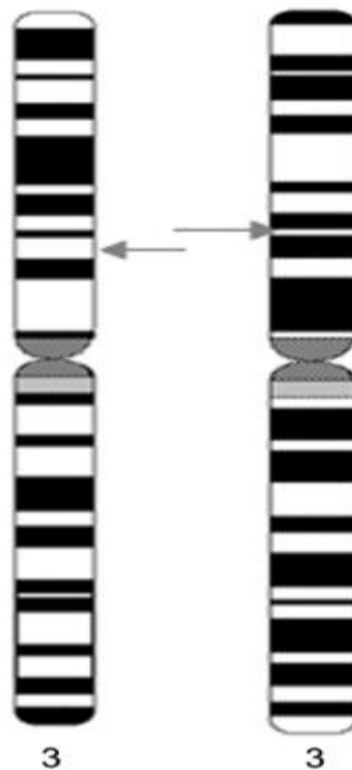


Figure 15: Chromosome Banding du chromosome 3: *G-banding associée aux régions riches en bases AT (chromosome de droite) et R-banding associée aux régions riches en bases GC (chromosome de gauche) (Figure extraite de Debatisse et al., 2006)*

Cependant, cette méthode présente certaines limites, notamment en termes de résolution. En effet, seuls les variants du nombre de copies d'une taille minimale de 5 à 10 Mb peuvent être détectés, et il est estimé qu'à peine 70 % des CNVs de 5 Mb sont effectivement identifiés. De plus, bien que cette technique permet de repérer la présence d'anomalies chromosomiques, elle ne fournit pas toujours d'informations précises sur les points de cassure ou les caractéristiques exactes des réarrangements. Cette imprécision complique l'interprétation des impacts fonctionnels de ces anomalies sur le génome, ainsi que sur les gènes et les voies biologiques qu'elles affectent.(74)

Puis avec le développement de sondes radioactives à la fin des années 60, la cytogénétique passe d'une analyse purement microscopique à une analyse plus précise de l'ADN. La technique d'hybridation *in situ* en fluorescence (*fluorescence in situ hybridization, FISH*), observée pour la première fois avec Pardue et Gall en 1969, repose aujourd'hui sur l'utilisation de sondes marquées par fluorescence qui s'hybrident spécifiquement à des séquences complémentaires d'ADN ou d'Acide Ribonucléique (ARN).(Figure 16)(75) Cette méthode permet de visualiser des régions chromosomiques précises avec une résolution d'environ 100 kb.(76)

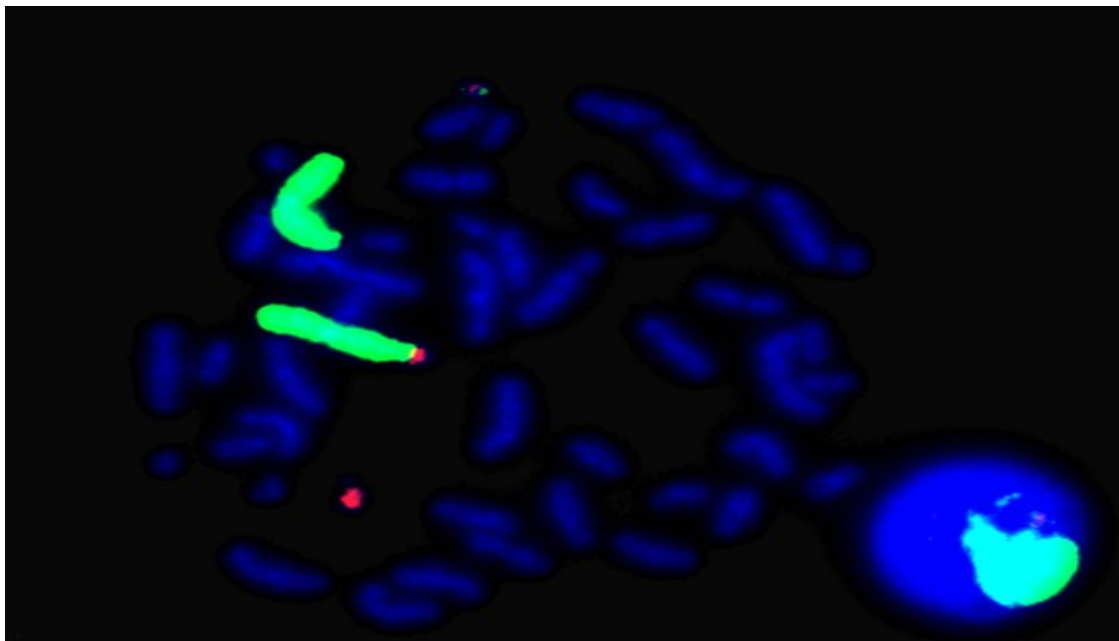


Figure 16: Hybridation In-Situ en Fluorescence démontrant une translocation t(1,22). (Figure extraite de Fauqueux et al.).

Contrairement au caryotype, qui offre une vue d'ensemble des chromosomes, la FISH cible des zones spécifiques, ce qui la rend particulièrement utile pour des analyses focalisées. Toutefois, son efficacité repose sur la connaissance préalable de l'anomalie à investiguer ou sur une forte suspicion, et nécessite la disponibilité de sondes adaptées. Cette technique va donc principalement pouvoir détecter les translocations interchromosomiques, les réarrangements intrachromosomiques (variants en nombres de copies) et anomalies plus complexes. Cependant, cette technique ne permet pas de détecter de petits réarrangements ni certaines duplications en tandem.(67,74,76)

L'un des principaux atouts de la FISH par rapport au caryotype est sa capacité à localiser précisément les *loci* impliqués dans des réarrangements chromosomiques, offrant ainsi des informations pour la caractérisation fine de ces anomalies.

3.2. Analyse chromosomique par puce ADN (CGH-Array/ACPA) :

Développée en 1992 par Kallioniemi *et al*, l'analyse chromosomique par puce ADN (ACPA ou aCGH; *array Comparative Genomic Hybridization*) est une technique avancée permettant de détecter des anomalies chromosomiques de nombre, souvent invisibles par caryotype ou FISH.(77) Cette méthode repose sur le mélange en quantités égales d'ADN isolé à partir d'un échantillon de référence et d'un échantillon de test, chacun étant marqué différemment : l'un avec de la biotine et l'autre avec de la digoxigénine. Ces ADN sont ensuite hybridés sur une puce contenant des fragments d'ADN fixés, correspondant à des régions spécifiques du génome, permettant ainsi la détection d'éventuels variants génomiques.(Figure 17)(67,74,76)

Les CNVs sont détectées en analysant les différences d'intensité de fluorescence entre les deux échantillons. Une surreprésentation ou une sous-représentation des séquences d'ADN dans l'échantillon de test reflète respectivement un gain ou une perte de copie. Grâce à une résolution comprise entre 25 et 200 kb, cette technologie permet de repérer des CNVs avec une sensibilité bien supérieure à celle du caryotype ou de la FISH.(67)

L'analyse par puce ADN est rapidement devenue le test de première intention dans le cadre clinique pour la détection de CNVs, en particulier chez les patients présentant des DI, des retards de développement ou des anomalies congénitales multiples. Son utilisation a significativement amélioré le rendement diagnostique, passant d'une détection de 3-5 % des cas à 15-20 %.(74)

Cependant, malgré ses avantages, l'analyse par puce ADN présente certaines limites. La localisation des points de cassure reste approximative, car elle dépend des sondes utilisées, et le positionnement précis des nouvelles copies n'est pas détectable. De plus, cette méthode ne permet pas de détecter efficacement les réarrangements chromosomiques équilibrés, tels que les inversions et les translocations, ni le mosaïcisme inférieur à 10-15 %, qui peuvent poser des défis supplémentaires pour le diagnostic.

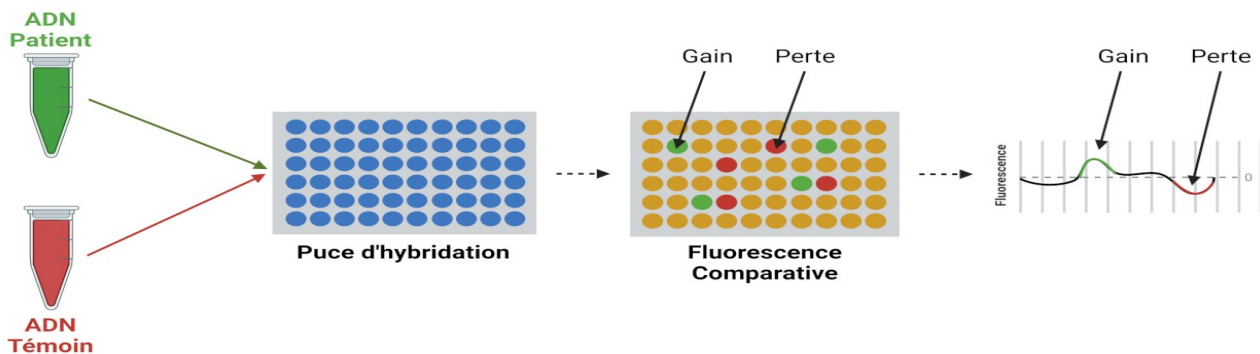


Figure 17: CGH-Array: Principe de la technique

3.3. Séquençage de nouvelles générations

Le séquençage de nouvelle génération, développé dans les années 2000, a révolutionné l'analyse génétique en permettant de séquencer l'intégralité du génome humain de manière rapide, précise et à grande échelle. Cette méthode repose sur la fragmentation de l'ADN en segments de 100 à 300 paires de bases, obtenus par des procédés mécaniques ou enzymatiques.(78,82) La terminologie de « nouvelle génération » largement utilisée depuis 2009, correspond au séquençage de deuxième génération actuel.

3.3.1. Méthode

La librairie préalablement préparée est chargée sur une matrice de séquençage appelée flow-cell. Parmi les technologies disponibles, Illumina est la plus couramment utilisée et sera détaillée ici.(79) Le séquençage Illumina, aussi appelé séquençage par synthèse, repose sur une technique appelée « amplification en pont », brevet publié en 1998 (WO1998044151A1). Les fragments d'ADN, dotés d'adaptateurs ligaturés à leurs extrémités, s'hybrident avec des oligonucléotides fixés sur la lame de verre (*flowcell*). Par cycles répétés d'amplification, chaque fragment donne naissance à des « clusters » clonaux composés d'environ 1 000 copies iden-

tiques. Grâce à leur disposition dense sur la lame, la *flowcell* permet de réaliser des millions de réactions parallèles.(Figure 18)(84)

Le séquençage se déroule en plusieurs cycles au cours desquels des nucléotides modifiés, spécifiques à chacune des quatre bases, sont incorporés dans les fragments en cours de synthèse. Chaque nucléotide est marqué par un fluorophore unique et agit comme un terminateur temporaire de synthèse. Une fois la fluorescence détectée, les terminateurs sont chimiquement débloqués, permettant le cycle suivant d'incorporation. Ce processus se répète sur plusieurs centaines de cycles, classiquement allant de 75 à 150, générant des lectures correspondant à chaque fragment d'ADN.(80)

L'utilisation de la détection par fluorescence permet une analyse rapide et précise grâce à l'imagerie directe, réalisée par des caméras. Ces fragments sont ensuite alignés sur un génome de référence afin d'identifier différentes anomalies génétiques, allant des grands variants de structure aux variants ponctuels telles que les SNVs.

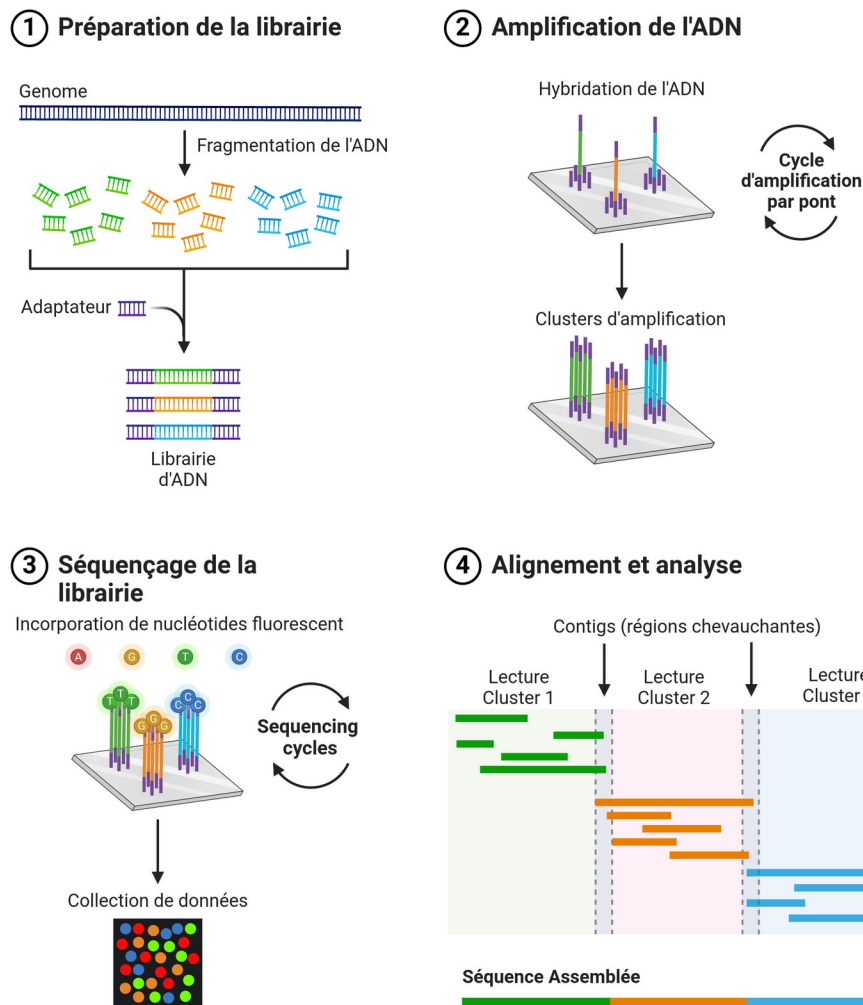


Figure 18: Principe du séquençage de Nouvelle Génération (NGS) par la technologie Illumina

Le NGS a considérablement accru le rendement diagnostique des maladies génétiques rares, en permettant la découverte de nouvelles mutations. Différentes approches sont possibles:(Figure 19) (74)

- **Panels ciblés de gènes** : Cette approche est utilisée lorsque des gènes spécifiques sont suspectés dans une pathologie. Le rendement diagnostique peut varier considérablement en fonction du trouble étudié. Par exemple, dans le cadre de la DI sporadique, ce rendement est d'environ 10 %.

- **Séquençage de l'exome** : Cette méthode permet de séquencer uniquement les régions codantes du génome. Le séquençage de l'exome a un rendement diagnostique de 35% dans le cadre des maladies rares mendéliennes, ce qui en fait une option puissante pour l'exploration génétique de ces pathologies.
- **Séquençage du génome complet** : Utilisé lorsque le séquençage de l'exome ne permet pas de détecter de variants, le séquençage du génome complet inclut l'analyse des régions non codantes du génome. Bien que plus coûteux, il permet d'identifier des anomalies supplémentaires avec un rendement additionnel de 3 à 5 % chez les patients avec un exome négatif. (81)

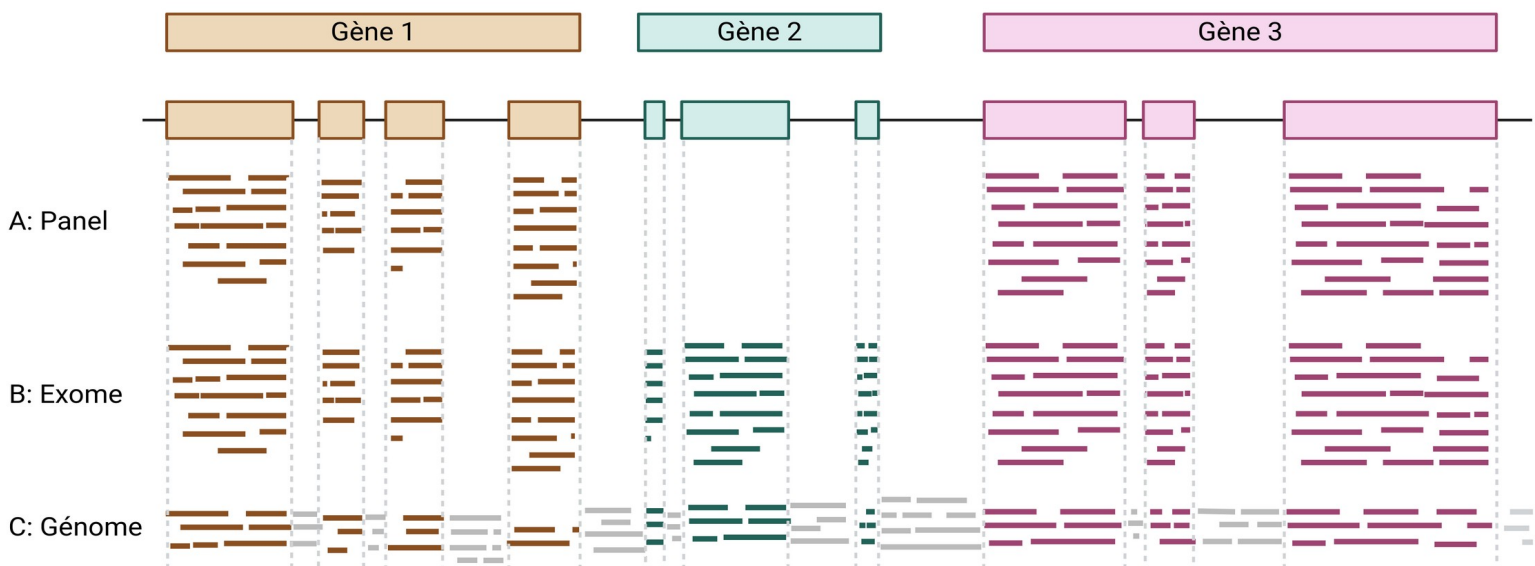


Figure 19: Alignement des fragments d'ADN du séquençage short-read dépendant du ciblage.

3.3.2. Détection des variants

Pour des réarrangements chromosomiques complexes, la détection dépend de la capture des points de cassure, ce qui est rare dans les panels ciblés ou même dans le séquençage d'exome.

Face au nombre croissant de résultats négatifs dans les séquençages d'exomes et aux rendements supplémentaires obtenus grâce aux séquençages de génome complet, les recommandations diagnostiques évoluent. Il est de plus en plus préconisé de privilégier le séquençage du génome complet comme première approche dans les analyses cliniques.(74)

Ces méthodes permettent de cartographier de manière globale les variants génétiques, telles que les SNPs et les variants en nombre de copies (CNV), notamment grâce à l'analyse de la profondeur de séquençage, qui est censé être homogène tout au long du génome.(Figure 19) (82) En effet, lorsqu'une région présente une couverture anormalement élevée ou basse par rapport au reste du génome, et qu'un multiplicateur significatif est observé entre les différentes régions, cela peut indiquer la présence d'un CNV.

De plus, le séquençage *paired-end* offre une opportunité supplémentaire d'analyser la taille d'insert ainsi que l'orientation des *reads* pairés. Chaque fragment d'ADN est séquencé aux deux extrémités, générant un *read* 1 (R1) qui lit la partie 5' et un *read* 2 (R2) qui lit la partie 3'. Ces *reads* sont normalement séparés par une distance de taille d'insert connue et adoptent une orientation inversée l'un par rapport à l'autre.(83)

Lors de l'alignement sur le génome de référence, la taille d'insert est comparée à la taille attendue.(Figure 20) Une taille anormale peut révéler la présence d'un variant structurel :

- Taille d'insert plus grande que prévu → Possible délétion
- Taille d'insert plus petite que prévu → Possible duplication en tandem

L'analyse de l'orientation des *reads* permet également d'identifier d'autres types de variants de structure, tels que :

- Inversions et duplication en tandem → lorsque l'orientation des *reads* ne suit plus le schéma classique
- Translocations → lorsque les *reads* appariés s'alignent sur deux chromosomes différents

Enfin, l'identification de *reads* chimériques, construction aberrante de *reads*, s'alignent sur des régions génomiques différentes, constitue un fort indicateur de la présence de réarrangements, comme des translocations équilibrées ou déséquilibrées, ainsi que des réarrangements chromosomiques plus complexes.(Figure 20) En théorie, chaque type de SV produit un motif caractéristique dans les lectures de séquençage, permettant en principe d'identifier la mutation sous-jacente. Cependant, en pratique, leur détection reste un défi en raison de plusieurs contraintes techniques inhérentes au NGS.

De plus, la forte présence de régions répétées ou hautement homologues constitue un défi supplémentaire. En effet, 52,5 % du génome humain est composé de séquences répétées.(84) Lorsque les points de cassure des variants se situent dans ces régions, ils deviennent particulièrement difficiles à résoudre avec les approches de séquençage standard, car l'alignement des *reads* y est complexe. Ces limitations peuvent entraîner plusieurs conséquences :

- Une mauvaise assignation des *reads*, compromettant ainsi la fiabilité de l'identification des SVs.
- Une sous-estimation de la complexité des réarrangements, certaines mutations restant totalement invisibles lorsqu'elles sont masquées par des régions répétitives mal assemblées.
- Des faux-positifs

Cette contrainte explique également pourquoi les duplications ne sont pas toujours entièrement caractérisées. Si le gain de copies peut être détecté, il demeure souvent impossible de déterminer précisément où la nouvelle copie est réinsérée dans le génome. Or, cette information est essentielle, car la réinsertion peut perturber des gènes adjacents ou des éléments régulateurs. Par conséquent, de nombreux variants identifiés par NGS sont classés comme variants de signification inconnue (VSI), compliquant leur interprétation clinique.

Par ailleurs, le NGS repose sur une étape d'amplification par PCR, susceptible d'introduire des biais. En effet, la PCR tend à amplifier préférentiellement certaines séquences selon leur teneur en GC. Toutefois, les régions extrêmement riches en GC posent souvent problème en raison de la formation de structures secondaires stables pouvant bloquer la polymérase, ce qui les rend sous-représentées et peut fausser l'évaluation de la profondeur de séquençage. (85) Une étude menée par Ebbert et al. a analysé dix génomes provenant de séquençage par synthèse avec une couverture moyenne de 37X et a révélé que plus de 36 000 régions (intergéniques et exoniques) étaient insuffisamment, voire pas du tout, séquencées. De plus, une liste de 76 gènes d'intérêt clinique a été établie, révélant des régions/exons difficiles à analyser avec cette technologie.(86)

Ces limites expliquent pourquoi les technologies d'analyses de lectures longues tel que le séquençage *long-read* (PacBio et Oxford Nanopore Technologie (ONT)) et l'*Optical Genome*

Mapping (Bionano), gagnent en popularité dans l'étude des variants de structure. Ces approches permettent en effet une meilleure résolution des points de cassure, en particulier dans les régions complexes du génome.

3.4. Cartographie Optique

La cartographie optique du génome est une approche pour analyser les grands génomes eucaryotes et est utilisé dans le génome humain dans le cadre de détection de variants complexes. Il consiste en l'analyse de brins d'ADN de haut poids moléculaire, de taille de plusieurs milliers de paires de bases atteignant une lecture moyenne de 200kb. L'ADN est préparé et étiqueté tel un code barre via un marquage avec un fluorophore par la méthyltransférase DLE-1 ayant un motif de reconnaissance de CTTAAG. Ces molécules sont ensuite alignés aux molécules de références via l'alignement des codes barres et permet d'analyser les variants de structures.(87)

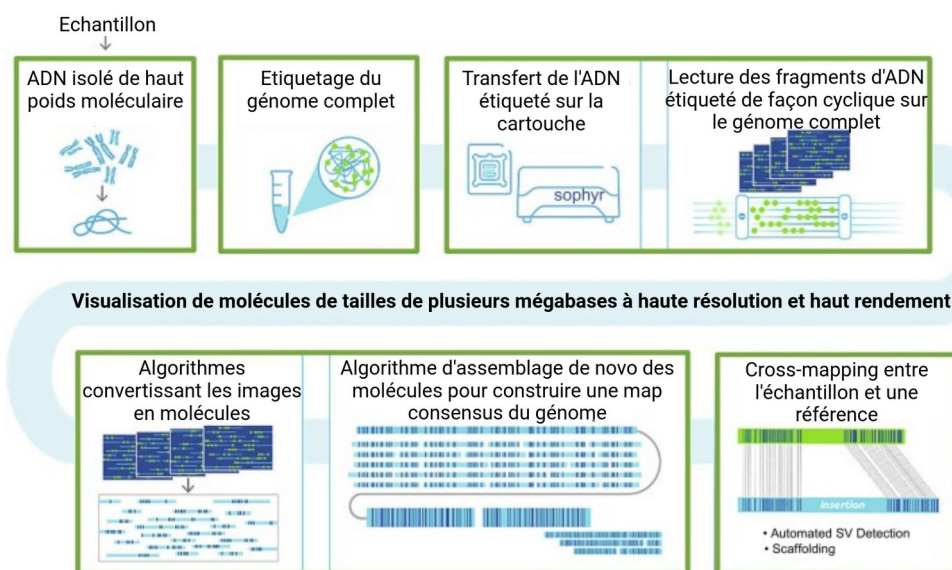


Figure 21: Principe de la Cartographie Optique du Génome: De la partie Wet-lab à la partie Bio-informatique (Figure adaptée de Delpu et al., 2021)

L'utilisation de longues molécules permet aux régions répétitives et à d'autres régions compliquées de cartographier plus facilement qu'avec des molécules courtes. Cela conduit à la création de cartes qui peuvent couvrir l'ensemble d'un bras d'un chromosome et permettant donc la détection d'insertions et de délétions aussi petites que 500pb, créant un décalage entre

les code barres. Les autres variants de structures doivent être tout de même de 30kb ou plus pour être détectable car ils sont dépendent de la présence des marques.(88)

Cependant, cette technique d'étude ne consiste pas en un séquençage et ne permet pas l'étude précise des variants de structure et de leur point de cassure.

3.5. Séquençage de troisième génération

Le séquençage de troisième génération, aussi appelé séquençage *long-read* (LRS), se distingue par sa capacité à séquencer des fragments d'ADN de grande taille (jusqu'à 1 Mb) en temps réel. Deux techniques principales sont utilisées:

- Séquençage en temps réel d'une seule molécule (SMRT) par PacBio.
- Séquençage Nanopore par Oxford Nanopore Technologies.

En séquençant des segments plus longs d'ADN, ces techniques peuvent capturer des informations manquantes dans les méthodes de *short-read*. (89)

3.5.1. Séquençage PacBio

Le premier séquenceur PacBio Science a été commercialisé en 2011. La technologie permet de séquencer des fragments de longueurs allant à 80kb et ayant une moyenne de 10 à 15kb. Les adaptateurs en forme d'épingles à cheveux se lient sur les fragments d'ADN double brin, permettant leur circularisation, ils sont appelés les *SMRT bells*.(90) Durant le séquençage, ces *SMRT bells* sont incorporés dans des chambres de séquençage de façon individuelle et l'ADN polymérase va être fixée sur ces adaptateurs et permettre de polymériser l'ADN complémentaire via incorporation de nucléotides fluorescents. Le signal fluorescent mais aussi les temps entre chaque incorporation vont permettre de faire la distinction entre chaque nucléotides incorporés mais aussi d'étudier les modifications épigénétiques.(Figure 22) (91)

Les fragments étant circulaires, ils peuvent être séquencés plusieurs fois afin d'augmenter la précision du séquençage.

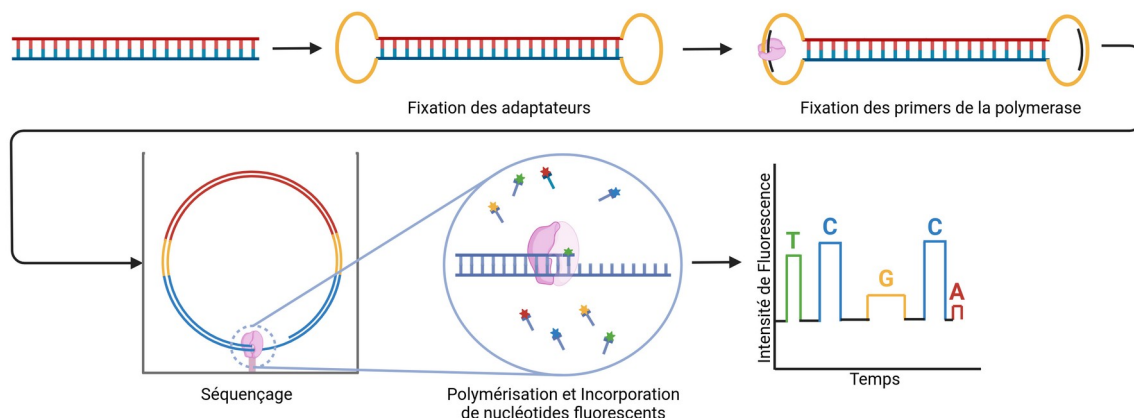


Figure 22: Principe du séquençage Long-Read par la technologie des SMRT bells de PacBio.

3.5.2. Oxford Nanopore Technologie

3.5.2.1. Origine du nanopore

La technologie des nanopores repose sur la détection de molécules individuelles en temps réel. Son origine remonte aux années 1950 avec le développement du compteur Coulter, un appareil conçu pour compter les globules rouges dans les fluides en détectant les variations de champ électrique induites par le passage de particules à travers un orifice microscopique.(92)

L'appareil développé par Wallace Coulter est constitué d'une sonde immergée dans un liquide à analyser, où deux électrodes génèrent un courant électrique.(93) Une paroi en verre perforée sépare ces électrodes, et l'aspiration du liquide à travers l'orifice permet aux particules de circuler une à une. Lorsqu'une molécule traverse l'orifice, elle modifie localement le champ électrique, entraînant une variation mesurable du courant. Cette variation, proportionnelle au volume de la particule, permet non seulement de la détecter, mais aussi de déterminer sa taille. L'appareil peut être calibré pour ne détecter que les particules dépassant un certain seuil, grâce à des solutions étalons contenant des particules de taille connue (grains de latex, pollen, spores...).(92,94) Pour être détectées, les molécules doivent traverser l'orifice une à une. La taille de cet orifice est donc adaptée en fonction du type de molécule analysé. Il doit être suffisamment petit pour permettre une obstruction partielle du courant électrique lors du passage d'une molécule, mais aussi assez large pour laisser celle-ci transiter sans entrave.

Initialement conçu pour l'hématologie, le compteur Coulter a vu son application élargie grâce à la réduction de la taille de l'orifice, passant du micron au nanomètre. Cette avancée a ouvert

la voie à la détection de molécules bien plus petites, comme l'ADN, l'ARN, les protéines ou encore certains métabolites. Toutefois, pour atteindre cette échelle nanométrique, il était nécessaire de concevoir un pore adapté, ce qui a conduit au développement des nanopores. (92,94)

Le concept des nanopores a été proposé par le professeur David Deamer, qui a envisagé qu'un canal protéique inséré dans une membrane lipidique pourrait être utilisé pour analyser les acides nucléiques. (Figure 23) (95,96) Peu après, il a démontré qu'un pore biologique pouvait remplacer l'orifice de Coulter. En particulier, il a montré que l' α -hémolysine, une toxine sécrétée par *Staphylococcus aureus*, pouvait perforer une bicouche lipidique et ainsi former des nanopores capables de laisser passer des molécules d'ADN et d'ARN.

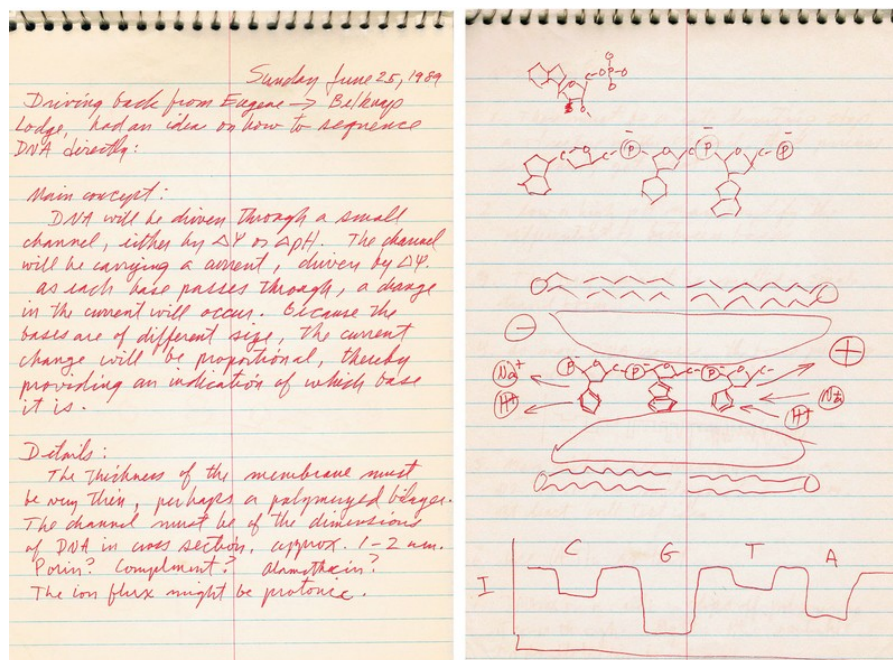


Figure 23: Carnet de David Deamer (1989): Deux pages dans lesquelles un premier draft du concept du séquençage Nanopore est visualisé. Le dessin présente 3 bases d'ADN simple brin transloquées au travers d'un nanopore via un voltage donnant un read-out électrique spécifique aux bases. (Figure extraite de Deamer et al., 2016)

En 1996, Branton et Deamer ont prouvé que ces nanopores permettaient non seulement la translocation des acides nucléiques, mais aussi leur détection et la mesure de leur longueur. (97,98)

En 2001, il a été démontré que les nanopores pouvaient identifier des brins d'ADN avec une résolution à l'échelle d'une seule base, distinguant des brins de 30 nucléotides différant d'une seule base.(99) Puis, en 2009, une avancée majeure a permis de différencier de manière fiable les quatre bases nucléotidiques.(100) Ces progrès ont été rendus possibles grâce à des modifications spécifiques des nanopores et à une analyse fine des variants du champ électrique induites par le passage des molécules.

Ainsi, la technologie des nanopores repose sur un principe simple mais puissant : détecter et analyser des molécules individuelles avec une précision extrême. Cette capacité a conduit au développement du séquençage par nanopores.

3.5.2.2. Type de nanopore

Les nanopores sont classés en trois catégories : les nanopores biologiques, les nanopores synthétiques solides et les nanopores hybrides.

- Les pores biologiques :

Les pores biologiques proviennent de protéines naturelles. Ces pores présentent des caractéristiques spécifiques, telles qu'une durée de vie relativement courte, une instabilité et une exigence d'un environnement particulier. Ils sont principalement produits par des bactéries, l' α -hémolysine étant l'un des exemples les plus connus. Cette protéine est thermiquement stable, fonctionne à des températures proches de 100°C et dans une plage de pH allant de 2 à 12. Cependant, sa région de sensibilité est relativement longue, atteignant jusqu'à 12 bases, ce qui peut influencer le passage des ions et affecter la précision de la détection.(Figure 24.A)(101)

Le MspA, produit par des mycobactéries, est un pore robuste capable de fonctionner à des températures avoisinant 80°C et dans une plage de pH allant de 0 à 14. Ce pore présente un intérêt particulier en raison de sa région de sensibilité plus réduite, ce qui améliore la précision dans la reconnaissance des quatre bases nucléotidiques. Contrairement à l' α -hémolysine, il permet une lecture plus fine des signaux électriques générés par le passage de l'ADN.(Figure 24.B)

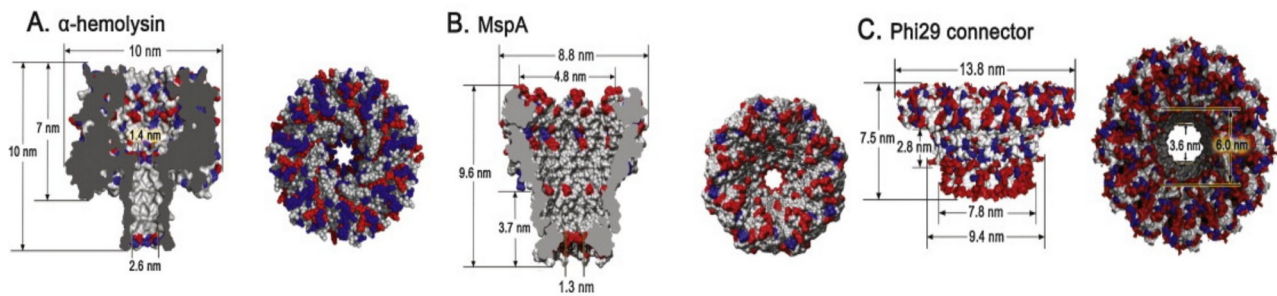


Figure 24: Structures de trois pores biologiques: (A) Toxine α -hémolysine de *Staphylococcus aureus*. (B) Porine MspA de *Mycobacterium smegmatis* (C) Canal connecteur du bactériophage Phi29 (Figure extraite de Haque et al., 2013)

Ces deux canaux, en raison de leur diamètre étroit, ne permettent que le passage de l'ADN simple brin.(101)

La protéine Phi29, capable de s'insérer dans une bicouche lipidique par un mécanisme de fusion, se distingue par sa grande robustesse. Son canal, plus large que ceux de l' α -hémolysine et du MspA, permet non seulement le passage de l'ADN simple brin, mais aussi celui de l'ADN double brin.(Figure 24.C)

- Les pores synthétiques à l'état solide :

Les pores synthétiques représentent une alternative intéressante aux pores biologiques en raison de leurs propriétés géométriques et dimensionnelles bien définies, de leur robustesse et de leur facilité de modification. Ces caractéristiques peuvent être précisément contrôlées, offrant ainsi une grande flexibilité dans leur conception et leur utilisation. Leur fabrication repose sur différentes techniques, telles que les réactions électrochimiques, la gravure laser ou encore la nanolithographie. Différents types de membranes peuvent être utilisées, notamment le silicium, l'aluminium, le bore ou encore le graphène. En raison de leur stabilité et de leur adaptabilité, ces pores sont les plus largement utilisés dans les applications modernes.(101)

- Les pores hybrides :

Les pores hybrides reposent sur l'insertion de protéines, telles que l' α -hémolysine, dans des membranes artificielles. Cette approche combine les avantages des pores biologiques, notamment leur capacité à interagir spécifiquement avec les molécules, avec la robustesse des membranes synthétiques. Elle permet ainsi d'obtenir des pores atomiquement stables tout en utili-

sant des supports moins fragiles que la bicouche lipidique, améliorant ainsi leur durabilité et leur compatibilité avec différentes conditions expérimentales.(101)

3.5.2.3. Principe

ONT propose une technologie de séquençage à molécule unique reposant sur l'utilisation de nanopores. Ce système fonctionne grâce à des cellules d'écoulement, appelées *flowcells*, constituées de deux compartiments remplis d'une solution ionique et séparés par une membrane contenant les nanopores. Le séquençage repose sur 3 étapes : la préparation de librairie, le séquençage, et le traitement bio-informatique.

Le processus débute par la préparation de la librairie de séquençage, qui repose sur l'utilisation d'ADN natif. Celui-ci subit d'abord une étape de réparation avant d'être lié à des adaptateurs spécifiques, eux-mêmes associés à des protéines motrices essentielles au bon déroulement du séquençage. Une fois la librairie prête, elle est chargée sur la *flowcell* via le pore d'entrée. Des tethers hydrophobes situés sur la *flowcell* capturent alors les molécules d'ADN et les dirigent vers la membrane où se trouvent les nanopores. Lorsque le complexe ADN-protéine motrice atteint un nanopore, la protéine se fixe et agit comme une hélicase, déroulant l'ADN double brin en un simple brin qu'elle transloque au travers du pore à une vitesse contrôlée.(Figure 25)(101)(103)

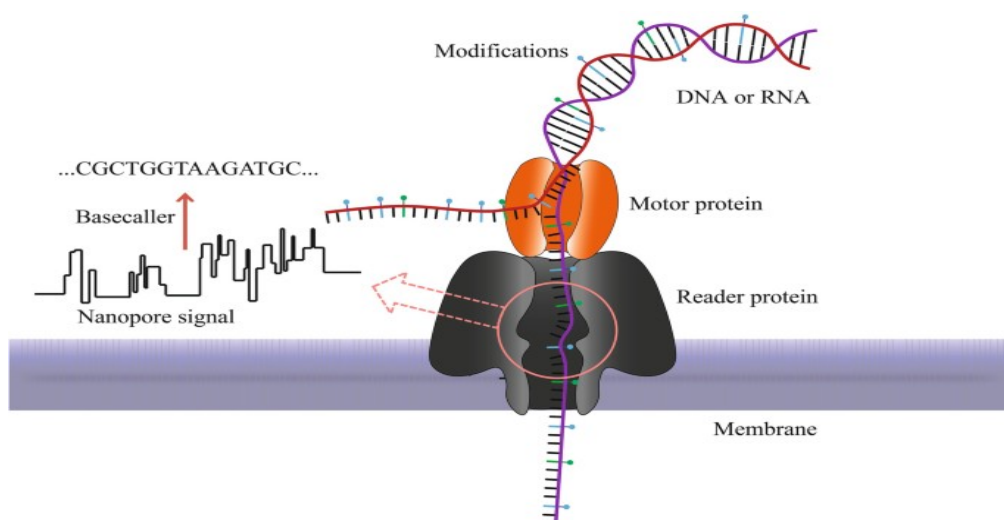


Figure 25: Principe du séquençage Long-Read par la technologie ONT: Schéma illustrant la translocation de l'ADN double-brin (ADNdb) en ADN simple-brin (ADNsb) à travers le nanopore, facilitée par le chargement du tether et représentation du principe de basecalling, permettant l'identification des bases nucléotidiques lors du passage de l'ADN à travers le nanopore. (Figures extraites de Ming He et al.)

Comme décrit précédemment avec le compteur Coulter, une tension constante est appliquée, générant un courant ionique à travers le nanopore. Les molécules d'ADN, chargées négativement, sont ainsi entraînées au travers du pore, ce qui perturbe le courant de manière caractéristique. Chaque modification du signal reflète une séquence nucléotidique spécifique et est enregistrée sous forme de fichiers bruts (FAST5 ou POD5). Ces données sont traitées en temps réel par des algorithmes bioinformatiques permettant de convertir les signaux en séquences nucléotidiques lisibles, générant ainsi des fichiers FASTQ. (103)

3.5.2.4. Spécificités

Grâce à l'utilisation d'ADN natif, cette technique de séquençage permet d'analyser des fragments d'ADN longs, avec une taille moyenne d'environ 10 kb et pouvant atteindre jusqu'à 1 Mb.(94) De plus, elle conserve les marques épigénétiques associées aux molécules d'ADN. Cette capacité offre un avantage considérable pour l'étude des modifications des bases, telles que les méthylations 5'-mC, 5'-hmC et 6mA, qui peuvent désormais être détectées et analysées avec une précision accrue, grâce à des algorithmes spécifiques.(104)

Cela permet non seulement de récupérer la séquence exacte des fragments d'ADN, mais aussi de collecter des informations sur les modifications épigénétiques associées. Ces dernières, jusqu'alors difficiles à détecter par des méthodes traditionnelles, peuvent maintenant être étudiées avec un degré de précision jamais atteint.(105)

De plus, comme mentionné précédemment, le séquençage *long-read* par technologie Nanopore présente la particularité d'être effectué en temps réel. Cette approche permet de suivre dynamiquement plusieurs paramètres, tels que la taille des *reads*, la qualité de séquençage ou encore l'activité des pores. L'un des développements majeurs rendus possibles par cette lecture en temps réel est l'*Adaptive Sampling*, une technique de ciblage bio-informatique en temps réel permettant l'enrichissement, ou la déplétion de régions spécifiques du génome. Ce ciblage repose sur l'implémentation d'un fichier BED, dans lequel sont définies les régions d'intérêt à enrichir ou à éviter.

Concrètement, lors de l'enrichissement, chaque fragment d'ADN est d'abord séquencé pendant environ une seconde (correspondant à ~400 pb). Cette séquence initiale est ensuite comparée aux régions cibles définies dans le fichier BED. Si le fragment appartient à l'une de ces régions, le séquençage se poursuit ; dans le cas contraire, la polarité du pore est inversée, in-

terrompant le séquençage et expulsant la molécule en cours. Elle offre ainsi la possibilité de réaliser un séquençage ciblé sans étapes d'enrichissement préalable, réduisant à la fois les coûts de séquençage et la taille des fichiers générés.(106–109)

Cependant, l'utilisation des nanopores présente encore certaines limitations qui sont activement prises en charge. En effet, la détection ne se fait pas base par base, mais par groupes de bases, ce qui peut entraîner des erreurs et expliquer que le séquençage Nanopore ait une précision d'environ 96.8 %, ce qui reste relativement faible comparé aux technologies NGS. (103,110)

Chaque nanopore possède des caractéristiques spécifiques, notamment des zones de détection de taille variable. Pour surmonter ces défis, plusieurs améliorations ont été apportées aux nanopores. Par exemple, la surface des membranes a été réduite pour accélérer la stabilisation du signal électrique, ce qui diminue le bruit de fond. De plus, la vitesse de translocation des molécules a été optimisée afin de garantir une mesure précise de chaque base. Enfin, les membranes ont été adaptées pour produire des segments de signal correspondant désormais à des blocs de 5 bases, permettant une détection plus fine et une distinction plus précise des bases individuelles. Ces améliorations contribuent à renforcer la précision globale des appels de bases et à rendre les nanopores encore plus performants pour l'analyse génomique.

ONT a continuellement amélioré ses nanopores, lançant régulièrement de nouvelles versions afin d'augmenter la précision et l'efficacité du séquençage. Ces versions incluent : R6 (juin 2014), R7 (juillet 2014), R7.3 (octobre 2014), R9 (mai 2016), R9.4 (octobre 2016), R10 (mars 2019) et R10.3 (janvier 2020).(103) L'objectif principal de ces améliorations est de perfectionner la précision des appels de bases.

Les protéines utilisées dans les nanopores des générations R6, R7, R7.3, R10 et R10.4 n'ont pas encore été divulguées par ONT. La version R9 a marqué un progrès important, avec une amélioration significative du rendement en séquençage atteignant environ 87 % de précision, contre seulement 64 % pour la génération R7. Cela a été rendu possible grâce à l'utilisation de la sous-unité de croissance dépendante de la Nanopore Curlin Sigma S (CSGG) d'*Escherichia coli*.

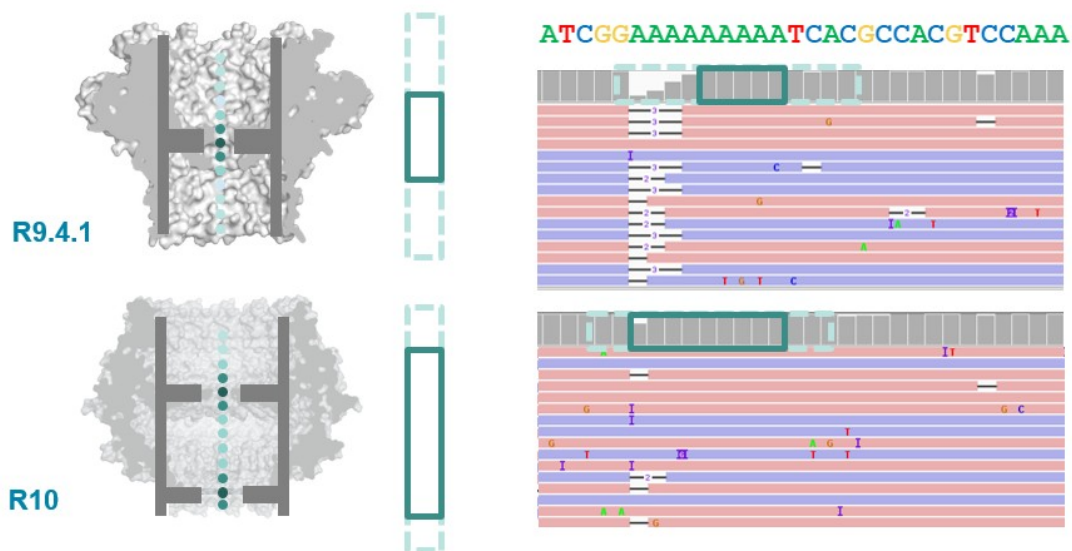


Figure 26: Comparaison structurelle des chimies de la technologie ONT (Figure extraite du site d'ONT)

La génération R10.3 (CsgG-FCP) a introduit un nanopore avec deux régions de sensibilité au lieu d'une seule, permettant une meilleure distinction des bases.(Figure 26) Plus récemment, la génération R10.4 a été lancée, bien que les détails de ses améliorations ne soient pas encore totalement divulgués. Cependant, ONT a annoncé que cette nouvelle version atteint désormais une précision d'appel des bases de 99 %.(111)

3.6. De la cytogénétique à la cytogénomique

L'évolution de la cytogénétique vers la cytogénomique a transformé notre façon d'explorer le génome. Autrefois limitée à l'observation des chromosomes au microscope, notre compréhension s'est affinée grâce aux technologies de séquençage et aux outils d'analyse avancés. Aujourd'hui, nous pouvons détecter des variants génétiques avec une précision à la base près, ouvrant la voie à des diagnostics moléculaires plus précis et à une meilleure prise en charge des maladies génétiques.

Cependant, chaque technique est associée à des spécificités de détection, à une résolution ainsi qu'à des limites. Les nouvelles technologies tendent donc à avoir une résolution à la base

unique, une précision de détection de 100 % mais aussi la capacité à détecter tous types de variants sans hypothèse préconçue.(Tableau 1)

Tableau 1: Comparaison des différentes techniques de cytogénétique

Technique	Coût	Nécessité de cellules ?	Résolution	Détection de CNVs ?	Détection de Variants Equilibrés ?	Mosaïcisme	Etude de la Séquence ?
Caryotype	+	Oui	5 à 10Mb	Oui	Oui	Non	Non
FISH	++	Oui	100kb	Oui	Uniquement les Translocations	Non	Non
ACPA/CGH-array	++	Non	25 et 200kb	Oui	Non	Oui mais élevé	Non
NGS	+++	Non	Résolution unique	Oui	Oui	Oui	Oui
LRS	++++	Non	Résolution unique	Oui	Oui	Oui	Oui
OGM	+++ /++++	Non	30kb	Oui	Oui	Non	Non

Ce qui nous importe principalement dans cette thèse est la détection et l'analyse des variants de structure à travers les différentes approches technologiques disponibles en routine diagnostique. La littérature récente tend à démontrer que les technologies émergentes permettent non seulement d'affiner le nombre réel de SVs détectés par patient, mais aussi de mettre en lumière les différences concrètes de performance entre ces approches. Dans ce contexte, les technologies de séquençage dites de troisième génération, en particulier le séquençage *long-read*, apparaissent comme particulièrement adaptées à l'étude fine des SVs. Leur capacité à générer des lectures longues et continues offre un avantage pour la résolution de variants complexes, souvent inaccessibles aux méthodes conventionnelles. Ces technologies se révèlent également pertinentes dans le cadre de réanalyses de génomes, notamment lorsqu'un SV potentiellement causal aurait pu être manqué par les outils de première intention. (Figure 27)

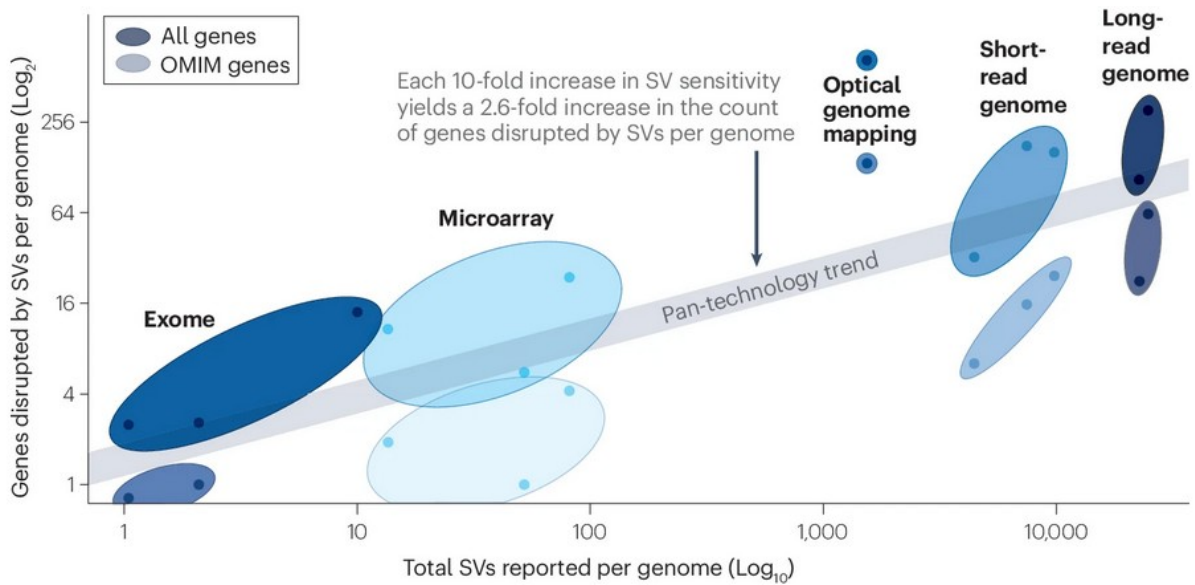


Figure 27: Rendement des analyses des variants de structure à travers les techniques d'étude du génome. (Figure extraite de Collins et al., 2025)

4. Utilisation du séquençage *long-read* pour l'étude des variants de structures

Comme expliqué précédemment, certains SVs présentent des points de cassure situés dans des régions difficiles à analyser par séquençage *short-read* (SRS ; NGS). Ces régions, souvent riches en séquences répétitives ou en structures complexes, peuvent donc masquer une partie de la complexité du génome, rendant leur détection et leur caractérisation difficiles, voire impossibles, avec les techniques traditionnelles. En conséquence, l'incapacité de capturer ces structures avec des lectures courtes peut entraîner une sous-estimation des événements génétiques responsables de certaines pathologies, en particulier dans des maladies comme le cancer ou les troubles neurodégénératifs, où les SVs jouent un rôle clé. En effet, depuis l'utilisation des nouvelles technologies de séquençage, la moyenne de 7400 Svs par génome a été revue à la hausse avec plus de 12000 SVs.(Figure 27)

Les lectures longues, issues de technologies comme PacBio ou ONT, se révèlent particulièrement avantageuses pour l'appel des SVs, car elles offrent la possibilité de couvrir des régions complexes et répétitives, mais également d'analyser des zones qui, autrement, seraient insuffisamment couvertes par le SRS. De cette manière, les lectures longues, notamment celles dépassant les 5 kb, permettent non seulement une meilleure cartographie des SVs, mais aussi la

détection de variants de structure de plus grande taille, souvent invisibles avec les lectures courtes. Grâce à cette approche, il devient possible de reconstruire plus fidèlement l'architecture du génome et d'améliorer ainsi la précision du diagnostic et la compréhension des mécanismes biologiques.

Stancu et ses collaborateurs, en 2017, ont proposé le LRS comme approche alternative pour la détection à l'échelle du génome des SVs cliniquement pertinents. Ils ont utilisé le séquençage du génome complet par la technologie ONT pour analyser deux patients présentant des anomalies congénitales causées par des réarrangements de *chromothripsis de novo*. Ils ont démontré que le LRS, même avec une faible couverture de séquençage (16X), a une qualité de détection supérieure par rapport au SRS.(112)

De plus, d'autres études ont aussi démontré les avantages potentiels du LRS par rapport au SRS dans le diagnostic clinique pour détecter les SVs pathogènes. Par exemple, une étude réalisée par Merker *et al*, a réussi à démontrer, via la technologie PacBio, la cause moléculaire d'un patient atteint du syndrome de Carney, alors que le séquençage du génome complet par SRS s'était révélé négatif. (113) De même, Miao *et al*. ont confirmé le diagnostic moléculaire d'un patient atteint d'un trouble de stockage du glycogène avec la technologie ONT, en identifiant un second variant dans un gène autosomique récessif.(114)

5. Interprétation des variants

En 2015, l'*American College of Medical Genetics and Genomics* (ACMG) et l'*Association For Molecular Pathology* (AMP) ont publié une directive qui fournit un cadre pour l'interprétation des variants de séquence. Cette directive est devenue une référence dans le domaine de la génétique clinique et a été adoptée par de nombreux laboratoires.

La directive a défini 28 critères, avec des codes qui traitaient des types de preuve. Chaque type de preuve s'est vu attribuer une classification bénigne (B) ou pathogène (P). Des règles de combinaison de ces critères ont également été prévues pour attribuer une pathogénicité pour chaque variant sur la base de critères appliqués.

Les lignes directrices ACMG/AMP ont été développées pour être largement applicables à de nombreux gènes, modèles de transmission et maladies. Cependant, certaines limites ont été

identifiées, notamment en ce qui concerne l'interprétation des variants dans les gènes présentant des effets de dosage ou dans les maladies multifactorielles.(115,116)

Le consortium ClinGen a été formé en 2013 pour développer des normes et des processus d'évaluation des gènes et des variants génomiques afin d'améliorer la validité clinique et l'utilité.

Plusieurs critères entrent en compte dans l'évaluation de la pathogénicité d'un variant :

- Les critères de fréquence dans la population : Les informations concernant la fréquence d'un variant dans la population peuvent être utilisées à la fois pour et contre la pathogénicité. Un variant survenant à une fréquence beaucoup plus élevée que prévu pour la maladie avec laquelle il est associé est un indicateur de sa nature bénigne. Il peut être considéré comme BA1, ce qui signifie qu'aucune évaluation supplémentaire n'est nécessaire. Par exemple, un variant trouvé à une fréquence de 5 % dans gnomAD chez des individus en bonne santé est très probablement bénin pour une maladie autosomique dominante rare dont la prévalence est inférieure à 1/10 000.
- Le type de variant et sa position : La connaissance du spectre des variantes dans un gène et de leur répartition selon la maladie associée est nécessaire. Le seul critère désigné avec un niveau de force très élevé pour la pathogénicité (PVS1) est défini comme une variant dans un gène où la perte de fonction est un mécanisme connu de la maladie. Par exemple, une mutation entraînant un codon stop prématuré dans *TP53*, un gène suppresseur de tumeur, a de fortes chances d'être pathogène. En revanche, un variant faux-sens dans une région non conservée d'un gène pourrait être considérée comme de signification inconnue.
- Les preuves fonctionnelles et familiales : La pathogénicité d'un variant peut être renforcée par des études fonctionnelles démontrant un effet délétère (critère PS3) ou par des études de ségrégation dans les familles. Par exemple, un variant observé chez plusieurs individus atteints dans une même famille et absent chez les individus non affectés apporte un argument supplémentaire en faveur de sa pathogénicité (critère PP1).

Grâce à ces critères, l'interprétation des variants gagne en précision, ce qui permet d'améliorer le diagnostic génétique et la prise en charge des patients. (Figure 28) Toutefois, l'évolution

des connaissances et l'accumulation de nouvelles données nécessitent des mises à jour régulières des classifications, comme en témoigne la réévaluation de certains variants initialement classés VSI qui sont désormais considérés comme pathogènes ou bénins.

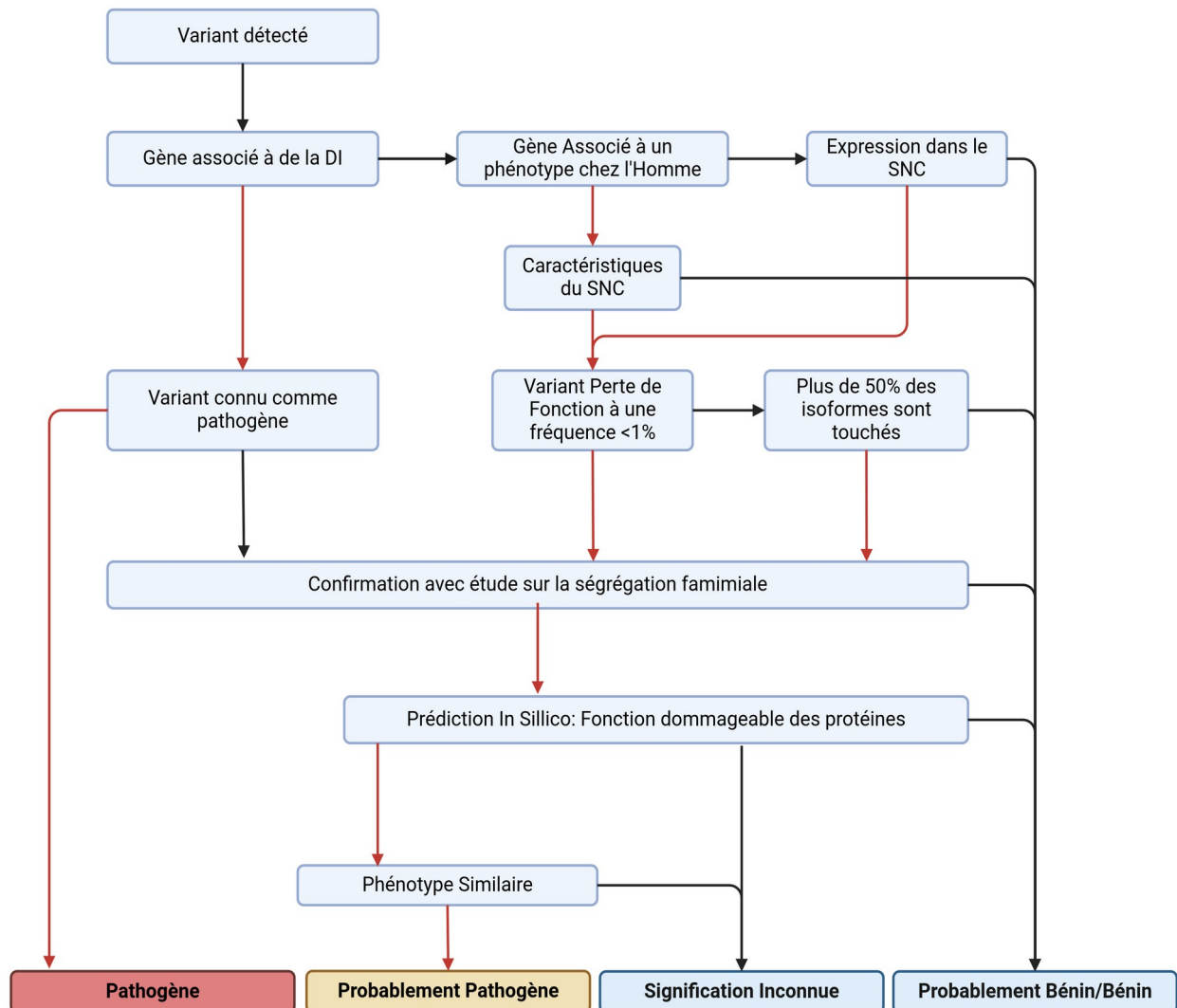


Figure 28: Stratégies pour la classification des variants. (flèches rouges : OUI ; flèche noire : NON) (Figure adaptée de Maia et al., 2021)

II/ Objectifs

La détection, la compréhension et l'analyse des variants de structure est un défi majeur en routine diagnostique moléculaire. L'approche par CGH-array en première intention est performante mais présente des limitations importantes concernant l'organisation réelle du génome lors de la détection des gains de copie, une résolution limitée et une non détection des variants équilibrés. Les méthodes plus résolutive, comme le séquençage *short-read*, sont limitées par la nature répétitive de l'ADN, qui complique l'interprétation de certains variants de structure, ainsi que par la complexité de réarrangements complexes. Ces défis, tant en détection qu'en caractérisation, nous amènent à nous interroger sur les apports des technologies émergentes, notamment le séquençage Nanopore. Nous cherchons donc à répondre aux questions suivantes :

- **La technologie Oxford Nanopore permet-elle une couverture suffisante de la région d'intérêt et du génome dans son ensemble ?**
- **Est-elle capable de caractériser structurellement les variants de structure, qu'ils soient simples ou complexes, et représente-t-elle une avancée significative par rapport aux technologies utilisées en routine ?**
- **Les algorithmes d'appel de variants de structure dédiés à cette technologie sont-ils suffisamment performants pour une détection fiable et précise, sans *a priori* ?**
- **Enfin, le séquençage Nanopore permet-il d'émettre des hypothèses fonctionnelles sur les variants de structure, d'en approfondir l'interprétation biologique et, potentiellement, d'en affiner la classification ?**

L'objectif principal de cette thèse est d'explorer les variants de structure, en mettant l'accent sur les gains de copies détectés par les technologies de diagnostic de routine chez les patients atteints de troubles neurodéveloppementaux. Pour ce faire, nous utilisons le séquençage *long-read* Nanopore pour évaluer ses performances, affiner la caractérisation de ces variants et analyser leurs implications diagnostiques.

L'enjeu est également de déterminer si cette technologie émergente peut améliorer le rendement diagnostique des patients atteints de troubles neurodéveloppementaux et si son adoption en routine est envisageable.(Figure 29)

III/ Matériel et Méthodes

1. Matériel

1.1. Matériel biologique

Différents types de matériel biologique ont été utilisés au cours de cette étude. L'ADN génomique a été extrait de manière automatisée à partir de sang périphérique, recueilli après obtention du consentement éclairé des patients. Certains ADN provenaient de patients externes, pour lesquels la méthode d'extraction et la méthode de conservation de l'ADN n'étaient pas identiques.

Des lignées lymphoblastiques immortalisées ont été établies à partir de sang périphérique collecté dans des tubes héparinés. L'ARN total a ensuite été extrait manuellement à partir de ces lignées cellulaires.

Cette étude a été conduite conformément aux principes éthiques de la Déclaration d'Helsinki et a reçu l'approbation du Comité de Protection des Personnes (CPP) « Île-de-France II ».

1.2. Matériels associés au séquençage Nanopore

1.2.1. Séquenceurs

Les technologies de séquençage *long-read* d'Oxford Nanopore Technologies (ONT) ont été employées à mesure de leur disponibilité au sein du Centre de Biologie Pathologie. Les échantillons ont ainsi été analysés sur trois types de plateformes distinctes :

- MinION Mk1C
- P2 Solo couplé à un GridION
- P2 Integrated (P2i)

1.2.2. Informatique et logiciels

L'analyse bioinformatique a évolué en parallèle des avancées technologiques du séquençage, s'adaptant notamment aux mises à jour successives des logiciels de *basecalling*. Les séquenceurs ont été pilotés par le logiciel MinKNOW, avec un *basecalling* réalisé initialement via

Guppy (v6.4.6), puis progressivement migré vers Dorado (v7.6.7), en fonction des versions disponibles au moment de l'analyse.

L'alignement des fichiers FASTQ au génome de référence a été effectué majoritairement avec Minimap2 (v2.24), outil rapide et optimisé pour le séquençage *long-read*.⁽¹¹⁷⁾ Pour certaines régions génomiques complexes ou particulièrement sujettes à des réarrangements, l'outil NGMLR (v0.2.7) a été préféré en raison de ses meilleures performances dans ce contexte.(Figure 29)

La qualité des données a été évaluée à l'aide de plusieurs outils complémentaires :

- Mosdepth
- Qualimap
- PycoQC

L'identification des variants de structure a reposé sur deux *pipelines* complémentaires :

- CuteSV (v2.1.1), optimisé pour les longues lectures avec de bonnes performances en sensibilité.⁽¹¹⁸⁾
- Sniffles2 (v2.0.7), particulièrement performant pour la détection dans des contextes à faible profondeur.⁽¹¹⁹⁾

Les fichiers VCFs générés servaient de base pour confirmer l'appel des variants précédemment identifiés par d'autres techniques utilisés en routine. Les fichiers VCFs n'ont pas été analysés en totalité ni filtrés.

L'analyse des données issues du séquençage Nanopore a été menée de manière collaborative :

- L'équipe de bioinformatique de la Plateforme de séquençage de l'Université de Lille a pris en charge le traitement initial des données brutes, en mobilisant une infrastructure performante, composée notamment de :
 - 2 GPU A6000

- 64 threads CPU
- 256 Go de RAM
- Une partie des analyses a été réalisée de manière autonome, sur une machine virtuelle dédiée, hébergée sur le cluster Gaia (Bi-Lille), disposant des ressources suivantes :
 - 12 vCPU
 - 90 Go de RAM

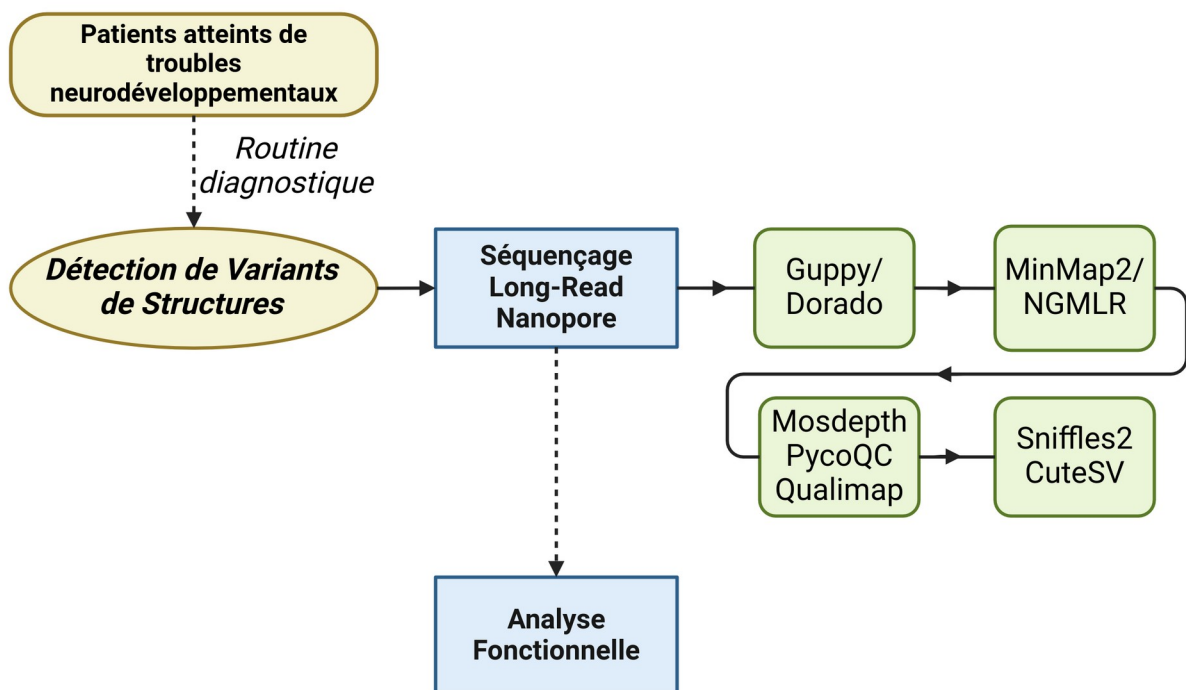


Figure 29: Protocole du projet de Thèse

2. Méthodes

Les principales méthodes employées tout au long de cette thèse sont présentées ci-dessous. Les techniques plus spécifiques ou utilisées de manière ponctuelle seront détaillées ultérieurement dans le contexte des thématiques concernées.

2.1. CGH-Array

Les échantillons de sang périphérique des patients ont été collectés dans des tubes EDTA. L'ADN génomique a été extrait à l'aide du système d'extraction robotisé CHEMAGIC-STAR (Hamilton, Reno, US-NV). Les expériences ont été réalisées sur de l'ADN génomique. Les puces Agilent Human Genome CGH 60K à oligonucléotides (Agilent, Santa Clara, CA, USA) ont été utilisées selon les protocoles du fabricant.

2.2. Librairie de séquençage ADN *Long-Read* Nanopore

Les bibliothèques d'ADN ont été préparées à l'aide du kit *Ligation Sequencing Kit V14* (Oxford Nanopore Technologies, ONT, Oxford, UK). Le séquençage a ensuite été réalisé selon le protocole standard sur un séquenceur d'ONT, utilisant une flow cell R10.4.1.

Après séquençage, le *basecalling* a été effectué à l'aide d'un *basecaller* en mode *High-Accuracy* (HAC).

Les lectures obtenues ont été alignées sur la dernière version du génome de référence humain GRCh38 à l'aide de MiniMap2 et/ou NGMLR.(117) Ces données alignées ont ensuite été utilisées pour les analyses en aval. La détection des variants structurales a été réalisée avec Sniffles2 et CuteSV.(118,119)

L'inspection visuelle des variants de structure dans les régions cibles a été effectuée à l'aide de logiciels tels que Integrative Genomics Viewer (IGV, version 2.10 à version 2.18) et Genome Ribbon (V1 et V2).(120)

2.3. Déroulement d'un *run* Nanopore sur ADN

Le déroulement des séquençages a été suivi en temps réel à l'aide du logiciel MinKNOW, qui fournit une visualisation dynamique des performances des *runs*. Les principaux indicateurs analysés étaient :

- Le nombre total de lectures (*reads*), réparties en :
 - lectures totales,

- lectures dites « pass » (Qscore > 9),
- lectures « fail » (Qscore < 9),
- lectures « skipped »
- La quantité totale de données générées (en gigabases, Gb)
- Le Qscore, indicateur de qualité des lectures
- Le N50, indicateur de la distribution de longueur des fragments
- L'état des pores : activité (en séquençage, disponibles, inactifs, non classés) et pore scan (répartition des statuts de fonctionnement)

Bien que d'autres paramètres tels que la température ou la vitesse de translocation soient accessibles, ils n'ont pas été retenus pour l'analyse, les *runs* s'étant déroulés dans des conditions satisfaisantes.

Les objectifs de couverture étaient de 30× pour le séquençage complet du génome (*Whole Genome Sequencing* ; WGS) (environ 100 Gb par échantillon) et de 10 Gb pour les expériences d'*adaptive sampling* ciblant 0,5 à 1 % du génome. Ces seuils sont théoriquement atteignables avec une *flowcell* unique par échantillon, un seul chargement de librairie et un *run* de 72 heures, sous réserve d'une qualité optimale de l'ADN et d'une préparation adéquate des librairies.

Un seuil minimal de couverture à 10× pour le WGS (soit environ 32 Gb) a été fixé pour inclure un échantillon dans les analyses, en accord avec les performances documentées du séquençage *long-read* à faible profondeur. La présence de *reads* chimériques a également constitué un critère déterminant de validation. Lorsque ces *reads* étaient détectés, aucun séquençage complémentaire n'a été jugée nécessaire, même si certaines métriques secondaires (qualité moyenne des lectures, homogénéité de la couverture) pouvaient présenter des variabilités.

2.3. Analyse qPCR ciblée

Une PCR quantitative a été réalisée à partir de l'ADN génomique extrait des patients, en utilisant le kit PowerUp SYBR Green (ThermoFisher Scientific, Waltham, MA, USA) et des amorces spécifiquement conçues pour quantifier le nombre de copies au niveau du *locus* variant. Le nombre de copies a été normalisé par rapport à un échantillon témoin traité dans les mêmes conditions que l'échantillon patient. Les niveaux relatifs d'amorces ont été calculés selon la méthode $2^{-\Delta\Delta Ct}$, avec normalisation sur un gène de référence. Une analyse statistique a été réalisée à l'aide de logiciels appropriés. Toutes les expériences ont été effectuées en triplicat.

2.4. Culture cellulaire

Les lignées cellulaires lymphoblastoïdes humaines (LCL) ont été établies par immortalisation des lymphocytes sanguins du sujet à l'aide du virus d'Epstein-Barr et maintenues dans un milieu RPMI 1640 (ThermoFisher Scientific) supplémenté avec 10 % de sérum bovin fœtal (FBS) (ThermoFisher Scientific) et 1 % de pénicilline-streptomycine (Life Technologies).

IV/ Résultats

1. Approche de séquençage et panorama des données

1.1. Metrics associé au séquençage

Au total, 15 patients ont été séquencés à l'aide de la technologie Oxford Nanopore. Parmi eux, 12 ont bénéficié d'un séquençage WGS, tandis que 3 ont été analysés par *Adaptive Sampling*, ciblant spécifiquement des régions d'intérêt. Les séquençages ont généré un volume total de 11,4 To de fichiers Pod5. La médiane des données de BAM obtenues était de 93,94 Gb pour les échantillons séquencés en WGS et de 8,48 Gb pour les échantillons séquencés par AS. La longueur moyenne des lectures était de 7934 pb (données sur 14 patients), avec un N50 des séquençages de WGS de 11,33 kb, ce qui témoigne d'une longueur suffisante pour outrepasser certains éléments répétés tels que les *SINEs* ou *LINEs*. En terme de profondeur de séquençage sur l'ensemble du génome, la médiane obtenue était de 26× pour les échantillons WGS (données sur 11 patients) et de 2,3× sur le génome total pour les échantillons AS (40X en moyenne sur la cible estimé *via* IGV).

Pour évaluer systématiquement la distribution de la couverture, les taux d'alignement et d'autres caractéristiques des lectures, les outils PycoQC, Mosdepth et Qualimap ont été utilisés. Les résultats de ces évaluations sont résumés dans le Tableau 2, qui présente pour chaque échantillon les caractéristiques du séquençage, les performances en termes de profondeur et de qualité, ainsi que les métriques post-alignement.

L'analyse de la concordance entre Sniffles2 et CuteSV a révélé une variation de performance en fonction des régions et des profils de variants. En moyenne, 66,6% (6/9) des variants détectés par Sniffles2 ont également été identifiés par CuteSV, tandis que 54,5% (6/11) des variants détectés par CuteSV ont été retrouvés par Sniffles2. Au total, parmi les 15 échantillons étudiés, un seul n'a pas été détecté par au moins un des deux outils. Ces résultats suggèrent que, bien que des divergences existent entre les deux outils, leur complémentarité pourrait être utile pour une détection plus exhaustive des variants de structure. Une analyse cependant plus détaillée des appels de variants par CuteSV et Sniffles2 sera présentée dans les sections suivantes, au fur et à mesure de la description des patients étudiés.

1.2. Spécificités du séquençage par *adaptive sampling*

Les données issues du séquençage par *Adaptive Sampling* présentent des métriques particulières, qui peuvent parfois prêter à confusion. Comme mentionné précédemment, cette méthode commence par séquençer tous les fragments sur environ 400 bases, avant d'évaluer si la séquence correspond à une région d'intérêt définie dans un fichier BED. Si le fragment fait parti de la région ciblée, le séquençage se poursuit intégralement ; dans le cas contraire, il est interrompu. Toutefois, les ~400 premières bases séquencées sont conservées et alignées sur le génome de référence, même si elles ne sont pas issues de la région ciblée.

Cette caractéristique explique pourquoi les données de profondeur de séquençage obtenues en *Adaptive Sampling* reflètent une couverture faible mais relativement uniforme sur l'ensemble du génome, constituée en majorité de *reads* peu informatifs. Les logiciels comme MinKNOW, ou même certains outils de contrôle qualité, ne permettent pas d'évaluer spécifiquement la profondeur obtenue sur les régions ciblées.

De façon similaire, le N50 fourni par MinKNOW est calculé sur l'ensemble des *reads* générés, incluant donc tous les fragments d'environ 400 bases qui ont été rejetés après l'amorçage du séquençage. Ce biais entraîne une sous-estimation de la taille des *reads* réellement pertinents pour l'analyse. Les outils de métriques post-alignement permettent cependant une distinction plus fine, offrant ainsi une estimation plus représentative de la taille moyenne des fragments utiles.

1.3. Impact des facteurs pré-analytiques sur les performances de séquençage

Nos résultats mettent en évidence l'importance de plusieurs paramètres pré-analytiques sur les performances du séquençage, en particulier les conditions d'extraction et de conservation de l'ADN, ainsi que la qualité de préparation et du chargement de la librairie. Ces éléments affectent directement la survie des pores actifs au cours du *run* et, par conséquent, la quantité totale de données générées.

L'échantillon UPN15 en fournit une illustration marquante. Issu d'un prélèvement externe, conservé selon des modalités différentes de celles des autres échantillons, il n'a permis de générer que 8 Gb de données. (Tableau 2) Ce rendement est comparable à celui obtenu pour les échantillons UPN1 et UPN2, mais utilisant des *flowcells* contenant environ cinq fois moins de

pores actifs, soulignant ainsi l'impact majeur de la qualité de l'ADN sur l'efficacité du séquençage.

La qualité de la librairie et son chargement sur la *flowcell* représentent également des facteurs critiques. L'échantillon UPN4, par exemple, a présenté un rendement particulièrement faible. (Tableau 2) L'analyse fine des rapports de séquençage et des paramètres de pore scan a révélé que, bien que 80 % des pores soient disponibles, seuls 20 % d'entre eux étaient effectivement en cours de séquençage (Figure 30). Cette inefficacité traduit un chargement insuffisant de la librairie, qui a non seulement limité le volume de données produites, mais a également compromis la durée de vie des pores, ces derniers devenant inactifs plus rapidement lorsqu'ils ne sont pas sollicités.

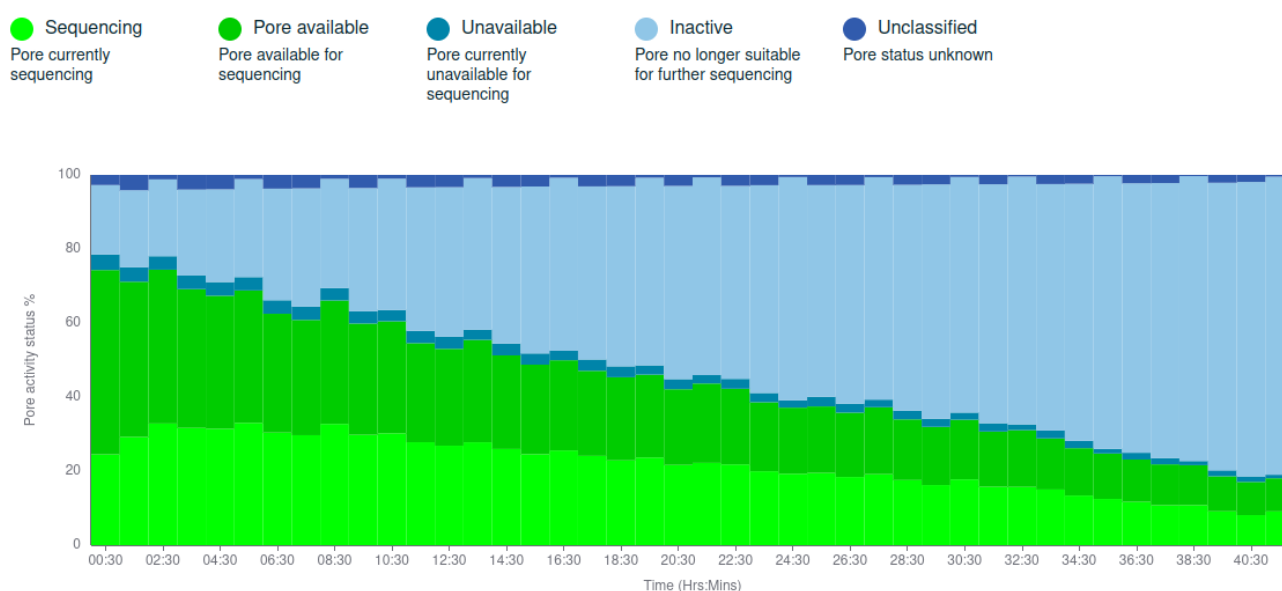


Figure 30: Profil d'activité des pores au cours du séquençage de l'échantillon UPN4

Ces observations soulignent l'importance cruciale d'un ADN de haute qualité et d'une librairie correctement préparée pour assurer le succès des *runs* de séquençage, tant sur le plan du rendement que de la stabilité des performances au cours du temps.

Toutefois, malgré des volumes de données parfois limités, l'objectif principal du projet a été atteint. En effet, comme illustré dans le Tableau 2 correspondant, l'ensemble des *runs* a permis d'identifier des *reads* chimériques informatifs, associés aux variants de structure spécifiques étudiés.

Tableau 2 : Caractéristiques de séquençage de quinze individus (UPN1–UPN15) présentant des variants de structure non résolus détectés lors de tests diagnostiques cliniques

Patient	Technique d'extraction ADN + qualité	Nom de la Flowcell	Type de la Flowcell	Versión de Guppy	Protocole de séquençage	WGS/AS ?	Taille à la TapeStation (bp)	Quantité de départ (µg)	Appareil	Nombre de pores au départ	Taille Pod5 (Go)	Quantité totale de read	Quantité totale	N50	Couverture >10X	Couverture >30X	Couverture médiane	%GC	Taille moyenne	Taille médiane	Nombre de reads supportant le variant	Durée du run	Nombre de canaux actifs	% de reads alignés	Appels des variants
UPN 1	Robotique	FAW79182	MinION	Guppy 6.4.6	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	AS	>60000	1	Mk1C	1650	NA	14,94M	12,6Gb	663b	5 %	0,20 %	3X (40X sur cible estimé manuellement)	40 %	5156	4999bp	5	47,33	511	99,30 %	CuteSV
UPN 2	Robotique	FAX91519	MinION	Guppy 7.0.9	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	AS	15452	3	GridION	1450	90	11,37M	6,6Gb	611b	1 %	0,2% %	2X (50X sur cible estimé manuellement)	40 %	5057	4999bp	11	22,99	511	99,30 %	CuteSV
UPN 3	Robotique	PAW56632	PromethION	Dorado 7.2.14	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	WGS	11873	3	P2i	6439	396	7,1M	44,97Gb	10,84kb	NA	NA	NA	NA	NA	NA	3	19,22	1609	98,50 %	No
UPN 4	Robotique	PAW10372	PromethION	Guppy 7.0.9	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	WGS	11881	3	P2s	7221	351	4,92M	34,77Gb	10,37kb	45,80 %	0,30 %	9X	40 %	8249	4999bp	10	41,79	2631	99,90 %	CuteSV + Sniffles
UPN 5	Robotique	PAQ46303	PromethION	Guppy 7.0.9	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	WGS	11969	3	P2s	7550	836	13,65M	92,86Gb	7,7kb	92,60 %	11,50 %	22X	40 %	6411	4999bp	Variant Complexe	34,85	2600	100 %	CuteSV (3 types de reads) + Sniffles (1 type de read)
UPN 6	Robotique	PAW04998	PromethION	Dorado 7.2.14	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	WGS	33896	3	P2s	8426	1203	13,18M	115,88Gb	16,07kb	94 %	76 %	35X	41 %	12529	9999bp	18	43,71	2625	99,90 %	CuteSV + Sniffles (pour la duplication)
UPN 7	Robotique	PAS57516	PromethION	Guppy 7.0.9	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	WGS	14063	3	P2s	8432	705	8,12M	68,07Gb	12,24kb	92 %	3 %	20X	41 %	9352	4999bp	16	69,45	2663	99,80 %	CuteSV + Sniffles
UPN 8	Robotique	PAW06528	PromethION	Dorado 7.2.14	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	WGS	24791	3	P2s	8467	1231	17,84M	119,26Gb	13,83kb	93 %	82 %	35X	41 %	11440	9999bp	10	43,72	2670	99,80 %	Sniffles
UPN 9	Robotique	PAU96371	PromethION	Dorado 7.2.14	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	WGS	25880	3	P2i	7620	865	13,54M	88,1Gb	14,42kb	93 %	21 %	25X	41 %	8318	4999bp	20	71,55	2733	99,90 %	CuteSV
UPN 10	Robotique	PAU00824	PromethION	Guppy 7.0.9	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	WGS	17430	3	P2s	7713	1008	11,99M	95,02Gb	10,82kb	93 %	42 %	29X	41 %	10603	9999bp	16	71,95	2653	99,80 %	CuteSV + Sniffles
UPN 11	Robotique	PAW63468	PromethION	Dorado 7.2.14	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	WGS	11873	3	P2i	6510	1869	19,32M	124,29Gb	8,26kb	93 %	61 %	32X	41 %	10170	9999bp	17 et 18	66,79	2561	99,60 %	Sniffles (2 types de reads)
UPN 12	Robotique	PAW06168	PromethION	Dorado 7.2.14	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	WGS	16093	3	P2i	7158	1180	13,12M	119,92Gb	11,96kb	94 %	74 %	34X	41 %	7141	4999bp	17 et 14	75,22	2564	99,80 %	CuteSV + Sniffles (pour les 2 callers : 2 types de reads)
UPN 13	Robotique	PAW56637	PromethION	Dorado 7.2.14	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	WGS	19751	3	P2i	5961	1008	11,03M	100,4Gb	11,93kb	93,00 %	40,00 %	28X	41,00 %	NA	NA	11	NA	NA	99,80 %	Sniffles
UPN 14	Robotique	PAS5759	PromethION	Guppy 7.0.9	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	WGS	11468	3	P2s	8180	645	14,27M	61,68Gb	8,13kb	90 %	1 %	18X	41 %	4896	2499bp	9	42,76	2692	99,60 %	CuteSV
UPN 15	ADN extérieur	PAW10221	PromethION	Guppy 7.0.9	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	AS	26323	3	P2s	8919	78,5	10,21M	8,49Gb	687b	1 %	0 %	2X (30X sur cible estimé manuellement)	40 %	12520	9999bp	7	19,29	2683	NA	CuteSV

2. Étude de variants récurrents, soumis au mécanisme de NAHR

2.1. Article 1 – Publié dans *Epilepsia: Long-Read Sequencing of Recurrent FGF12 Duplications in Epilepsy: Insights into Structural Mechanisms and Aberrant Isoforms*

2.1.1. Introduction

Le facteur de croissance des fibroblastes 12 (FGF12), codé par *FGF12* (anciennement désigné *FHF1*), appartient à la famille des facteurs homologues des fibroblastes (FHF), qui comprend *FHF1* (*FGF12*), *FHF2* (*FGF13*), *FHF3* (*FGF11*) et *FHF4* (*FGF14*).⁽¹²¹⁾ *FGF12* est majoritairement exprimé dans le système nerveux en développement et mature, les tissus musculaires cardiaques et squelettiques, ainsi que dans l'épithélium olfactif.⁽¹³¹⁾ *FGF12* génère deux principaux transcrits. Le transcrit long, désigné A, code une protéine comportant une séquence de type signal de localisation nucléaire (NLS), permettant sa localisation nucléaire. En revanche, le transcrit court, désigné B, code une protéine dépourvue de cette séquence, ce qui entraîne sa localisation cytoplasmique.^(121,123) *FGF12* interagit avec des canaux sodiques voltage-dépendants (Nav), notamment Nav1.2 (SCN2A), Nav1.5 (SCN5A) et Nav1.6 (SCN6A), et module les cinétiques d'inactivation rapide de ces canaux.^(122,124)

Des variants faux-sens hétérozygotes pathogènes récurrents de *FGF12* ont été associés à des encéphalopathies développementales et épileptiques autosomiques dominantes (MIM #617166), et ont été rapportés dans plus de vingt publications.^(125–129) Le phénotype associé comprend un retard du développement, une déficience intellectuelle (DD/ID), une hypotonie axiale, une ataxie et une épilepsie précoce pharmacorésistante.⁽¹³⁰⁾ Des études électrophysiologiques ont démontré que ces variants récurrents ont un effet de gain de fonction, entraînant un décalage dépolarisant de l'inactivation rapide dépendante du voltage du canal Nav1.6.^(125,131) Outre les variants ponctuels, des duplications partielles de *FGF12* associées à un phénotype similaire ont été rapportées chez six patients.^(132–134) Ces duplications sont supposées produire un transcrit plus long, associé à un mécanisme de perte de fonction.^(132,135)

Nous rapportons ici deux nouveaux patients porteurs d'une duplication partielle récurrente de *FGF12*. Grâce au séquençage *long-read* de l'ADN, nous avons entièrement résolu le variant

structurel sous-jacent, sans hypothèse préalable, et démontré qu'il s'agissait d'une duplication en tandem. En complément, nous avons étudié les transcrits aberrants contenant *FGF12* à l'aide du séquençage *long-read* de l'ADN complémentaire (ADNc) pleine longueur. Nous fournissons également une description détaillée des phénotypes cliniques et électrocliniques de la maladie. Cette étude met en lumière le rôle central du LRS dans l'élucidation des variants de structure et apporte de nouvelles perspectives sur leurs conséquences fonctionnelles.

2.1.2. Méthodes spécifiques

2.1.2.1. Déclaration éthique

Les données cliniques ont été recueillies et un consentement éclairé a été obtenu pour les analyses génétiques. Des prélèvements sanguins ont été réalisés chez les individus atteints ainsi que chez leurs parents, après obtention des consentements éclairés. Les analyses ont été effectuées dans un cadre diagnostique, conformément aux règles bioéthiques du droit français. Cette étude a été approuvée par le Comité de Protection des Personnes (CPP), référence n°2023-A00473-42. Un consentement écrit et éclairé a été obtenu auprès de l'ensemble des participants ou de leurs représentants légaux.

2.1.2.2. Séquençage des amplicons d'ADNc

La synthèse d'ADNc a été réalisée à l'aide du kit *SuperScript III First-Strand Synthesis System* (ThermoFisher Scientific, Waltham, MA, USA), suivie d'une RT-PCR utilisant la polymérase *NEBNext High-Fidelity* (New England Biolabs, Ipswich, MA, USA). Plusieurs amorces ont été conçues pour amplifier et analyser à la fois les transcrits wild-type (WT) et aberrants, en fonction de l'hypothèse formulée. L'objectif était d'amplifier la jonction de la duplication en tandem dans les transcrits de *FGF12*. Pour les transcrits aberrants, des amorces ciblant les exons 4 à 3 ont été utilisées, tandis que pour les transcrits WT, les amorces ciblaient les exons 3 à 6 ; ces dernières permettaient également l'amplification des transcrits aberrants (liste des amorces disponible sur demande, Figure_Annexe 1). Les produits PCR amplifiés ont été préparés sous forme de bibliothèques pour le séquençage Nanopore, puis chargés sur un *flowcell* PromethION. Le *basecalling* a été effectué avec Guppy (version 6.4.6) en mode *High-Accuracy* (HAC). Les lectures obtenues ont été alignées sur le génome de référence humain GRCh38 à l'aide de Minimap2, avec des paramètres adaptés à l'alignement des

données transcriptomiques (splice-aware). Tous les alignements chimériques ont été cartographiés sur le génome à l'aide de l'outil BLAST, afin de déterminer leur origine génomique, ce qui a permis de distinguer les transcrits WT des transcrits aberrants.(136)

2.1.2.3. Analyse des points de cassures

Des amorces de séquençage ont été conçues afin d'amplifier les points de cassure spécifiques chez les deux patients. Une première PCR d'environ 6 000 pb a été réalisée à l'aide de la polymérase *Expand Long Template* (Roche, Bâle, Suisse). Celle-ci a été suivie d'une PCR nichée utilisant la polymérase *NEBNext High-Fidelity 2X* (New England Biolabs, Ipswich, MA, USA). Les produits PCR ont été extraits d'un gel d'agarose à 2 %, purifiés à l'aide d'un kit de purification PCR (Macherey-Nagel), puis séquencés par la méthode de Sanger (liste des amorces disponible sur demande).

2.1.3. Résultats

2.1.3.1. Description clinique et caractéristiques des EEG de UPN1

UPN1 est né de parents en bonne santé, non apparentés, après une grossesse sans complication. Le développement psychomoteur précoce était normal (marche acquise à 14 mois, premiers mots à 12-13 mois, bonnes interactions sociales)(Tableau 3). Aucun trait dysmorphique n'a été observé. L'épilepsie a débuté brutalement à l'âge de 18 mois, avec des états de mal épileptique moteurs généralisés récurrents non fébriles (trois états de mal en un mois) et des crises fronto-temporales à bilatéralisation rapide. Les premiers EEG, réalisés dans un contexte de crises multiples quotidiennes, montraient une activité de fond désorganisée et ralentie (3,5 Hz en moyenne, prédominance delta à 1,5 Hz entrelacée d'activité thêta à 4–5 Hz), avec une amplitude très élevée atypique (150–400 microvolts) sans gradient antéro-postérieur en état d'éveil, et de rares pointes focales, souvent diffuses sur un hémisphère, droit ou gauche (Figure 31A). En sommeil, l'activité delta diminuait, tandis que l'activité thêta augmentait, et des fuseaux de sommeil bilatéraux étaient présents (Figure 31B).

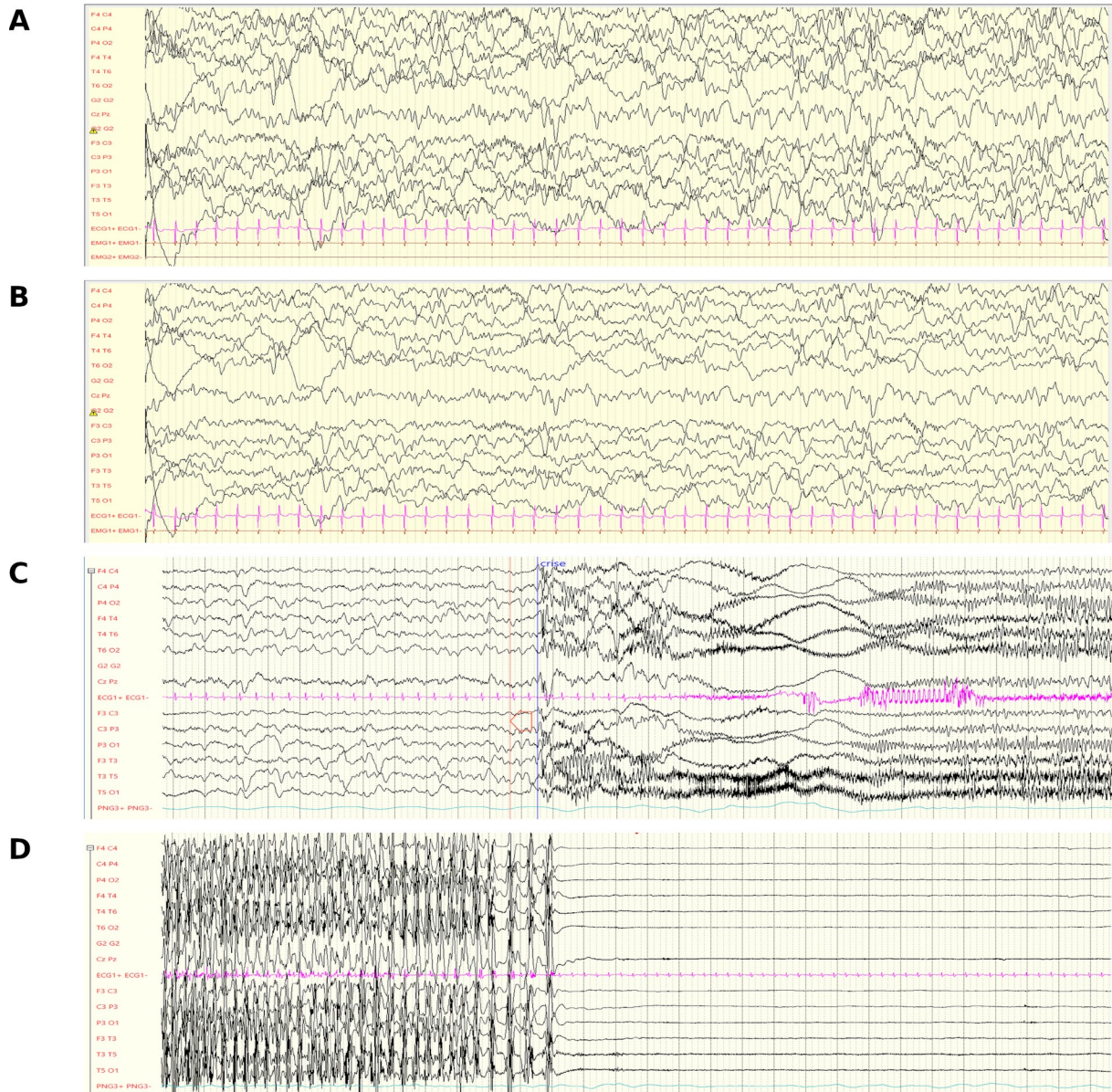


Figure 31: EEG de UPN1 au début de l'épilepsie (Phase aiguë 1), à l'âge de 18 mois, en état d'éveil (1A), pendant le sommeil (1B) et lors des crises (1C-D). (A) En éveil, activité anormalement lente, désorganisée et de haute amplitude, entre 200 et 400 microvolts, avec des pointes multifocales rares. Réglages des filtres : 0,53 – 70 Hz, filtre coupe-bande à 50 Hz ; sensibilité : 20 $\mu\text{V}/\text{mm}$. (B) Pendant le sommeil, activité delta réduite à 1,5 Hz, enrichie en θ à 4–5 Hz, présence de fuseaux de sommeil bilatéraux. Réglages des filtres : 0,53 – 70 Hz, filtre coupe-bande à 50 Hz ; sensibilité : 15 $\mu\text{V}/\text{mm}$. (C-D) Crises tonico-cloniques bilatérales, suivies d'une suppression de l'activité EEG pendant une minute. Réglages des filtres : 0,53 – 70 Hz, filtre coupe-bande à 50 Hz ; sensibilité : 25 $\mu\text{V}/\text{mm}$.

Le monitoring EEG a enregistré des crises tonico-cloniques bilatérales, durant 1 à 1,5 minute, comportant un spasme myoclonique, une phase tonique, puis une phase clonique bilatérale (Figure 31C-D), ainsi que des crises à début fronto-temporal avant bilatéralisation. Quelques

semaines après le début de l'épilepsie, le patient a présenté une régression psychomotrice sévère, incluant une perte complète du langage oral, un trouble relationnel majeur (absence de contact visuel, absence de sourire, de rire ou de pleurs), un syndrome pyramidal, une perte de la marche et une ataxie. L'IRM cérébrale initiale, l'analyse du liquide cébrospinal et les examens immunologiques et métaboliques étaient normaux. Aucun signe d'atteinte extra-neurologique n'a été observé, et l'évaluation cardiaque était normale. Compte tenu de la sévérité du tableau clinique, un traitement par immunoadsorption a été initié à l'âge de 20 mois, dans l'hypothèse d'une encéphalite séronégative. Cette approche ayant permis une amélioration partielle (récupération de la marche), elle a été suivie d'un traitement par perfusions de rituximab pendant 18 mois.

Après la phase aiguë initiale, l'épilepsie a évolué avec des salves mensuelles de crises tonico-cloniques bilatérales, toniques, toniques focalisées, ainsi que de nombreux épisodes de crises focales complexes durant 3 à 5 jours. Les Figure 32A-B illustrent une crise focale d'une minute avec un début temporel droit, enregistrée à l'âge de 3 ans. Les EEG montrent une hypersynchronie lente prolongée à l'éveil durant plusieurs minutes, avec un rythme delta généralisé et une altération du contact (Figure 32C), une augmentation des ondes triphasiques diffuses en période d'éveil, ainsi que des pointes multifocales plus fréquentes. Parmi les 29 EEG réalisés sur une période de 5 ans, un seul montrait une amplitude normale avec une activité de fond subnormale et de rares activités paroxystiques en sommeil, enregistré après une période de deux mois sans crise. Cette amélioration transitoire n'était associée à aucune modification thérapeutique. L'évolution de l'IRM cérébrale a révélé une atrophie cérébrale et cérébelleuse modérée (Figure_Annexe 2). Un TEP cérébrale a mis en évidence un hypométabolisme marqué et symétrique des régions associatives frontales, pariétales et temporales.

À l'âge de 5 ans, des crises myocloniques et des absences atypiques sont apparues, accompagnées d'une augmentation marquée des pointes-ondes généralisées subcontinues à l'EEG. La fréquence des crises toniques et tonico-cloniques bilatérales s'est accrue, survenant plusieurs fois par jour. L'activité de fond était caractérisée par des pointes généralisées isolées ou des salves de pointes-ondes à 2,5 Hz, associées à des absences atypiques et à des myoclonies concomitantes (Figure 32D), entrelacées d'une activité delta bi- et triphasique. L'état neurologique du patient s'est fortement détérioré, avec une perte complète des acquisitions motrices, une absence totale d'autonomie pour l'alimentation, et une réduction drastique du temps d'éveil quotidien, sans aucune réponse aux stimulations.

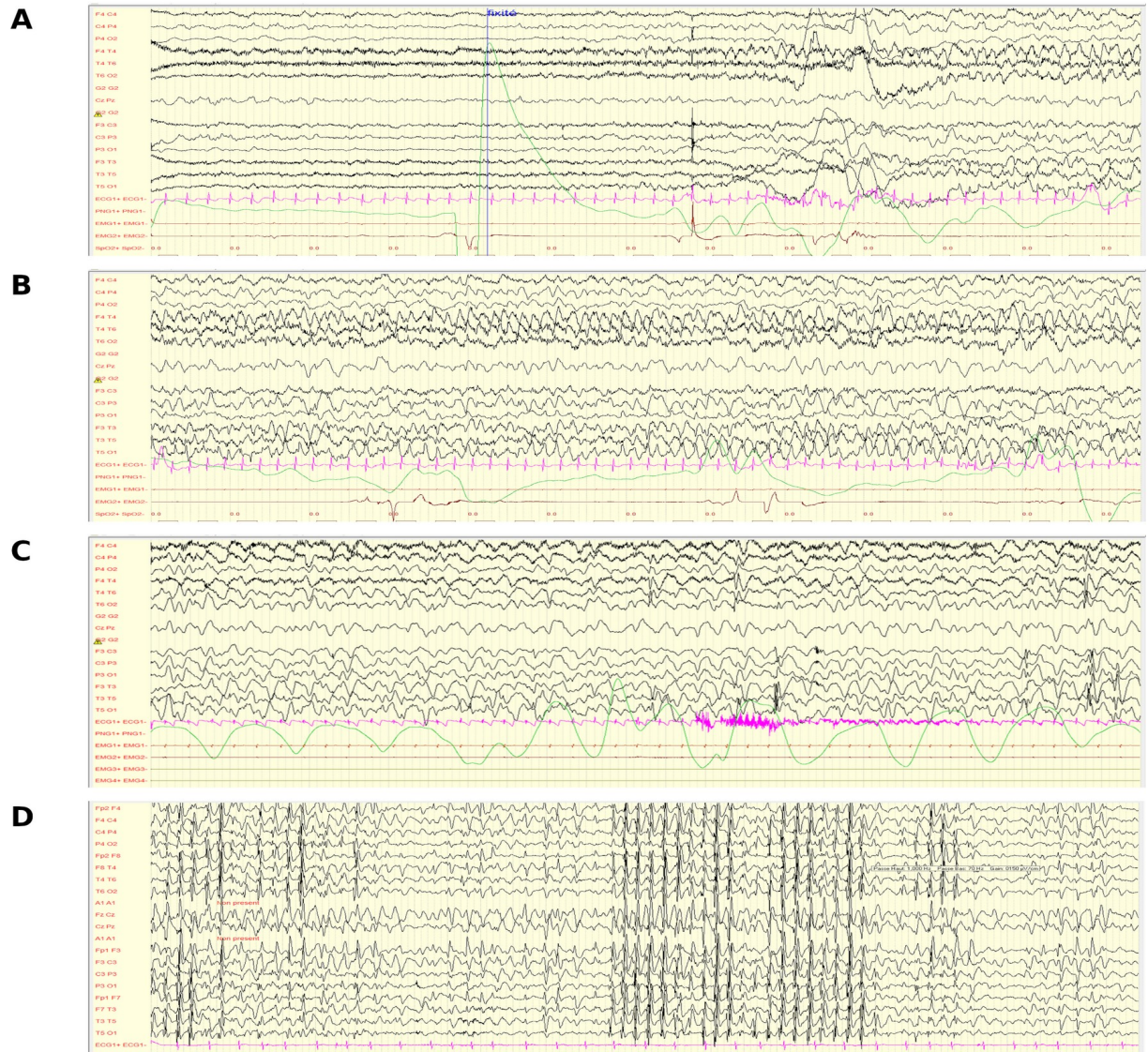


Figure 32 : EEG de UPN1 à l'âge de 3 ans pendant les crises (phase 2 « lune de miel » avec crises intermittentes) (2A-B), un mois plus tard (2C), et lors du dernier suivi à 5 ans (phase 3 d'aggravation). (A-B) Crises tonico-clo-niques bilatérales suivies d'une suppression de l'activité EEG pendant une minute. Réglages des filtres : 0,53 – 70 Hz, filtre coupe-bande à 50 Hz ; sensibilité : 25 μ V/mm. (C) Hypersynchronie lente au réveil durant plusieurs minutes, activité delta généralisée rythmique, avec altération du contact pendant cette période sur plusieurs EEG successifs et augmentation des ondes triphasiques diffuses. Réglages des filtres : 0,53 – 70 Hz, filtre coupe-bande à 50 Hz ; sensibi-lité : 10 μ V/mm. (D) Tracé de fond d'encéphalopathie diffuse riche en activité delta bi- et triphasique (cercle bleu), avec pointes généralisées isolées ou salves de pointes-ondes à 2,5 Hz (en rouge), associées à des absences atypiques et des myoclonies concomitantes. Réglages des filtres : 0,53 – 70 Hz, filtre coupe-bande à 50 Hz ; sensibilité : 10 μ V/mm.

Depuis le début de l'épilepsie, plusieurs traitements antiépileptiques ont été essayés. Le régime cétogène, les vitamines B6, B8, ainsi que le cannabidiol (10 mg/kg/j) se sont révélés in-efficaces. La lévétiracétam (55 mg/kg/j), le valproate de sodium (30 mg/kg/j), la lamotrigine

(mal tolérée à 1 mg/kg/j), le clonazépam (0,15 mg/kg/j), le clobazam (0,15 mg/kg/j), le topiramate (mal toléré à 2 mg/kg/j) et le phénobarbital (5 mg/kg/j) ont montré une efficacité partielle et transitoire. Des bloqueurs des canaux sodiques ont été introduits après le diagnostic génétique à l'âge de 5 ans et 3 mois. La carbamazépine (20 mg/kg/j) a temporairement réduit l'activité épileptique, sans amélioration des anomalies de fond à l'EEG, et a perdu son efficacité après quelques mois. La phénytoïne a permis de contrôler efficacement les états de mal tonico-cloniques bilatéraux, mais son utilisation prolongée a entraîné une aggravation de l'ataxie et des myoclonies. Au dernier suivi (à 6 ans et demi), le patient était traité par phénobarbital et felbamate (Figure_Annexe 3A). Il restait alité, peu réactif, et présentait des crises toniques et tonico-cloniques bilatérales plusieurs fois par jour.

2.1.3.2. Description clinique et caractéristiques des EEG de UPN2

UPN2 est le troisième enfant d'un couple non consanguin, sans antécédent familial d'épilepsie. Il est né à terme, eutrophique, après une grossesse sans complications. L'épilepsie a débuté à l'âge de 5 mois, par des crises généralisées caractérisées par une hypotonie et une aréactivité durant environ 2 minutes, suivies d'un retour rapide à la normale. Il a rapidement présenté des crises récurrentes avec déviation de la tête et mouvements cloniques des membres supérieurs. L'EEG initial a révélé un rythme de base modérément ralenti, associé à des ondes lentes et des pointes dans les régions temporales pendant le sommeil (Figure 33A). Malgré le début de l'épilepsie, il a acquis les premières étapes du développement moteur (position assise à 9 mois, marche à 14 mois)(Tableau 3). Une régression du langage a été observée à 15 mois dans un contexte d'otite séreuse, suivie d'une absence de progression du langage et d'un isolement social croissant. Un diagnostic de trouble du spectre autistique (TSA) sévère a été confirmé à l'âge de 4 ans. Le patient a acquis la marche autonome, sans signes de syndrome pyramidal ni d'ataxie. Il est bien intégré dans un établissement médico-éducatif spécialisé. Les IRM cérébrales réalisées à 6 mois et à 2 ans étaient normales.

Il présentait 1 à 2 crises par mois au cours des premières années de vie. Les traitements anti-épileptiques administrés ont été peu efficaces (valproate de sodium, lévétiracétam, oxcarbazépine, lamotrigine et topiramate). Vers l'âge de 3 ans, l'EEG a révélé un ralentissement global avec une activité delta désorganisée et de voltage anormalement élevé (Figure 33B).

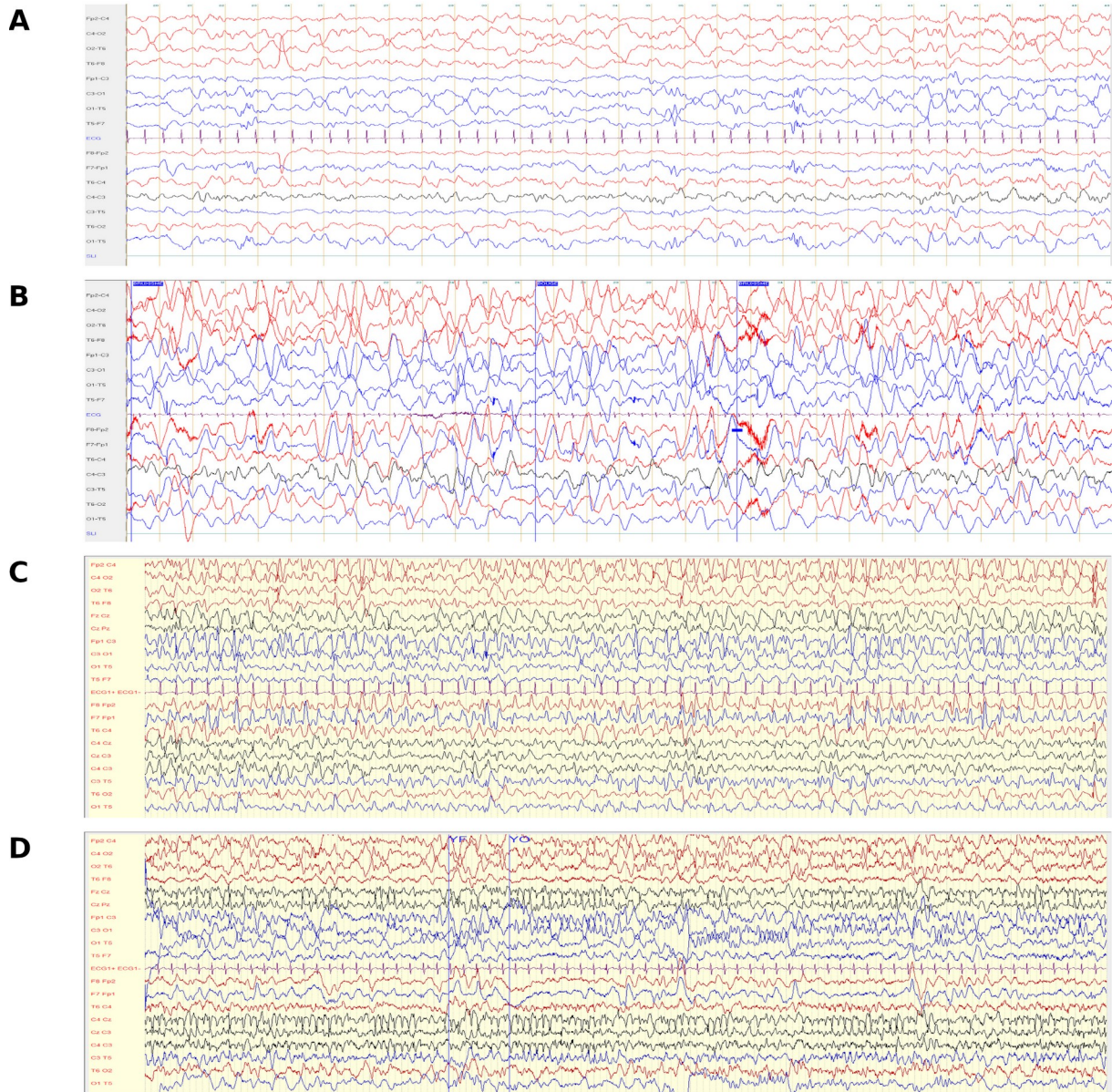


Figure 33 : EEG de UPN2. (A) Phase aiguë 1 à l'âge de 6 mois, en sommeil. Ondes lentes et pointes dans les régions temporales et occipitales gauches. Réglages des filtres : 0,53 – 70 Hz, filtre coupe-bande à 50 Hz ; sensibilité : 20 $\mu\text{V/mm}$. (B) À 4 ans. Tracé de fond lent, désorganisé, avec ondes delta de haut voltage. Réglages des filtres : 0,53 – 70 Hz, filtre coupe-bande à 50 Hz ; sensibilité : 15 $\mu\text{V/mm}$. (C) À 6 ans. Tracé de fond lent, désorganisé et de haut voltage avec ondes bi- et triphasiques, salves continues de pointes-ondes dans les régions frontales. Réglages des filtres : 0,53 – 70 Hz, filtre coupe-bande à 50 Hz ; sensibilité : 20 $\mu\text{V/mm}$. (D) À 6 ans. Tracé de fond lent et désorganisé avec décharges rythmiques thêta pointues localisées au vertex central. Réglages des filtres : 0,53 – 70 Hz, filtre coupe-bande à 50 Hz ; sensibilité : 10 $\mu\text{V/mm}$.

À 6 ans, il a présenté une augmentation significative de la fréquence des crises, principalement des épisodes tonico-cloniques bilatéraux pouvant être prolongés, ainsi que plusieurs types de crises focales. L'activité de fond à l'EEG s'est dégradée et montrait un *pattern* domi-

nant d'encéphalopathie delta, avec des ondes delta bi- et triphasiques fréquentes en plus de décharges bilatérales de pointes-ondes frontales (Figure 33C). Les modifications ultérieures du traitement n'ont pas apporté d'amélioration significative (lacosamide, clonazépam, phénytoïne, pérampandol, topiramate, cannabidiol, régime cétogène), voire ont aggravé l'activité épileptique (clobazam) (Figure_Annexe 2B). Les EEG de suivi ont continué à montrer une activité persistante désorganisée, avec des crises asymptomatiques localisées aux régions centrales droites et médianes (Figure 33D). Son niveau de conscience s'est progressivement altéré. Au dernier suivi, à l'âge de 6 ans et demi, le patient présente un retard intellectuel profond, une autonomie limitée dans les activités quotidiennes, et des troubles du langage, ne prononçant que quelques mots.

À l'âge de 7 ans, le patient a présenté un épisode de deux mois d'état de mal épileptique réfractaire, avec un diagnostic suspecté mais non confirmé d'encéphalite limbique. La prise en charge a nécessité une admission en unité de soins intensifs et un traitement par échanges plasmatiques. L'état de mal myoclonique a été résolu après l'administration orale de phénytoïne et de zonisamide, avec un retour progressif et complet à l'état de base du patient : marche autonome, présence de stéréotypies, absence de langage oral, et expression émotionnelle par le sourire et des vocalisations. Aucun signe de syndrome pyramidal n'a été observé.

2.1.3.3. Caractérisation des réarrangements structuraux par séquençage LRS

Les deux patients présentaient un gain *de novo* de taille minimale de 536 kb au sein de *FGF12*, localisé sur le *locus* 3q28q29, tel que précédemment détecté par analyse CGH sur puce : arr[GRCh38] 3q28q29(192182695_192719315)x3. Un séquençage ciblé par la technologie *Long-Read Sequencing* a été réalisé à l'aide de la méthode d'*Adaptive Sampling*, en ciblant la région chr3:163,000,000–195,000,000, qui inclut *FGF12* parmi plus de 120 gènes. Après enrichissement, la duplication ciblée a atteint une couverture d'environ 35× chez les deux patients, tandis que la couverture moyenne sur l'ensemble du génome était de 2 à 3× (Tableau 2). Le séquençage a généré 12,6 Gb et 6,6 Gb de données, correspondant respectivement à 14,9 millions et 11,37 millions de lectures pour les UPN1 et UPN2. La taille moyenne des lectures supportant le variant de gain de copies impliquant *FGF12* était de 15 508 pb pour UPN1 et de 7 948 pb pour UPN2.

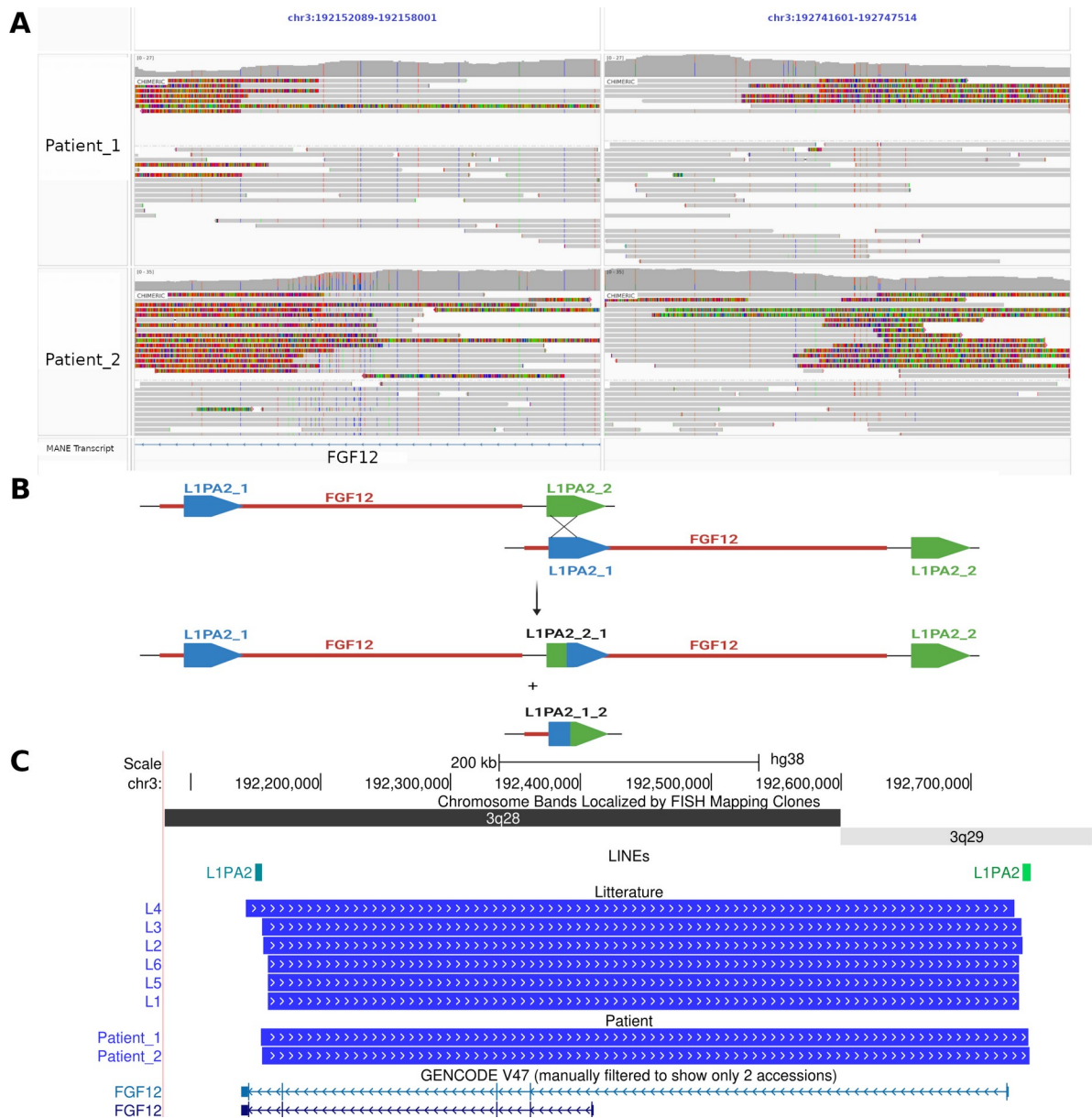


Figure 34 : Étude des réarrangements d'ADN englobant le gène *FGF12* (A) Visualisation des lectures issues du séquençage long recouvrant les points de cassure des duplications de *FGF12* chez les UPN1 et UPN2, affichées avec l'Integrated Genome Viewer (IGV). (B) Représentation schématique du mécanisme de réparation par recombinaison homologue non allélique (NAHR), localisée précisément sur les éléments LINE L1PA2 aux points de cassure de la duplication de *FGF12*. (C) Vue d'ensemble des duplications identifiées chez des patients de la littérature (L1 à L6) et chez les patients de notre étude (UPN1 et UPN2), ainsi que la localisation des éléments L1PA2 dans le Genome Browser UCSC (Université de Californie, Santa Cruz) (chr3:192080000-193020000).

Les coordonnées précises des duplications identifiées étaient les suivantes : NC_000003.12:g.192153910_192743885dup chez UPN1, et NC_000003.12:g.192155125_192745081dup chez UPN2 (Figure 34A). Par ailleurs, la détection de variants de structure à l'aide de CuteSV a identifié la duplication

chez les deux patients, confirmant ainsi ces événements. Les points de cassure étaient situés au sein des éléments répétitifs *LINE* *L1PA2-A* et *L1PA2-B*, respectivement en chr3:192,149,316–192,155,328 et chr3:192,739,214–192,745,242. Ces éléments *LINEs* partagent une homologie de séquence de 98,14 %. L'alignement de lectures chimériques a révélé une duplication tandem dans les deux cas, avec un point de cassure 5' localisé dans l'intron terminal, commun aux deux isoformes de *FGF12*, et un point de cassure 3' situé en aval de la région 5'UTR, dans une région intergénique (Figure 34B–C). L'implication des éléments *L1PA2-A* et *L1PA2-B* chez les deux patients suggère un mécanisme de recombinaison homologue non allélique (NAHR) (Figure 34B). Ce mécanisme, ainsi que la présence des duplications, ont été confirmés par le séquençage Sanger des points de cassure (Figure_Annexe 4), révélant des séquences homologues de 73 pb chez UPN1 et de 53 pb chez UPN2.

2.1.3.4. Séquençage LRS des isoformes majeures A et B de *FGF12*

L'analyse des ADNcs a été réalisée sur les isoformes A et B de *FGF12*. Les deux transcrits WT partagent les exons terminaux, désignés exons 3 à 6, mais diffèrent par leurs exons initiaux : le transcrit sauvage A contient l'exon 1A, tandis que le transcrit sauvage B contient les exons 1B et 2. Les exons 1A, 1B et 2 forment ensemble les exons initiaux du transcrit (Initial Transcript Exons, ITE).

L'ARN a été extrait à partir de lignées cellulaires lymphoblastoïdes (LCL) dérivées de UPN1, puis rétro-transcrit en fragments d'ADNc. Des amorces PCR ont été conçues pour amplifier d'éventuels transcrits aberrants, et la bibliothèque d'amplification a été séquencée par technologie de séquençage *long-read*. L'analyse des séquences des derniers exons a permis d'identifier deux types d'isoformes aberrantes en plus des isoformes WT (Figure 35A).

La première isoforme aberrante A correspond à une fusion des séquences codantes des exons 5 et 3, avec saut de l'exon 6, conduisant à la séquence ITE-3-4-5-3-4-5-6 (Figure 35B).

La seconde isoforme aberrante B est caractérisée par l'insertion de 123 nucléotides issus de la région intergénique chr3:192,738,861-192,738,983, désignée « INS », insérée après l'exon 5 puis suivie de l'exon 3, formant la séquence ITE-3-4-5-INS-3-4-5-6 (Figure 35B).

Les protéines issues des isoformes sauvages A et B comptent respectivement 243 et 181 acides aminés. La protéine dérivée de l'isoforme aberrante A basée sur l'isoforme B est prédite à 320 acides aminés, sans arrêt prématuré en raison de la fusion en phase des exons 5 et 3

(Figure 35C). En revanche, la protéine issue de l'isoforme aberrante B est prédite tronquée, avec un codon STOP prématuré (Figure 35C).

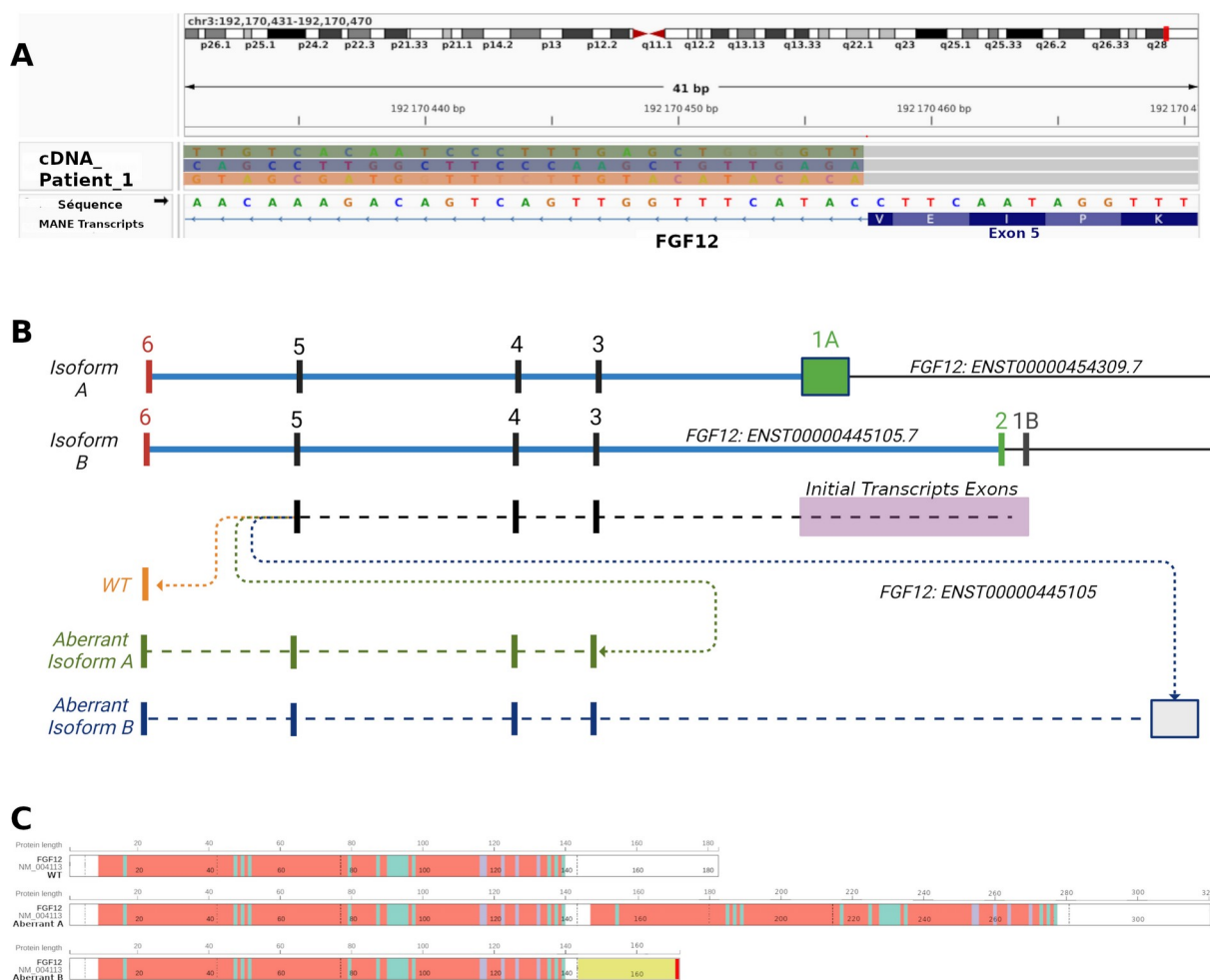


Figure 35: Étude de l'impact sur les transcrits ARN de FGF12 (A) Visualisation des transcrits aberrants en séquençage long de UPN1 au niveau de l'exon 5, affichée avec IGV. (B) Modélisation des séquences des isoformes des transcrits sauvages (WTA et WTB) et des transcrits aberrants, basée sur la visualisation des séquences des transcrits. (C) Modélisation des séquences des protéines sauvages et aberrantes (carré jaune : séquence intergénique ; carré rouge : codon STOP).

2.1.4. Discussion

Les courants sodiques résurgents jouent un rôle clé dans la régulation de l'excitabilité des neurones centraux et périphériques. Dans cette étude, nous avons utilisé le séquençage *long-read* pour explorer l'organisation structurale de microduplications intragéniques récurrentes dans *FGF12* et évaluer leurs effets transcriptionnels chez deux patients atteints d'encéphalo-

pathie développementale et épileptique (DEE), présentant un phénotype chevauchant et des anomalies EEG spécifiques.

Des microduplications récurrentes de *FGF12*, dont les points de cassure sont localisés dans les éléments répétitifs *LINE L1PA2-A* et *L1PA2-B*, ont été précédemment rapportées chez des patients DEE via analyse par CGH-array, puis confirmées par séquençage Sanger (Tableau 3). (132–134,137) Le séquençage *long-read* a permis de caractériser précisément la réorganisation structurale complète, correspondant à une duplication en tandem entre les deux répétitions intercalées et paralogues *L1PA2-A* et *L1PA2-B*, auparavant seulement suspectée. (132–134) Contrairement au séquençage génomique *short-read*, le *long-read* offre une approche globale et non biaisée pour analyser l'organisation bidimensionnelle du génome, soutenant ainsi la formulation d'hypothèses physiopathologiques. (109,138) Cette réorganisation est probablement médiée par un mécanisme de recombinaison homologue non allélique (NAHR), exploitant l'homologie de séquence entre les répétitions *L1PA2-A* et *L1PA2-B* orientées dans le même sens. Ce mécanisme, impliqué dans d'autres régions instables du génome flanquées d'éléments répétitifs, peut conduire à des délétions ou à des duplications. (139) La duplication récurrente de *FGF12* observée ici présente des points de cassure hautement conservés, avec des régions homologues ne différant que par quelques bases, ce qui renforce l'hypothèse d'un événement médié par NAHR. (132–134,137)

En tant que membre de la sous-famille des FHF (*Fibroblast Homologous Factors*), *FGF12* possède au moins deux sites d'initiation de la transcription, à l'origine de deux principaux isoformes. Le transcrit A (ENST00000454309.7; NM_021032) code l'isoforme A de *FGF12*, qui présente une région N-terminale plus longue que l'isoforme B, codée par le transcrit B (ENST0000044105.7; NM_004113). Sur le plan structural, les protéines FHF partagent un repliement en tréfoil β , comprenant une région C-terminale conservée masquant une surface hydrophobe du tréfoil. (122,123,140,141) Ce motif structural contient également des séquences spécifiques permettant l'interaction directe avec les extrémités cytoplasmiques des canaux sodiques voltage-dépendants (Nav), modulant ainsi leur activité. Ces interactions renforcent l'excitabilité neuronale en provoquant un décalage dépolarisant du seuil d'inactivation rapide des canaux Nav. (122,123,141) *FGF12* contient des résidus de surface essentiels pour l'interaction avec les canaux Nav. L'isoforme A présente notamment un motif N-terminal conservé de 15 à 20 acides aminés, agissant comme une particule de blocage du canal ouvert. Ce motif participe aux mécanismes d'inactivation rapide et prolongée des canaux Nav, en stabilisant un

état inactivé de longue durée.(131,150) Une variante faux-sens pathogène, p.(Arg144His), identifiée dans l'isoforme A, a été rapportée comme responsable d'encéphalopathie développementale et épileptique (DEE). Cette substitution altère l'interaction entre FGF12 et les canaux Nav, induit un décalage dépolarisant de l'inactivation du canal, et entraîne un effet de gain de fonction.(123,125,135)

Les duplications récurrentes de *FGF12* identifiées dans cette étude incluent le motif central ainsi que les résidus clés nécessaires à l'interaction avec les canaux sodiques Nav, notamment Lys9, Ile11, Arg52, Tyr93 et Val95 (selon la nomenclature de l'isoforme A de *FGF12*). (135) Ces duplications sont associées à deux transcrits aberrants, chacun issu du site d'initiation de transcription respectif des transcrits de type sauvage A ou B. Le premier transcrit aberrant, dérivé du transcrit A sauvage, est composé de huit exons, selon la séquence suivante : exon 1A-3-4-5-3-4-5-6, sans codon stop prématuré prédit. Le second transcrit aberrant, dérivé du transcrit B sauvage, comprend sept exons ainsi qu'une région intergénique (INS) : exon 1B-2-3-4-5-INS-3-4-5-6 (Figure 35). Des travaux antérieurs utilisant la RT-PCR et le séquençage Sanger à partir de sang et de follicules pileux chez un patient porteur d'une duplication de *FGF12* avaient suggéré la présence d'un long transcrit aberrant contenant tous les exons en phase de lecture.(142,144). Cette hypothèse est confortée par nos résultats, qui confirment par séquençage *long-read* de l'ADNc la présence de ce même long transcrit, correspondant au transcrit aberrant A. Ce transcrit pourrait potentiellement être entièrement traduit, donnant naissance à une protéine hypothétique plus longue comportant deux domaines β -trefoil, dans la mesure où les exons 5 et 3 sont maintenus en phase et ne génèrent pas de décalage du cadre de lecture. Bien que la séquence protéique n'ait pas été analysée, nous suspectons qu'une telle conformation structurale aberrante pourrait entraîner une instabilité de la protéine et sa dégradation. Nos résultats confirment également l'existence d'un second transcrit aberrant, caractérisé par l'insertion d'une séquence intergénique hors cadre, menant à un codon stop prématuré.(133) Il convient également de souligner que *FGF12* peut être affecté à la fois par des variants de structure récurrents et non récurrents. Par exemple, un patient porteur de variants de structure bi-alléliques, incluant une délétion de l'exon 6 et une duplication des exons 3–4 de *FGF12*, a récemment été décrit.(142) Dans cette étude, le séquençage *long-read* a permis de résoudre des variants bi-alléliques de petite taille. Le patient présentait des crises focales motrices évoluant vers des crises toniques-cloniques bilatérales et un état de mal épileptique, avec une réponse apparente au traitement. Ces observations illustrent la grande hétérogénéité des réarran-

gements liés à *FGF12* et soulignent l'intérêt du séquençage *long-read* pour la caractérisation de variants de structure complexes et cliniquement pertinents.

Des variants faux-sens pathogènes ont été associés à un effet gain de fonction sur les canaux sodiques Nav, tandis que les duplications ont été associées à une perte de fonction. Malgré ces mécanismes opposés, les deux altérations entraînent un phénotype identique de DEE (encéphalopathie développementale et épileptique), ne différant que par l'âge d'apparition des crises. Les variants gain de fonction abaissent le seuil de déclenchement des canaux Nav, induisant une hyperexcitabilité neuronale et des crises précoces.(125,135) En revanche, les duplications perte de fonction ne déclenchent pas d'hyperexcitabilité immédiate et sont associées à un début des crises légèrement plus tardif. La protéine allongée aberrante hypothétique pourrait déplacer la dépendance en voltage de l'inactivation des canaux Nav vers des potentiels plus hyperpolarisés, altérant ainsi leur activité et renforçant l'hypothèse d'un mécanisme de perte de fonction.(135) Dans l'hippocampe, les fibres moussues transmettent les signaux des cellules granulaires du gyrus denté vers la région CA3. Ces fibres ciblent à la fois des cellules pyramidales excitatrices et des interneurons inhibiteurs GABAergiques. Ces derniers, en libérant le neurotransmetteur GABA, jouent un rôle essentiel dans la suppression de l'activité neuronale, participant à la régulation de l'excitation et au maintien de l'équilibre du réseau neuronal.(143–145) Les canaux Nav, notamment Nav1.2, sont fortement exprimés dans les cellules granulaires du gyrus denté, qui projettent vers la région CA3. Fait intéressant, les fibres moussues établissent des connexions environ 50 fois plus fréquemment avec des interneurons inhibiteurs spécifiques appelés cellules à panier qu'avec les cellules pyramidales excitatrices. Une perte de fonction des canaux Nav diminue l'efficacité de la signalisation inhibitrice, ce qui augmente l'hyperexcitabilité dans la région CA3.(143–145) Cette excitabilité altérée pourrait expliquer l'apparition plus tardive des crises (âge moyen d'apparition à 19 mois pour les duplications, contre 19 jours pour les variants faux-sens), un délai qui correspond à l'expression de ces cellules dans un cerveau plus mature, tel qu'observé chez les patients porteurs d'une duplication de *FGF12*.(125,132–135) Ainsi, comme précédemment suspecté, les patients présentant une duplication de *FGF12* présentaient un début des crises plus tardif, accompagné d'une stagnation ou d'une régression apparente du développement après leur apparition.(132)

La majorité des patients porteurs de variants faux-sens pathogènes dans *FGF12* présentaient une épilepsie pharmacorésistante, avec un EEG montrant principalement des anomalies multi-

focales non spécifiques. Lorsque les données EEG étaient disponibles, la moitié des patients présentaient une activité de fond ralentie, des pointes interictales focales ou multifocales, et plus rarement des schémas discontinus de type suppression-burst. Les EEG ictaux étaient rapportés comme focaux (avec une propagation rapide vers les régions bilatérales) ou généralisés, avec une activité rapide de bas voltage suivie de pointes diffuses et d'une suppression post-ictale.(129,130) À l'inverse, chez les patients présentant des duplications de *FGF12*, nous avons observé un profil EEG plus complexe, avec certaines spécificités communes (Tableau 3). D'abord, au début de l'épilepsie, les deux patients présentaient une altération marquée de l'électrogénèse compatible avec une encéphalopathie sévère, suggérant une étiologie inflammatoire. En effet, tous deux présentaient une activité très lente et de fort voltage, comme celle décrite dans les encéphalites.(146) L'un de nos patients, comme d'autres rapportés dans la littérature, a reçu un traitement immunomodulateur.(137) Chez ces patients traités, une évolution en trois phases a été observée. La première phase correspond à une tempête de crises accompagnée d'une régression cognitive (Figure 36). La phase 2 est une « phase de lune de miel », caractérisée par une réponse partielle au traitement, une période sans crises, et une amélioration notable des performances cognitives. Enfin, la phase 3 se manifeste par une aggravation de l'épilepsie, avec un état de mal non convulsif, une obnubilation et des myoclonies, accompagnés à l'EEG de pointes-ondes diffuses sous-continuelles et de nombreuses ondes triphasiques, évoluant vers un état grabataire.(132,133,137) Les deux patients rapportés dans cette étude ont présenté, au début de l'épilepsie, des crises toniques-cloniques bilatérales et des crises focales temporales. Cette localisation temporale est cohérente avec l'hypothèse d'un déséquilibre entre les signaux excitateurs et inhibiteurs dans ces régions.(143) La cinétique d'apparition et de progression des crises dans les duplications de *FGF12* pourrait suggérer une évolution similaire à celle observée dans les pertes de fonction de Nav1.2 (*SCN2A*), avec un début plus tardif que les formes à gain de fonction de Nav1.2.(143) Le parallèle avec les altérations de Nav1.2 pourrait également s'étendre aux approches thérapeutiques. Tout comme les patients porteurs de variants faux-sens pathogènes, ceux porteurs de duplications ont montré une résistance marquée aux traitements antiépileptiques, malgré de nombreuses stratégies thérapeutiques. Bien que la phénytoïne ait démontré une efficacité partielle chez les patients porteurs de duplications de *FGF12*, son effet global était limité et n'a pas permis d'amélioration clinique significative.(130,132,133)

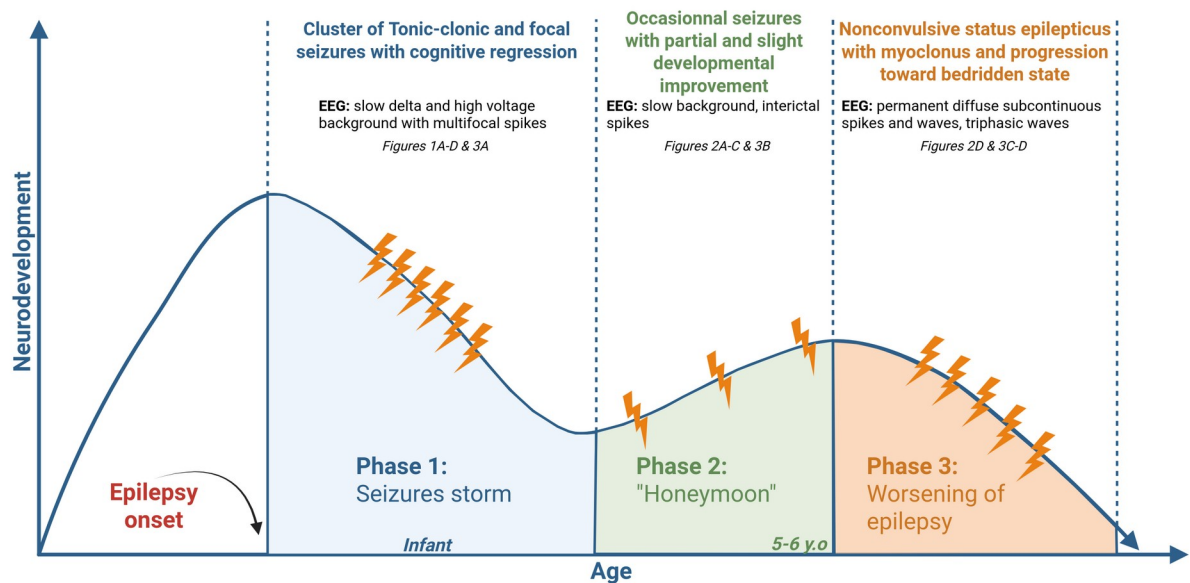


Figure 36 : Évolution en trois phases de l'épilepsie et du développement neurodéveloppemental. Après un développement initial, les patients connaissent une phase de crises convulsives intenses (Phase 1), suivie d'une amélioration transitoire (Phase 2), puis d'un déclin ultérieur marqué par un état de mal épileptique non convulsif et une détérioration motrice (Phase 3).

Nous avons fourni la première caractérisation approfondie des duplications de *FGF12* en combinant une approche de séquençage *long-read* sur l'ADN avec une analyse sur l'ADNc, permettant de révéler la réorganisation structurale complète induite par des mécanismes de recombinaison homologue non-allélique (NAHR) ainsi que son impact sur les transcrits. Cela a conduit à l'identification de deux isoformes aberrantes distinctes : une isoforme A allongée et une isoforme B avec décalage du cadre de lecture. Les conséquences électrocliniques des duplications de *FGF12* semblent plus spécifiques qu'on ne le pensait jusqu'à présent, comparées aux variants faux-sens pathogènes ponctuels, avec une régression clinique observée après le début des crises et des profils EEG distinctifs à l'apparition et au cours de l'évolution, suggérant que ces duplications pourraient être suspectées sur la base des données EEG.

Tableau 3: Tableau clinique des patients nouvellement rapportés et précédemment publiés présentant la duplication récurrente de 500 kb du gène FGF12

	Patient 1	Patient 2	Shi et al. 2017	Verheyen et al. 2020	Abraham et al. 2024	Willemsen et al. (p1) 2020	Willemsen et al. (p2) 2020	Willemsen et al. (p3) 2020	Duplication Synthesis	Missense Variant Synthesis - Trivisano et al.
Patient age at last consultation	6 y.o.	6y.o.	8 y.o.	7 y.o.	22 y.o.	10 y.o.	5 y.o.	30 y.o.	[6-22 y.o.]	[1mhs - 33 y.o.]
Genomic variant	3q28q29(chr3:192,153,010-192,743,885)	3q28q29(chr3:192,154,622-192,744,509)	3q28q29(chr3:192159189-192736896)	3q28q29(chr3:192156263-192739147)	3q28q29(chr3:192155384-192738956)	3q28q29(chr3:192142300-192733325)	3q28q29(chr3:192159179-192736896)	3q28q29(chr3:192159179-192736896)	normal 4, DD 2	
de novo	de novo	de novo	de novo	de novo	de novo	de novo	de novo	de novo		
Age at diagnosis	2 y.o.	5 months	NA	NA	16 y.o.	NA	NA	NA		
Neurological impairment										
Intellectual impairment	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes (100%)	Yes (87.5%)
Early Psychomotor development	Normal	Normal	Normal	Normal	Normal	Developmental delay	Developmental delay	Developmental delay		
Sitting age	NA	9 months	NA	NA	NA	NA	NA	NA		
Walking age	15 months	15 months	NA	NA	NA	NA	NA	NA		
Age of first words	12 months	2 words at 18 months (dad, mom)	NA	NA	NA	NA	NA	NA		
Current language level: oral (words, sentences), written	Absence of language	Few words	Few words	Single words	No speech (before her death at 22)	No speech	Reduced vocabulary, drooling	Speech difficulties	Abnormal speech : 100%	
Autonomy in daily life: cleanliness, eating, dressing	Lack of autonomy	Lack of autonomy	Lack of autonomy, ataxic walking	Loss of practical hand skills, independent walk	Loss the ability to speak and eat and wheelchair	Wheelchair dependant, independently walk	Walk independently at 5 y.o	Care supervisor for her daily activities, limiting w	Reduced autonomy : 100%	
IQ?	Profound ID	NA	Profound ID	NA	Severe ID	NA	Mild to moderate	Mild to moderate	Mild to Profound	Mild to Severe
ASD or autistic traits?	Yes	NA	NA	NA	NA	Yes, level 3	NA	NA	Yes (23%), No (7%), NA (70%)	Yes (43.7%), No (25 %), Other (18.7%)
Neurological signs	Global Hypotonia	NA	Aphasia, Ataxia	hand apraxia	NA	Ataxia	Uncoordinated gait	Unsteadiness		
Behavioral issues	Poor eye contact	short attention span	NA	Inappropriate laughing, and crying	NA	NA	NA	NA		
Sleep disturbance	Yes	no	Yes	Yes	NA	NA	NA	NA		
long-term treatment	DCI, BZD, PB and CZP	2 types of LTG	PHT	LTG, LEV, TPM, PHT and BZD	PHT, VPA and GBP	VPA, BZD, LTG and TPM	VPA and CBZ	VPA and CBZ	DCI (1/8), BZD (3/8), PB (1/8), CZP (1/8), LTC(4/8), PHT (2/8), LEV (1/8), TPM (2/8), VPA (4/8), CBZ (1/8), GBP (1/8)	
Seizures										
Febrile seizure: yes/no	No	no	No	No	No	Yes, 13 months	No	No	No (7/8)	No (100%)
Febrile seizure number	NA	0	0	0	0	0	0	0	NA	NA
Febrile seizure: start/end age	18 months	5 months	3 y.o.	4 y.o.	22 months	15 months	1 month	19 months [1 mhs- 4y.o.]		
Age of onset										
Seizure type	recurrent non-febrile status epilepticus	absence, then clonias	Tonic clonic, brief tonic or complex partial s	Generalized, tonic, and tonic-clonic	Provoked generalized convulsion	Tonic-clonic, atonic, tonic, myoclonic, auto	Cluster of Generalized tonic-clonic	Generalized tonic-clonic		TS (7/16), FTS(8/16), FDTCSQ(16), MSQ(2/16), PSC(16), SS(1/16), SJ(1/16), ESE(2/16)
Status epilepticus	yes	yes	na	yes	yes	Yes	Yes	Yes	yes (3/8)	
Generalized epilepsy yes/no	Yes	yes	yes	Yes	Yes	Yes	Yes	Yes	Yes (100%)	Yes (44%)
Partial epilepsy yes/no	Yes	yes (tempor occipital left)	Yes	NA	No	No	No	No	Yes (07%)	Yes (07%)
Drug resistance	yes	yes	Yes	Yes	No	Yes	No	Yes	Yes (75%)	Yes (44%)
current treatment	CZP DCI PB BZD	LTG 600mg 2-day TPM 2-day CZP	PHT	LTG, LEV, TPM, CLB	No	VPA, BZD, LTG, TPM	VPA	VPA, CBZ		
Frequency of seizure	Multiday	1 per month	Daily seizure at start	NA	Depends	NA	NA	regular breakthrough seizures	frequent 50 %	Frequent (31%), infrequent (0%), No (31%), NA (13%), twice (13%), monthly (8 %)
Treatments already used	VPA, LTG, CZP	Lacosamid, BZD	PB, BZD, VPA, and Potassium bromide	VPA, PHT, CBZ, plasmapheresis	PHT, VPA, GBP	NA	VPA	NA	PB (13/16/13), VPAB(16), PHT(11/16), GVG(5/16), TPM (6/16), CZP (1/16), PNC(16), LEV (9/16), RCM(2/16), PER(16), VNS(3/16), PREG(1/16), CBZ(7/16), LTG(2/16), RUP(1/16), CLB(2/16), RSM(1/16), ZNS(1/16), KME(1/16), RET(1/16), COC(1/16), LCM (1/16), CLZ(1/16), steroids (1/16), PLP (2/16), folic acid (1/16), biotine (1/16)	
Effective treatment			PHT	PHT, LTG, LEV, TPM, CLB	PHT, VPA, GBP	VPA, BZD, LTG et TPM	VPA	VPA and CBZ	PHT (3/8), VPA (4/8), LTG (2/8), LEV (1/8), BZD (2/8), TPM (2/8), CBZ (1/8) and GBP (1/8)	
Aggravating treatment	No	NA	NA	NA	NA	NA	No	NA	non reported	
Regression of acquisition	Yes at 20 months	Yes	Yes	Yes	Yes, 7 y.o	Yes, 3y.o.	No	No	yes (6/8 75%)	
EEG first seizure Background trace	Disorganised background activity	Diffuse delta hypersynchrony of sleep onset, ample, then polymorphic delta rhythm overlaid by sleep spindles, slow islandic spikes	NA	NA	NA	NA	At 13 months: Normal EEG	NA		Interictal : Slow (56 %), Normal (31 %), no description (23%)
EEG first seizure paroxysmal abnormalities	few focal spikes of variable localization, often diffuse over one hemisphere	left temporo-occipital paroxysmal wave spikes	NA	NA	NA	NA	At 13 months: Normal EEG	NA		Interictal : Multifocal SW (62 %), diffuse SW (6 %), Focal SW (6 %) and Suppression Burst (6%)
EEG follow-up (age) Background trace	background of diffuse encephalopathy rich in bi- and triphasic delta activity		Severe suppression of background activity	Background suppression when NCSE. Ex-sleep background, diffuse slowing	At 8y.o: Normal EEG	At 3 y.o: Slowed background activity	na	NA	Abnormal background 4/5	Slow (64 %), Normal (14 %), No description (22 %)
EEG follow-up (age) paroxysmal anomalies	isolated generalized spikes or bursts at 2.5Hz, with atypical absence and concomitant myoclonus		Paroxysmal bursts	Periodic bursts of slow wave activity (NCSE) or Primary generalized and bicoceptual epileptic discharges	At 8y.o: Normal EEG	At 3y.o: Multifocal epileptic discharges	At 2 y.o: Generalized epileptic activity	NA	Paroxysmal abnormality 5/6	Multifocal SW (64 %), Focal SW (21 %), generalized and focal paroxysmal in temporal regions (17%)
MRI										
MRI abnormalities : description and picture if possible	At 4y.o: Cortico-subcortical atrophy	2018: Normal MRI 2024 : Nodular lesion of the same signal as the 5mm cortex of the white matter of the left cerebellar peduncle. Probable gray matter heterotopia	At 8 y.o: Mild cerebral and cerebellar atrophy	NA	At 18 y.o : T2 hyperintensity in the left posterior periventricular white matter consistent with a perivascular space	NA	Mild prominence of the subarachnoid space in the frontal regions bilaterally otherwise normal	Normal		Cerebellar Atrophy (19 %), Normal (38 %), Mild Cerebral Atrophy (6%), Temporal Sclerosis/Cerebellar Atrophy (6%), Cerebral Atrophy (19%), Mild Chiari (6%) and Mild enlargement of lateral ventricle (6%)
Pyramidal signs, spasticity, hyperreflexia	Progressive spasticity of the 4 members	no	No	NA	NA	NA	NA	NA		
Other										
Height (Standard deviation)	At 6 y.o: -1SD	-0.5SD	At birth: -1.2 SD	NA	NA	NA	NA	NA		
Weight (Standard deviation)	At 6 y.o: M	-0.5SD	At birth: -0.5 SD	NA	NA	NA	NA	NA		
Head circumference (Standard deviation)	At 6 y.o: -1SD	NA	At birth: -0.4 RT	NA	NA	NA	NA	NA		
Appetite disorders: food selection, small appetite	No	no	Dysphagia when seizures occurred	NA	NA	Progressive feeding difficulties, reflux	NA	NA		
Enteral nutrition, gastrostomy	Gastrostomy for nutritional indications (12)	no	Gastrostomy	Gastrostomy at one moment	Gastrostomy	Tube feeding (PEG)	NA	NA		
Infection disorders	critical blindness	no	NA	NA	NA	NA	NA	NA	05-anti	
Deafness	NA	NA	Hypoaacusis, transsympatric aerators -- then	NA	NA	NA	NA	NA		
Other ?	Constipation	NA	Dyspnea	No	No	Upper airway infection, Tachycardia during	Dysmorphic features	No		

ID: Intellectual Disability; NA: Not available information; y.o.: years old
NCSE: Non convulsivant status epilepticus

2.2. Etude d'un variant récurrent lié à l'X: IDS

2.2.1. Méthode complémentaire

PCR long-range

Des amorces spécifiques ont été conçues afin d'amplifier et de vérifier l'un des points de cassure associés à l'inversion du gène IDS, ainsi que la séquence sauvage correspondante, chez le patient UPN3, sa mère et un témoin sain. Des longues PCR (long-range PCR) ont été réalisées à l'aide de la polymérase Advantage GC Genomic LA (Takara Bio, Shiga, Japon). Les produits PCR ont ensuite été séparés par électrophorèse sur un gel d'agarose à 2 %.

2.2.2. Résultats

2.2.2.1. Présentation clinique et familiale du patient UPN3

Le patient UPN3 de sexe masculin a été diagnostiqué cliniquement et biochimiquement comme porteur d'une mucopolysaccharidose de type II (MPS II, maladie de Hunter, MIM #309900). Il présentait un phénotype compatible de la maladie(147), incluant notamment une dysmorphie faciale avec des traits grossiers, les mains trapues et courtes avec une brachydactylie des premiers, deuxièmes et troisièmes doigts dans les deux mains, des troisièmes phalanges très courtes et des coudes peu fléchis. Il présentait aussi une hépatomégalie globale de contour émoussé et de structure homogène, une splénomégalie globale et une néphromégalie. Les dosages de glycosaminoglycane urinaire étaient augmentés à 68.6mg/g créatinine, avec un profil électrophorétique évocateur d'une accumulation de dermatane sulfate et héparane sulfate, typique de cette pathologie.

Ce syndrome est un trouble génétique récessif lié à l'X qui touche principalement les hommes associé à des variants localisés dans le gène *IDS*.(147) Après une première analyse moléculaire du centre de référence de diagnostic moléculaire de la mucopolysaccharidose de type II avec un sanger et un panel centré sur le gène associé ainsi que d'autres techniques tels qu'une CGH-array et le séquençage d'un panel 500 gènes (incluant *IDS*), aucune anomalie moléculaire n'a pu être identifiée. Le patient UPN3 est décédé à l'âge de 9 ans, sans que le diagnostic moléculaire ait pu être formellement établi.

Dans ce contexte, bien que le diagnostic clinique ait permis d'envisager la mise en place d'une prise en charge thérapeutique, l'absence de confirmation moléculaire posait une limite majeure pour le conseil génétique. *IDS* étant localisé sur le chromosome X, une transmission maternelle était envisageable, rendant l'identification du variant causal indispensable pour estimer précisément le risque de récurrence. Cette problématique est devenue particulièrement urgente lorsque lorsqu'une grossesse a été détectée chez la tante maternelle du patient. En l'absence de variant identifié, un risque de transmission ne pouvait être exclu, justifiant la nécessité de poursuivre les investigations moléculaires afin de localiser l'anomalie responsable de la pathologie.

2.2.2.2. Diagnostic moléculaire d'une mucopolysaccharidose de type II par séquençage *Long-Read*

Comme mentionné précédemment, différentes approches diagnostiques ont été mises en œuvre, incluant notamment le séquençage d'un panel de gènes comprenant *IDS*, sans mise en évidence de variant pathogène. Dans ce contexte, un LRS-WGS a été réalisé afin d'explorer d'éventuelles anomalies de structure non détectées par les approches de routine. L'analyse a été orientée principalement vers les variants de structure, les variants ponctuels ayant été considérés comme déjà explorés.

Le séquençage a permis de générer au total environ 45Gb égale à une profondeur de 13X et donc environ 6-7X sur le chromosome X (Tableau 2). Dans un premier temps, l'inspection des résultats issus des outils de détection de variants de structure Sniffles2 et CuteSV, ainsi que l'analyse visuelle à l'aide de l'outil IGV, n'a permis d'identifier ni de variant de structure significatif, ni regroupement de lectures chimériques suggérant une anomalie localisée. Toutefois, dans un second temps, l'utilisation de l'outil de visualisation Genome Ribbon (120), ainsi qu'une option spécifique d'IGV (« *Link Supplementary Alignment* »), a mis en évidence plusieurs lectures s'alignant sur des brins opposés au niveau de *IDS*. Trois lectures chimériques, compatibles avec un événement d'inversion, ont ainsi été détectées. (Figure 37A-B) Néanmoins, ces lectures ne s'alignent pas exactement aux mêmes points de rupture et semblent de qualité incertaine. Les cassures impliqueraient les régions situées aux environs de chrX:149,486,089–149,489,234 (GRCh38 ; intron 7 de *IDS*) et chrX:149,520,144–149,529,146 (GRCh38 ; *EOLA1-DT*).

Une analyse plus fine de ces régions a mis en évidence, dans l'intron 7 du gène *IDS*, une séquence de 1 620 pb présentant une homologie de 96,2 % avec une séquence localisée sur le brin opposé au point de cassure identifié dans *EOLA1-DT*. (Figure 37C) Cette forte similarité pourrait expliquer la difficulté d'alignement des lectures chimériques, en raison de la présence de ces régions homologues proches.

Afin de confirmer plus précisément la présence de cette inversion, une PCR long-range a été réalisée, contournant les régions homologues, de manière à amplifier à la fois l'allèle sauvage et l'allèle inversé. Cette PCR a mis en évidence la présence de l'inversion à l'état hémizygote chez le patient UPN3, et a confirmé l'absence de l'allèle sauvage. (Figure 37D-E) Par ailleurs, l'étude de l'ADN maternel a permis d'identifier cette même inversion à l'état hétérozygote, confirmant la transmission maternelle de l'événement. (Figure 37D-E)

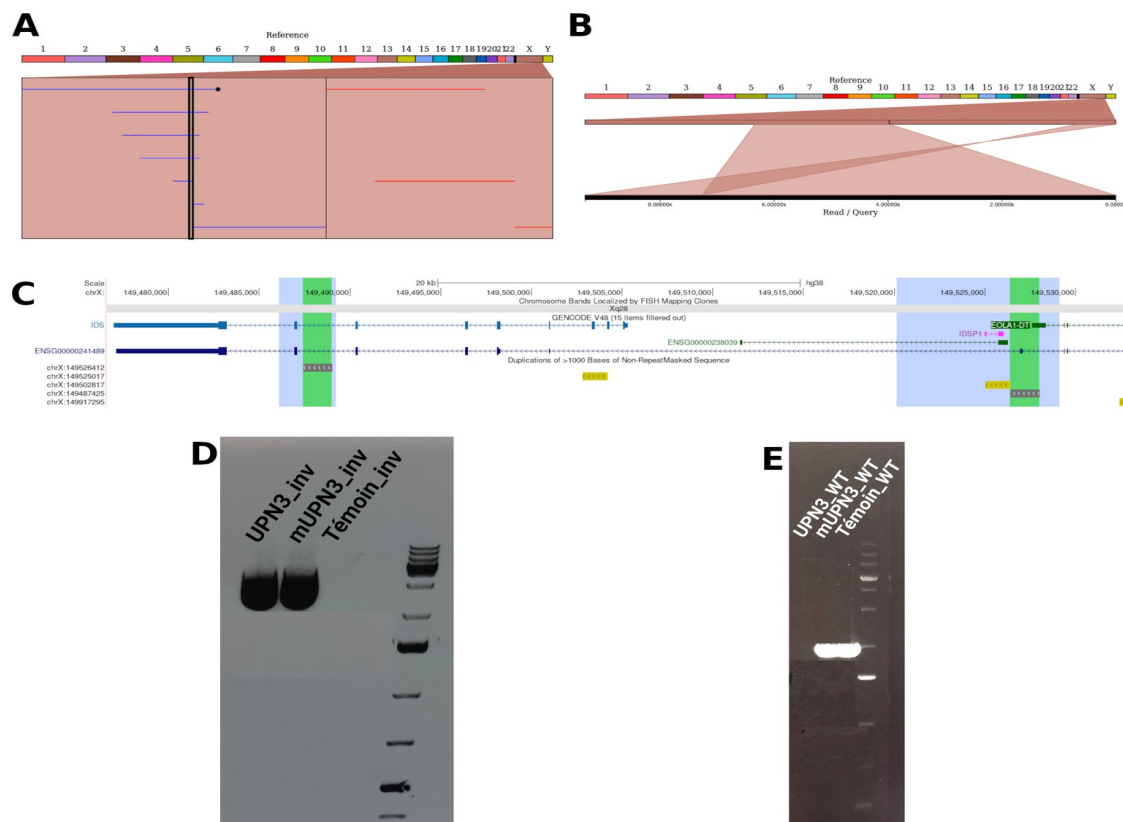


Figure 37: Diagnostic moléculaire de UPN3. (A) Visualisation des reads de séquençage Nanopore via l'outil Ribbon-Genome (B) Visualisation de l'alignement d'un read chimérique spécifique (C) Visualisation sur UCSC genome browser de l'alignement de la région d'incertitude (en bleu) et de la région homologue correspondante (en vert). (D) Électrophorèse des produits de PCR détectant l'inversion du gène *IDS*. (E) Électrophorèse des produits de PCR amplifiant la séquence de type sauvage (WT) de *IDS*.

2.2.3. Conclusion

Le séquençage *long-read* a permis d'identifier des *reads* chimériques dans des régions répétées du génome, caractérisées par un score d'alignement très faible. Bien que les données obtenues restent approximatives et ne puissent être considérées comme totalement concluantes en l'absence de seconde technique de validation, elles ont permis de formuler une hypothèse moléculaire cohérente avec le diagnostic clinique initial.

Il est important de noter que le séquençage *long-read* présente certaines limitations, notamment dans des régions génomiques complexes, mais celles-ci peuvent être atténuées par diverses stratégies, telles qu'une augmentation de la couverture ou l'utilisation de méthodes de *basecalling* plus performantes. Dans ce cas, la profondeur de séquençage sur le chromosome X était d'environ 6–7×, une couverture relativement faible qui souligne néanmoins la sensibilité de cette technologie pour la détection des variants de structure. Une couverture plus élevée aurait très probablement amélioré la visualisation et l'interprétation des *reads* chimériques. De plus, le *basecalling* a été effectué en mode *high-accuracy*, et non en mode *super high-accuracy*, ce qui a pu limiter la qualité des appels dans les régions problématiques. La forte homologie entre les séquences impliquées a également pu constituer un obstacle à un alignement précis.

Dans ce contexte, bien que l'approche *long-read* n'ait pas permis une détection non biaisée de l'inversion, elle s'est révélée précieuse pour explorer et soutenir une hypothèse moléculaire ciblée, fondée sur les données cliniques. Cette hypothèse a pu être confirmée par une technique indépendante, à savoir la PCR.

L'association des données issues du séquençage *long-read* et de la PCR a donc permis de confirmer la présence d'une inversion hémizygote emportant *IDS*, en concordance avec le diagnostic clinique. Cette inversion, déjà décrite dans la littérature, présente un point de cassure situé dans l'intron 7 de *IDS* et dans une région de l'ancien pseudogène *IDS-2*, aujourd'hui pseudogène nommé *IDSP1*, et implique une séquence homologue d'environ 1,6 kb. (148) Ce séquençage a tout de même permis de localiser plus précisément le point de cassure au sein du pseudogène *EOLA1-DT*, et non dans *IDSP1*, bien que ces deux gènes présentent une duplication segmentaire. Toutefois, la duplication observée dans *IDSP1* n'est pas similaire à celle retrouvée dans l'intron 7 du gène *IDS*.

Ainsi, ces résultats mettent en évidence un variant probablement récurrent, vraisemblablement induit par un mécanisme de recombinaison homologue non allélique (NAHR), facilité par la présence de deux régions homologues inversées.

Sur la base de cette observation, des amorces spécifiques ont été conçues pour permettre la confirmation moléculaire de cette inversion récurrente, ouvrant la voie au développement d'un test diagnostique ciblé pour le syndrome de Hunter. Cette stratégie permettra notamment de tester la présence de l'inversion chez la tante maternelle du patient. Et enfin, l'identification de cette inversion dans d'autres cas cliniques non encore confirmés moléculairement pourrait permettre de déterminer son caractère récurrent et sa pertinence en tant que marqueur diagnostique du syndrome.

3. Étude de variants non récurrents

3.1. Article 2: Publié dans CSBJ : *Combining long-read DNA and RNA sequencing to enhance molecular understanding of structural variations leading to copy gains*(149)

3.1.1. Introduction

Les variants de structure sont des réarrangements génomiques à grande échelle qui affectent des segments d'ADN de 50 paires de bases ou plus. Les SVs sont classés en deux grands types selon leur effet sur le dosage de l'ADN : les variants déséquilibrés et les variants équilibrés. Les variants déséquilibrés, également appelés variants du nombre de copies, incluent les duplications et les délétions qui entraînent des gains et des pertes d'ADN, perturbant le dosage génétique ou l'environnement régulateur. En revanche, les variants équilibrés, tels que les inversions et les translocations équilibrées, ne modifient pas la quantité totale d'ADN mais réorganisent la structure génomique. Ces réarrangements peuvent affecter l'expression ou la régulation des gènes, en particulier lorsque les points de cassure perturbent des gènes fonctionnellement critiques.(23,34,150)

Une caractérisation précise des SVs est essentielle pour un diagnostic exact. Leur détection repose généralement sur des techniques de cytogénétique conventionnelle et moléculaire, incluant le caryotypage, l'hybridation *in situ* en fluorescence (FISH), les puces d'hybridation génomique comparative et le séquençage short-read.(23) Cependant, ces approches présentent

des limites importantes, notamment pour résoudre les réarrangements génomiques complexes (*Complex Genomic Rearrangement* ; CGRs) et les SVs situés dans des régions hautement répétées du génome. Ces difficultés compliquent la caractérisation précise de certaines duplications, ce qui est pourtant essentiel pour évaluer leurs effets potentiels sur les gènes ou les éléments régulateurs voisins.(151) En conséquence, certains SVs identifiés par SRS ou CGH peuvent rester classés comme des variants de signification inconnue, ce qui pose des défis importants en termes d'interprétation clinique et de prise de décision.(152)

Dans ce contexte, l'utilisation des technologies de séquençage *long-read* pour caractériser des événements complexes a permis de faire progresser de manière significative la compréhension des SVs.(109,153) Les *reads* générées par les technologies LRS varient de 10 à 150 kb, permettant de capturer des SVs de longueur complète ou quasi complète. En réduisant les biais de séquençage et en résolvant l'organisation des régions complexes au niveau de l'ADN et de l'ADNc, les stratégies LRS sont aujourd'hui considérées comme un élément clé de la cytogénomique.(154)

Dans cette étude, nous avons utilisé le LRS à la fois sur l'ADN et l'ADNc afin de caractériser pleinement deux cas de cytogénétique structurale non résolus : un gain de deux copies du gène *ZMYM2* et un gain de copie du gène *REER*. *ZMYM2* code un régulateur transcriptionnel à doigts de zinc impliqué dans la promotion et le maintien de l'identité cellulaire.(155) Les variants *ZMYM2* avec perte de fonction monoallélique provoquent un syndrome craniofacial autosomique dominant avec anomalies rénales et cardiaques (NECRC, MIM #619522). De plus, certaines données suggèrent un mécanisme dominant négatif potentiel via une fuite du mécanisme de dégradation des ARNm non-sens (NMD).(156) *REER* code pour un corégulateur des récepteurs nucléaires impliqué dans la régulation du signal de l'acide rétinoïque.(157) Les variants pathogènes de *REER* sont associés à un trouble neurodéveloppemental autosomique dominant caractérisé ou non par des anomalies cérébrales, oculaires et cardiaques (MIM #605226). La plupart des variants pathogènes dans *REER* entraînent une perte de fonction, bien que des phénotypes sévères aient également été rapportés avec des variants faux-sens spécifiques.(157,158)

Notre objectif était donc d'intégrer une approche combinée ADN et ADNc par LRS afin de préciser la classification de variants de structure précédemment désignés comme des variants de signification inconnue.

3.1.2. Matériels et méthodes supplémentaires

Des échantillons de sang périphérique ont été prélevés chez les individus affectés et leurs parents après obtention du consentement éclairé. Cette étude a été approuvée par le comité d'éthique du Comité de Protection des Personnes (CPP), référence #2023-A00473–42. Un consentement éclairé écrit a été obtenu de tous les participants ou de leurs représentants légaux.

3.1.2.1. Descriptions cliniques

UPN4 est un garçon de 2 ans, né de parents non apparentés. L'enfant est né à 39 semaines de gestation après une grossesse durant laquelle une cardiopathie congénitale avait été suspectée, mais non confirmée après la naissance. À l'âge de 15 mois, son poids était de 9,7 kg (-1 DS), sa taille de 75 cm (-2 DS) et son périmètre crânien de 44 cm (-3 DS). Les caractéristiques cliniques incluent une hypotonie axiale débutant à 2 mois, des troubles de l'alimentation, un retard de développement modéré, une microcéphalie postnatale et une cryptorchidie. L'analyse CGH a identifié un gain de deux copies en 13q12.11 englobant les gènes *PSPC1*, *ZMYM5* et *ZMYM2*.

UPN5 est une fille de 11 ans, née de parents non apparentés, après une grossesse à terme sans complication. À l'âge de 10 ans, elle pesait 41 kg (+1 DS), mesurait 1,46 m (+2 DS) et avait un périmètre crânien de 55 cm. Elle présentait un retard de langage, ses premiers mots apparaissant vers l'âge de 3 ans. Elle présentait également un retard global et une déficience intellectuelle modérée. Un gain de copie en 1p36.23 englobant les gènes *RERE* et *ENO1* a été initialement identifié par CGH.

3.1.2.2. Cartographie optique du génome (OGM)

L'ADN de poids moléculaire ultra-élevé a été isolé à partir de sang total de patient selon le protocole Bionano Prep SP pour sang humain congelé (Bionano Genomics, San Diego, CA, USA). En résumé, l'ADN a été marqué à l'aide de l'enzyme DLE-1 au niveau du motif CT-TAAG, et la structure de fond a été colorée en suivant le protocole Direct Label and Stain (DLS) de Bionano Genomics. L'ADN marqué a été quantifié à l'aide du test Promega Quanti-

fluor HS, puis chargé sur une puce Saphyr pour linéarisation et imagerie sur l'instrument Saphyr.

3.1.2.3. Librairie de séquençage ADNc *Long-Read* Nanopore

L'ARN a été extrait des LCL à l'aide du kit *PureLink RNA Minikit* (Invitrogen, Carlsbad, CA, USA). Pour l'expérience de séquençage ONT, le protocole de séquençage direct de l'ADNc (SQK-LSK114 kit) a été suivi conformément aux instructions du fabricant.

En résumé, pour chaque patient, 2 à 4 µg d'ARN total ont été rétrotranscrits à l'aide d'amorces spécifiques au patient, suivis d'un changement de brin afin de produire un ADNc pleine longueur. Les bibliothèques préparées ont ensuite été chargées sur des *flowcells* PromethION, une par patient, pour le séquençage.

Le *basecalling* a été effectué avec Guppy (version 6.4.6) en mode *High Accuracy* (HAC). Les lectures obtenues ont été alignées sur le génome de référence humain GRCh38 à l'aide de Minimap2, avec des paramètres adaptés aux données transcriptomiques.⁽¹¹⁷⁾ Une RT-qPCR a été réalisée pour étudier la distribution des différents transcrits.

3.1.2.4. Analyse cytogénétique conventionnelle

Le caryotype a été réalisé par bande RHG sur des étalements en métaphase préparés à partir de sang périphérique de la patiente UPN5 selon les procédures standard. Une analyse FISH a été effectuée sur des noyaux en métaphase et en interfase provenant de lymphocytes sanguins à l'aide des clones fosmides G248P82408E3, G248P86196F8, G248P82014D9 et G248P89339F11 pour UPN5.

3.1.2.5. Séquençage Hi-C

Le séquençage Hi-C a été réalisé pour UPN5 sur des LCLs à l'aide du kit Arima HiC pour lignées cellulaires de mammifères (Arima Genomics, Carlsbad, CA, USA). En bref, 10 millions de cellules ont été réticulées avec 2 % de formaldéhyde, lysées et digérées avec les enzymes fournies dans le kit Arima. Les fragments obtenus ont été marqués à la biotine, ligaturés, puis la réticulation a été inversée. Les fragments d'ADN marqués à la biotine ont ensuite été fragmentés à une taille d'environ 400 pb à l'aide d'un instrument Covaris S220 (Covaris LLC,

Woburn, MA, USA) et de tubes AFA. Les fragments marqués ont ensuite été enrichis de manière sélective avec les billes fournies dans le kit. L'ADN a ensuite été traité selon les étapes de réparation du protocole Arima à l'aide du kit *NEBNext Ultra II DNA Library Prep* (New England Biolabs, Ipswich, MA, USA). Les librairies de séquençage ont été préparées par ligation d'adaptateurs et d'index à l'aide du kit *NEBNext Multiplex Oligos for Illumina* (New England Biolabs). La purification des produits PCR et la sélection de taille ont été réalisées avec les billes Agencourt AMPure XP (Beckman Coulter, A63881). Les librairies ont été soumises à un séquençage profond (~300 millions de lectures) en *run paired-end* 150 pb sur un séquenceur NextSeq550 (Illumina, San Diego, CA, USA). Les lectures appariées issues du séquençage ont été alignées à l'aide du Burrows-Wheeler Aligner (BWA-MEM 0.7.17).(159) Les paramètres ont été définis selon la documentation de HiCEXplorer 3.7.3.(160) En bref, les lectures appariées du Hi-C ont été alignées séparément avec des paramètres définissant une forte pénalité d'extension des gaps (-E50) et une pénalité de clipping en 5' et 3' à zéro (-L0). Comme recommandé, le score de correspondance a été fixé à 1 (-A1) et la pénalité de *mis-match* à 4 (-B4). Les alignements ont été convertis de fichiers SAM à BAM puis triés à l'aide de Sambamba.(161) Pour l'exploration des données, des matrices de contact Hi-C ont été produites à différentes résolutions à l'aide de HiCEXplorer. Les résolutions de 50 kb et 1 Mb ont été sélectionnées respectivement pour la visualisation des contacts locaux/TADs et des contacts interchromosomiques.

3.1.3. Résultats

3.1.3.1. Le séquençage *long-read* permet une caractérisation à l'échelle nucléotidique d'une triplication impliquant le gène *ZMYM2*

Un gain de deux copies en 13q12.11, englobant les gènes *PSPC1*, *ZMYM5* et *ZMYM2*, a initialement été identifié par CGH chez UPN4 (Figure 38A). Ce résultat a ensuite été confirmé par qPCR ciblée et considéré comme *de novo* (Figure_Annexe 5A). Un séquençage *long-read* utilisant la technologie Nanopore a été réalisé pour caractériser cette réorganisation. L'ensemble des données de séquençage est présenté dans le Table_Annexe 4. L'analyse visuelle de la profondeur de lecture et les résultats des outils de détection de variants de structure ont confirmé la présence du gain de copies (Figure 38B). Des lectures chimériques ont permis d'identifier les points de cassure aux positions chr13:19 680 196–20 033 950, chaque gain de copie s'étendant sur 353 755 pb et étant organisé selon une configuration tandem (Figure

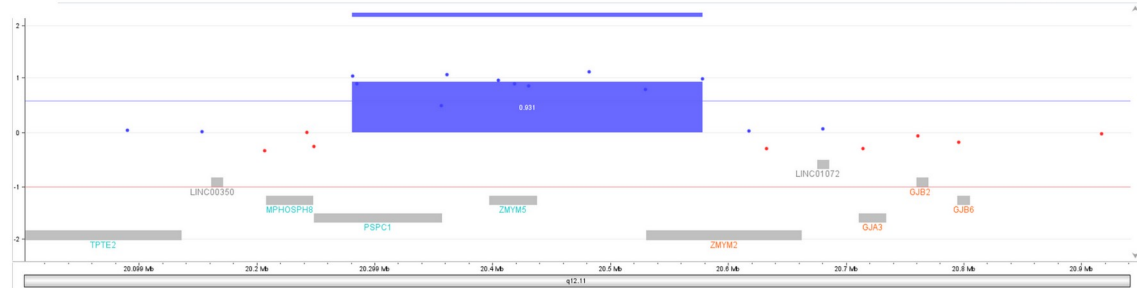
38C). Cette duplication inclut l'intégralité des gènes *PSPC1* et *ZMYM5*, tandis que le gène *ZMYM2* est tronqué au niveau de l'intron 10 (Figure 38D-E). Ce gain de copies n'a pas été rapporté dans les bases de données gnomAD CNV ou DGV. Si *ZMYM2* est un gène déjà associé à une pathologie, aucune altération de *PSPC1* ou *ZMYM5* n'a été rapportée à ce jour. De plus, les scores prédictifs de sensibilité aux gains ou pertes de copies n'indiquaient pas de potentiel pathogène significatif pour *ZMYM5* ou *PSPC1*.(162)

L'analyse des séquences chimériques a révélé les jonctions en amont et aval du gain de copies, avec une microhomologie de 3 pb au niveau des points de cassure, validée par séquençage Sanger (Figure_Annexe 5B). Par ailleurs, l'absence d'éléments répétitifs similaires dans la région et aux points de cassure suggère qu'un mécanisme de recombinaison homologue non allélique (NAHR) est peu probable (Table_Annexe 1 et Table_Annexe 2).

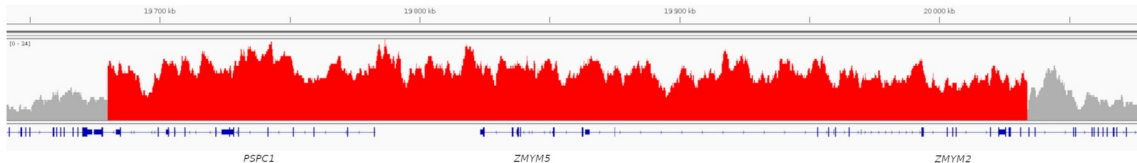
La caractérisation fine de la structure obtenue par LRS a confirmé la nature en tandem de la réorganisation. Cette analyse a révélé deux allèles WT de *ZMYM2*, en plus de deux allèles tronqués contenant les exons 1 à 10. Au vu des scores prédictifs issus de l'outil pTriplo et de l'absence de cas précédemment rapportés présentant un gain de copies affectant ces gènes, aucune investigation complémentaire de l'impact sur *PSPC1* et *ZMYM5* n'a été jugée nécessaire.

L'analyse sur ADNc par séquençage *long-read* Nanopore a confirmé la présence simultanée de transcrits complets de *ZMYM2* et de transcrits tronqués anormaux s'arrêtant à proximité de l'exon 10 avec des signaux de polyadénylation (Figure 38E). Une RT-qPCR a été réalisée pour quantifier la répartition entre les transcrits complets et tronqués, révélant un ratio de 50:50 chez le patient. Ce résultat suggère un décalage du cadre de lecture (*frameshift*) induit par le variant de structure de type triplication impliquant *ZMYM2* (Figure 38F-G).

A Comparative genomic hybridization assay of 13q12.11 locus



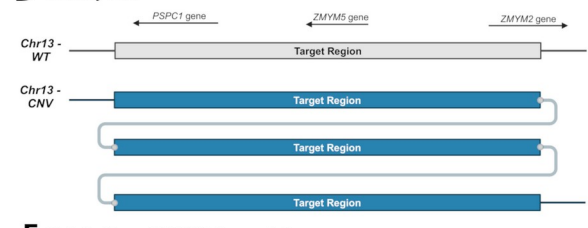
B Sequencing depth of Long-Read Sequencing data



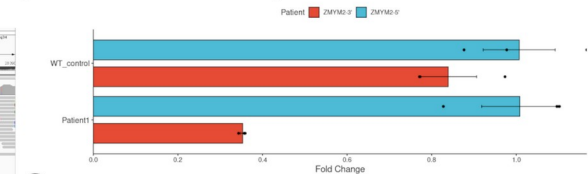
C Ribbon Visualization of chimeric reads observed in the 13q12.11 locus



D Subway Plot



F Distribution of ZMYM2 transcripts



G Plot of total ZMYM2 transcripts

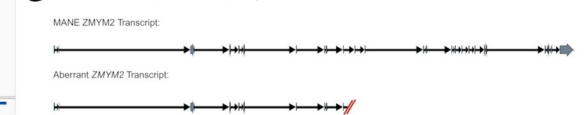


Figure 38: Analyse moléculaire du gain de deux copies chez UPN4. (A) L'analyse par hybridation génomique comparative (CGH) révèle un gain de 0,3 Mb englobant l'intégralité des gènes PTPN12 et ZMYM5, ainsi qu'une duplication partielle du gène ZMYM2. (B) Le profil de couverture issu du séquençage long-read Nanopore de l'ADN montre un gain de deux copies couvrant les gènes PTPN12, ZMYM5 et les exons 1 à 10 de ZMYM2, avec identification précise des points de cassure. (C) La visualisation des données de séquençage long-read via Ribbon Genome Viewer met en évidence un événement de triPLICATION en tandem. (D) Schéma des allèles sauvage et tripliqué, avec mise en évidence des régions dupliquées (zones encadrées). (E) Visualisation dans IGV (Integrated Genome Viewer) des données de séquençage long-read de l'ADN et de l'ADNc, montrant les points de cassure génomiques et leur impact sur les transcrits de ZMYM2. (F) Analyse par RT-qPCR quantifiant les transcrits totaux (barres bleues, ciblant la région 5' de ZMYM2) et les transcrits de type sauvage (barres rouges, ciblant la région 3' de ZMYM2). (G) Représentation schématique des transcrits ZMYM2 issus de l'annotation MANE (Matched Annotation from NCBI and EMBL-EBI) et des transcrits aberrants.

3.1.3.2. Le séquençage *long-read* révèle un variant structurel hautement complexe impliquant le gène *RERE*

L'analyse CGH a mis en évidence un gain de copies en 1p36.23 d'une taille de 499 kb, incluant les gènes *RERE* et *ENO1* UPN5 (Figure 39A). Un examen plus approfondi a révélé la présence de trois sondes consécutives au sein de cette région présentant un ratio logarithmique d'environ 1, suggérant un second gain de copies inclus dans le premier, aboutissant à une configuration DUP-TRP-DUP. Le réarrangement a ensuite été confirmé par qPCR ciblée et considéré comme *de novo*. Ce gain de nombre de copies spécifique n'a pas été rapporté auparavant dans les bases de données gnomAD CNV ni DGV.(163,164) Un séquençage *long-read* utilisant la technologie Nanopore a été réalisé afin de caractériser plus finement le réarrangement. L'ensemble des données de séquençage est présenté dans la Table_Annexe 4. L'analyse de la profondeur de couverture a confirmé la structure déséquilibrée DUP-TRP-DUP, avec des segments de 87 189 pb, 213 819 pb et 246 772 pb, nommés « B », « C » et « D », respectivement, conformément aux résultats de la CGH (Table_Annexe 1). Trois CNV de plus petite taille ont également été identifiés : un gain de copies en amont du segment « B », nommé « A », de taille de 1 871 pb, et deux autres gains situés 7 Mb en aval, au locus 1p36.21, de tailles de 46 495 pb et 1 034 pb, nommés « E » et « F », respectivement (Figure 39B ; Figure_Annexe 6 et Figure_Annexe 7). Ces variants ont été validés par qPCR, avec des valeurs de fold-change de 1,5, compatibles avec une duplication, et de 2,0, compatibles avec une triplication. (Figure_Annexe 8A). Pour plus de clarté, nous utiliserons désormais la désignation A-B-C-D-E-F pour chacun des segments génomiques (Figure 39C ; Figure_Annexe 7). L'analyse fine des lectures chimériques a identifié plusieurs organisations impliquant ces différents segments, notamment une interaction entre les loci 1p36.23 et 1p36.21, ainsi qu'entre ces deux loci et la région 22q11.2 (Figure 39C). Ces observations suggèrent un réarrangement de structure plus complexe.

Dans cette configuration, deux hypothèses de réarrangement étaient envisageables : une insertion sur le chromosome 22 ou une translocation réciproque. Des analyses cytogénétiques complémentaires ont donc été réalisées pour trancher entre ces deux scénarios. Le caryotype et la peinture chromosomique ont mis en évidence la présence du 22pter sur le 1pter et du 22qter sur le 1qter, soutenant l'hypothèse d'une translocation réciproque (Figure_Annexe 8B-C). L'OGM et la FISH ont également été réalisés avec des sondes ciblant les régions 'B', 'C', 'D' et 'E'. L'OGM a identifié huit points de cassure, tandis que les régions 'A' et 'F' n'ont pas été

détectées en raison de leur petite taille, et cette technique a aussi confirmé la translocation t(1;22) impliquant une région hautement répétée du *locus* 22q11.2 (Figure_Annexe 8D–F). Ainsi, la variation structurale initialement identifiée comme une duplication du gène *REER* s’est avérée impliquer des chromosomes dérivés 1 et 22 résultant d’une translocation t(1;22), avec deux régions complexes déséquilibrées aux deux points de cassure incluant des segments génomiques de la région *REER*.

La configuration du réarrangement indique la présence d’un allèle sauvage de *REER* et de deux allèles aberrants dérivés des chromosomes der(1) et der(22). L’analyse des transcrits par LRS a révélé la présence d’un transcrit de pleine longueur ainsi que deux transcrits aberrants distincts (Figure 39D-E). Le premier transcrit aberrant correspond à un alignement avec le gène *REER* jusqu’à l’exon 10, poursuivie sur le chromosome 22, correspondant au dérivé der(1). Le second transcrit aberrant s’étend jusqu’à l’exon 12 du gène *REER*, puis s’aligne avec l’exon 3 du gène *FHAD1*, suivi des exons 2 à 7 de l’ARN non codant *FAM230G*, correspondant au chromosome der(22). Il est à noter que la séquence terminale de l’exon 12 de *REER* et la séquence initiale de l’exon 3 de *FHAD1* sont phasées, suggérant la possible formation d’un transcrit de fusion impliquant *REER*, *FHAD1* et *FAM230G* (Figure 39E). Une RT-qPCR a été réalisée pour quantifier la proportion relative des transcrits de pleine longueur et des transcrits aberrants, révélant un ratio de 50:50 chez le patient (Figure 39F). Ces résultats suggèrent donc un décalage du cadre de lecture induit par la réorganisation complexe impliquant *REER* sur der(1) et der(22).

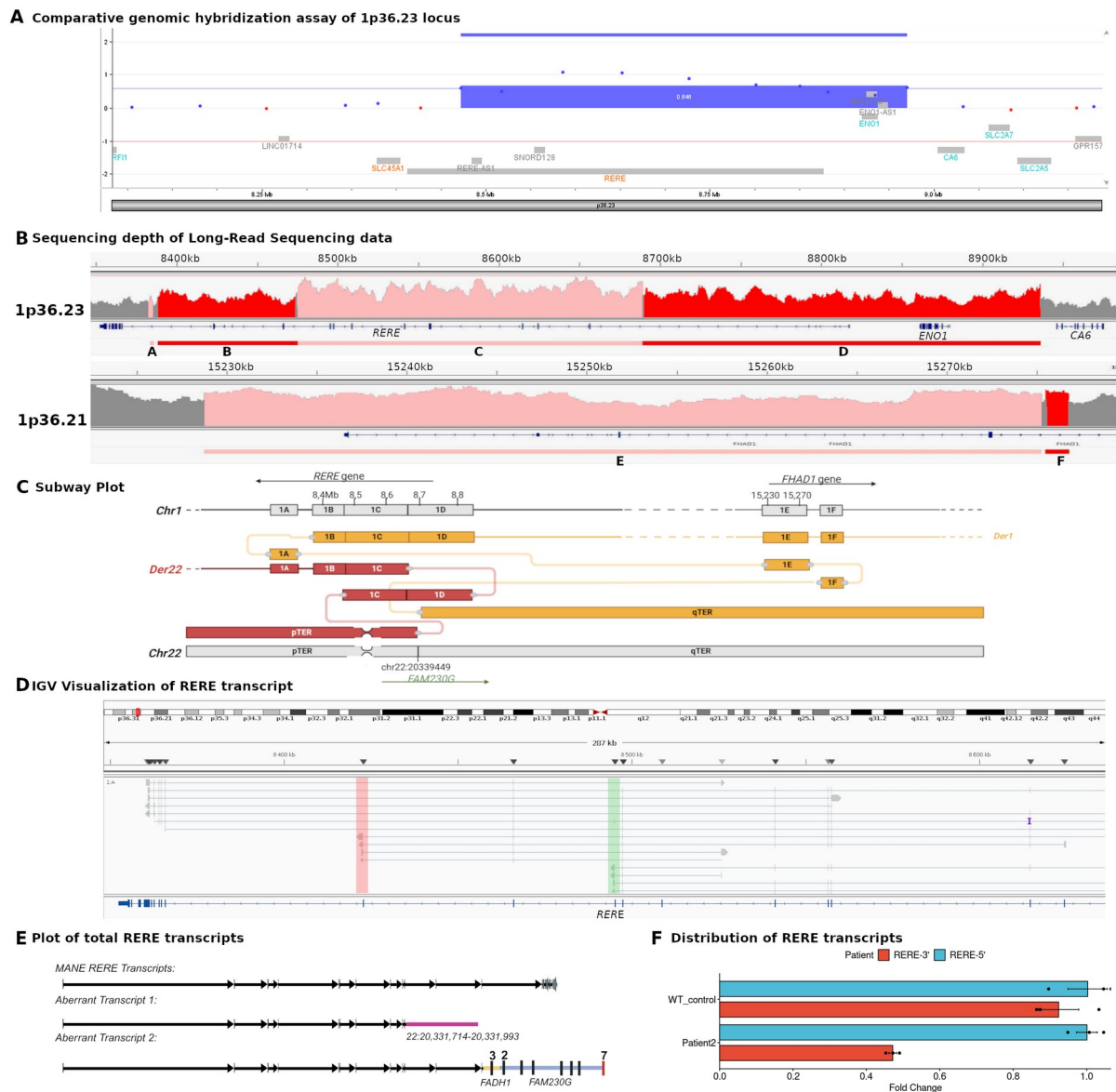


Figure 39: Analyse moléculaire du réarrangement complexe du gène *RERE* chez UPN5. (A) Analyse par hybridation génomique comparative identifiant un gain de 0,5 Mb incluant l'ensemble du gène *ENO1* et une duplication partielle du gène *RERE*. (B) Profil de couverture issu du séquençage ADN long-read Nanopore, montrant des gains de copies, désignés A à F. (C) Schéma récapitulatif (« subway plot ») représentant l'ensemble du réarrangement complexe observé chez le patient. Les structures grises représentent les chromosomes de type sauvage, les segments jaunes correspondent au chromosome dérivé 1, et les segments rouges au chromosome dérivé 22. (D) Visualisation IGV des données de séquençage d'ADNc Nanopore, montrant à la fois le transcript complet de type sauvage et des transcrits aberrants de *RERE* s'arrêtant aux exons 10 et 12. La barre verte indique l'exon 10 de *RERE*, et la barre rouge l'exon 12. (E) Schéma des transcrits MANE et aberrants du gène *RERE*. (F) Analyse par RT-qPCR quantifiant l'ensemble des transcrits (barres bleues, ciblant la région 5' de *RERE*) et les transcrits de type sauvage (barres rouges, ciblant la région 3' de *RERE*).

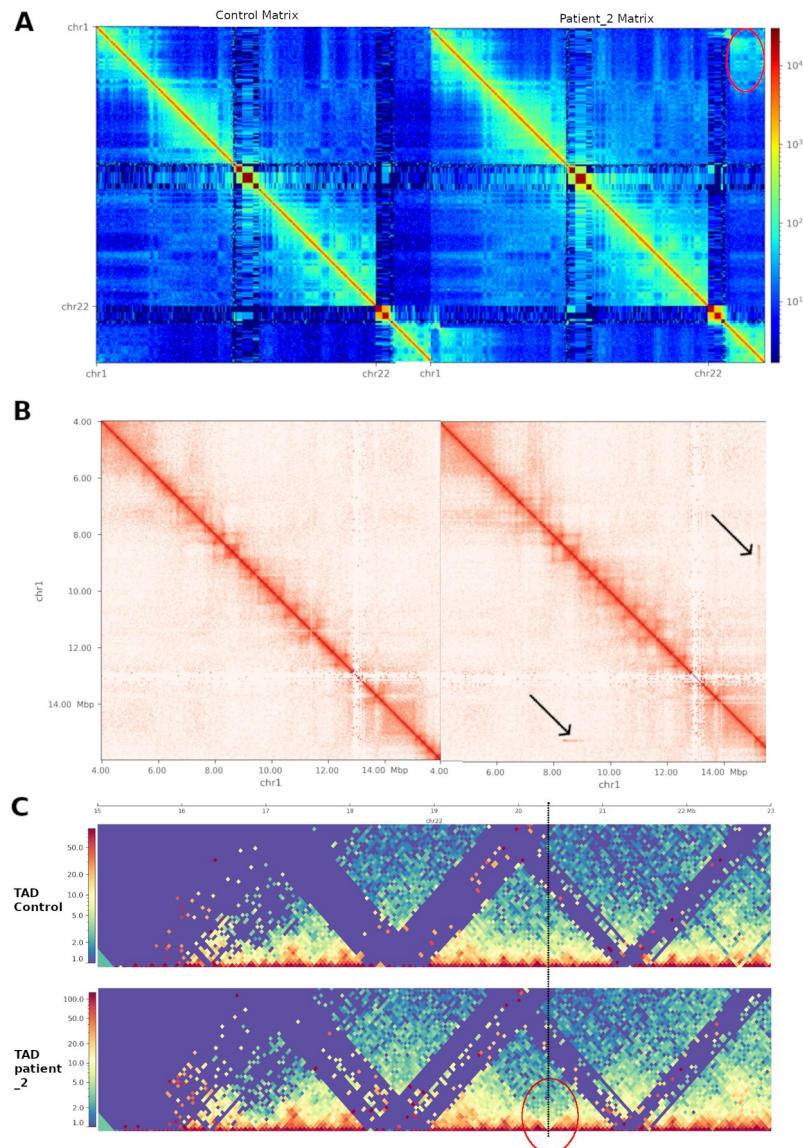


Figure 40: Analyse du génome 3D ciblant la réorganisation structurale de UPN5. (A) Cartes de contact cis Hi-C des chromosomes 1 et 22 comparant un témoin de type sauvage (WT) et de UPN5. Les cadres rouges mettent en évidence des contacts interchromosomiques ectopiques observés entre les chromosomes 1 et 22 chez UPN5 **(B)** Carte de contact cis Hi-C du chromosome 1 chez le témoin WT et UPN5 Les flèches noires indiquent des contacts intrachromosomiques ectopiques au sein du chromosome 1 chez UPN5 **(C)** Carte des domaines topologiquement associés (TAD) du chromosome 22 (chr22:15,000,000–23,000,000 ; hg38) dérivée des données Hi-C pour le témoin WT et l'UPN5. Des altérations dans la structure des TAD sont visibles chez UPN5 par rapport au témoin WT.

Enfin, un séquençage Hi-C a été réalisé afin d'étudier l'effet de ce réarrangement complexe sur l'organisation tridimensionnelle des chromosomes 1 et 22. Des contacts interchromosomiques entre les chromosomes 1 et 22 ont été observés, cohérents avec l'hypothèse de translo-

cation (Figure 40A). Des interactions intrachromosomiques ont également été identifiées sur le chromosome 1 entre les régions 1p36.23 et 1p36.21 (Figure 40B). Compte tenu de la désorganisation de la région 1p36, nous avons exploré la possibilité de nouvelles interactions *enhancer*-gène. Bien que la translocation semble altérer l'architecture de la chromatine, son effet sur les interactions *enhancer*-promoteur semble limité ou négligeable (Figure 40C). Aucune nouvelle interaction *enhancer*-promoteur significative n'a été observée, suggérant une perturbation régulatrice minimale (données non montrées).

2.1.4. Discussion

La caractérisation précise des variants de structure, en particulier des événements complexes impliquant des gains de copies, demeure l'un des principaux défis en cytogénomique. L'émergence de nouvelles technologies de séquençage, notamment le séquençage *long-read* de l'ADN et de l'ADNc, a considérablement amélioré la capacité à élucider l'organisation et les conséquences des SVs (109,153,154,165). Dans cette étude, les approches LRS ont permis la caractérisation de deux CGRs, permettant l'identification précise des points de cassure et la définition claire de la nature des SV.

Le LRS a permis une analyse complète pour UPN4, caractérisant précisément le CNV comme une triplication tandem avec un point de cassure situé dans l'intron 10 du gène *ZMYM2*. Cette approche a localisé la triplication et ses points de cassure à l'échelle nucléotidique, identifiant une microhomologie de 3 pb aux deux points de cassure (Table_Annexe 2, Figure_Annexe 5B). Cette microhomologie, dans le contexte de gains de copies, fournit des éléments clés sur les mécanismes possibles impliqués et suggère fortement l'implication de processus basés sur la réplication, tels que la réplication induite par cassure avec microhomologie (MMBIR) ou le blocage de fourche suivi de changement de matrice (FoSTeS). Ces mécanismes auraient utilisé la microhomologie de 3 pb comme matrice de réparation lors de l'événement de réarrangement.(34,37,44) Ces processus peuvent entraîner des effondrements répétés de la réplication, produisant des variants de structure plus complexes, comme la triplication observée (Figure 38D). Cette triplication en tandem de *ZMYM2* diffère des gains de copies médiés par recombinaison homologue non allélique (NAHR)(23), tels que les triplications de type I, qui ne comportent ni duplications flanquantes ni inversions, ou de type II, qui comprennent typiquement des inversions et des duplications flanquantes.(23,166) Le mécanisme de NAHR ne s'ap-

plique pas ici en raison de l'absence d'éléments répétés dans la région (Table_Annexe 1 et Table_Annexe 2).

Dans des publications antérieures, une microcéphalie a été rapportée chez 4 patients sur 14, des troubles neurodéveloppementaux chez 9 sur 14, une hypotonie chez 3 sur 14, des troubles de l'alimentation chez 2 sur 14, et une cryptorchidie chez 1 patient. UPN4, qui présentait un retard de développement, une cryptorchidie et une microcéphalie, évoquait ce phénotype. (156) Le LRS sur l'ADNc du patient UPN4 a permis d'examiner les conséquences de la réorganisation de *ZMYM2* sur les profils de transcription. Notamment, nous avons observé une répartition 50:50 entre les transcrits sauvages et les transcrits tronqués dans l'échantillon de LCL, suggérant la production d'une protéine *ZMYM2* aberrante. Nous émettons l'hypothèse que les transcrits tronqués de *ZMYM2* échappent à la dégradation du mécanisme de NMD. Comme le suggèrent Connaughton et al., des variants pathogènes tronquants similaires ont été identifiés dans une famille présentant un phénotype comparable. À l'aide d'un test BRET (Bioluminescence Resonance Energy Transfer), qui mesure les interactions protéine-protéine via un transfert d'énergie d'une luciférase à un fluorophore, les auteurs ont démontré que certains variants tronquants pouvaient perturber l'interactome de *ZMYM2*. Si ces variants permettent aux transcrits d'échapper à la NMD, ils pourraient altérer les facteurs de transcription liant l'ADN, les corépresseurs, et des protéines impliquées dans la régulation et l'organisation de la chromatine. (156) Ces altérations pourraient avoir des effets en aval sur des facteurs clés du développement de l'appareil urinaire et contribuer aux malformations urinaires observées chez ce patient. (167) En particulier, des variants tronqués de *ZMYM2* ont été montrés comme interagissant partiellement avec FOXP1, un facteur de transcription impliqué dans les pathologies neurodéveloppementales et rénales. (168) Ces résultats suggèrent que les protéines tronquées de *ZMYM2* pourraient perturber l'interactome natif en entrant en compétition avec les protéines WT, contribuant ainsi au phénotype du patient.

Le séquençage du génome de UPN5 par LRS a révélé un CGR avec plusieurs gains de copies, des inversions intercalées, et une translocation réciproque entre les chromosomes 1 et 22, avec des points de cassure au sein des gènes *RERE* et *FHAD1* (Figure 40C). Le LRS a permis d'identifier précisément tous les gains de copies, leur organisation, et les localisations exactes des points de cassure, y compris leur séquence à la base près. (108,109,113) Malgré la complexité des points de cassure, souvent localisés dans des régions répétées, des microhomolo-

gies ont été identifiées entre les régions génomiques impliquées (Table_Annexe 1 et Table_Annexe 2).

Nous proposons un mécanisme en deux étapes à l'origine de ce réarrangement. Dans une première étape, une réparation par MMBIR/FoSTeS aurait provoqué une chromoanasythèse sur le chromosome 1, conduisant aux CNVs observée. Des microhomologies de 3 à 8 pb auraient pu servir de matrice de réparation. La microhomologie entre les segments 'C' et 'F' n'a pas pu être directement observée, probablement en raison de la perte de la lecture chimérique correspondante, suggérant que cette interaction a été altérée par des événements de réparation ultérieurs.(34,37,44)

Dans une seconde étape, deux cassures double-brin (DSBs) sur les chromosomes 1 et 22 auraient déclenché la formation de la translocation réciproque.(169) Au point de cassure du chromosome 22, une région contenant plusieurs copies d'une séquence de type A – caractérisée par des segments riches en AT et très variables – a été identifiée (Figure_Annexe 9A). (170) Cette région tend à former des structures secondaires cruciformes, ce qui favorise l'instabilité génomique et rend la région plus susceptible aux DSBs (Figure_Annexe 9B).(171) De manière similaire, la cassure sur le chromosome 1 pourrait avoir été induite par une inversion près de la région 'B', susceptible de générer des structures secondaires favorisant l'instabilité génomique.(172)

Une fois ces DSBs survenues, les mécanismes cellulaires de réparation ont été engagés pour restaurer l'intégrité du génome. Sur la base des séquences identifiées aux points de cassure, la jonction d'extrémités non homologues (NHEJ) semble être le mécanisme le plus probable, facilitant la fusion entre 1pTER/22pTER et 1qTER/22qTER (Figure 40C, Figure_Annexe 9C). Ce processus de réparation expliquerait la translocation réciproque et fournit un modèle plausible pour le réarrangement complexe observé. Ce mécanisme en deux temps aurait pu se produire à différents stades du développement : pré-méiotique, méiotique ou post-zygotique. Ces étapes sont critiques pour l'émergence de telles altérations, bien que le timing précis d'apparition reste difficile à connaître.

L'analyse transcriptomique et structurale a permis d'évaluer l'impact de ce réarrangement complexe sur l'expression de *RERE*. Des variants pathogènes monoalléliques de *RERE* sont associés à un trouble neurodéveloppemental caractérisé par des anomalies cérébrales, ocu-

laïres et cardiaques (MIM #605226).(158) Ces variants sont principalement de type perte de fonction. Cependant, des phénotypes sévères ont aussi été associés à certains variants faux-sens localisés dans le domaine atrophine-1, suggérant un possible mécanisme dominant négatif.(157,158) Les données cumulées montrant un transcrit sauvage aux côtés de deux transcrits tronqués aberrants de *RERE* (Figure 40D-E) soutiennent fortement l'haploinsuffisance comme mécanisme pathogène principal.

L'analyse par RT-qPCR, révélant un ratio 50:50 entre les transcrits sauvages et aberrants, met en évidence plusieurs observations intéressantes. Le premier transcrit aberrant, comprenant les exons 1 à 10 de *RERE* et s'alignant sur le chromosome 22, semble être dégradé par la voie NMD.(173) Le second transcrit aberrant, incluant les exons 1 à 12 de *RERE*, l'exon 3 de *FHAD1* et les exons 2 à 7 du lncRNA *FAM230G*, pourrait avoir une pertinence fonctionnelle. Ce transcrit de fusion suggère la possibilité d'une protéine chimérique contribuant au phénotype observé. Notamment, des transcrits de fusion impliquant des lncRNAs ont déjà été décrits comme codant des protéines fonctionnelles en cancérologie. La répartition équimolaire entre ce transcrit et le transcrit sauvage renforce sa possible implication fonctionnelle.(174) En cohérence avec la littérature, les patients porteurs de variants *RERE* menant à une haploinsuffisance présentent souvent un phénotype plus modéré et moins spécifique, comme observé chez UPN5.(157,158) De façon surprenante, ce SV complexe, impliquant plusieurs CNVs, ainsi que des inversions et une translocation, semble avoir un impact limité sur l'environnement régulateur du *locus RERE*, n'induisant qu'un effet dérégulateur modeste. Bien que les données Hi-C suggèrent une réorganisation possible des domaines topologiquement associés (Figure 40C), aucun événement de détournement d'*enhancer* n'a été observé.(64,175,176)

Dans les deux cas rapportés, le LRS a significativement amélioré notre compréhension des CNVs et des CGRs. Bien que ces SV soient survenus *de novo*, ils pourraient être transmis à la descendance. Leur caractérisation détaillée est donc essentielle pour le conseil génétique, permettant une meilleure évaluation du risque reproductif. Cette technologie a permis d'établir un diagnostic moléculaire probable chez ces patients en errance diagnostique. Néanmoins, l'analyse des données LRS dans un contexte diagnostique reste complexe, avec de nombreux défis bioinformatiques. Les outils actuels de détection de SV échouent souvent à identifier tous les points de cassure et à restituer la complexité complète des réarrangements.(177) Les deux outils utilisés ici, CuteSV et Sniffles2, ont identifié respectivement uniquement la duplication de UPN4 et une seule inversion chez UPN5.(118,119) De plus, CuteSV a détecté une

inversion entre les régions 'A' et 'B' et une délétion entre 'A' et 'E'. La détection de cette « délétion » n'est pas surprenante, car elle reflète le contact observé entre ces deux régions. Cependant, ce variant est mal nommé : il ne s'agit pas d'une véritable délétion mais d'une insertion d'un segment dupliqué, ce que l'algorithme n'a pas pris en compte. Cette divergence souligne les limites actuelles des algorithmes de détection, notamment pour les CGRs. Des facteurs tels que la complexité du SV, la qualité du séquençage et la profondeur de lecture jouent un rôle déterminant. Par exemple, une couverture faible ou une région génomique complexe peut empêcher la résolution des points de cassure, entraînant une sous-détection des réarrangements.(177) Le développement d'algorithmes intégrant efficacement la profondeur, la consistance d'alignement, et le contexte de séquence sera crucial pour exploiter tout le potentiel du séquençage *long-read* en diagnostic. Ces avancées pourraient améliorer la détection des points de cassure et la caractérisation fine des réarrangements. Cependant, la curation manuelle et la validation expérimentale complémentaire demeurent indispensables pour garantir une interprétation précise et pertinente cliniquement des SVs identifiés.

Le LRS a démontré sa polyvalence en permettant à la fois une analyse du génome et une caractérisation du transcriptome. Contrairement aux méthodes classiques, il permet le séquençage direct de transcrits complets sans amplification préalable par PCR, préservant les caractéristiques natives de l'ARN et limitant les biais.(178,179) Cette capacité a été essentielle pour identifier précisément les transcrits tronqués et de fusion chez les deux patients, fournissant des informations clés sur l'impact fonctionnel des SV sur l'expression génique. Les techniques SRS ne permettent pas de couvrir directement de grandes régions génomiques, s'appuyant plutôt sur une reconstitution statistique des séquences. Elles ont donc des difficultés à détecter avec précision des événements de fusion ou des réarrangements complexes.(180) L'OGM, bien qu'étant un outil puissant pour détecter les grands SVs et visualiser intuitivement les réarrangements à l'échelle du génome, présente une limite de résolution autour de 5 kb, ce qui peut empêcher la détection de petits événements, comme cela a été observé dans notre étude avec les régions 'A' et 'F'. Dans une étude comparant OGM et LRS par technologie Nanopore, l'OGM détectait 74 % des jonctions dans les cas complexes, tandis que le LRS en identifiait 100 %.(153)

Nos résultats soulignent l'importance d'intégrer le LRS d'ADN et de ADNc dans le diagnostic clinique des maladies rares. Ces approches apportent des informations indispensables sur la nature des SVs, permettent d'émettre des hypothèses sur leurs mécanismes de formation et

leurs effets pathogènes, et améliorent la classification des variants de signification inconnue. En renforçant la précision diagnostique, cette stratégie globale soutient une prise en charge plus personnalisée et efficace des patients concernés.

3.2. Article 3 – En soumission: Deciphering Structural Mechanisms of Copy Gain Variants With Long-Read Sequencing

3.2.1. Introduction

Les variants de structure sont des altérations du génome affectant des segments d'au moins 50 paires de bases et sont généralement classés en variants déséquilibrés ou équilibrés. Les SVs déséquilibrés, également appelés variants du nombre de copies, en particulier les duplications, jouent un rôle important dans les troubles génomiques, mais leur caractérisation précise reste un défi.(23,34) Les puces à hybridation génomique comparative et le séquençage *short-read* sont largement utilisés pour détecter les duplications, mais ces méthodes présentent des limites. L'approche par CGH fournit des informations sur les changements de dosage, mais ne permet pas de déterminer la configuration structurelle des duplications, notamment pour savoir si elles sont en tandem, dispersées ou complexes. De même, le SRS peut avoir des difficultés à résoudre les points de cassure et à différencier des duplications structurellement distinctes, en particulier dans les régions répétitives ou très homologues.(151) Par conséquent, les approches conventionnelles laissent souvent les duplications non résolues sur le plan structural, ce qui complique leur interprétation clinique et la compréhension des mécanismes sous-jacents.

Les technologies de séquençage avancées, telles que le séquençage *long-reads*, offrent une solution prometteuse en permettant une caractérisation structurelle plus précise des variants de structure ou variants du nombre de copies au niveau nucléotidique. Contrairement aux approches de séquençage *short-reads*, le LRS génère des lectures étendues (>10 kb) qui réduisent les biais de séquençage et permettent la résolution complète des SV. Cette technologie permet une analyse plus précise des points de cassure et offre une meilleure compréhension des régions génomiques complexes, améliorant ainsi la classification des variants .(109,153)

Dans cette étude, nous avons appliqué le LRS à l'ADN de sept patients afin de caractériser des variants de gain de copies détectés précédemment par puces CGH et/ou séquençage à

short-reads, d'élucider les mécanismes de réparation sous-jacents à leur formation, et d'identifier des éléments structuraux susceptibles d'aider à définir leur rôle dans la pathogénicité.

3.2.2. Méthodes spécifiques

Séquençage *short-read*

Les échantillons de UPN9 ont été préparés avec le kit Illumina DNA Prep Enrichment (Illumina). Un séquençage exome en lecture appariée de 150 pb a été réalisé sur une plateforme Illumina NovaSeq 6000, avec une couverture moyenne ciblée de 150X, analysée selon les meilleures pratiques GATK. Les lectures ont été alignées sur la version 38 du génome humain de la Genome Reference Consortium (GRCh38) et les variants annotés à l'aide de Variant Effect Predictor (181) en combinaison avec le *pipeline* interne ANATOLE2. Les CNV ont été appelés avec l'outil CNV-DRAGEN (Illumina). Les bibliothèques génomiques en trio de UPN 11 et de ses parents ont été préparées sans amplification (module NEBNext Ultra II End-repair/A-tailing, New England Biolabs) et séquencées en mode lecture appariée sur la plateforme NovaSeq6000 (Illumina). Les lectures ont été alignées sur la référence humaine GRCh38. Les CNV et SV ont été appelés à l'aide de MANTA et ClinSV.(182,183)

3.2.3. Résultats

3.2.3.1 Le séquençage *long-reads* (LRS) identifie des duplications en tandem associées à des mécanismes de réparation par microhomologie

UPN6 est un garçon de 11 ans présentant une déficience intellectuelle, un état de mal épileptique, des comportements auto-agressifs, des stéréotypies et des mouvements de battement des mains. L'analyse par puce CGH a révélé une duplication *de novo* en 21q22.3 ainsi qu'une délétion *de novo* en 20p13, toutes deux confirmées par qPCR ciblée. Si la délétion en 20p13 est probablement le facteur principal du phénotype en raison de la disruption du gène *CSNK2A1*, impliqué dans le syndrome neurodéveloppemental d'Okur-Chung (MIM #617062), le gain de copie en 21q22.3 nécessite une investigation plus approfondie. Ce gain de copie englobe plusieurs gènes morbides référencés dans OMIM (*TSPEAR*, *ITGB2*), suggérant un rôle potentiel dans la modulation de la présentation clinique (Tableau 2). Le LRS a été réalisé, générant 115,88 Gb de données avec une couverture du génome de 35X et une longueur N50 des lec-

tures de 16,07 kb. La délétion hétérozygote terminale sur chr20:1,608,153 a été observée dans les *reads* mais non détectée par les logiciels d'appel de variants. Un gain de copie de 403 327 pb a été identifié par les logiciels Sniffles2 et CuteSV et caractérisé comme une duplication en tandem sur chr21:44,591,282-44,994,608, révélant deux copies adjacentes de la région 21A (Figure 41A et Figure_Annexe 10A-B). L'analyse des points de rupture et de séquence a révélé que les points de cassure de la duplication se situent au sein du gène *TSPEAR* et à proximité du gène *SLX9* (Figure 41B). Une microhomologie de 3 pb, "GAG", a été identifiée à la jonction du point de cassure (Figure 41C). Faute de données supplémentaires informatives, la pertinence pathogénique de ce variant de gain de copie n'a pas pu être clarifiée davantage. Il reste donc classé comme variant de signification inconnu, sans que sa contribution potentielle au phénotype du patient puisse être exclue.

UPN7 est un garçon de 6 ans présentant un retard du développement neurodéveloppemental, une fente labio-palatine, une surdité bilatérale, une hypospadias, un cryptorchidisme et une ectrodactylie (Tableau 2). L'analyse par puce CGH a identifié un gain de copie hérité de la mère au locus Xp22.5, englobant les gènes *STAG2* et *XIAP*, confirmé par qPCR ciblé. Ce gain de copie recouvre partiellement des duplications pathogènes déjà décrites. L'ectrodactylie n'ayant jamais été rapportée chez les patients porteurs de cette duplication, nous avons ré-analysé ce gain de copie par séquençage *long-reads* afin d'identifier d'éventuelles structures additionnelles. Le LRS a généré 68,07 Gb de données avec une couverture du génome de 20X et une longueur N50 des lectures de 12,24 kb. Un gain de copie de 328 128 pb a été identifié par les logiciels Sniffles et CuteSV sur chrX:123,875,069-124,203,250 et caractérisé comme une duplication en tandem, révélant deux copies adjacentes de la région XA (Figure 41D et Figure_Annexe 10C-D). Les points de cassure sont situés dans l'intron 1 du gène *XIAP* et dans une région intergénique entre les gènes *STAG2* et *SH2D1A* (Figure 41E). Une microhomologie de 2 pb, "CT", a été identifiée à la jonction du point de cassure (Figure 41F). Cette duplication reste classée comme pathogène et explique partiellement le phénotype du patient. Toutefois, elle n'explique pas l'ectrodactylie, qui pourrait résulter d'un autre facteur génétique.

UPN8 est une fille de 14 ans présentant une déficience intellectuelle légère, une obésité, une hypothyroïdie, des troubles du comportement, un raccourcissement des mains et des pieds, ainsi qu'une hyperlaxité des coudes (Tableau 2). L'analyse par puce CGH a détecté un gain de copie *de novo* en 6q16.3, confirmé par qPCR ciblé. Ce gain de copie, incluant le gène *GRIK2*, a initialement été classé comme variant de signification inconnue. Le séquençage

long-reads a généré 119,26 Gb de données, couvrant le génome avec une profondeur de 35X et une longueur N50 des lectures de 13,83 kb. Un gain de copie de 917 638 pb a été détecté par Sniffles2, avec des points de cassure en chr6:101,752,983-102,670,620. Le variant a été caractérisé comme une duplication en tandem, révélant deux copies adjacentes de la région 6A (Figure 41Get Figure_Annexe 10E-F). La duplication s'étend de l'intron 7 à la région 3'-UTR du gène *GRIK2*. Une microhomologie de 3 pb, "GAA", a été identifiée à la jonction du point de cassure (Figure 41I). Cette microhomologie a été confirmée par séquençage Sanger (données non montrées). Toutefois, l'analyse par LRS sur ADN n'a pas permis de conclure quant au caractère bénin ou pathogène de ce variant.

UPN9 est un garçon de 14 ans présentant un syndrome cérébelleux, caractérisé par un élargissement modéré des scissures vermiennes, un retard de langage et des troubles de la coordination motrice (Tableau 2). Le séquençage de l'exome a identifié une duplication *de novo* en 5q31.3 (Figure S1H en supplément). Ce gain de copie, classé comme variant de signification inconnue, englobe le gène *PURA*, connu pour être impliqué dans un trouble neurodéveloppemental avec insuffisance respiratoire néonatale, hypotonie et difficultés alimentaires (#MIM 616158). Le séquençage longue lecture a généré 88,1 Gb de données, avec une couverture de l'appel de variants par CuteSV, a permis de localiser un gain de copie de 215 355 pb, avec des points de cassure en chr5:140,028,373-140,243,707, et de le caractériser comme une duplication en tandem, révélant deux copies adjacentes de la région 5A (Figure 41J-Ket Figure_Annexe 10G-H). Une microhomologie de 4 pb, "AGTG", a été identifiée à la jonction du point de rupture (Figure 41L). Cette microhomologie a été confirmée par séquençage Sanger (données non montrées). Toutefois, l'analyse par LRS sur ADN n'a pas permis de conclure sur le caractère bénin ou pathogène de ce variant.

Ces quatre cas illustrent des duplications en tandem associées à des microhomologies aux points de cassure, soulignant une possible implication de mécanismes réparateurs spécifiques dans la genèse de ces variants.

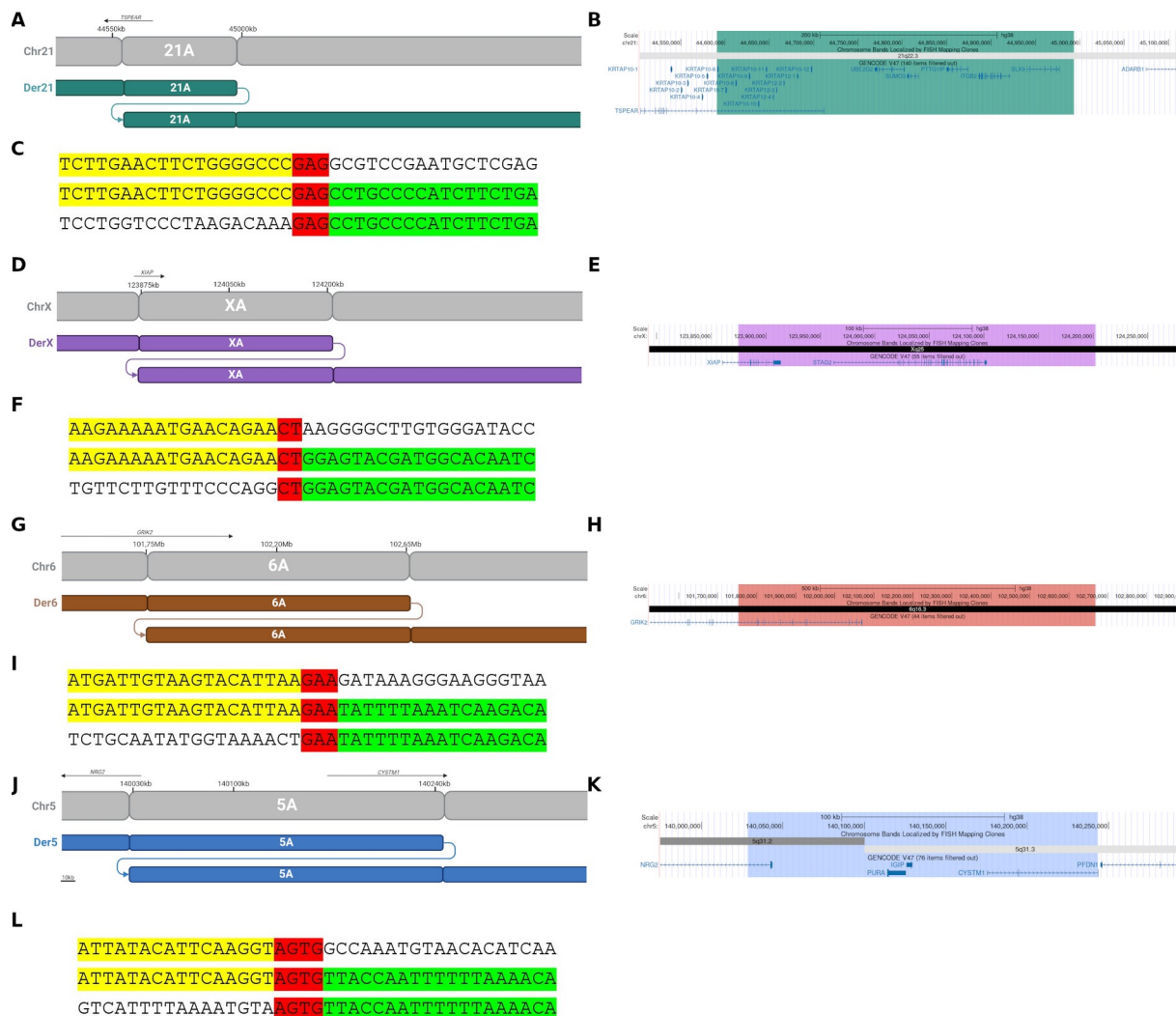


Figure 41 : Analyse moléculaire des variants de gain de copies chez UPN6-9, mettant en évidence des points de cassure associés à des microhomologies. (A, D, G, J) Représentations en “subway plot” des variants de gain de copies en tandem situés sur les chromosomes 21 (450 kb), X (325 kb), 6 (900 kb) et 5 (210 kb), chez UPN6, UPN7, UPN8 et UPN9 respectivement. (B, E, H, K) Visualisations dans le navigateur génomique UCSC des points de cassure localisés dans les gènes *TSPEAR*, *XIAP* (englobant *STAG2*), *GRIK2* et *NRG2/CYSTM1*, chez les mêmes patients respectivement. (C, F, I, L) Analyses des points de cassure mettant en évidence des microhomologies de 3 pb, 2 pb, 3 pb et 4 pb, respectivement chez UPN6, UPN7, UPN8 et UPN9.

3.2.3.2. Le séquençage *long-reads* révèle un mécanisme de réparation marqué par des cicatrices d'insertion aux points de rupture

UPN10 est une fille de 15 ans présentant une déficience intellectuelle modérée (acquisition de la position assise à 10 mois, marche à 28 mois), des traits dysmorphiques, une hypotonie et une hyperlaxité (Tableau 2). L'analyse par puce CGH a mis en évidence une triPLICATION de

novo au locus 21q22.13, confirmée par qPCR ciblée. Ce gain de copie, classé comme variant de signification inconnue et incluant des gènes morbides OMIM, a justifié une exploration complémentaire pour mieux évaluer son rôle potentiel pathogénique. Le séquençage long-reads a généré 95,02 Gb de données, avec une couverture génomique de 29X et une longueur N50 des lectures de 10,82 kb. L'approche combinée avec les outils Sniffles et CuteSV a permis d'identifier deux points de rupture associés, caractérisant structurellement deux gains de copies consécutifs formant une triplication tandem de 482601 pb en chr21:37,230,352-37,712,952 (Figure 42A, Figure_Annexe 11A-B). L'analyse des points de rupture a révélé que ceux-ci étaient localisés dans l'intron 5 du gène *VPS26C* et dans l'intron 3 du gène *KCNJ6*, englobant le gène *DYRK1A* (Figure 42B). Par ailleurs, une insertion de 26 pb a été identifiée à la jonction des points de rupture, suggérant un mécanisme de formation complexe (Figure 42C).

Cette observation illustre une triplication en tandem complexe, associée à une insertion au point de cassure, suggérant l'intervention de mécanismes réparateurs atypiques dans la formation du variant.

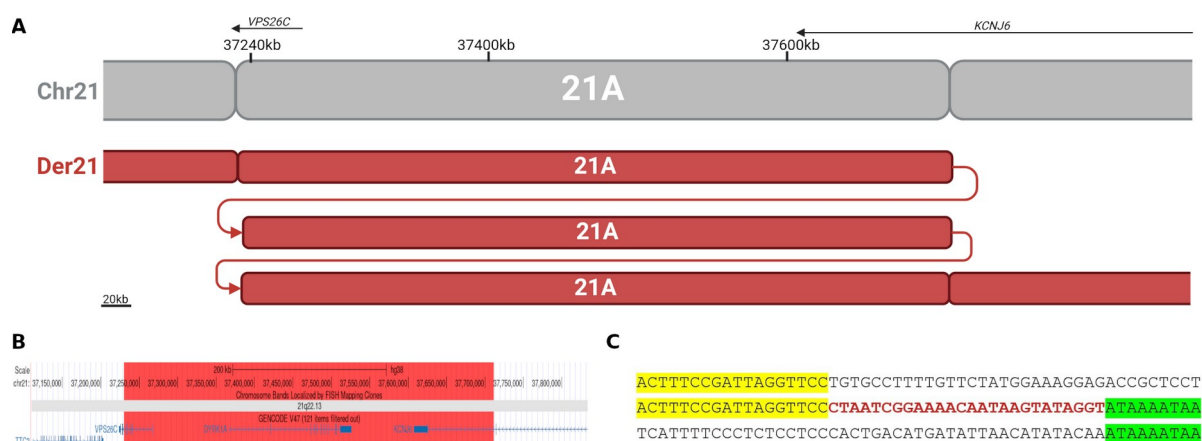


Figure 42 : Analyse moléculaire des variants de gain de copies chez UPN10, mettant en évidence des cicatrices d'insertion aux points de cassure. (A) Subway plot illustrant un variant de gain de deux copies en tandem, localisé sur le chromosome 21, d'environ 460 kb, chez le patient UPN10. (B) Visualisation du variant dans le navigateur génomique UCSC chez UPN10, montrant un point de cassure situé dans les gènes *VPS26C* et *KCNJ6*. (C) Analyse des points de cassure du variant chez UPN10, mettant en évidence une insertion de 26 pb.

3.2.3.3. Le séquençage long-reads permet une analyse approfondie des variants de structure complexes et des mécanismes de réparation

UPN11 est un garçon de 12 ans présentant une déficience intellectuelle et des troubles du comportement, incluant un trouble du spectre autistique et un comportement auto-agressif (Tableau 2). Le séquençage du génome a mis en évidence deux gains de copies rapprochés et inversés au locus Xp11.22, d'origine maternelle (Données supplémentaires). Le séquençage longue lecture a généré 124,29 Gb de données, avec une couverture du génome de 32X et une longueur médiane des lectures (N50) de 8,26 kb. En combinaison avec l'outil Sniffles, deux inversions chevauchantes ont été identifiées: l'une s'étendant sur 290 689 pb (chrX:52 930 637–53 221 326) et l'autre sur 664 895pb (chrX:53 054 017–53 718 912). En parallèle, deux duplications rapprochées ont été détectées sur la base de l'augmentation locale de la profondeur de lecture : la première entre chrX:52 930 637–53 054 016 et la seconde entre chrX:53 221 326–53 718 911. La première duplication est localisée dans une région intergénique et dans l'intron 1 du gène *GPR173*, tandis que la seconde se situe dans l'intron 1 du gène *KDM5C* et dans une région intergénique incluant les gènes *SMC1A* et *HUWE1* (Figure 43A-B). La caractérisation structurale de ces variants indique qu'ils résultent d'un même événement complexe de réarrangement génomique, formant une configuration DUP/INV-NML-DUP. Cette structure implique que ce réarrangement comprend une duplication inversée et qu'elle contient une seconde duplication qui semblent être en contact (Figure 43A et Figure_Annexe 11C-D). L'analyse des points de cassure a révélé qu'une microhomologie de 2 pb, "GC", est partagée entre la fin de la duplication 3B et la fin de la duplication 3A, tandis qu'aucune microhomologie ni insertion/délétion n'a été détectée au point de cassure entre le début des duplications 3A et 3B (Figure 43C-D). Cet événement complexe entraînant une duplication d'une région connue pour sa triplosensibilité, contenant *SMC1A* et *HUWE1*, a été classé comme pathogène.

taires ont été réalisées afin de mieux en caractériser la structure et d'évaluer ses implications potentielles. Le séquençage *long-reads* a généré 119,92 Gb de données, avec une couverture du génome de 32× et une longueur médiane des lectures de 11,96 kb. En combinaison avec les outils Sniffles et CuteSV, l'approche LRS a confirmé la présence du réarrangement complexe et permis sa caractérisation précise, révélant une configuration DUP-TRP/INV-DUP (Figure_Annexe 11E-H). Les points de cassure ont été localisés aux positions chr3:157 023 828–167 491 869 (14 468 042 pb) pour 3A, chr3:167 491 870–167 662 642 (170 773 pb) pour 3B, et chr3:167 662 643–184 386 326 (16 723 684 pb) pour 3C (Figure 43E-F). L'analyse des points de cassure a révélé l'absence de microhomologie entre la fin de 3C et la fin de 3A (*reads* en faveur de l'inversion), tandis qu'une microhomologie de 3 pb ("TTC") a été identifiée entre le début des segments 3A et 3B (*reads* en faveur de l'inversion) (Figure 43G-H). En raison de la taille globale du réarrangement déséquilibré, ce CNV a été classé comme pathogène.

L'ensemble de ces analyses démontre que, même si le variant a été automatiquement classé pathogène de par sa grande taille, le séquençage *long-reads* permet non seulement la détection précise de variants complexes, mais aussi une caractérisation fine des points de cassure. L'identification de signatures moléculaires distinctes dans ces réarrangements souligne la diversité des mécanismes à l'origine de ces événements génomiques complexes.

3.2.4. Discussion

Dans cette étude, nous avons utilisé le séquençage *long-reads* pour caractériser des variants de structure de type gain de copies chez sept patients, soulignant son intérêt pour la résolution de réarrangements génomiques complexes restant partiellement élucidés par les approches conventionnelles telles que le séquençage *short-reads* ou les CGH-array. Le LRS permet une résolution plus fine des points de cassure, identifie des signatures de microhomologie et détecte des architectures structurales complexes, améliorant significativement la classification des variants. La caractérisation précise de ces variations structurelles a des implications cliniques directes, en permettant une évaluation plus fiable de leur caractère pathogène. Notre étude met en évidence l'importance d'approches de séquençage complémentaires pour obtenir de nouvelles informations contribuant à un diagnostic moléculaire complet en cytogénomique.

L'analyse par séquençage *long-reads* a permis d'identifier quatre duplications en tandem, une triplification en tandem, une DUP-TRP/INV-DUP, ainsi qu'un réarrangement complexe de type DUP/INV-NML-DUP. Bien que cette étude se soit principalement concentrée sur la caractérisation structurale et la résolution des points de cassure, les informations supplémentaires fournies par le LRS ont permis de réévaluer la classification de plusieurs variants. En effet, la cartographie précise des points de cassure et de l'architecture structurale a facilité la réinterprétation de certains variants de nombre de copies, une étape essentielle pour leur évaluation clinique.

Parmi les quatre duplications en tandem, aucune n'a pu être reclassée sur la seule base des données structurales. Par exemple, la duplication observée chez UPN6, bien qu'elle englobe des gènes morbides listés dans OMIM, demeure un variant de signification inconnue. La duplication identifiée chez UPN7, initialement classée comme pathogène, a été réexaminée en raison de son lien potentiel avec le phénotype d'ectrodactylie. Toutefois, la caractérisation structurale n'a pas permis de conclure à une corrélation phénotypique élargie.

De même, la duplication observée chez UPN8 est restée classée comme VSI. Néanmoins, compte tenu du phénotype spécifique de ce patient, notamment l'obésité, et de la localisation de la région dupliquée, nous émettons l'hypothèse d'une implication possible de *SIM1*. Ce gène est impliqué dans l'obésité et l'hypopituitarisme, justifiant une exploration plus approfondie de son contexte régulateur et de l'impact potentiel sur l'architecture locale de la chromatine.(184)

Chez UPN9, les données de LRS n'ont pas permis de reclasser la duplication. Toutefois, l'implication de *PURA*, un gène associé à des troubles du développement neurologique incluant une hypotonie néonatale, une insuffisance respiratoire et des troubles de l'alimentation rend ce variant digne d'intérêt, et des études fonctionnelles ultérieures pourraient s'avérer informatives.(185)

La triplification en tandem englobant *DYRK1A* chez le patient 5 a été reclassée comme probablement pathogène, en accord avec la sensibilité au dosage de ce gène dans les phénotypes neurodéveloppementaux et compte tenu du caractère *de novo* du variant.

L'analyse du réarrangement complexe DUP/INV-NML-DUP chez UPN11 a conduit à sa reclassification pathogène en raison de l'implication de *HUWE1*, un gène associé au retard intellectuel lié à l'X.(186)

Bien que les analyses fonctionnelles aient été hors du périmètre de cette étude, la résolution structurale apportée par le LRS met en évidence des *loci* candidats pour des investigations futures.

Notre étude apporte des éclairages mécanistiques sur la formation des variants de structure, en soulignant le rôle des processus de réparation de l'ADN et de la réplication. Le séquençage *long-reads* a permis la détection de caractéristiques des points de cassure, telles que des microhomologies et des insertions, révélant ainsi les empreintes de mécanismes de réparation distincts.

La jonction classique non homologue (c-NHEJ) ligature directement les extrémités d'ADN et introduit fréquemment de petites insertions/délétions. La jonction médiée par microhomologie (MMEJ), une variante de la NHEJ, utilise de courtes séquences homologues (5–25 pb) et entraîne souvent des délétions ou des réarrangements complexes. Les mécanismes basés sur la réplication, tels que le *fork stalling and template switching* (FoSTeS) ou la réplication induite par cassure médiée par microhomologie (MMBIR), se produisent durant la réplication, générant des insertions, duplications ou triplications via des changements de matrice via des microhomologies. Contrairement à la recombinaison homologue non allélique (NAHR), ces processus aboutissent à des réarrangements variables et non récurrents .(23,34,37,44)

Dans notre cohorte, des microhomologies aux points de cassure ont été détectées dans 57 % des cas (4/7), notamment dans les duplications en tandem des UPN6 à UPN9 (2–7 pb), cohérent avec des mécanismes de réparation médiés par microhomologie.(44) Toutefois, il reste difficile de déterminer si ces SV résultent de mécanismes de recombinaison comme c-NHEJ/MMEJ ou de processus basés sur la réplication tels que FoSTeS/MMBIR.(187) Bien que FoSTeS/MMBIR soient typiquement associés à des réarrangements complexes, ils peuvent également générer des duplications simples avec microhomologies, ce qui complique leur distinction par rapport aux événements NHEJ ou MMEJ.(187,188)

UPN10 présente une cicatrice au niveau des points de cassure, caractérisée par l'insertion de 26 pb (1/7, 14 %). Ce type d'insertion est généralement associé à des mécanismes de répara-

tion à fort taux d'erreur, tels que FoSTeS/MMBIR, capables de générer des variants de structure complexes comme des triplications..(34,44)

Chez UPN11 et UPN12 (29 %), des réarrangements complexes impliquant quatre points de cassure (DUP-TRP/INV-DUP et DUP-NML-DUP/INV) ont révélé des signatures mixtes: une jonction avec microhomologie, indicative de FoSTeS/MMBIR, et une autre sans microhomologie, compatible avec NHEJ. Ces observations suggèrent que plusieurs mécanismes de réparation peuvent agir séquentiellement, avec une initiation par des processus basés sur la réplication, suivie d'une résolution par NHEJ sans insertion ni délétion (Figure_Annexe 12A-B). (189)

Pour étudier la complexité structurale des CNVs, le séquençage *long-reads* a été appliqué à sept patients présentant des gains de copie de signification inconnue ou des phénotypes discordants. Le LRS a atteint une couverture moyenne du génome de 28,5X avec une longueur moyenne de lecture de 9 936 pb, détectant des variants de structure avec environ 16 lectures supports.

Bien que ces variants de structure aient déjà été identifiés par des techniques conventionnelles, nous avons choisi d'utiliser des logiciels de détections de variants de structure (*SV callers*) pour évaluer s'ils pouvaient être repérés uniquement via le LRS et sans *a priori*. Deux outils parmi les meilleurs ont été sélectionnés : Sniffles2 et CuteSV. (177,190–192) Pour assurer la comparabilité, nous avons harmonisé les paramètres afin que les deux outils détectent des SVs d'au moins 30 pb avec un support minimal de 8 lectures. La plupart des réglages par défaut ont été conservés, mais quelques ajustements étaient nécessaires: le paramètre par défaut de CuteSV “--maxsize 100000” n'était pas compatible avec les variants de grande taille, nous l'avons donc modifié en “--maxsize -1” avec “--min_support 8” ; pour Sniffles2, nous avons utilisé “--minsvlen 30” et “--minsupport 8”.

Sniffles2 a détecté 75 % (6/8) des variants, manquant celui de UPN9 et la délétion de UPN6, tandis que CuteSV en a détecté 71 % (5/7), manquant celui de UPN8 et la délétion de UPN 6 (la détection de UPN11 a échoué). La taille médiane des variants détectés par Sniffles2 était de 532 936 pb, contre 357 360 pb pour CuteSV. Sans modification du paramètre “--maxsize” de CuteSV, aucun variant n'aurait été détecté, ce qui souligne l'importance du réglage des paramètres.

La détection pourrait être améliorée en utilisant un autre outil d'alignement, comme NGMLR, qui, selon la littérature, améliore la détection des SVs malgré une vitesse de calcul moindre et un taux d'erreur plus élevé. Sniffles reste l'outil de détection le plus performant mais sa sensibilité diminue avec une couverture de séquençage plus faible.(177,193) Plus surprenant, Sniffles2 n'a pas détecté le variant de UPN9 malgré 20 lectures supports, suggérant la nécessité d'ajuster certains paramètres. Plusieurs études recommandent ainsi de réduire le seuil minimal de support à trois lectures et de modifier les paramètres de regroupement dans CuteSV.

Notre évaluation met en évidence les limites des paramètres par défaut, en particulier pour les SVs de grande taille. Si Sniffles2 a de très peu surpassé CuteSV, l'optimisation des paramètres reste cruciale pour améliorer la sensibilité. Les efforts futurs devraient se concentrer sur l'affinement des *pipelines* de détection et des stratégies de séquençage afin d'améliorer la détection, la classification et l'interprétation des SV, au bénéfice de la recherche comme du diagnostic clinique.

3.3. Étude de variants associés à *MED13L*

3.3.1. Contextualisation

Le syndrome *MED13L* (MIM #616789) est causé par des variants pathogènes monoalléliques de *MED13L*, qui code pour une sous-unité du complexe Mediator, un régulateur clé des processus transcriptionnels. *MED13L* partage sa localisation fonctionnelle au sein du module kinase du complexe Mediator avec son paralogue *MED13*. Ces deux protéines présentent une région régulatrice spécifique centrée autour de la thréonine 326, contenant un motif de dégradation phosphorylé, également appelé phosphodégron.

L'étude du spectre clinique des anomalies liées à *MED13L* a conduit à l'identification majoritaire de variants de perte de fonction, associés à un trouble du neurodéveloppement caractérisé par une déficience intellectuelle modérée, un retard de langage et des traits dysmorphiques, parfois accompagnés de malformations cardiaques congénitales. De plus, deux variants de structure impliquant des translocations avec un point de cassure intragénique dans *MED13L* ont été rapportés. Les patients porteurs de ces réarrangements présentaient des anomalies cardiaques, une microcéphalie, ainsi qu'un retard du développement associé à une déficience in-

telle. Ces observations soutiennent l'idée que les variants de structure affectant directement la structure de *MED13L* peuvent être délétères.

Dans ce contexte, nous avons identifié deux patients présentant des gains de copies partiels du locus *MED13L*, détectés par analyse CGH. Contrairement aux cas décrits précédemment, ces variants sont des duplications, et non des translocations, et les phénotypes associés sont relativement peu spécifiques. Une caractérisation fine de ces duplications était donc nécessaire afin d'évaluer leur impact potentiel sur *MED13L* et de déterminer si tout variant structurel affectant ce gène est nécessairement pathogène.

3.3.2. Résultats

3.3.2.1. Présentation clinique des patients UPN13 et UPN14

UPN13 est un garçon de 2 ans présentant un retard de développement et une hypertonie. L'analyse par puce CGH a révélé un gain de copies d'origine paternelle au locus 12q24.21, englobant le gène *MED13L*, observation confirmée par qPCR.

UPN14, âgé de 10 ans, présente quant à lui une déficience intellectuelle sévère. Une analyse similaire a mis en évidence un gain de copies au même locus, incluant également *MED13L* et hérité par un des membres de sa famille.

Compte tenu des données de la littérature avancée précédemment, et des phénotypes observés chez UPN13 et UPN14, l'impact pathogène potentiel de ces duplications mérite d'être interrogé. Dans ce contexte, une analyse plus approfondie de ces variants a été réalisée par LRS-WGS.

3.3.2.2. Étude de l'impact biologique des gains de copies associés à *MED13L*

Afin d'évaluer l'impact potentiel du gain de copies observé chez UPN13 et UPN14, un LRS-WGS a été réalisé. Celui-ci a permis de générer 100,4 Gb de données, correspondant à une couverture moyenne du génome de 28X, avec une longueur médiane de lecture de 11,93 kb pour UPN13 et 61,68 Gb de données, avec une couverture moyenne de 18X et une longueur N50 de 8,13 kb pour UPN14 (Tableau 2).

L'analyse des données de UPN13 a mis en évidence un gain de copies de 308 355 pb, détecté par l'outil Sniffles2 entre les positions chr12:115,949,420-116,257,774 (GRCh38), caractérisé comme une duplication en tandem. Les points de cassure étaient localisés dans l'intron 1 de *MED13L* en 5', et en région intergénique à proximité de la région 3'-UTR du même gène. (Figure 44.A-E) Une séquence de microhomologie de 7 pb ("TATTTTTT") a été identifiée au niveau de la jonction. (Figure 44.C) Les données suggèrent donc une duplication partielle de *MED13L* ne contenant pas le premier exon codant.

L'analyse des données de UPN14 à l'aide de l'outil CuteSV a permis d'identifier deux points de cassure définissant une duplication en tandem de 11 347 pb, localisée entre les positions chr12:115,949,781-115,961,127 (GRCh38) (Figure 44.B-E). Une insertion de 2 pb a été détectée au niveau de la jonction de cassure, et confirmée par séquençage Sanger (Figure 44.D). Le point de cassure proximal était situé en région intergénique entre les gènes *TBX3* et *MED13L*, tandis que le point de cassure distal se trouvait dans la région 3'-UTR de *MED13L*.

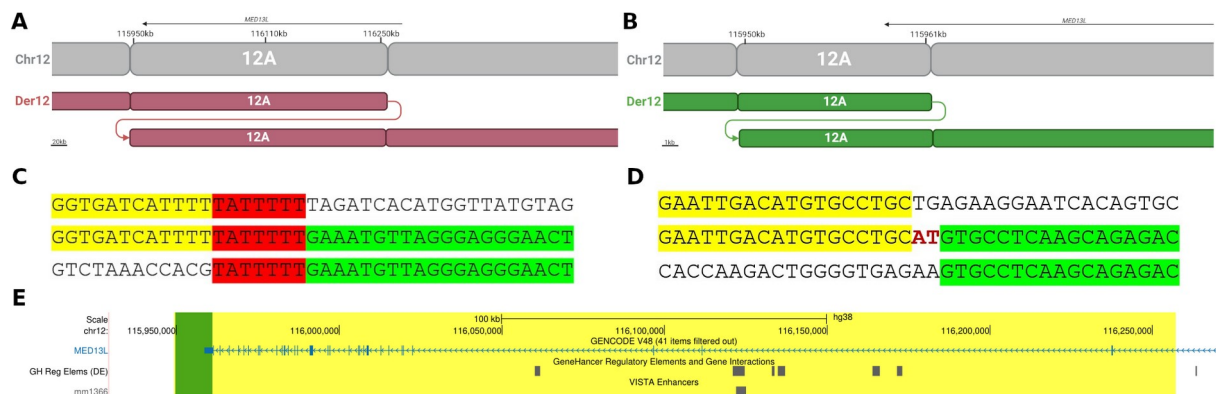


Figure 44: Analyses moléculaires des gains de copies impliquant *MED13L*. (A–B) Représentation par subway plots des duplications en tandem identifiées chez les patients UPN13 et UPN14. (C–D) Analyse des points de cassure aux jonctions des duplications pour chacun des patients. (E) Visualisation sur le navigateur UCSC de la séquence de type sauvage (WT) de *MED13L* et des régions concernées par les duplications. (en jaune : segment dupliqué chez le patient UPN13 ; en vert, celui du patient UPN14).

Ces deux analyses ont permis de mettre en évidence des gains de copies emportant partiellement *MED13L* ne causant pas de modification sur la structure biologique de *MED13L* laissant donc deux copies intactes du gène. Le gène n'étant pas connu pour être triplosensible, en l'absence d'effet délétère manifeste sur sa structure, ainsi qu'au regard de l'héritage du variant, ces gains de copie ont été reclassé comme probablement bénigne.

3.3.3. Conclusion

Le séquençage *long-read* Nanopore a permis une caractérisation fine de la structure des gains de copies initialement détectés par CGH-array. Ces derniers, bien que facilement identifiables, restent peu informatifs en l'absence de données sur la localisation précise et l'orientation de la copie supplémentaire, ce qui justifie leur classification initiale comme variants de signification inconnue.

L'analyse par LRS-WGS a révélé, chez les deux patients, des duplications en tandem impliquant *MED13L* de manière partielle : de l'exon 2 à l'exon 31 chez UPN14, et uniquement la région 3'-UTR pour UPN15. Cette organisation en tandem implique que la copie supplémentaire n'altère pas la structure du gène sur l'allèle de référence, ce qui soutient leur reclassement comme variants probablement bénins.

Par ailleurs, l'étude fine des points de cassure a permis de mieux comprendre les mécanismes d'émergence et de réparation de ces duplications. Chez UPN13, une microhomologie de 7 pb a été identifiée à la jonction, suggérant un mécanisme de réparation dépendant de la réplication, tel que FoSTeS ou MMBIR mais aussi potentiellement des mécanismes comme NHEJ/MMEJ. En revanche, chez UPN14, la présence d'une insertion de 2 pb est plus en faveur d'un mécanisme de type NHEJ, non conservatif, caractérisé par la présence de cicatrices d'insertion.

Bien que les deux patients présentent un gain de copies impliquant *MED13L*, les caractéristiques moléculaires de ces variants diffèrent nettement : la taille des duplications, la localisation des points de cassure ainsi que les mécanismes de réparation impliqués ne sont pas les mêmes. Ces observations soulignent l'hétérogénéité structurelle de ces duplications et confirment qu'il s'agit de variants non récurrents associés à *MED13L*.

3.4. Étude d'un variant associé à *GPSM2*

3.4.1. Méthode spécifique

Les prédictions de l'impact sur l'épissage des variants introniques ont été réalisées à l'aide de NetGene2(194), un outil *in silico* reposant sur des réseaux de neurones artificiels, permettant

d'identifier et de localiser de manière prédictive des sites donneurs et accepteurs d'épissage dans une séquence donnée.

3.4.2. Résultats

3.4.2.1. Présentation clinique du patient UPN15 et explorations par génome *Short-Read*

UPN15 est un garçon né de parents non apparentés, présentant un phénotype syndromique associée à une surdité neurosensorielle profonde, une hydrocéphalie, une agénésie du corps calleux, une gyration corticale anormale, une dysplasie cérébelleuse ainsi qu'un hypertélorisme.

Dans le cadre de l'exploration étiologique de cette surdité, un séquençage du génome complet en trio (patient et ses deux parents) a été réalisé par technologie *short-read*.

Cette analyse a permis d'identifier une variation hétérozygote, de transmission paternelle, correspondant à une délétion de 4 paires de bases localisée dans l'intron 13 du gène *GPSM2* (chr3:108922579-108922582, GRCh38). Les outils de prédiction *in silico* (tels que SpliceAI et SpiP) suggèrent fortement une abolition du site donneur d'épissage. Cette variation est extrêmement rare dans la population générale, avec une fréquence allélique de $1,371 \times 10^{-6}$ dans la base de données gnomAD, et est rapportée comme pathogène dans ClinVar (ID: 419262).

GPSM2 est responsable du syndrome de Chudley-McCullough (MIM #604213), un syndrome autosomique récessif caractérisé par une surdité neurosensorielle d'apparition précoce associée à une combinaison de malformations cérébrale, notamment une agénésie partielle du corps calleux, une dysplasie cérébelleuse et la présence de kystes arachnoïdiens.(195) Le phénotype clinique étant très spécifique, il oriente fortement vers ce diagnostic dès la description initiale.

Le mode de transmission étant récessif, la pathogénicité de la variation identifiée ne peut être pleinement interprétée qu'en présence d'un second variant pathogène sur l'autre allèle. Or, aucun autre variant n'a été détecté par les algorithmes d'appel de variant. Néanmoins, au vu de la forte concordance phénotypique avec le syndrome de Chudley-McCullough, l'hypothèse d'un second variant passé inaperçu reste hautement probable.

Dans ce contexte, une inspection manuelle des données issues du séquençage génomique *short-read* a été réalisée. Cette analyse a révélé, au niveau de l'intron 10 de *GPSM2*, la présence de lectures chimériques suggérant une possible translocation impliquant le chromosome 9. Ces lectures chimériques ont été observées chez le patient ainsi que chez sa mère. Toutefois, en raison du manque de précision inhérents à la technologie *short-read*, les éléments recueillis ne permettaient pas de conclure quant à la nature exacte de cette réorganisation ni à son éventuelle pathogénicité. Afin de mieux caractériser cet événement, un séquençage par technologie *long-read* a été envisagé.

3.4.2.2. Diagnostic moléculaire du syndrome Chudley-McCullough par séquençage Long-Read

Caractérisation moléculaire du syndrome Chudley-McCullough

Dans ce contexte, un LRS-WGS a été réalisé afin de caractériser plus précisément le variant de structure suspecté à partir des données *short-read*. Le séquençage a été effectué par *adaptive sampling*, ciblant environ 0,5 % du génome, correspondant à la région englobant *GPSM2* (chr3:100,000,000–116,000,000, GRCh38). Cette approche a permis de générer un total de 8,49 Gb de données, correspondant à une couverture moyenne d'environ 30× sur la région ciblée, et de 2× sur le reste du génome (Tableau 2).

L'analyse des données obtenues a permis, dans un premier temps, de confirmer la présence de la délétion pathogène de 4 pb identifiée dans l'intron 13 de *GPSM2* (Figure 45.A). Par ailleurs, les outils de détection de variants de structure, en particulier CuteSV, ainsi qu'une inspection visuelle approfondie, ont mis en évidence une insertion d'environ 2,5 kb localisée dans l'intron 10 du même gène (Figure 45.B). L'analyse des lectures chimériques a révélé que cette insertion provenait d'une région du chromosome 9 et était inversée par rapport à sa séquence d'origine (chr9:87,757,558-87,759,515, GRCh38). (Figure 45.C-D).

Ces résultats ont donc permis de détecter une insertion inversée dans *GPSM2* de transmission maternelle, confirment les observations issues du génome *short-read*, et illustrent les limites de cette technologie pour la détection et la caractérisation fine de certains variants de structure, notamment en raison de la faible taille des lectures générées.

SVA et prédiction in silico

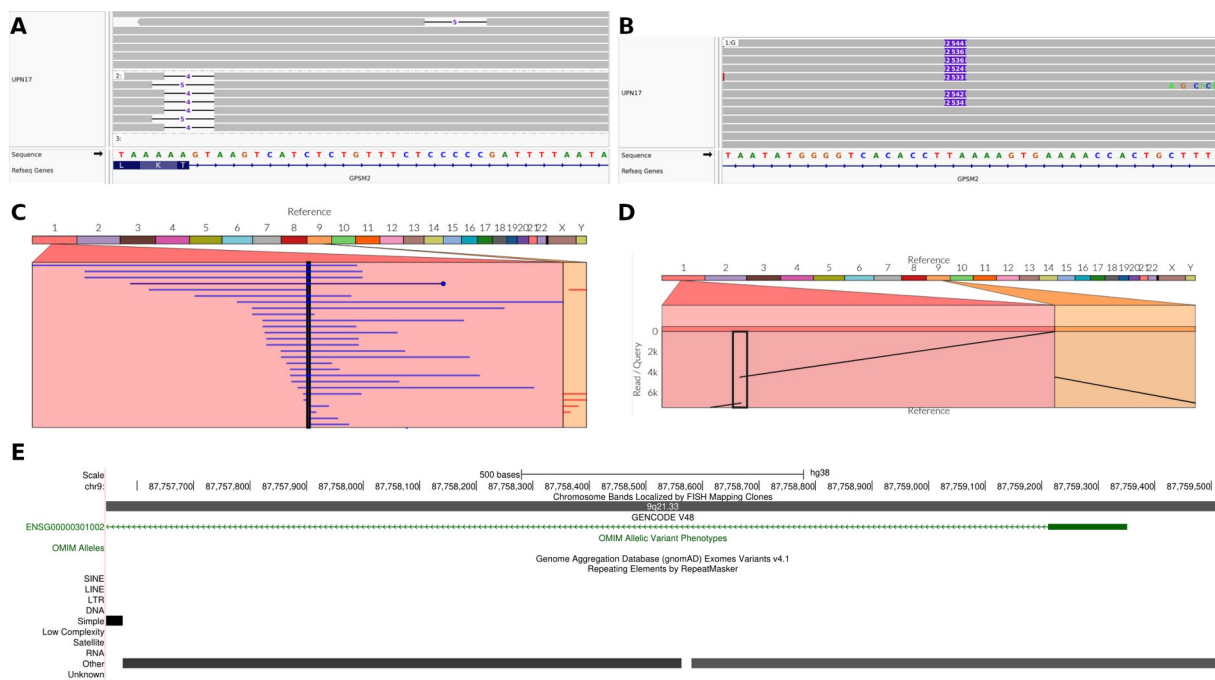


Figure 45: Analyse des variants affectant le gène *GPM2* chez UPN15. (A) Visualisation dans IGV de la délétion de 4 pb, localisée dans l'intron 13 de *GPM2* et héritée du père. (B) Visualisation dans IGV d'une insertion d'environ 2 500 pb dans l'intron 10 de *GPM2*. (C) Alignement des lectures au niveau de l'insertion dans l'intron 10, visualisé avec Ribbon genome browser. (D) Lecture chimérique suggérant l'insertion d'un segment inversé, visualisée avec Ribbon genome browser. (E) Visualisation dans UCSC Genome Browser de la région du chromosome 9, mettant en évidence des séquences répétées caractéristiques des éléments SVA.

L'analyse des *reads* issues du séquençage long-read a permis de caractériser finement l'insertion. Dans la région du chromosome 9, des séquences répétées caractéristiques ont été mises en évidence. Leur analyse a révélé que l'insertion correspond à un rétrotransposon de type SVA (SINE-VNTR-Alu), une famille de rétrotransposons non autonomes, spécifiques des hominidés et connue pour son implication dans diverses pathologies humaines (Figure 45.E). (196,197)

Compte tenu des caractéristiques des SVAs, nous avons émis l'hypothèse qu'il pourrait perturber l'expression ou l'épissage des transcrits de *GPM2*, entraînant un effet potentiellement pathogène. Les SVAs peuvent influencer la transcription des gènes, comme documenté dans la littérature. Une insertion intronique peut modifier l'épissage ou la stabilité des ARN messagers, capter des exons en amont, initier la transcription depuis l'intron ou en interne, et pro-

duire des transcrits de longueurs variées selon le signal polyA utilisé, impactant ainsi la régulation normale du gène hôte.

Afin d'évaluer cette possibilité, une analyse *in silico* de l'épissage a été réalisée à l'aide de l'outil NetGene2. Comparée à la séquence de l'intron 10 de l'allèle de référence, l'analyse prédit l'émergence d'un nouveau site donneur d'épissage à proximité de la région insérée, ainsi qu'un site accepteur situé environ 1200 pb en aval (Figure 46.A-C). Ces éléments suggèrent la création d'un nouvel exon cryptique à l'intérieur de l'intron 10, susceptible d'induire

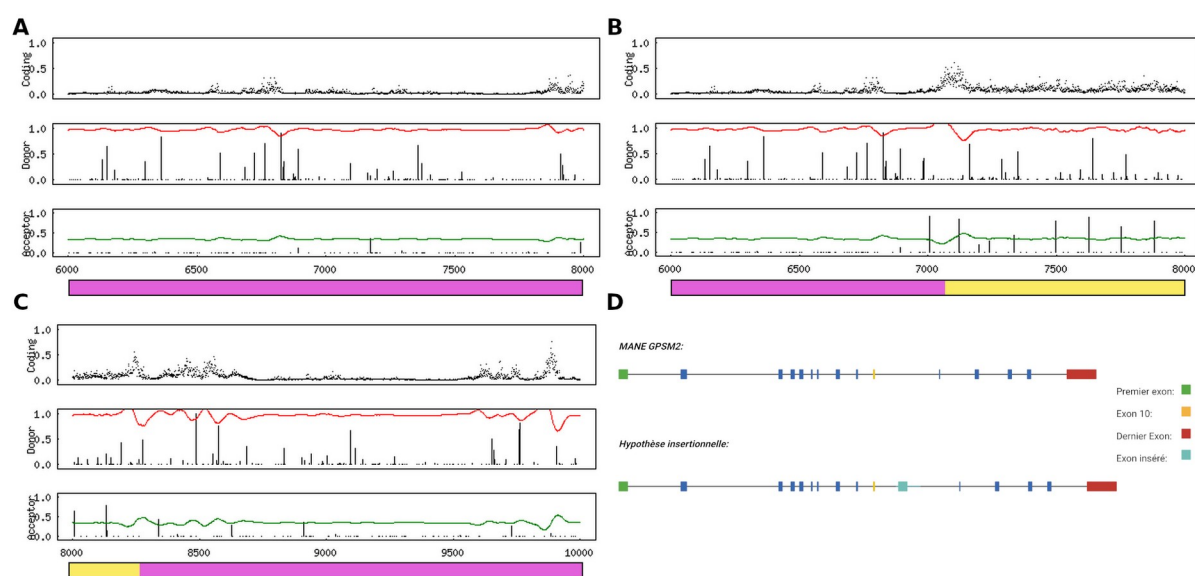


Figure 46: Prédiction *in silico* de l'impact de l'insertion sur l'épissage de GPSM2. (A) Prédiction des sites d'épissage dans la séquence WT à l'aide de l'outil NetGene2. **(B)** Mise en évidence d'un site donneur d'épissage potentiel dans la séquence du patient UPN15. **(C)** Identification d'un site accepteur d'épissage potentiel. **(D)** Schéma illustrant la potentielle formation d'un nouvel exon dans l'intron 10 de GPSM2. un décalage du cadre de lecture et, par conséquent, d'altérer la structure du transcrit de GPSM2 (Figure 46.D).

Compte tenu du phénotype de UPN15, hautement évocateur du syndrome de Chudley-McCullough, et de l'hypothèse d'un décalage du cadre de lecture induit par l'insertion inversée, cette dernière a été interprétée comme probablement pathogène.

3.4.3. Conclusion

Le séquençage *long-read* a permis de palier aux problèmes associés à la taille des fragments séquencés par séquençage par synthèse. En effet, à cause de cette limite, le diagnostic moléculaire

laire du patient n'a pas pu être fait. UPN15 présente un phénotype très spécifique et similaire permettant une hypothèse de diagnostique du syndrome Chudley-McCullough. Ce séquençage *short-read* a donc confirmé en partie cette hypothèse via détection d'un variant hétérozygote associé à une délétion de 4pb dans l'intron 13 de *GPSM2*. Mais ce gène étant connu pour être récessif, la seconde partie confirmant le diagnostic moléculaire était manquante.

L'inspection visuelle des données de SRS a permis de poser une hypothèse de translocation avec le chromosome 9. Cependant à cause de la taille trop petite des fragments de séquençage, la caractérisation structurelle de ce variant n'a pas pu être fait.

Le séquençage *long-read* a permis de caractériser structurellement le variant identifié, qui s'est révélé être une insertion inversée d'un pseudo-exon dans l'intron 10 de *GPSM2* (Figure 45.B-D). Ces données ont également permis une analyse fine du mécanisme à l'origine du réarrangement. En effet, la présence de SVA suggère l'implication du mécanisme de *Target-Primed Reverse Transcription* (Figure 45G). En effet, les SVAs (SINE-VNTR-Alu) exploitent la machinerie des LINE-1 pour leur insertion. Elle repose sur un mécanisme appelé reverse transcription amorcée sur le site cible (*Target-Primed Reverse Transcription*). L'endonucléase de L1 clive l'ADN génomique riche en A/T (AATTTT dans ce cas), générant une extrémité 3'-hydroxyle utilisée comme amorce pour la synthèse d'ADNc à partir de l'ARN polyadénylé du SVA. Ce processus aboutit à une insertion accompagnée d'une duplication du site cible (TSD) de 13bp dans cette insertion et d'une queue poly-A, signatures caractéristiques de ce mécanisme. (Figure_Annexe 13). (198)

V/ Discussion

1. Caractérisations structurelle et moléculaire de variants de structure

1.1. Analyse par LRS : Vers un meilleur taux diagnostique ?

La majorité des variants de structure étudiés dans notre cohorte correspondait à des gains de copies. Ce type de variant, fréquemment identifié en routine diagnostique, est souvent complexe à interpréter, notamment en raison du manque d'information concernant l'organisation structurale.

Les données de la littérature confirment ces difficultés : avec la technologie CGH-array, entre 10 et 15 % des CNVs demeurent classés comme de signification incertaine, en grande partie à cause de la résolution limitée de la méthode et des incertitudes liées à l'interprétation. Dans une étude rapportant que 15,7 % des profils CGH révélaient des gains de copies classés de signification inconnus, une simple réanalyse des données avec des outils d'annotation actualisés a permis de reclasser près de 40,6 % de ces variants, soulignant l'importance des efforts de ré-interprétation pour améliorer le rendement diagnostique.(199)

Ces données, associées aux limites des technologies *short-read*, montrent que de nombreux gains de copies ne sont pas caractérisés de manière précise avec ces techniques. En effet, les biais techniques pour le SRS, notamment liés à la présence de séquences répétées ou à la faible *mappabilité* de certaines régions du génome de référence, empêchent une détection fiable des CNVs. Ainsi, un grand nombre de variants mal identifiés ou initialement classés comme de signification inconnue restent aujourd'hui à réinterpréter, en particulier à la lumière des nouvelles approches de séquençage *long-read*.

Dans ce contexte, nous avons entrepris une réanalyse de 12 gains de copie identifiés chez 15 patients de notre cohorte (UPN4-UPN11 et UPN13-UPN15), initialement classées comme VSI ou dont le lien avec le phénotype restait incertain. Grâce à la technologie LRS, nous avons pu affiner considérablement la description de ces variants, en accédant à une résolution structurelle plus fine. Cette approche a permis de localiser précisément les points de cassure et de déterminer la nature exacte des réarrangements sous-jacents, révélant dans certains cas des configurations inattendues ou complexes pouvant contribuer à la pathogénicité du variant.

Au total, cette réanalyse a permis de proposer une hypothèse diagnostique et de reclasser 25 % des variants (3/12 ; UPN10-UPN11 et UPN 15) comme probablement pathogènes ou

pathogènes. Deux duplications (16,7 %, UPN13 et UPN14) ont été reclassées comme probablement bénignes, tandis que quatre autres (33,3 %, UPN6 à UPN9) n'ont pas pu être reclassées avec certitude. (Table_Annexe 5)

Par ailleurs, deux duplications supplémentaires (UPN4 et UPN5) ont pu être reclassées comme probablement pathogènes grâce à une approche fonctionnelle fondée sur l'analyse de l'ADNc par séquençage *long-read*. En effet, le LRS permet de mieux comprendre l'impact des variants de structure sur les gènes affectés et a permis de confirmer l'effet délétère suspecté sur les bases de données génomiques.

Ainsi, en combinant l'analyse structurale du génome à une approche transcriptomique, au total 42,7 % des gains de copies initialement classés comme VSI ont pu être reclassés comme probablement pathogènes. Dans les cas non résolus, l'étude de l'impact potentiel sur l'architecture tridimensionnelle du génome, en particulier via des perturbations des interactions régulatrices à distance, représente une piste complémentaire à explorer pour mieux comprendre la signification fonctionnelle de ces duplications. De plus, pour les patients pour lequel le variant avait déjà été classé pathogène que cela soit au vu de sa taille ou par rapport aux données déjà existantes dans les bases de données (UPN1-UPN2, UPN7 et UPN12), les données ont tout de même permis de préciser le réarrangement complet mais aussi d'en préciser le point de cassure et de pouvoir préciser le mécanisme de survenues. (Table_Annexe 5)

En complément de l'analyse des gains de copies, une inversion impliquant le gène *IDS* a également été identifiée (UPN3). Cette découverte a non seulement permis de confirmer le diagnostic clinique de syndrome de Hunter, précédemment suspecté, mais également de détecter un variant de structure qui n'avait pas été mis en évidence par les analyses antérieures.

Ces performances pourraient encore s'améliorer avec le développement d'outils d'annotation dédiés aux variants de structure. En effet, l'essor des technologies *long-read* a récemment relancé l'intérêt pour ces variants, entraînant le développement de nouvelles approches bio-informatiques. L'un des premiers outils disponibles, *AnnotSV*, permet l'annotation fonctionnelle des SVs, mais il se heurte à des limites pratiques : le grand nombre de variants détectés par génome (souvent plus de 10 000 SVs par fichier VCF) engendre des fichiers très volumineux et complexes à exploiter.(200) Leur analyse nécessite un important travail de pré-tri et de filtrage pour aboutir à un fichier interprétable. D'autres outils sont en cours de développement et

montrent des perspectives intéressantes, comme *LRGenotate*, spécifiquement conçu pour l'annotation des SVs dans un contexte de maladies rares, et qui semble offrir des performances prometteuses.(201)

1.2. Que nous aurait révélé le séquençage short-read ?

Associés à ces résultats et au rendement diagnostique globalement encourageant, se pose la question de la légitimité de recourir d'emblée au séquençage *long-read*, notamment dans les cas où un SRS n'a pas été réalisé en amont. En effet, parmi les 15 patients de notre cohorte, 60 % des gains de copies ont été initialement détectés par CGH-array, 6,7 % par séquençage ciblé (panel ou exome), et 26,6 % par WGS-SRS.

À la lumière des recommandations actuelles en matière de diagnostic moléculaire, il est raisonnable de penser qu'une partie des variants identifiés par CGH-array (soit environ 60 %) auraient pu, en théorie, être accessibles à une première analyse via le WGS-SRS.

Toutefois, nos résultats suggèrent que cette approche n'aurait pas toujours été suffisante. En effet, chez le patient UPN15, le séquençage *long-read* a permis de détecter un variant non appelé avec les outils d'analyse SRS. Par ailleurs, cette technologie a apporté des informations structurales précises dans des cas complexes (notamment UPN5 et UPN11), pour lesquels la résolution des points de cassure ou l'orientation des duplications n'a pas pu être atteinte avec un séquençage *short-read*.

Enfin, dans le cas du patient UPN3, porteur d'une inversion impliquant *IDS*, il est peu probable que ce variant ait pu être identifié par WGS-SRS. Les points de cassure étaient localisés dans des régions riches en séquences répétées, plus précisément au sein de duplications segmentaires, qui compromettent fortement la qualité d'alignement avec les lectures courtes.

Par ailleurs, les gains de copies détectés par CGH-array dans cette cohorte présentaient, pour la plupart, des tailles importantes ou des configurations laissant supposer une complexité structurale sous-jacente. Dans ce contexte, un SRS aurait probablement apporté peu d'informations supplémentaires sur la structure fine des variants, ou bien des données partielles nécessitant de toute façon une confirmation par une autre technique. À l'inverse, l'approche LRS a permis, dès le départ, d'envisager une résolution complète des événements structuraux.

De plus, l'objectif principal de cette étude était précisément d'évaluer la plus-value du séquençage *long-read* dans des cas où les variants étaient classés de signification inconnus ou soulevaient encore des interrogations. Dans cette optique, il était donc pertinent de se tourner directement vers le LRS, indépendamment des techniques utilisées en amont, afin d'illustrer son potentiel diagnostique dans un large éventail de configurations.

Enfin, l'ajout systématique d'un SRS en amont aurait non seulement allongé les délais de diagnostic, mais également augmenté les coûts, sans garantie d'une interprétation complète des variants identifiés. Néanmoins, dans les situations où le variant détecté semble de faible impact ou lorsqu'aucun gène connu pour être impliqué en pathologie n'est concerné, un SRS préalable peut rester pertinent. C'est une idée qui sera débattue par la suite.

1.3. Les gains de copies : une complexité sous-jacente ?

L'analyse des variants de notre cohorte s'inscrit dans un objectif plus spécifique : mieux caractériser la complexité structurale que peuvent présenter certains gains de copies, au-delà de ce que suggère leur profil initial. Cette complexité est d'ailleurs bien documentée dans la littérature. Comme évoqué précédemment, plusieurs études portant sur les duplications impliquant le gène *MECP2* ont montré que, bien que la majorité de ces variants correspondent à des duplications simples en tandem, des architectures plus complexes telles que des DUP-NML-DUP ou des DUP-TRP-DUP ont également été identifiées.(26,202)

Si certains indices de cette complexité peuvent être suggérés lors de l'analyse CGH-array, ils passent néanmoins parfois totalement inaperçus. Cela peut s'expliquer par plusieurs facteurs : la taille du variant, une répartition sous-optimale des sondes sur la région concernée, ou encore les paramètres algorithmiques de détection, qui nécessitent un certain nombre de sondes déviantes pour appeler un variant.

Une étude de Shen *et al.* illustre bien cette problématique : une duplication de plusieurs exons du gène *DMD* a été identifiée par MLPA chez un patient, ainsi que chez sa mère. L'analyse par séquençage *long-read* a ensuite révélé que cette duplication s'accompagnait en réalité d'une inversion, soulignant la complexité de l'événement et l'intérêt du LRS pour en élucider la structure précise. (203)

Nos données illustrent également la complexité structurelle que peuvent présenter certains gains de copies, ainsi que les limites des approches de diagnostic moléculaire conventionnelles pour les caractériser. Sur les 15 patients de notre cohorte, quatre variants (UPN5 et UPN11-UPN12 et UPN15) présentaient une architecture complexe qui avait initialement été mal interprétée. Par exemple, le variant de UPN5, d'abord visualisé comme une duplication incluant *RERE*, avait ensuite été décrit par les algorithmes de détection comme deux inversions et une délétion, avant qu'une analyse plus approfondie ne révèle une chromoanasyndrome associée à une translocation. De même, le variant d'UPN11, interprété comme deux duplications distinctes, s'est révélé correspondre à un réarrangement de type DUP-NML-DUP avec une inversion associée. Le cas d'UPN12, considéré comme une duplication simple, a été requalifié en motif DUP-TRP-DUP. Enfin, pour UPN15, une hypothèse initiale de translocation a été posée au profit d'une insertion inversée.

Ces exemples montrent combien certains réarrangements peuvent être sous-estimés ou mal caractérisés par les outils classiques, et à quel point le séquençage longue lecture permet une résolution fine de ces structures. Cette meilleure compréhension structurelle ouvre la voie à de nouvelles hypothèses sur l'impact biologique de ces variants, en particulier sur les gènes impliqués ou les mécanismes régulateurs à distance.

1.4. Étude associée aux mécanismes d'apparitions

L'analyse approfondie des gains de copies, et plus largement des SVs, a également permis d'explorer les mécanismes impliqués dans leur formation. Contrairement à une approche populationnelle, qui permet d'évaluer la récurrence d'un variant en identifiant des événements de taille et de localisation similaires chez des individus non apparentés, l'étude fine des points de cassure offre une perspective complémentaire. En examinant les séquences présentes aux jonctions, il est en effet possible d'identifier les signatures moléculaires laissées par les mécanismes de réparation de l'ADN, et ainsi de proposer une hypothèse mécanistique pour chaque événement.

Dans notre cohorte, 20 % des variants (3/15) se sont révélés récurrents, avec des points de cassure suggérant des mécanismes de NAHR médiés par la présence de longues régions d'homologie : des éléments répétitifs de type *LINEs* pour UPN1 et UPN2, et des duplications segmentaires pour UPN3. Les 80 % restants correspondaient à des variants non récurrents, pour

lesquels l'analyse des jonctions a mis en évidence la présence de micro-homologies ou d'insertions, en faveur de mécanismes de réparation de type MMBIR/FoSTeS ou NHEJ. (Table_Annexe 5)

Ces résultats illustrent la diversité des mécanismes en jeu dans la formation des gains de copies, et soulignent l'importance d'une analyse à haute résolution pour mieux les caractériser. En combinant l'étude fine des points de cassure à l'interprétation structurale des variants, il devient possible de mieux comprendre leur origine, de distinguer les événements récurrents des événements uniques.

2. Le séquençage long-read par la technologie Nanopore

2.1. Mise en œuvre du séquençage génomique complet sur *flowcell* Nanopore

L'un des objectifs de cette thèse, au-delà de l'étude des variants de structure, était également de démontrer la faisabilité technique d'un séquençage ciblé et génomique complet par technologie *long-read* Nanopore. Pour cela, un protocole spécifique a été développé, basé sur une approche d'un patient par *flowcell*. L'objectif était d'atteindre les performances annoncées par la technologie Nanopore, soit entre 100 et 200 Gb de données générées par *flowcell*, avec un N50 d'environ 25 kb pour l'ADN génomique, ce qui correspond à une profondeur de séquençage théorique de 31 à 62X. Ce rendement est en principe obtenu via plusieurs lavages et rechargements de la *flowcell* afin de débloquer les pores bloqués au cours du séquençage.

Dans notre cas, compte tenu de la quantité limitée d'ADN disponible pour certains patients, nous avons opté pour une stratégie plus conservative, consistant à effectuer un seul chargement par *flowcell*, avec un temps de séquençage de 72 heures. Cette approche, a permis d'atteindre en moyenne une profondeur de 26X sur les 11 WGS analysés, soit un rendement légèrement en deçà à la cible initiale de 30X. Néanmoins, cette couverture s'est révélée suffisante pour la détection et la caractérisation des variants de structure. Par ailleurs, plusieurs études antérieures ont montré que des profondeurs de séquençage bien moindres (4X) pouvaient déjà permettre la détection efficace des variants de structure, y compris par des algorithmes d'appel des variants de structure.(204)

La taille des fragments d'ADN, et donc la valeur du N50, est un paramètre essentiel dans les protocoles de séquençage *long-read*, car elle conditionne directement la capacité à reconstruire des variants de structure complexes. Dans notre étude, la cible initiale était d'obtenir un N50 compris entre 25 et 30 kb, correspondant à de véritables longs fragments d'ADN génomique. Pour atteindre cet objectif, plusieurs stratégies ont été envisagées, en tenant compte des contraintes techniques et de disponibilité du matériel.

Un premier protocole proposé par ONT reposait sur une extraction manuelle d'ADN à haut poids moléculaire à l'aide d'un kit spécifique de Qiagen. Cependant, cette méthode n'a pas été retenue dans notre laboratoire, où les extractions sont effectuées de manière automatisée. Nous avons ensuite tenté un chargement direct de l'ADN extrait, sans étape de fragmentation préalable. Cette approche s'est révélée sous-optimale : les fragments, trop longs, provoquaient un blocage fréquent des pores, nécessitant de nombreux lavages de la *flowcell* et des rechargements répétés de *librairies*, avec un impact notable sur le rendement et ne correspondant pas à notre stratégie conservatrice.

D'autres protocoles de fragmentation, comme l'utilisation du Megaruptor(205), ont été décrits dans la littérature pour générer de longs fragments. Cet équipement n'étant pas disponible, nous nous sommes donc orientés vers un protocole recommandé par ONT ciblant une taille de fragments plus petite d'environ 10 kb, basé sur l'utilisation des Covaris g-TUBE. Cette méthode nous a permis d'obtenir, pour nos données de WGS, une taille de fragment moyenne de 11,33 kb.

À l'issue de l'analyse de nos données, des pistes d'optimisation ont été identifiées pour améliorer ce protocole. En particulier, plusieurs études, telles que celles de Miller *et al.* ou de De Clercq *et al.*, ont également utilisé les g-TUBE mais rapportent des N50 plus élevés, autour de 15 à 20 kb.(109,153,206) Des tests sont donc actuellement en cours dans notre laboratoire afin d'augmenter le N50 des fragments et d'optimiser ainsi le rendement et la qualité des données générées.

2.2. Faut-il privilégier l'*Adaptive Sampling* ?

Comme présenté dans les résultats, les approches de séquençage ont différé selon les patients et les variants étudiés. Trois des quinze patients de la cohorte (UPN1, UPN2 et UPN15) ont été analysés par séquençage *long-read* ciblé (*Targeted Long-read Sequencing* ; T-LRS), per-

mettant une caractérisation fine de leurs variants de structure. À partir de ces cas, s'est posée la question de la pertinence d'utiliser la technique d'*adaptive sampling* comme alternative au séquençage génomique complet.

Le WGS génère en effet des volumes de données particulièrement importants, ce qui soulève des problématiques de stockage. Dans notre projet, plus de 11 To de données ont été produits, incluant les fichiers bruts Pod5, sans compter les fichiers d'alignement (BAM) ou d'appel des variants (VCF). Bien que les fichiers Pod5 puissent théoriquement être supprimés après le *basecalling*, ils ont été conservés, compte tenu de la nouveauté de la technologie, des évolutions fréquentes du *basecaller* Dorado, et des nouveaux algorithmes d'appel de méthylation qui exploitent ces fichiers bruts.

L'*adaptive sampling* apparaît donc comme une alternative intéressante pour réduire les volumes de données générées, tout en ciblant spécifiquement les régions d'intérêt. Cette stratégie s'est révélée efficace pour détecter les variants de structure lorsque le ciblage représentait seulement 0,5 à 1 % du génome. Néanmoins, cette méthode présente une limite majeure : elle ne permet pas toujours une analyse fine lorsque les variants présentent une structure complexe. C'est ce que nous avons observé chez le patient UPN5, dont le variant n'a pas pu être entièrement caractérisé par *adaptive sampling* (donnée non présentée), en raison d'un manque de lectures informatives. Un LRS-WGS a alors dû être réalisé en complément.

En effet, comme discuté précédemment, les variants de structure peuvent présenter une grande complexité. Or, l'*adaptive sampling* repose sur une sélection spécifique de régions à enrichir, définies dans un fichier BED. Si des lectures informatives se situent en dehors de ces régions, elles peuvent être perdues, compromettant ainsi l'interprétation des données. Dans le contexte d'une étude exploratoire, où aucun *a priori* sur l'architecture du variant n'est possible, cela représente une limite. En revanche, lorsque des hypothèses structurales peuvent être formulées à partir d'analyses préliminaires (comme un caryotype ou une puce CGH), il devient possible d'optimiser le fichier BED en conséquence. C'est le cas, par exemple, de l'étude de Miller *et al.* dans laquelle l'intégration de données cytogénétiques a permis de cibler avec précision les régions d'intérêt et, grâce au T-LRS, de reconstruire entièrement un réarrangement complexe.(109)

Enfin, l'intérêt économique de cette méthode est à nuancer. Bien que le ciblage d'un faible pourcentage du génome permette d'envisager un séquençage sur une *flowcell* de type MinION, dont la capacité est inférieure à celle d'une PromethION, la différence de coût reste peu significative. À l'unité, une *flowcell* MinION est proposée à environ 800 \$, contre 955 \$ pour une PromethION (prix catalogue ONT). L'écart de prix, relativement faible, ne justifie pas systématiquement le recours à l'*adaptive sampling*, d'autant plus si un nouveau séquençage est nécessaire en cas d'échec d'analyse.

C'est pourquoi nous avons principalement choisi d'axer notre étude sur des analyses en WGS, plutôt que sur l'*adaptive sampling*. Néanmoins, cette approche ciblée présente un réel intérêt dans certains contextes : notamment pour le développement de panels de gènes ou l'étude de variants spécifiques retrouvés chez plusieurs patients, qui pourraient être séquencés et multiplexés sur une même *flowcell*. Une étude menée par Miller *et al.* illustre d'ailleurs la pertinence de cette stratégie : en ciblant par T-LRS des patients atteints d'un syndrome de Werner non résolu, les auteurs ont pu identifier avec succès des variants de structure auparavant manqués par les approches conventionnelles.(206) Dans ce cadre, l'*adaptive sampling* permettrait non seulement de cibler les régions d'intérêt, mais aussi d'optimiser les coûts et la gestion des données, en autorisant le séquençage de plusieurs échantillons sur une seule *flowcell*.

2.3. Vers une mise en place en routine diagnostique ?

Au vu de nos résultats, la pertinence de l'utilisation du LRS, en WGS ou en *adaptive sampling*, dans un cadre diagnostique de routine peut être discutée. Il ne s'agit probablement pas d'une technique de première intention, mais plutôt d'un outil complémentaire pertinent dans des contextes bien définis, par exemple en cas de suspicion de syndrome génétique spécifique ou, comme dans notre étude, pour affiner l'interprétation de variants de structure initialement classés comme de signification inconnue et pertinents.

Une question importante émerge alors : le LRS pourrait-il, à terme, remplacer le SRS ? Si cette perspective fait l'objet de nombreux débats, il est important de souligner que, dans une approche sans *a priori* sur le génome et non biaisée par des stratégies de capture, la détection des SNVs et des INDELs reste un enjeu majeur. Or, ces petits variants représentent encore, à elles seules, environ 65 % et 22 % des variants pathogènes identifiés dans les troubles du neurodéveloppement.(47)

Dans ce contexte, le séquençage des génomes par technologie SRS reste à ce jour une approche privilégiée, en particulier en raison de son coût plus accessible notamment pour les analyses en trio avec les parents (Tableau 1) ainsi que pour sa meilleure précision en *basecalling*. Une étude comparative récente sur l'appel de variants entre les technologies SRS et LRS a mis en évidence une différence notable en faveur du SRS concernant la détection des INDELs, ces derniers étant significativement moins bien appelés dans les données LRS. Les INDELs représentant une source importante d'erreurs, cette limitation peut avoir un impact non négligeable sur la précision diagnostique. Néanmoins, cette même étude confirme notre étude sur l'avantage du LRS pour la détection des variants de structure, en particulier dans les régions répétées du génome, confirmant son intérêt complémentaire dans certaines situations cliniques.(207)

L'appel des SVs repose sur des algorithmes dont la performance est en constante amélioration, mais qui nécessitent encore une optimisation fine. Comme discuté dans notre troisième article portant sur les duplications, le réglage des paramètres d'algorithmes tels que *cuteSV* ou *Sniffles2* peut avoir un impact significatif sur la sensibilité et la spécificité de détection. Le design d'un *pipeline* bioinformatique adapté à la détection de grands réarrangements est donc essentiel pour garantir des résultats robustes. Le choix et la configuration des outils influencent directement la capacité à détecter certains types de variants, et par conséquent, la performance diagnostique. Dans notre étude, bien que l'utilisation conjointe de *cuteSV* et *Sniffles2* se soit révélée suffisante compte tenu du ciblage spécifique des données analysées, une stratégie davantage optimisée aurait sans doute permis d'améliorer la détection et l'interprétation de certains événements complexes. De nouveaux outils récemment développés, tels que *DeBreak*, *Svision* ou encore *pbsv*, apparaissent comme des alternatives prometteuses et mériteraient d'être évalués dans des contextes diagnostiques variés, notamment dans la perspective de *pipelines* conçus pour l'analyse de génomes complets sans *a priori*. (191,208)

Cependant, la détection des SVs n'est pas exclusivement accessible par le LRS. Une part importante des variants détectés dans notre étude sont des duplications en tandem, qui peuvent, dans certains cas, être détectées par les technologies SRS. Comme évoqué précédemment, certains variants de structure complexes, en particulier ceux situés dans des régions répétées du génome, sont difficilement détectables en SRS, ce qui souligne la valeur ajoutée du LRS dans ces contextes. Toutefois, dans des cas moins complexes, le SRS reste une approche informative et pertinente pour la détection des variants de structure.

À ce jour, le LRS peut donc être considéré comme une technologie complémentaire au SRS.

Cela dit, cette limitation tend à s'atténuer avec l'évolution constante de la technologie Nanopore, dont les améliorations portent à la fois sur l'architecture des nanopores afin d'affiner la détection des bases et sur le développement continu de leur *basecaller*, Dorado. Ainsi, la question d'un remplacement partiel ou total du SRS par le LRS reste ouverte, en particulier dans les années à venir, bien que le coût des réactifs et du stockage constitue encore un facteur limitant majeur dans le choix de la technologie.

Malgré ces considérations, la technologie Nanopore a d'ores et déjà été mise en œuvre en routine diagnostique à Lille mais pas pour les indications spécifiques évoquées précédemment. En effet, ses applications ne se limitent pas à ces seuls contextes. Comme mentionné en introduction, l'un des atouts du LRS est également la possibilité d'accéder à des données de méthylation. Les lectures longues permettent aussi de séquencer efficacement des amplicons issus de PCR long-range, mais aussi d'accéder au *phasing* haplotypique. Le LRS est donc aujourd'hui utilisée en routine pour ces applications.

Ces applications ont déjà fait l'objet de plusieurs études. Par exemple, dans le cas d'un gène soumis à empreinte parentale, Paschal *et al.* ont utilisé le *phasing* haplotypique pour démontrer qu'une délétion identifiée était localisée sur l'allèle maternel, apportant ainsi une preuve diagnostique.(209) Dans une autre étude plus axée sur la méthylation, LaFlamme *et al.* ont mis en évidence une hyperméthylation du bras long du chromosome 13, probablement induite par une translocation t(X;13), et susceptible d'expliquer le syndrome d'encéphalopathies épileptique et développementale observé chez la patiente.(210) La possibilité d'accéder au *phasing* est particulièrement précieuse lorsque l'ADN parental n'est pas disponible, notamment pour l'interprétation de pathologies récessives ou l'analyse de gènes soumis à empreinte. Le LRS constitue ainsi un outil puissant pour enrichir l'interprétation des variants dans des contextes génétiques complexes.

Il ne faut pas négliger que le séquençage long-read représente également une alternative pertinente dans le cadre de projets fonctionnels, en complément d'autres approches. Bien que sa mise en œuvre en routine puisse s'avérer complexe, notamment en raison de sa dépendance à la qualité des résultats, les données obtenues sont loin d'être négligeables. Cette technologie pourrait d'ailleurs constituer un atout majeur pour établir des collaborations avec des labora-

toires de recherche en vue de confirmer certains résultats. Comme nous avons pu le démontrer, les analyses transcriptomiques réalisées à partir de l'ADNc des patients UPN4 et UPN5, ainsi que l'étude d'amplicons des transcrits de UPN1 et UPN2, ont permis une caractérisation fine de l'impact des variants sur les transcrits. Ces approches ont également contribué à une compréhension plus approfondie des conséquences biologiques associées.

2.4. Qu'en est-il de la seconde approche de séquençage Long-Read : PacBio ?

Les analyses par séquençage *long-read* se multiplient, tant dans les publications scientifiques que dans les laboratoires, et les deux principales technologies : Oxford Nanopore Technologies et PacBio HiFi y sont régulièrement utilisées. Chaque équipe de recherche tend à privilégier l'une ou l'autre en fonction de ses objectifs. Par exemple, l'équipe de Miller *et al.* s'appuie principalement sur la technologie Oxford Nanopore pour ses travaux (109,206,209,210), tandis que celle de Hoischen *et al.* se tourne davantage vers la technologie HiFi de PacBio. (211,212)

Ces deux approches reposent sur des principes techniques fondamentalement différents. La technologie Nanopore permet le séquençage de fragments d'ADN simple brin de très grande taille allant jusqu'à plusieurs mégabases en lecture unique. En revanche, la technologie PacBio repose sur un séquençage circulaire de fragments généralement limités à ~60 kb, qui sont lus plusieurs fois afin de générer une séquence consensus de haute fidélité (HiFi).

Ces différences se traduisent par des avantages complémentaires. Nanopore offre des lectures exceptionnellement longues, ce qui facilite la résolution de réarrangements très complexes ou de régions peu *mappables*. À l'inverse, PacBio HiFi se distingue par une précision de lecture supérieure, grâce au séquençage multiple d'un même fragment. Cette complémentarité a été bien mise en évidence dans l'étude comparative de Mahmoud *et al.*(213), qui souligne que le choix de la technologie dépend avant tout du contexte et des objectifs du projet. Par exemple, pour l'identification de variants de structure complexes, les longues lectures de Nanopore peuvent être déterminantes. En revanche, pour une détection plus précise des SNV ou INDELS, notamment en comparaison avec Illumina, PacBio HiFi reste préférable, bien que les deux technologies montrent des performances inférieures à celles du *short-read*. Concernant

les variants de structure, les deux technologies *long-read* offrent des performances très robustes, souvent supérieures aux approches traditionnelles. (214)

Dans notre étude, nous avons choisi d'utiliser la technologie Oxford Nanopore, principalement en raison de sa capacité à générer des lectures globalement plus longues (même si le N50 initialement visé n'a pas été atteint), mais également pour des raisons économiques et pratiques. Les séquenceurs Nanopore, conçus comme des appareils de paillasse, présentent en effet un avantage en termes de flexibilité d'installation, contrairement aux plateformes PacBio qui nécessitent une infrastructure plus lourde. Par ailleurs, le coût global d'utilisation d'un système ONT demeure nettement inférieur à celui d'un système PacBio, ce qui rend cette technologie plus accessible pour de nombreux laboratoires.

Cette différence de coût peut néanmoins s'expliquer en partie par le fait que PacBio fournit une suite logicielle complète et bien intégrée pour l'analyse et l'annotation des données. À l'inverse, ONT propose sa propre solution logicielle, EPI2ME, qui permet certaines étapes d'analyse, mais dont les performances et les fonctionnalités restent plus limitées à ce jour. Cependant, l'écosystème ONT bénéficie d'un développement rapide, et de nombreuses alternatives open source ou tierces se développent pour pallier ces limitations. Des outils de visualisation comme IGV ou Ribbon Genome Browser, notamment, s'imposent progressivement comme des supports fiables pour l'exploration et l'interprétation des données issues de Nanopore, en particulier pour la visualisation de variants de structure complexes.

VI/ Conclusion

Les travaux présentés dans cette thèse ont permis d'apporter des éléments de réponse solides aux objectifs formulés. Ce projet visait à évaluer la pertinence du séquençage *long-read* par ONT pour l'étude des variants de structure, en particulier les gains de copies détectés par les techniques de routine diagnostique, et à en analyser les implications en termes de caractérisation, d'interprétation fonctionnelle et de potentiel diagnostique.

L'ensemble des données générées a confirmé que le séquençage *long-read* offre une couverture suffisante, tant au niveau global du génome qu'au sein des régions ciblées. Il s'est avéré capable de révéler des réarrangements complexes passés inaperçus avec les approches conventionnelles, et d'en fournir une description structurale fine, qu'il s'agisse de variants simples ou complexes. Dans plusieurs cas, cette résolution accrue a permis une modélisation précise de l'organisation des points de cassure, mettant en évidence des structures telles que des insertions, des inversions ou des événements de type chromoanasynthese.

Les algorithmes d'appel de variants utilisés dans le cadre de cette étude, bien que nécessitant encore des ajustements méthodologiques, se sont révélés globalement robustes. Leur combinaison a permis une détection fiable sans recourir à des hypothèses préalables, ce qui constitue un atout dans le contexte de l'analyse génomique non ciblée.

Au-delà de la détection, le recours à la technologie Nanopore a permis d'enrichir l'interprétation fonctionnelle de plusieurs variants. Des hypothèses mécanistiques ont pu être formulées à partir des données structurales obtenues, conduisant à une réévaluation de la classification clinique de certains variants identifiés.

Ces résultats soulignent l'intérêt du séquençage *long-read* pour améliorer la résolution et la compréhension des variants de structure, en complément des approches actuellement utilisées en diagnostic. Ils suggèrent que cette technologie pourrait contribuer à augmenter le rendement diagnostique chez les patients atteints de troubles du neurodéveloppement, et qu'elle constitue une piste sérieuse pour faire évoluer les pratiques de routine à moyen terme.

VII/ Bibliographie

- 1.....Minchin S, Lodge J. Understanding biochemistry: structure and function of nucleic acids. Essays in Biochemistry. 11 oct 2019;63(4):433.
- 2.....Pathak I, Bordoni B. Genetics, Chromosomes. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 [cité 5 nov 2024]. Disponible sur: <http://www.ncbi.nlm.nih.gov/books/NBK557784/>
- 3.....Health (US) NI of, Study BSC. Understanding Human Genetic Variation. In: NIH Curriculum Supplement Series [Internet] [Internet]. National Institutes of Health (US); 2007 [cité 7 nov 2024]. Disponible sur: <https://www.ncbi.nlm.nih.gov/books/NBK20363/>
- 4.....Lu YF, Goldstein DB, Angrist M, Cavalleri G. Personalized Medicine and Human Genetic Diversity. Cold Spring Harbor Perspectives in Medicine. sept 2014;4(9):a008581.
- 5.....Consortium T 1000 GP. A map of human genome variation from population scale sequencing. Nature. 28 oct 2010;467(7319):1061.
- 6.....Lee CE, Singleton KS, Wallin M, Faundez V. Rare Genetic Diseases: Nature's Experiments on Human Development. iScience. 1 mai 2020;23(5):101123.
- 7.....Amberger JS, Hamosh A. Searching Online Mendelian Inheritance in Man (OMIM): A Knowledgebase of Human Genes and Genetic Phenotypes. Curr Protoc Bioinformatics. 27 juin 2017;58:1.2.1-1.2.12.
- 8.....Buckley M, Terwagne C, Ganner A, Cubitt L, Brewer R, Kim DK, et al. Saturation genome editing maps the functional spectrum of pathogenic VHL alleles. Nat Genet. juill 2024;56(7):1446-55.
- 9.....D'haene E, Vergult S. Interpreting the impact of noncoding structural variation in neurodevelopmental disorders. Genetics in Medicine. 25 sept 2020;23(1):34.
- 10.....Maia N, Sá MJN, Melo-Pires M, Brouwer APM de, Jorge P. Intellectual disability genomics: current state, pitfalls and future challenges. BMC Genomics. 20 déc 2021;22:909.
- 11.....Crocq MA, Guelfi JD. DSM-5: manuel diagnostique et statistique des troubles mentaux. 5e éd. Issy-les-Moulineaux: Elsevier Masson; 2015.
- 12.....Vasudevan P, Suri M. A clinical approach to developmental delay and intellectual disability. Clinical Medicine. déc 2017;17(6):558-61.
- 13.....Lejeune J, Turpin R, Gautier M. [Chromosomal diagnosis of mongolism]. Arch Fr Pediatr. 1959;16:962-3.
- 14.....Bulayeva K, Lesch KP, Bulayev O, Walsh C, Glatt S, Gurganova F, et al. Genomic Structural Variants Are Linked With Intellectual Disability. Journal of neural transmission (Vienna, Austria : 1996). 28 janv 2015;122(9):1289.

- 15.....Nesta AV, Tafur D, Beck CR. Hotspots of Human Mutation. *Trends in Genetics*. août 2021;37(8):717-29.
- 16.....Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 4 janv 2018;46(D1):D1062-7.
- 17.....Warshawsky I. Chapter 1 - Molecular Biology Basics for the Pathologist. In: Tubbs RR, Stoler MH, éditeurs. *Cell and Tissue Based Molecular Pathology* [Internet]. Philadelphia: Churchill Livingstone; 2009 [cité 8 nov 2024]. p. 3-9. Disponible sur: <https://www.sciencedirect.com/science/article/pii/B9780443069017500067>
- 18.....Haraksingh RR, Snyder MP. Impacts of Variation in the Human Genome on Gene Regulation. *Journal of Molecular Biology*. 1 nov 2013;425(21):3970-7.
- 19.....Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, et al. The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Res*. 5 janv 2013;23(5):749-61.
- 20.....Queremel Milani DA, Tadi P. Genetics, Chromosome Abnormalities. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 [cité 8 nov 2024]. Disponible sur: <http://www.ncbi.nlm.nih.gov/books/NBK557691/>
- 21.....Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. *Nature*. mai 2020;581(7809):444-51.
- 22.....Collins RL, Talkowski ME. Diversity and consequences of structural variation in the human genome. *Nat Rev Genet*. 21 janv 2025;1-20.
- 23.....Weckselblatt B, Rudd MK. Human structural variation: mechanisms of chromosome rearrangements. *Trends Genet*. oct 2015;31(10):587-99.
- 24.....Zarrei M, Burton CL, Engchuan W, Young EJ, Higginbotham EJ, MacDonald JR, et al. A large data resource of genomic copy number variation across neurodevelopmental disorders. *NPJ Genom Med*. 7 oct 2019;4:26.
- 25.....Wilch ES, Morton CC. Historical and Clinical Perspectives on Chromosomal Translocations. In: Zhang Y, éditeur. *Chromosome Translocation* [Internet]. Singapore: Springer; 2018 [cité 8 nov 2024]. p. 1-14. Disponible sur: https://doi.org/10.1007/978-981-13-0593-1_1
- 26.....Pehlivan D, Bengtsson JD, Bajikar SS, Grochowski CM, Lun MY, Gandhi M, et al. Structural variant allelic heterogeneity in MECP2 duplication syndrome provides insight into clinical severity and variability of disease expression. *Genome Medicine*. 18 déc 2024;16(1):146.
- 27.....Pellestor F, Gaillard J, Schneider A, Puechberty J, Gatinois V. Chromoanagenesis, the mechanisms of a genomic chaos. *Seminars in Cell & Developmental Biology*. 1 mars 2022;123:90-9.

- 28.....Nazaryan-Petersen L, Bjerregaard VA, Nielsen FC, Tommerup N, Tümer Z. Chromothripsis and DNA Repair Disorders. *J Clin Med*. 25 févr 2020;9(3):613.
- 29.....Pellestor F, Gatinois V. Chromoanasythesis: another way for the formation of complex chromosomal abnormalities in human reproduction. *Human Reproduction*. 1 août 2018;33(8):1381-7.
- 30.....Pellestor F, Gatinois V. Chromoanagenesis: a piece of the macroevolution scenario. *Molecular Cytogenetics*. 28 janv 2020;13(1):3.
- 31.....Mehta A, Haber JE. Sources of DNA Double-Strand Breaks and Models of Recombinational DNA Repair. *Cold Spring Harbor Perspectives in Biology*. sept 2014;6(9):a016428.
- 32.....Ciccio A, Elledge SJ. The DNA Damage Response: Making it safe to play with knives. *Molecular cell*. 22 oct 2010;40(2):179.
- 33.....Scully R, Panday A, Elango R, Willis NA. DNA double strand break repair pathway choice in somatic mammalian cells. *Nature reviews Molecular cell biology*. 1 juill 2019;20(11):698.
- 34.....Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet*. avr 2016;17(4):224-38.
- 35.....Bursted B, Zamariolli M, Bellucco FT, Melaragno MI. Mechanisms of structural chromosomal rearrangement formation. *Molecular Cytogenetics*. 14 juin 2022;15(1):23.
- 36.....Park SS, Stankiewicz P, Bi W, Shaw C, Lehoczy J, Dewar K, et al. Structure and evolution of the Smith-Magenis syndrome repeat gene clusters, SMS-REPs. *Genome Res*. mai 2002;12(5):729-38.
- 37.....Currall BB, Chiang C, Talkowski ME, Morton CC. Mechanisms for Structural Variation in the Human Genome. *Curr Genet Med Rep*. 1 juin 2013;1(2):81-90.
- 38.....Neira-Fresneda J, Potocki L. Neurodevelopmental Disorders Associated with Abnormal Gene Dosage: Smith–Magenis and Potocki–Lupski Syndromes. *Journal of Pediatric Genetics*. 28 sept 2015;4(3):159.
- 39.....Piwowarczyk P, Massalska D, Obodzińska I, Gawlik Zawiślak S, Bijok J, Kucińska-Chahwan A, et al. Prenatal diagnosis of Emanuel syndrome – case series and review of the literature. *Journal of Obstetrics and Gynaecology*. 3 oct 2022;42(7):2615-20.
- 40.....Emanuel BS, Zackai EH, Medne L. Emanuel Syndrome. In: Adam MP, Feldman J, Mirzaa GM, Pagon RA, Wallace SE, Amemiya A, éditeurs. *GeneReviews®* [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cité 8 nov 2024]. Disponible sur: <http://www.ncbi.nlm.nih.gov/books/NBK1263/>
- 41.....Frontiers | Polymerases and DNA Repair in Neurons: Implications in Neuronal Survival and Neurodegenerative Diseases [Internet]. [cité 19 mars 2025]. Disponible sur: <https://www.frontiersin.org/journals/cellular-neuroscience/articles/10.3389/fncel.2022.852002/full>

- 42.....Corneo B, Wendland RL, Deriano L, Cui X, Klein IA, Wong SY, et al. Rag mutations reveal robust alternative end joining. *Nature*. sept 2007;449(7161):483-6.
- 43.....Pathways of DNA double-strand break repair during the mammalian cell cycle - PubMed [Internet]. [cité 19 févr 2025]. Disponible sur: <https://pubmed-ncbi-nlm-nih-gov.ressources-electroniques.univ-lille.fr/12897142/>
- 44.....Bahrambeigi V, Song X, Sperle K, Beck CR, Hijazi H, Grochowski CM, et al. Distinct patterns of complex rearrangements and a mutational signature of microhomeology are frequently observed in PLP1 copy number gain structural variants. *Genome Med*. 9 déc 2019;11:80.
- 45.....Segar MW, Sakofsky CJ, Malkova A, Liu Y. MMBIRFinder: A Tool to Detect Microhomology-Mediated Break-Induced Replication. *IEEE/ACM Trans Comput Biol Bioinform*. 2015;12(4):799-806.
- 46.....Carvalho CMB, Ramocki MB, Pehlivan D, Franco LM, Gonzaga-Jauregui C, Fang P, et al. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat Genet*. 2 oct 2011;43(11):1074-81.
- 47.....Sanchis-Juan A, Megy K, Stephens J, Armirola Ricaurte C, Dewhurst E, Low K, et al. Genome sequencing and comprehensive rare-variant analysis of 465 families with neurodevelopmental disorders. *Am J Hum Genet*. 3 août 2023;110(8):1343-55.
- 48.....E S, G P, V M, Ž M, A S, V Ž, et al. Molecular and Functional Characterisation of a Novel Intragenic 12q24.21 Deletion Resulting in MED13L Haploinsufficiency Syndrome. *Medicina (Kaunas, Lithuania)* [Internet]. 29 juin 2023 [cité 21 févr 2025];59(7). Disponible sur: <https://pubmed.ncbi.nlm.nih.gov/37512036/>
- 49.....Leroy C, Jacquemont ML, Doray B, Lamblin D, Cormier Daire V, Philippe A, et al. La duplication Xq25 : rôle crucial du gène *STAG2* dans cette nouvelle cohésinopathie. *Morphologie*. 1 déc 2015;99(327):169.
- 50.....Malan V, Romana S. Republication de « Anomalies chromosomiques et leur diagnostic en pathologie constitutionnelle ». *Journal de Pédiatrie et de Puériculture*. 1 déc 2024;37(6):409-31.
- 51.....van Dijk FS, Huizer M, Kariminejad A, Marcelis CL, Plomp AS, Terhal PA, et al. Complete *COL1A1* allele deletions in osteogenesis imperfecta. *Genetics in Medicine*. 1 nov 2010;12(11):736-41.
- 52.....Elgin SCR, Reuter G. Position-Effect Variegation, Heterochromatin Formation, and Gene Silencing in *Drosophila*. *Cold Spring Harb Perspect Biol*. août 2013;5(8):a017780.
- 53.....Yeung KS, Chee YY, Luk HM, Kan ASY, Tang MHY, Lau ET, et al. Spread of X inactivation on chromosome 15 is associated with a more severe phenotype in a girl with an unbalanced t(X; 15) translocation. *Am J Med Genet A*. oct 2014;164A(10):2521-8.

- 54.....Kleinjan DJ, van Heyningen V. Position effect in human genetic disease. *Hum Mol Genet*. 1998;7(10):1611-8.
- 55.....Hamamoto K, Fukaya T. Molecular architecture of enhancer–promoter interaction. *Current Opinion in Cell Biology*. 1 févr 2022;74:62-70.
- 56.....Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. *Nature reviews Genetics*. avr 2013;14(4):288.
- 57.....Croft B, Ohnesorg T, Hewitt J, Bowles J, Quinn A, Tan J, et al. Human sex reversal is caused by duplication or deletion of core enhancers upstream of SOX9. *Nat Commun*. 14 déc 2018;9(1):1-10.
- 58.....Claringbould A, Zaugg JB. Enhancers in disease: molecular basis and emerging treatment strategies. *Trends in Molecular Medicine*. 1 nov 2021;27(11):1060-73.
- 59.....Mohanta TK, Mishra AK, Al-Harrasi A. The 3D Genome: From Structure to Function. *International Journal of Molecular Sciences*. 27 oct 2021;22(21):11585.
- 60.....Dekker J, Heard E. Structural and Functional Diversity of Topologically Associating Domains. *FEBS letters*. 5 sept 2015;589(20 0 0):2877.
- 61.....Spielmann M, Lupiáñez DG, Mundlos S. Structural variation in the 3D genome. *Nat Rev Genet*. juill 2018;19(7):453-67.
- 62.....Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*. 13 oct 2016;538(7624):265-9.
- 63.....Nmezi B, Giorgio E, Raininko R, Lehman A, Spielmann M, Koenig MK, et al. Genomic deletions upstream of lamin B1 lead to atypical autosomal dominant leukodystrophy. *Neurol Genet*. 24 janv 2019;5(1):e305.
- 64.....Melo US, Jatzlau J, Prada-Medina CA, Flex E, Hartmann S, Ali S, et al. Enhancer hijacking at the ARHGAP36 locus is associated with connective tissue to bone transformation. *Nat Commun*. 11 avr 2023;14:2034.
- 65.....Tjio JH, Levan A. The Chromosome Number of Man. *Hereditas*. 1956;42(1-2):1-6.
- 66.....Ferguson-Smith MA. History and evolution of cytogenetics. *Mol Cytogenet*. 20 mars 2015;8:19.
- 67.....Balajee AS, Hande MP. History and evolution of cytogenetic techniques: Current and future applications in basic and clinical research. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*. 1 déc 2018;836:3-12.
- 68.....Caspersson T, Zech L, Johansson C. Differential binding of alkylating fluorochromes in human chromosomes. *Exp Cell Res*. juin 1970;60(3):315-9.

- 69.....Caspersson T, Zech L, Johansson C, Modest EJ. Identification of human chromosomes by DNA-binding fluorescent agents. *Chromosoma*. 1 juin 1970;30(2):215-27.
- 70.....Drets ME, Shaw MW. Specific Banding Patterns of Human Chromosomes. *Proc Natl Acad Sci U S A*. sept 1971;68(9):2073-7.
- 71.Verma RS, Lubs HA. A simple R banding technic. *Am J Hum Genet*. janv 1975;27(1):110-7.
- 72.....Debatisse Mi, Achkar ,Eliane El, and Dutrillaux B. Common Fragile Sites Nested at the Interfaces of Early and Late-Replicating Chromosome Bands : Cis Acting Components of the G2/M Checkpoint? *Cell Cycle*. 16 mars 2006;5(6):578-81.
- 73.....Karyotyping | Learn Science at Scitable [Internet]. [cité 9 janv 2025]. Disponible sur: <http://www.nature.com/scitable/topicpage/karyotyping-for-chromosomal-abnormalities-298>
- 74.....Berisha SZ, Shetty S, Prior TW, Mitchell AL. Cytogenetic and molecular diagnostic testing associated with prenatal and postnatal birth defects. *Birth Defects Res*. 1 mars 2020;112(4):293-306.
- 75.....Pardue ML, Gall JG. MOLECULAR HYBRIDIZATION OF RADIOACTIVE DNA TO THE DNA OF CYTOLOGICAL PREPARATIONS. *Proc Natl Acad Sci U S A*. oct 1969;64(2):600-4.
- 76.....Liehr T. Molecular Cytogenetics in the Era of Chromosomics and Cytogenomic Approaches. *Front Genet* [Internet]. 13 oct 2021 [cité 9 janv 2025];12. Disponible sur: <https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2021.720507/full>
- 77.....Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*. 30 oct 1992;258(5083):818-21.
- 78.....Qin D. Next-generation sequencing and its clinical application. *Cancer Biol Med*. févr 2019;16(1):4-10.
- 79.....Reinartz J, Bruyins E, Lin JZ, Burcham T, Brenner S, Bowen B, et al. Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Brief Funct Genomic Proteomic*. févr 2002;1(1):95-104.
- 80.....Slatko BE, Gardner AF, Ausubel FM. Overview of Next Generation Sequencing Technologies. *Curr Protoc Mol Biol*. avr 2018;122(1):e59.
- 81.....Muñoz-Barrera A, Rubio-Rodríguez LA, Díaz-de Usera A, Jáspez D, Lorenzo-Salazar JM, González-Montelongo R, et al. From Samples to Germline and Somatic Sequence Variation: A Focus on Next-Generation Sequencing in Melanoma Research. *Life*. nov 2022;12(11):1939.
- 82.....Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. *Brief Funct Genomics*. sept 2015;14(5):305-14.

- 83.....Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. *Genome Biology*. 20 nov 2019;20(1):246.
- 84.....Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 15 févr 2001;409(6822):860-921.
- 85.....Meyer CA, Liu XS. Identifying and Mitigating Bias in Next-Generation Sequencing Methods for Chromatin Biology. *Nat Rev Genet*. nov 2014;15(11):709-21.
- 86.....Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol*. 20 mai 2019;20(1):97.
- 87.....Delpu Y, Barseghyan H, Bocklandt S, Hastie A, Chaubey A. Chapter 8 - Next-generation cytogenomics: High-resolution structural variation detection by optical genome mapping. In: Liehr T, éditeur. *Cytogenomics* [Internet]. Academic Press; 2021 [cité 21 mars 2025]. p. 123-46. Disponible sur: <https://www.sciencedirect.com/science/article/pii/B9780128235799000229>
- 88.....Dremsek P, Schwarz T, Weil B, Malashka A, Laccone F, Neesen J. Optical Genome Mapping in Routine Human Genetic Diagnostics—Its Advantages and Limitations. *Genes (Basel)*. 8 déc 2021;12(12):1958.
- 89.....Hiatt SM, Lawlor MJ, Handley LH, Ramaker RC, Rogers BB, Partridge EC, et al. Long-read genome sequencing for the molecular diagnosis of neurodevelopmental disorders. *HGG Adv*. 8 avr 2021;2(2):100023.
- 90.....Kj T, Cs C, Dr R, Js E, Sw T. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic acids research* [Internet]. août 2010 [cité 20 févr 2025];38(15). Disponible sur: <https://pubmed.ncbi.nlm.nih.gov/20571086/>
- 91.....Rhoads A, Au KF. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*. 1 oct 2015;13(5):278-89.
- 92.....Graham MDon. The Coulter Principle: A history. *Cytometry Part A*. 2022;101(1):8-11.
- 93.....Graham MDon. The Coulter principle: Imaginary origins. *Cytometry Part A*. 2013;83(12):1057-61.
- 94.....Feng Y, Zhang Y, Ying C, Wang D, Du C. Nanopore-based Fourth-generation DNA Sequencing Technology. *Genomics Proteomics Bioinformatics*. févr 2015;13(1):4-16.
- 95.....Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. *Nat Biotechnol*. 6 mai 2016;34(5):518-24.
- 96.....Wang Y, Yang Q, Wang Z. The evolution of nanopore sequencing. *Frontiers in Genetics* [Internet]. 2015 [cité 6 févr 2023];5. Disponible sur: <https://www.frontiersin.org/articles/10.3389/fgene.2014.00449>

- 97.....Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*. 26 nov 1996;93(24):13770-3.
- 98.....Akeson M, Branton D, Kasianowicz JJ, Brandin E, Deamer DW. Microsecond Time-Scale Discrimination Among Polycytidylic Acid, Polyadenylic Acid, and Polyuridylic Acid as Homopolymers or as Segments Within Single RNA Molecules. *Biophysical Journal*. 1 déc 1999;77(6):3227-33.
- 99.....S H, S C, H B. Sequence-specific detection of individual DNA strands using engineered nanopores. *Nature biotechnology* [Internet]. juill 2001 [cité 20 févr 2025];19(7). Disponible sur: <https://pubmed.ncbi.nlm.nih.gov/11433274/>
- 100.....Stoddart D, Heron AJ, Mikhailova E, Maglia G, Bayley H. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America*. 20 avr 2009;106(19):7702.
- 101.....Haque F, Li J, Wu HC, Liang XJ, Guo P. Solid-State and Biological Nanopore for Real-Time Sensing of Single Chemical and Sequencing of DNA. *Nano Today*. févr 2013;8(1):56-74.
- 102.....Lin B, Hui J, Mao H. Nanopore Technology and Its Applications in Gene Sequencing. *Biosensors*. 30 juin 2021;11(7):214.
- 103.....Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol*. nov 2021;39(11):1348-65.
- 104.....Lin B, Hui J, Mao H. Nanopore Technology and Its Applications in Gene Sequencing. *Biosensors (Basel)*. 30 juin 2021;11(7):214.
- 105.....Montel F. Séquençage de l'ADN par nanopores - Résultats et perspectives. *Med Sci (Paris)*. 1 févr 2018;34(2):161-5.
- 106.....Payne A, Holmes N, Clarke T, Munro R, Debebe B, Loose M. Nanopore adaptive sequencing for mixed samples, whole exome capture and targeted panels [Internet]. *bioRxiv*; 2020 [cité 8 nov 2022]. Disponible sur: <https://www.biorxiv.org/content/10.1101/2020.02.03.926956v1>
- 107.....Zhang H, Li H, Jain C, Cheng H, Au KF, Li H, et al. Real-time mapping of nanopore raw signals. *Bioinformatics*. 12 juill 2021;37(Suppl_1):i477-83.
- 108.....Nakamichi K, Huey J, Sangermano R, Place EM, Bujakowska KM, Marra M, et al. Targeted long-read sequencing enriches disease-relevant genomic regions of interest to provide complete Mendelian disease diagnostics. *JCI Insight*. 12 sept 2024;9(20):e183902.
- 109.....Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, et al. Targeted long-read sequencing identifies missing disease-causing variation. *Am J Hum Genet*. 5 août 2021;108(8):1436-49.

- 110.....Nyaga DM, Tsai P, Gebbie C, Phua HH, Yap P, Le Quesne Stabej P, et al. Benchmarking nanopore sequencing and rapid genomics feasibility: validation at a quaternary hospital in New Zealand. *npj Genom Med*. 8 nov 2024;9(1):1-13.
- 111.....Sanderson ND, Kapel N, Rodger G, Webster H, Lipworth S, Street TL, et al. Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. *Microb Genom*. 10 janv 2023;9(1):mgen000910.
- 112.....Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat Commun*. 6 nov 2017;8(1):1326.
- 113.....Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med*. janv 2018;20(1):159-63.
- 114.....Miao H, Zhou J, Yang Q, Liang F, Wang D, Ma N, et al. Long-read sequencing identified a causal structural variant in an exome-negative case and enabled preimplantation genetic diagnosis. *Hereditas*. 2018;155:32.
- 115.....Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. mai 2015;17(5):405-24.
- 116.....Harrison SM, Biesecker LG, Rehm HL. Overview of specifications to the ACMG/AMP variant interpretation guidelines. *Curr Protoc Hum Genet*. sept 2019;103(1):e93.
- 117.....Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 15 sept 2018;34(18):3094-100.
- 118.....Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology*. 3 août 2020;21(1):189.
- 119.....Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol*. oct 2024;42(10):1571-80.
- 120.....Nattestad M, Aboukhalil R, Chin CS, Schatz MC. Ribbon: intuitive visualization for complex genomic variation. *Bioinformatics*. 20 avr 2021;37(3):413-5.
- 121.....Olsen SK, Garbi M, Zampieri N, Eliseenkova AV, Ornitz DM, Goldfarb M, et al. Les facteurs homologues du facteur de croissance des fibroblastes (FGF) partagent une homologie structurelle mais non fonctionnelle avec les FGF *. *Journal of Biological Chemistry*. 5 sept 2003;278(36):34226-36.
- 122.....Goldfarb M. FIBROBLAST GROWTH FACTOR HOMOLOGOUS FACTORS: EVOLUTION, STRUCTURE, AND FUNCTION. *Cytokine Growth Factor Rev*. avr 2005;16(2):215-20.

- 123.....Biadun M, Karelus R, Krowarsch D, Opalinski L, Zakrzewska M. FGF12: biology and function. *Differentiation*. 25 nov 2023;100740.
- 124.....Hanada Y, Nakamura Y, Ozono Y, Ishida Y, Takimoto Y, Taniguchi M, et al. Fibroblast growth factor 12 is expressed in spiral and vestibular ganglia and necessary for auditory and equilibrium function. *Sci Rep*. 31 juill 2018;8(1):11491.
- 125.....Siekierska A, Isrie M, Liu Y, Scheldeman C, Vanthillo N, Lagae L, et al. Gain-of-function FHF1 mutation causes early-onset epileptic encephalopathy with cerebellar atrophy. *Neurology*. 7 juin 2016;86(23):2162.
- 126.....Guella I, Huh L, McKenzie MB, Toyota EB, Bebin EM, Thompson ML, et al. De novo FGF12 mutation in 2 patients with neonatal-onset epilepsy. *Neurol Genet*. 10 nov 2016;2(6):e120.
- 127.....Villeneuve N, Abidi A, Cacciagli P, Mignon-Ravix C, Chabrol B, Villard L, et al. Heterogeneity of FHF1 related phenotype: Novel case with early onset severe attacks of apnea, partial mitochondrial respiratory chain complex II deficiency, neonatal onset seizures without neurodegeneration. *Eur J Paediatr Neurol*. sept 2017;21(5):783-6.
- 128.....Takeguchi R, Haginoya K, Uchiyama Y, Fujita A, Nagura M, Takeshita E, et al. Two Japanese cases of epileptic encephalopathy associated with an FGF12 mutation. *Brain Dev*. sept 2018;40(8):728-32.
- 129.....Velíšková J, Marra C, Liu Y, Shekhar A, Park DS, Iatckova V, et al. Early onset epilepsy and sudden unexpected death in epilepsy with cardiac arrhythmia in mice carrying the early infantile epileptic encephalopathy 47 gain-of-function FHF1(FGF12) missense mutation. *Epilepsia*. juill 2021;62(7):1546-58.
- 130.....Trivisano M, Ferretti A, Bebin E, Huh L, Lesca G, Siekierska A, et al. Defining the phenotype of FHF1 developmental and epileptic encephalopathy. *Epilepsia*. juill 2020;61(7):e71-8.
- 131.....Paprocka J, Jezela-Stanek A, Koppolu A, Rydzanicz M, Kosińska J, Stawiński P, et al. FGF12p.Gly112Ser variant as a cause of phenytoin/phenobarbital responsive epilepsy. *Clin Genet*. sept 2019;96(3):274-5.
- 132.....Willemsen MH, Goel H, Verhoeven JS, Braakman HMH, de Leeuw N, Freeth A, et al. Epilepsy phenotype in individuals with chromosomal duplication encompassing FGF12. *Epilepsia Open*. 9 mai 2020;5(2):301-6.
- 133.....Shi RM, Kobayashi T, Kikuchi A, Sato R, Uematsu M, An K, et al. Phenytoin-responsive epileptic encephalopathy with a tandem duplication involving FGF12. *Neurol Genet*. févr 2017;3(1):e133.
- 134.....Abraham A, Ramsey K, Belnap N, Szelinger S, Jepsen W, Balak C, et al. FGF12 copy number variant associated with epileptic encephalopathy. *Clin Genet*. juill 2024;106(1):114-5.

- 135.....Seiffert S, Pendziwiat M, Bierhals T, Goel H, Schwarz N, Ven A van der, et al. Modulating effects of FGF12 variants on NaV1.2 and NaV1.6 being associated with developmental and epileptic encephalopathy and Autism spectrum disorder: A case series. *eBioMedicine* [Internet]. 1 sept 2022 [cité 20 août 2024];83. Disponible sur: [https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964\(22\)00416-9/fulltext](https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(22)00416-9/fulltext)
- 136.....Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 15 déc 2009;10:421.
- 137.....Verheyen S, Speicher MR, Ramler B, Plecko B. Childhood-onset epileptic encephalopathy due to FGF12 exon 1-4 tandem duplication. *Neurol Genet*. oct 2020;6(5):e494.
- 138.....Mantere T, Kersten S, Hoischen A. Long-Read Sequencing Emerging in Medical Genetics. *Front Genet*. 2019;10:426.
- 139.....Szafranski P, Kośmider E, Liu Q, Karolak JA, Currie L, Parkash S, et al. LINE- and Alu-containing genomic instability hotspot at 16q24.1 associated with recurrent and nonrecurrent CNV deletions causative for ACDMPV. *Hum Mutat*. déc 2018;39(12):1916-25.
- 140.....Xiao Y, Theile JW, Zybura A, Pan Y, Lin Z, Cummins TR. A-type FHF's mediate resurgent currents through TTX-resistant voltage-gated sodium channels. *Elife*. 20 avr 2022;11:e77558.
- 141.....Goldfarb M. Fibroblast growth factor homologous factors: canonical and non-canonical mechanisms of action. *The Journal of Physiology*. 2024;602(17):4097-110.
- 142.....Ohoi S, Miyauchi A, Osaka H, Lourenco CM, Arakaki N, Sengoku T, et al. Biallelic structural variations within FGF12 detected by long-read sequencing in epilepsy. *Life Sci Alliance*. août 2023;6(8):e202302025.
- 143.....Wolff M, Johannesen KM, Hedrich UBS, Masnada S, Rubboli G, Gardella E, et al. Genetic and phenotypic heterogeneity suggest therapeutic implications in SCN2A-related disorders. *Brain*. 1 mai 2017;140(5):1316-36.
- 144.....Kaplan MR, Cho MH, Ullian EM, Isom LL, Levinson SR, Barres BA. Differential control of clustering of the sodium channels Na(v)1.2 and Na(v)1.6 at developing CNS nodes of Ranvier. *Neuron*. avr 2001;30(1):105-19.
- 145.....Acsády L, Kamondi A, Sík A, Freund T, Buzsáki G. GABAergic cells are the major postsynaptic targets of mossy fibers in the rat hippocampus. *J Neurosci*. 1 mai 1998;18(9):3386-403.
- 146.....Galovic M, Schmitz B, Tettenborn B. C15EEG in Inflammatory Disorders, Cerebrovascular Diseases, Trauma, and Migraine. In: Schomer DL, Lopes da Silva FH, Schomer DL, Lopes da Silva FH, éditeurs. *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields* [Internet]. Oxford University Press; 2017 [cité 20 juin 2025]. p. 0. Disponible sur: <https://doi.org/10.1093/med/9780190228484.003.0015>

- 147.....Verma S, Pantoom S, Petters J, Pandey AK, Hermann A, Lukas J. A molecular genetics view on Mucopolysaccharidosis Type II. *Mutation Research/Reviews in Mutation Research*. 1 juill 2021;788:108392.
- 148.....Bondeson ML, Dahl N, Malmgren H, Kleijer WJ, Tønnesen T, Carlberg BM, et al. Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. *Hum Mol Genet*. avr 1995;4(4):615-21.
- 149.....Fauqueux J, Meneboo JP, Caumes R, Thomès L, Yahya EA, Thuillier C, et al. Combining long-read DNA and RNA sequencing to enhance molecular understanding of structural variations leading to copy gains. *Comput Struct Biotechnol J*. 2025;27:1732-40.
- 150.....Schuy J, Grochowski CM, Carvalho CMB, Lindstrand A. Complex genomic rearrangements: an underestimated cause of rare diseases. *Trends Genet*. nov 2022;38(11):1134-46.
- 151.....Newman S, Hermetz KE, Weckselblatt B, Rudd MK. Next-Generation Sequencing of Duplication CNVs Reveals that Most Are Tandem and Some Create Fusion Genes at Break-points. *The American Journal of Human Genetics*. 5 févr 2015;96(2):208-20.
- 152.....Boycott KM, Rath A, Chong JX, Hartley T, Alkuraya FS, Baynam G, et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *Am J Hum Genet*. 4 mai 2017;100(5):695-705.
- 153.....De Clercq G, Vantomme L, Dewaele B, Callewaert B, Vanakker O, Janssens S, et al. Full characterization of unresolved structural variation through long-read sequencing and optical genome mapping. *Sci Rep*. 25 nov 2024;14(1):1-12.
- 154.....Damaraju N, Miller AL, Miller DE. Long-Read DNA and RNA Sequencing to Streamline Clinical Genetic Testing and Reduce Barriers to Comprehensive Genetic Testing. *J Appl Lab Med*. 3 janv 2024;9(1):138-50.
- 155.....Owen DJ, Aguilar-Martinez E, Ji Z, Li Y, Sharrocks AD. ZMYM2 controls human transposable element transcription through distinct co-regulatory complexes. *Elife*. 7 nov 2023;12:RP86669.
- 156.....Connaughton DM, Dai R, Owen DJ, Marquez J, Mann N, Graham-Paquin AL, et al. Mutations of the Transcriptional Corepressor ZMYM2 Cause Syndromic Urinary Tract Malformations. *Am J Hum Genet*. 1 oct 2020;107(4):727-42.
- 157.....Fregeau B, Kim BJ, Hernández-García A, Jordan VK, Cho MT, Schnur RE, et al. De Novo Mutations of RERE Cause a Genetic Syndrome with Features that Overlap Those Associated with Proximal 1p36 Deletions. *Am J Hum Genet*. 5 mai 2016;98(5):963-70.
- 158.....Jordan VK, Fregeau B, Ge X, Giordano J, Wapner RJ, Balci TB, et al. Genotype–phenotype correlations in individuals with pathogenic RERE variants. *Hum Mutat*. mai 2018;39(5):666-75.
- 159.....Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. arXiv; 2013 [cité 6 mai 2025]. Disponible sur: <http://arxiv.org/abs/1303.3997>

- 160.....Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun*. 15 janv 2018;9(1):1-15.
- 161.....Sambamba: fast processing of NGS alignment formats | Bioinformatics | Oxford Academic [Internet]. [cité 6 mai 2025]. Disponible sur: <https://academic-oup-com.ressources-electroniques.univ-lille.fr/bioinformatics/article/31/12/2032/214758>
- 162.....Riggs ER, Andersen EF, Cherry AM, Kantarci S, Kearney H, Patel A, et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet Med*. févr 2020;22(2):245-57.
- 163.....Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*. janv 2024;625(7993):92-100.
- 164.....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*. janv 2014;42(Database issue):D986-992.
- 165.....Fujimoto A, Wong JH, Yoshii Y, Akiyama S, Tanaka A, Yagi H, et al. 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*. 29 avr 2021;13(1):65.
- 166.....Shimajima K, Mano T, Kashiwagi M, Tanabe T, Sugawara M, Okamoto N, et al. Pelizaeus-Merzbacher disease caused by a duplication-inverted triplication-duplication in chromosomal segments including the PLP1 region. *Eur J Med Genet*. juin 2012;55(6-7):400-3.
- 167.....Estruch SB, Graham SA, Quevedo M, Vино A, Dekkers DHW, Deriziotis P, et al. Proteomic analysis of FOXP proteins reveals interactions between cortical transcription factors associated with neurodevelopmental disorders. *Hum Mol Genet*. 1 avr 2018;27(7):1212-27.
- 168.....Wu ST, Feng Y, Song R, Qi Y, Li L, Lu D, et al. Foxp1 Is Required for Renal Intercalated Cell Differentiation and Acid-Base Regulation. *J Am Soc Nephrol*. 1 mai 2024;35(5):533-48.
- 169.....Lieber MR, Gu J, Lu H, Shimazaki N, Tsai AG. Nonhomologous DNA end joining (NHEJ) and chromosomal translocations in humans. *Subcell Biochem*. 2010;50:279-96.
- 170.....Delias N. A family of long intergenic non-coding RNA genes in human chromosomal region 22q11.2 carry a DNA translocation breakpoint/AT-rich sequence. *PLoS One*. 2018;13(4):e0195702.
- 171.....Correll-Tash S, Lilley B, Salmons Iv H, Mlynarski E, Franconi CP, McNamara M, et al. Double strand breaks (DSBs) as indicators of genomic instability in PATRR-mediated translocations. *Hum Mol Genet*. 25 févr 2021;29(24):3872-81.

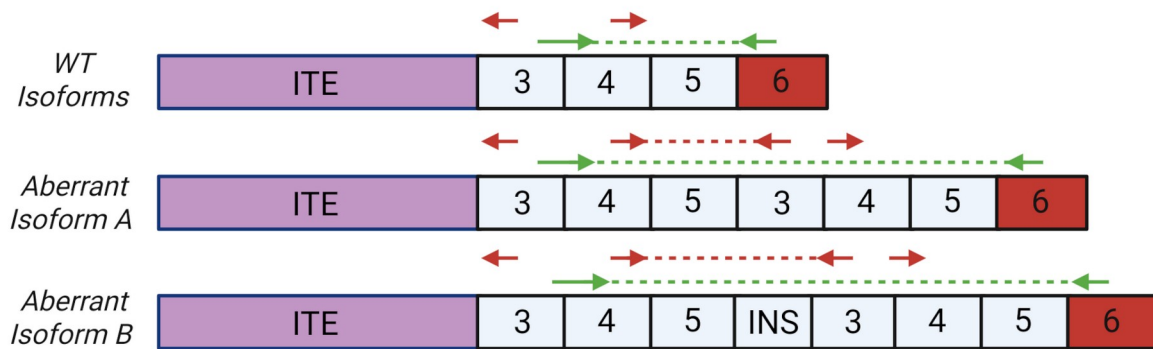
- 172.....Ait Saada A, Guo W, Costa AB, Yang J, Wang J, Lobachev KS. Widely spaced and divergent inverted repeats become a potent source of chromosomal rearrangements in long single-stranded DNA regions. *Nucleic Acids Res.* 8 mai 2023;51(8):3722-34.
- 173.....Hug N, Longman D, Cáceres JF. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res.* 29 févr 2016;44(4):1483-95.
- 174.....Sánchez-Marín D, Silva-Cázares MB, Porras-Reyes FI, García-Román R, Campos-Parra AD. Breaking paradigms: Long non-coding RNAs forming gene fusions with potential implications in cancer. *Genes Dis.* mai 2024;11(3):101136.
- 175.....Wang X, Xu J, Zhang B, Hou Y, Song F, Lyu H, et al. Genome-wide detection of enhancer-hijacking events from chromatin interaction data in re-arranged genomes. *Nat Methods.* juin 2021;18(6):661-8.
- 176.....Ramani V, Shendure J, Duan Z. Understanding Spatial Genome Organization: Methods and Insights. *Genomics Proteomics Bioinformatics.* févr 2016;14(1):7-20.
- 177.....Helal AA, Saad BT, Saad MT, Mosaad GS, Aboshanab KM. Benchmarking long-read aligners and SV callers for structural variation detection in Oxford nanopore sequencing data. *Sci Rep.* 14 mars 2024;14(1):6160.
- 178.....Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet.* oct 2020;21(10):597-614.
- 179.....Vollmers C, Penland L, Kanbar JN, Quake SR. Novel Exons and Splice Variants in the Human Antibody Heavy Chain Identified by Single Cell and Single Molecule Sequencing. *PLOS ONE.* 22 janv 2015;10(1):e0117050.
- 180.....Nattestad M, Goodwin S, Ng K, Baslan T, Sedlazeck FJ, Rescheneder P, et al. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res.* août 2018;28(8):1126-35.
- 181.....McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol.* 6 juin 2016;17(1):122.
- 182.....Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics.* 15 avr 2016;32(8):1220-2.
- 183.....Minoche AE, Lundie B, Peters GB, Ohnesorg T, Pinese M, Thomas DM, et al. ClinSV: clinical grade structural and copy number variant detection from whole genome sequencing data. *Genome Med.* 25 févr 2021;13(1):32.
- 184.....Gonsalves R, Aleck K, Newbern D, Shaibi G, Kapadia C, Oatman O. Severe early onset obesity and hypopituitarism in a child with a novel SIM1 gene mutation. *Endocrinol Diabetes Metab Case Rep.* 6 oct 2020;2020:20-0042, EDM200042.

- 185.....Taniguchi N, Watanuki K, Nakato D, Takenouchi T, Kosaki K, Koga H. PURA-related neurodevelopmental disorders: a systematic review on genotype-phenotype correlations. *J Med Genet.* 26 févr 2025;62(3):191-8.
- 186.....Santos-Rebouças CB, Boy R, Fernandes GNS, Gonçalves AP, Abdala BB, Gonzalez LGC, et al. A novel Xp11.22 duplication involving HUWE1 in a male with syndromic intellectual disability and additional neurological findings. *European Journal of Medical Genetics.* 1 avr 2023;66(4):104716.
- 187.....Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet.* juill 2009;41(7):849-53.
- 188.....McVey M, Lee SE. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet.* nov 2008;24(11):529-38.
- 189.....Grochowski CM, Bengtsson JD, Du H, Gandhi M, Lun MY, Mehaffey MG, et al. Inverted triplications formed by iterative template switches generate structural variant diversity at genomic disorder *loci*. *Cell Genomics.* 10 juill 2024;4(7):100590.
- 190.....Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol.* oct 2024;42(10):1571-80.
- 191.....Yuan N, Jia P. Comprehensive assessment of long-read sequencing platforms and calling algorithms for detection of copy number variation. *Brief Bioinform.* 10 sept 2024;25(5):bbae441.
- 192.....Dierckxsens N, Li T, Vermeesch JR, Xie Z. A benchmark of structural variation detection by long reads through a realistic simulated model. *Genome Biology.* 15 déc 2021;22(1):342.
- 193.....Dierckxsens N, Li T, Vermeesch JR, Xie Z. A benchmark of structural variation detection by long reads through a realistic simulated model. *Genome Biology.* 15 déc 2021;22(1):342.
- 194.....Senthamizhan V, Ravindran B, Raman K. NetGenes: A Database of Essential Genes Predicted Using Features From Interaction Networks. *Front Genet.* 2021;12:722198.
- 195.....Blauen A, Stutterd CA, Stouffs K, Dumitriu D, Deggouj N, Lockhart PJ, et al. Chudley-McCullough Syndrome: A Recognizable Clinical Entity Characterized by Deafness and Typical Brain Malformations. *J Child Neurol.* 1 févr 2021;36(2):152-8.
- 196.....Hancks DC, Kazazian H. SVA retrotransposons: Evolution and genetic instability. *Semin Cancer Biol.* août 2010;20(4):234-45.
- 197.....Pfaff AL, Singleton LM, Kõks S. Mechanisms of disease-associated SINE-VNTR-Alus. *Exp Biol Med (Maywood).* mai 2022;247(9):756-64.
- 198.....Tripathi M, Chauhan U, Tripathi P, Agrawal S. Role of Alu Element in Detecting Population Diversity. *Int J Hum Genet.* 1 mars 2008;8:61-74.

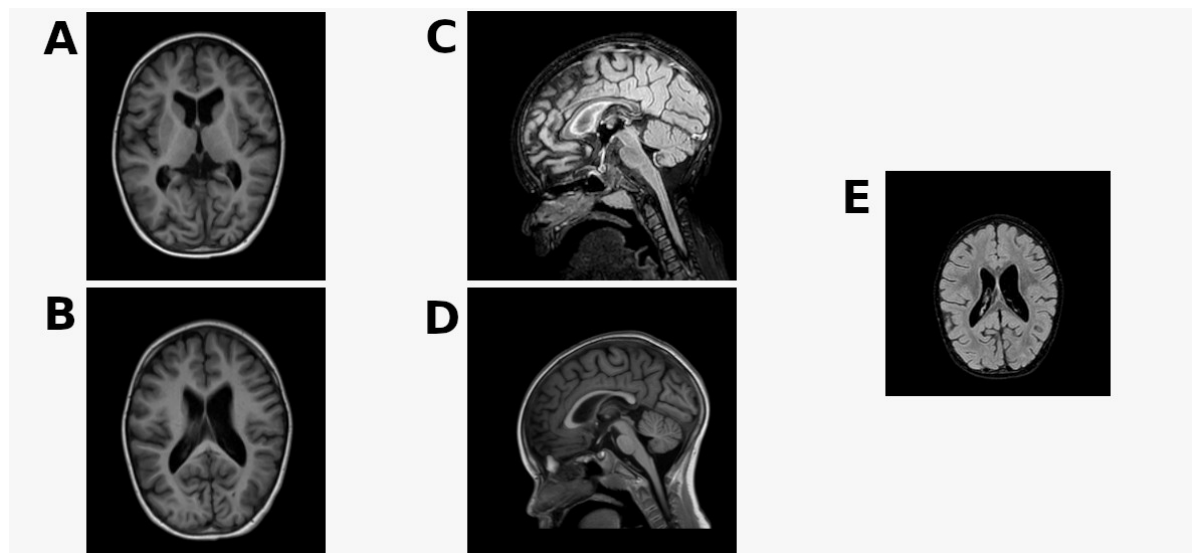
- 199.....Ravel JM, Renaud M, Muller J, Becker A, Renard É, Remen T, et al. Clinical utility of periodic reinterpretation of CNVs of uncertain significance: an 8-year retrospective study. *Genome Med.* 23 mai 2023;15:39.
- 200.....Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, et al. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics.* 15 oct 2018;34(20):3572-4.
- 201.....Negi S, Stenton SL, Berger SI, Canigiula P, McNulty B, Violich I, et al. Advancing long-read nanopore genome assembly and accurate variant calling for rare disease detection. *The American Journal of Human Genetics.* févr 2025;112(2):428-49.
- 202.....Carvalho CMB, Zhang F, Liu P, Patel A, Sahoo T, Bacino CA, et al. Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. *Hum Mol Genet.* 15 juin 2009;18(12):2188-203.
- 203.....Shen J, Ding T, Sun X, Yang J, Zhang Y, Wang J, et al. Comprehensive analysis of genomic complexity in the 5' end coding region of the DMD gene in patients of exons 1–2 duplications based on long-read sequencing. *BMC Genomics.* 19 mars 2024;25:292.
- 204.....Leung HCM, Yu H, Zhang Y, Leung WS, Lo IFM, Luk HM, et al. Detecting structural variations with precise breakpoints using low-depth WGS data from a single oxford nanopore MinION flowcell. *Sci Rep.* 16 mars 2022;12(1):1-8.
- 205.....Kovaka S, Fan Y, Ni B, Timp W, Schatz MC. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat Biotechnol.* avr 2021;39(4):431-41.
- 206.....Miller DE, Lee L, Galey M, Kandhaya-Pillai R, Tischkowitz M, Amalnath D, et al. Targeted long-read sequencing identifies missing pathogenic variants in unsolved Werner syndrome cases. *J Med Genet.* nov 2022;59(11):1087-94.
- 207.....Kosugi S, Terao C. Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. *Hum Genome Var.* 17 avr 2024;11(1):1-10.
- 208.....Liu Z, Xie Z, Li M. Comprehensive and deep evaluation of structural variation detection pipelines with third-generation sequencing data. *Genome Biol.* 15 juill 2024;25:188.
- 209.....Paschal CR, Zalusky MPG, Beck AE, Gillentine MA, Narayanan J, Damaraju N, et al. Concordance of Whole-Genome Long-Read Sequencing with Standard Clinical Testing for Prader-Willi and Angelman Syndromes. *The Journal of Molecular Diagnostics.* 1 mars 2025;27(3):166-76.
- 210.....LaFlamme CW, Rastin C, Sengupta S, Pennington HE, Russ-Hall SJ, Schneider AL, et al. Diagnostic utility of DNA methylation analysis in genetically unsolved pediatric epilepsies and CHD2 epismature refinement. *Nat Commun.* 6 août 2024;15:6524.
- 211.....Pauper M, Kucuk E, Wenger AM, Chakraborty S, Baybayan P, Kwint M, et al. Long-read trio sequencing of individuals with unsolved intellectual disability. *Eur J Hum Genet.* avr 2021;29(4):637-48.

- 212.....Höps W, Weiss MM, Derks R, Galbany JC, Ouden A den, van den Heuvel S, et al. HiFi long-read genomes for difficult-to-detect, clinically relevant variants. *Am J Hum Genet.* 6 févr 2025;112(2):450-6.
- 213.....Mahmoud M, Huang Y, Garimella K, Audano PA, Wan W, Prasad N, et al. Utility of long-read sequencing for All of Us. *Nat Commun.* 29 janv 2024;15:837.
- 214.....Harvey WT, Ebert P, Ebler J, Audano PA, Munson KM, Hoekzema K, et al. Whole-genome long-read sequencing downsampling and its effect on variant calling precision and recall. *bioRxiv.* 4 mai 2023;2023.05.04.539448.

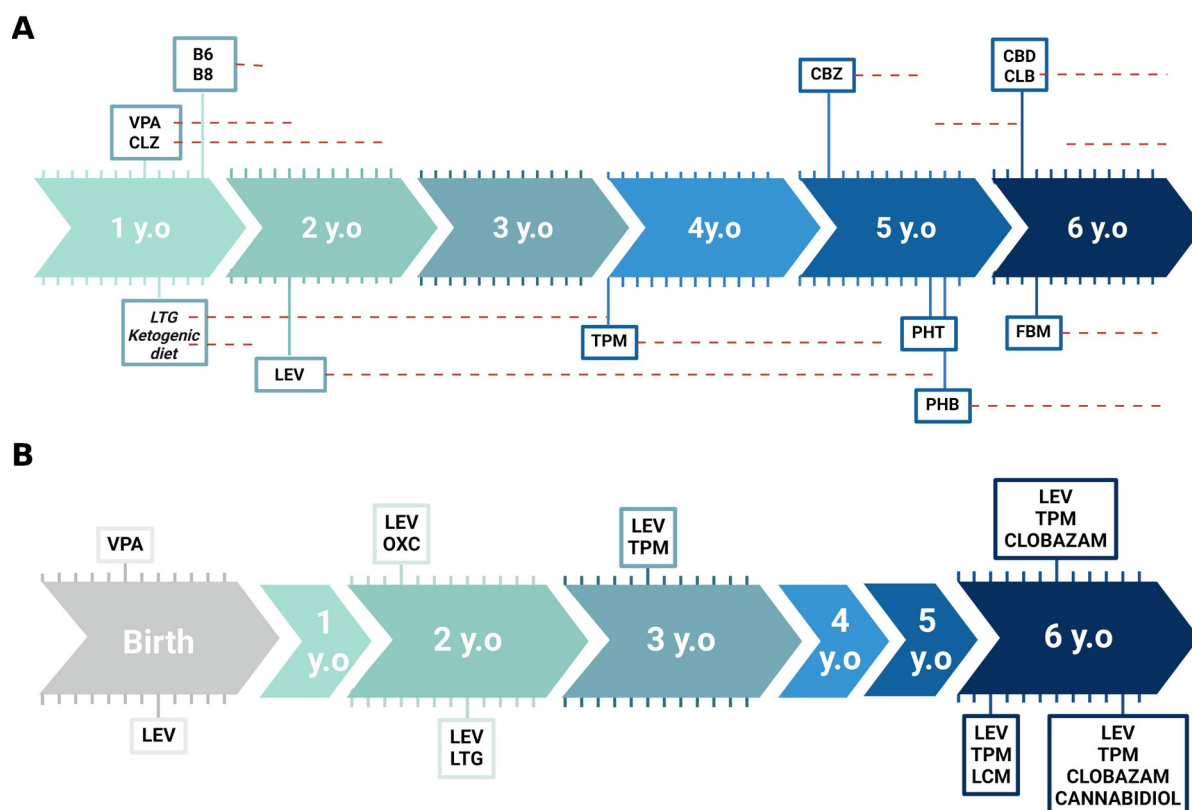
VIII/ Annexes



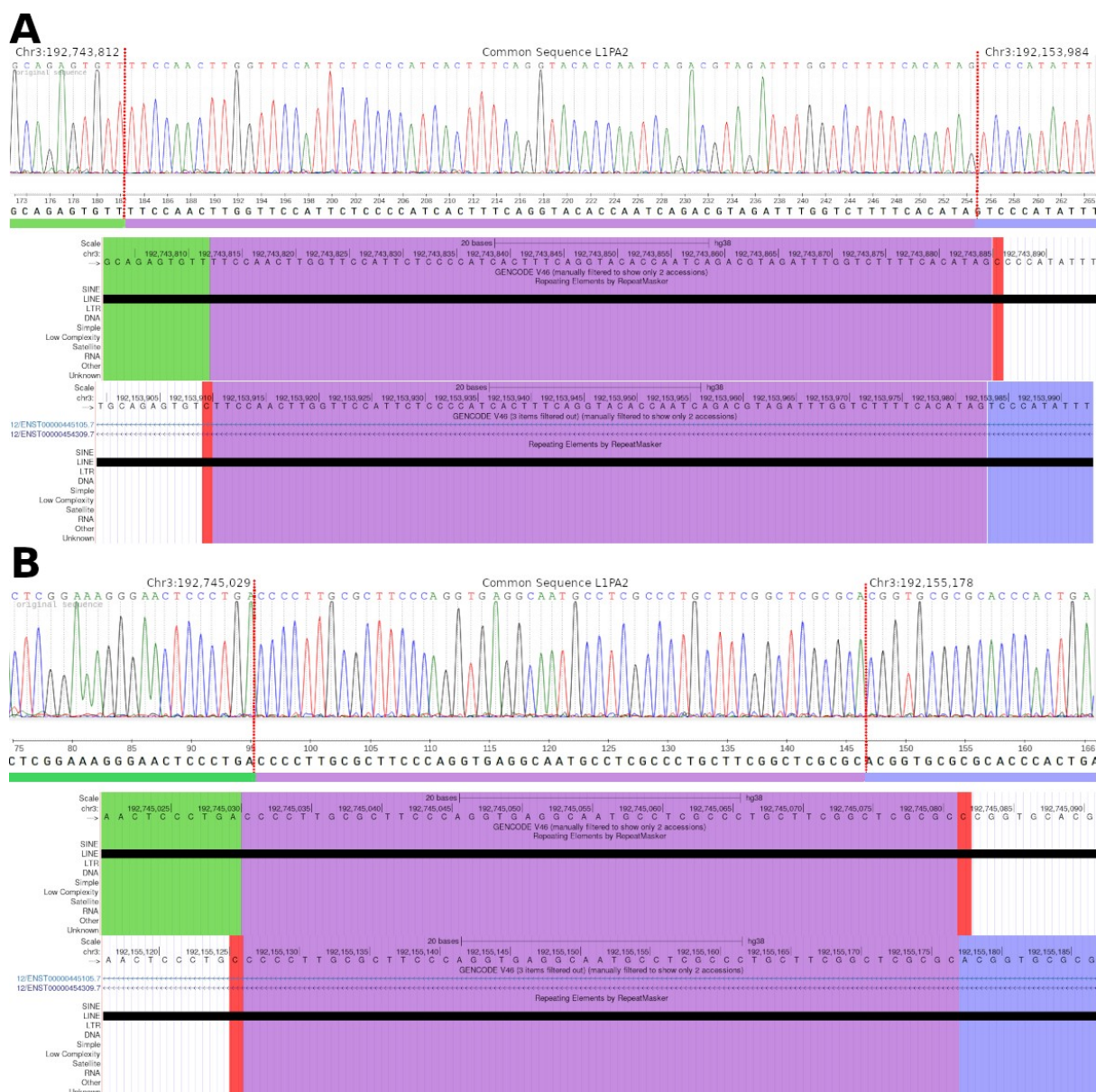
Figure_Annexe 1 : Positions des amorces pour l'amplification des transcrits de FGF12. Les flèches vertes indiquent les amorces couvrant les exons 4 à 3 (orientation inverse), amplifiant spécifiquement uniquement les transcrits aberrants.



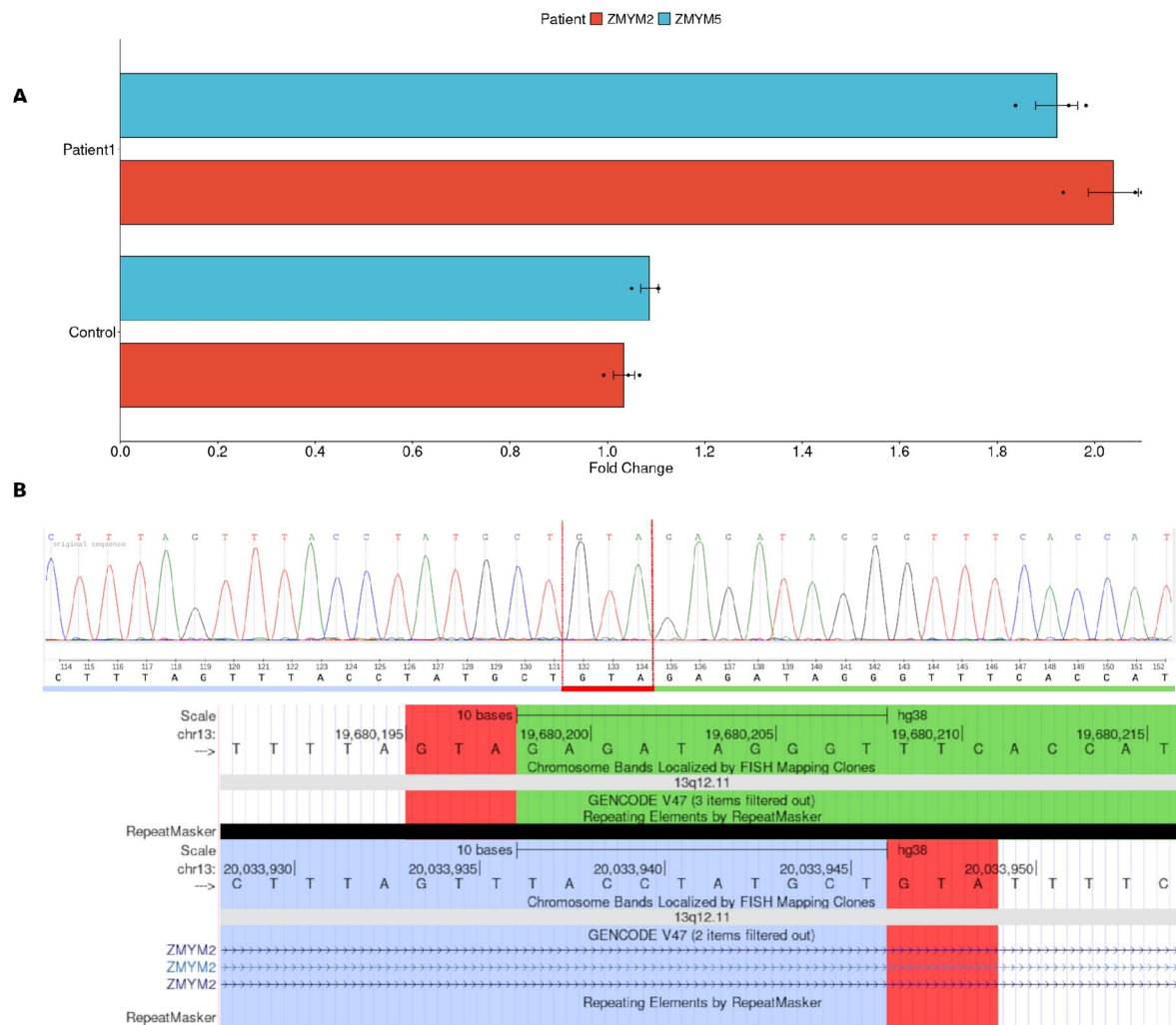
Figure_Annexe 2 : IRM cérébrale de UPN1 à l'âge de 4 ans. (A-B) Coupes axiales en pondération T1. (C-D) Coupes sagittales en pondération T1. (E) Image en séquence FLAIR.



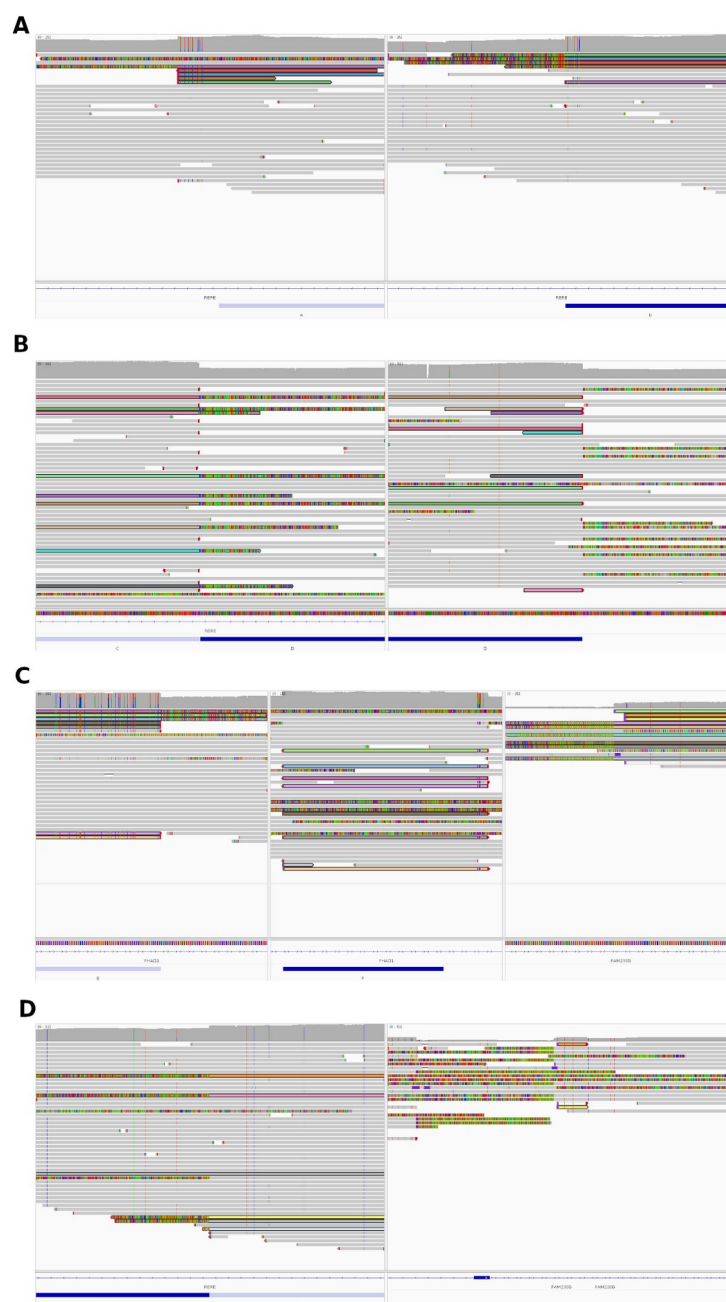
Figure_Annexe 3 : Chronologie des traitements thérapeutiques. (A) Chronologie des traitements thérapeutiques pour UPN1. Les lignes rouges en pointillés indiquent la durée de chaque traitement. (B) Chronologie des traitements thérapeutiques pour UPN2.



Figure_Annexe 4 : Séquençage Sanger des points de cassure de la duplication en tandem. (A) Séquençage Sanger du point de cassure de la duplication en tandem chez UPN, montrant une séquence homologue de 73 pb. (B) Séquençage Sanger du point de cassure de la duplication en tandem chez UPN2, montrant une séquence homologue de 53 pb. Légende des couleurs : le vert indique la fin de la duplication, le bleu indique le début de la duplication, le violet représente la séquence commune des deux éléments L1PA2, et le rouge désigne la séquence non homologue entre les deux éléments L1PA2.



Figure_Annexe 5 : Analyse moléculaire complémentaire du gain de deux copies chez le patient UPN4. (A) qPCR de la région dupliquée chez un témoin sain (WT) et le UPN4, confirmant la présence d'un gain de deux copies chez le UPN4. (B) Séquençage Sanger du point de cassure chez le UPN4 montrant une homologie de 3 pb entre les séquences en amont et en aval.



Figure_Annexe 6 : Visualisation dans Integrated Genome Viewer (IGV) de l'ADN du patient UPN5 à partir des données de séquençage long-read Nanopore. (A) Visualisation IGV des lectures couvrant les régions A et B, montrant une inversion. **(B)** Visualisation IGV des lectures couvrant les régions C et D, montrant une inversion. **(C)** Visualisation IGV des lectures couvrant les régions E, F et le chromosome 22, montrant un remaniement complexe. **(D)** Visualisation IGV des lectures couvrant les régions B, C et le chromosome 22, montrant un remaniement interchromosomique.

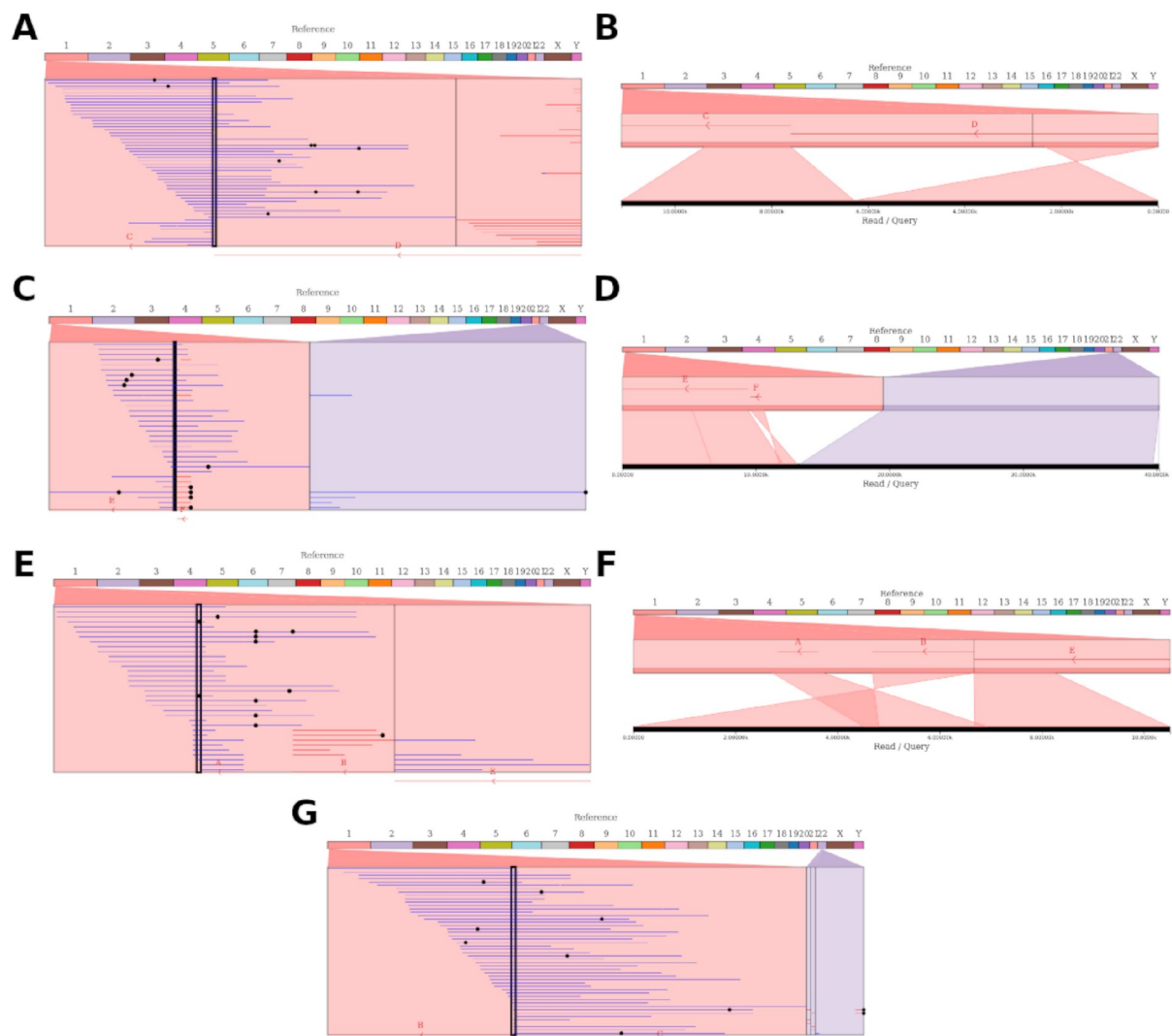


Figure Annexe 7: Visualisation dans Ribbon Genome Viewer de l'ADN du patient UPN5 à partir des données de séquençage long-read Nanopore. (A) Visualisation Ribbon des lectures couvrant les régions C et D, montrant une inversion. (B) Visualisation Ribbon d'une lecture unique couvrant les régions C et D. (C) Visualisation Ribbon des lectures couvrant les régions E, F et le chromosome 22, montrant un remaniement complexe. (D) Visualisation Ribbon d'une lecture unique couvrant les régions E, F et le chromosome 22. (E) Visualisation Ribbon des lectures couvrant les régions A, B et E, montrant un remaniement complexe. (F) Visualisation Ribbon d'une lecture unique couvrant les régions A, B et E. (G) Visualisation Ribbon des lectures couvrant les régions B et C, montrant une inversion.

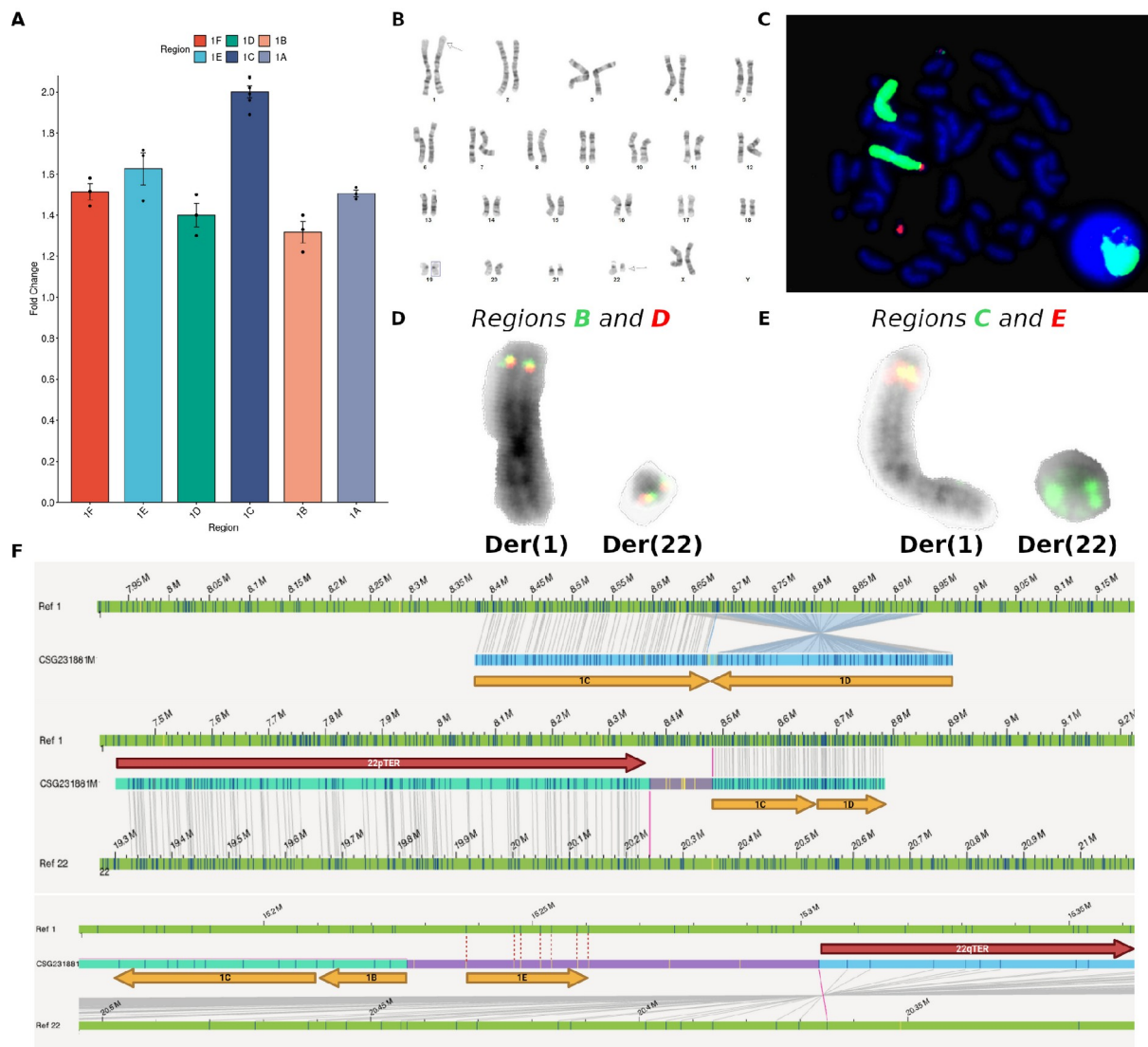
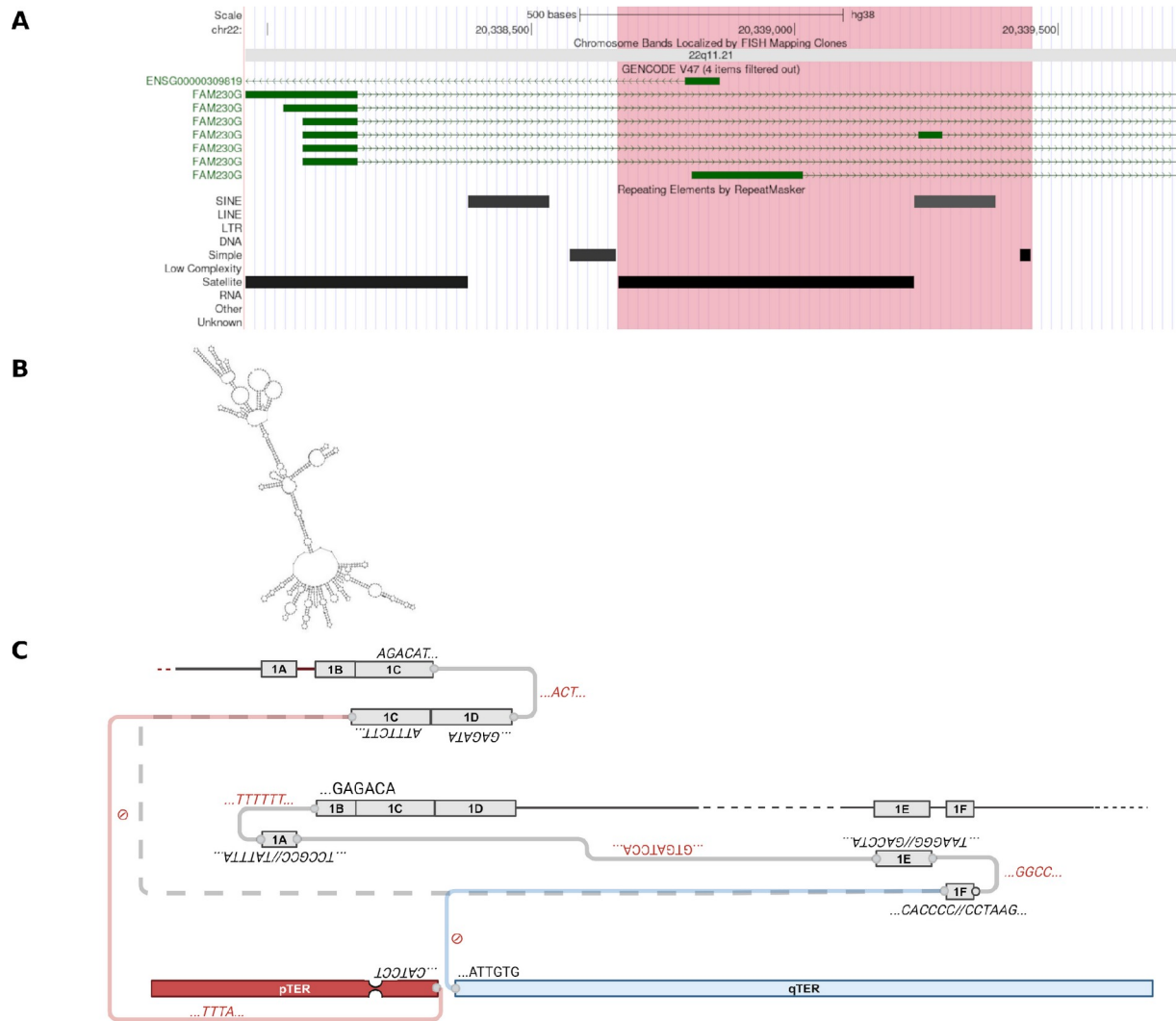


Figure Annexe 8: Analyse moléculaire complémentaire du gain de copie chez le patient UPN5. (A) Analyse par qPCR des régions dupliquées chez le patient UPN5, confirmant la présence de gains de cinq copies et d'un gain de deux copies. (B) Caryotype des cellules du patient UPN5 montrant une translocation réciproque entre les chromosomes 1 et 22. (C) Chromosome painting des chromosomes de UPN5, confirmant la translocation réciproque entre les chromosomes 1 et 22. (D) Visualisation par FISH des régions B et D sur les chromosomes dérivés 1 et 22. (E) Visualisation par FISH des régions C et E sur les chromosomes dérivés 1 et 22. (F) Analyse par Optical Genome Mapping (OGM) illustrant des molécules réarrangées, confirmant la présence de remaniements génomiques complexes.



Figure_Annexe 9: Analyse structurale et mécanistique chez le patient UPN5. (A) Visualisation via UCSC Genome Browser des éléments répétitifs situés à proximité du point de cassure sur le chromosome 22. L'élément surligné correspond à une structure de séquence connue localisée près du point de cassure. **(B)** Prédiction *in silico* de la structure secondaire de l'élément surligné. L'analyse informatique suggère un motif potentiel de repliement de cette séquence répétitive. **(C)** Représentation schématique du mécanisme proposé à l'origine de la chromoanagenèse et de la translocation chez le patient UPN5.

Table_Annexe 1: Positions précises des points de cassure des régions CNV identifiées chez les deux patients.

Genes	5' repeating element	3' repeating element	Variant	Number of reads supporting	Called by Callers
PSPC1; ZMYM5; ZMYM2	AluSx	None	Tandem Triplication		10 Sniffles (Duplication)
RERE	Repeat(AAAAT)n	AluY	Duplication		10 No
RERE	AluY	AluSz	Duplication		7 No
RERE	AluSz	None	Triplication		22 No
RERE; ENO1	None	AluJb	Inverted Duplication		19 Sniffles (Inversion)
FHAD1	AluY	None	Duplication		11 No
FHAD1	None	None	Duplication		8 No
NA		PATRR	Translocation		4 No
NA		PATRR	Translocation		6 No

Table_Annexe 2: Résumé de l’homologie et des éléments mécanistiques entre les régions réarrangées chez les deux patients

Patient	Region Rearranged	SVs formed	Chimeric Sequence	Microhomology	junction DEL/INS	proposed formation mechanism
UPN4	3' TG/5' TG	Tandem Duplication	CTTTAGTTTACCTATGCT GT AGAGATAGGGTTTCACCAT	3bp - "GTA"	No	FoSTeS/MMBIR
	5' _1B/5' _1A	Inverted Inserted duplication	ATTTTATTTTATTTA TTTTT GAGACAGA	6bp - "TTTTTT"	No	FoSTeS/MMBIR
UPN5	3' _1A/5' _1E	Inserted duplication	CCACTTTGGCGGAT TGGATCAC AGGTCAGGAGA	8bp - "TGGATCAC"	Yes - G>T	FoSTeS/MMBIR
	3' _1E/3' _1F	Inverted Inserted duplication	GTCAC CGCCGGCC CTTAGGCTATACTTTGGC	4bp - "GGCC"	No	FoSTeS/MMBIR
	5' _1F/22qTER	Translocation	AGTTTTGTGAGGATGGGGTGCACTTTTGTAGTAGAGA	No	211 bp INS	NHEJ
	3' _1C/3' _1D	Inversion	CTTTGATGAAGACAT ACT GCACCCAGCAAACCTT	3bp - "ACT"	No	FoSTeS/MMBIR
	5' _1C/22pTER	Translocation	CGTTGAAGGATGCAGGATG TAAAA AGAAAT	5bp - "TAAAA"	No	NHEJ

TG: Target Region

RED: Microhomology

GREEN: Substitution not seen in dbSNP base

BLUE: dbSNP (<1%)

Table_Annexe 3: Amorces spécifiques utilisées pour les expériences de PCR et de qPCR.

Patient	fragment	primer_forward	primer_reverse	Type of PCR
UPN4	ZMYM2	ggcatctcagaaggggaagtg	tctcaggaggtcttcgttcaa	qPCR
	ZMYM2	ttggtcatccagcttgcct	gctggagcagaaattgaagg	qPCR
	ZMYM2-5'	aatcctgattcttgatctcca	aacaaaagaggtgagctgacct	RT-qPCR
	ZMYM2-3'	acaggagtatttgttgaagt	tcttcaactcgagagaatattgacc	RT-qPCR
UPN5	1A	atgtgaccagcgttcctgag	aacagcatagacccaaggcc	qPCR
	1B	cctgagacatgagatcggcc	ttgactgtttgagaggccc	qPCR
	1C	gctctgcctgctaactggat	agccattctgtgaggaaccg	qPCR
	1C	actggcggtagatgtctcct	tgagaatcccttgccgtgg	qPCR
	1D	attccctaacacctcccca	cggccttcctgcttctta	qPCR
	1E	cagagaaggtggctgggaag	ggggaactacaggtgacagc	qPCR
	1F	ctagcgtgggtgtcctgaag	gcgtgctcctcttacctcc	qPCR
	RERE_exon2/3	tcaaactggtccacaactcc	tcaaactggtccacaactcc	RT-qPCR
	RERE_exon18/19	cagccagtcagctaggttct	ctcttctggccagcttga	RT-qPCR

Table_Annexe 4: Mesures de la qualité de la préparation de librairie et du séquençage long-read chez les deux patients.

Patient	Sequencing Protocol	DNA/RNA source	DNA/RNA processing	tape station size (bp)	start prep (µg)	kit	flow cell device	load1 (fmol)	pores	start sequencin	yield (Gbp)	coverage	N50 (bp)	length_mean (bp)	length_median (bp)	quality_mean	quality_median	number of reads	
UPN4	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	Blood	Shearing (Covaris g-tube - 4000RPM x 1min)	11881		SQK-3LSK114	R10.4	PROM	33		8242	34,27	9,7X	9629	6793	5885	18,9	19,4	4,37M
	Ligation sequencing V14 - Direct cDNA sequencing (SQK-LSK114)	Cell Line (LCLs)	Full-lenght cDNA reverse-transcription	NA		SQK-4,785LSK114	R10.4	PROM	<30fmol		6731	17,22	2,1X	1906	1353	926	11,1	11,9	9,87M
UPN5	Ligation Sequencing gDNA V14 - human sample (N50 30 kb) on PromethION (SQK-LSK114)	Blood	Shearing (Covaris g-tube - 4000RPM x 1min)	11970		SQK-3LSK114	R10.4	PROM	44		7550	92,86	23X	7727	7020	6558	18,7	19,2	13,65M
	Ligation sequencing V14 - Direct cDNA sequencing (SQK-LSK114)	Cell Line (LCLs)	Full-lenght cDNA reverse-transcription	NA		SQK-3LSK114	R10.4	PROM	<20fmol		7423	13,68	2,7X	1252	923	626	10,9	11,8	12,06M

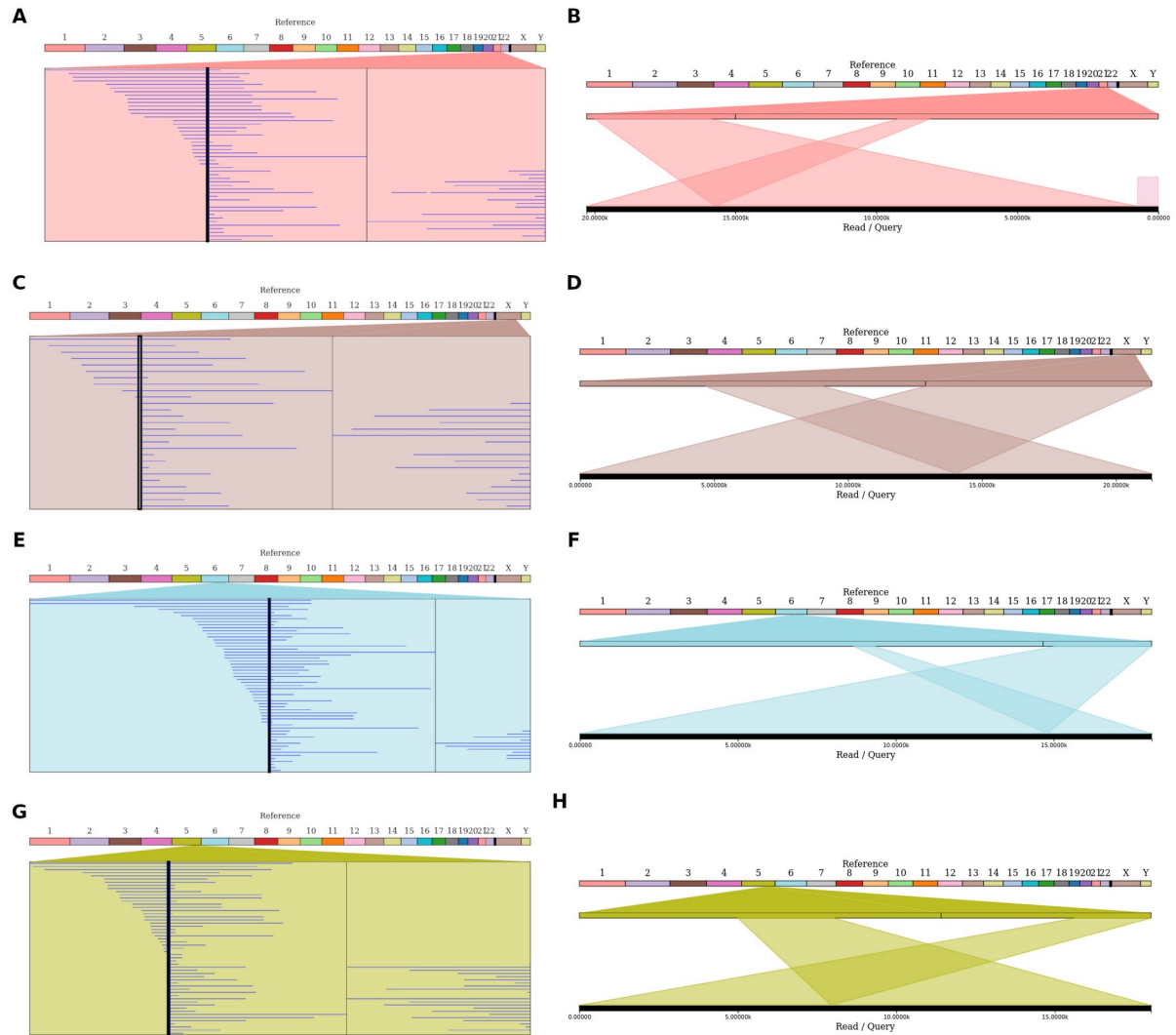
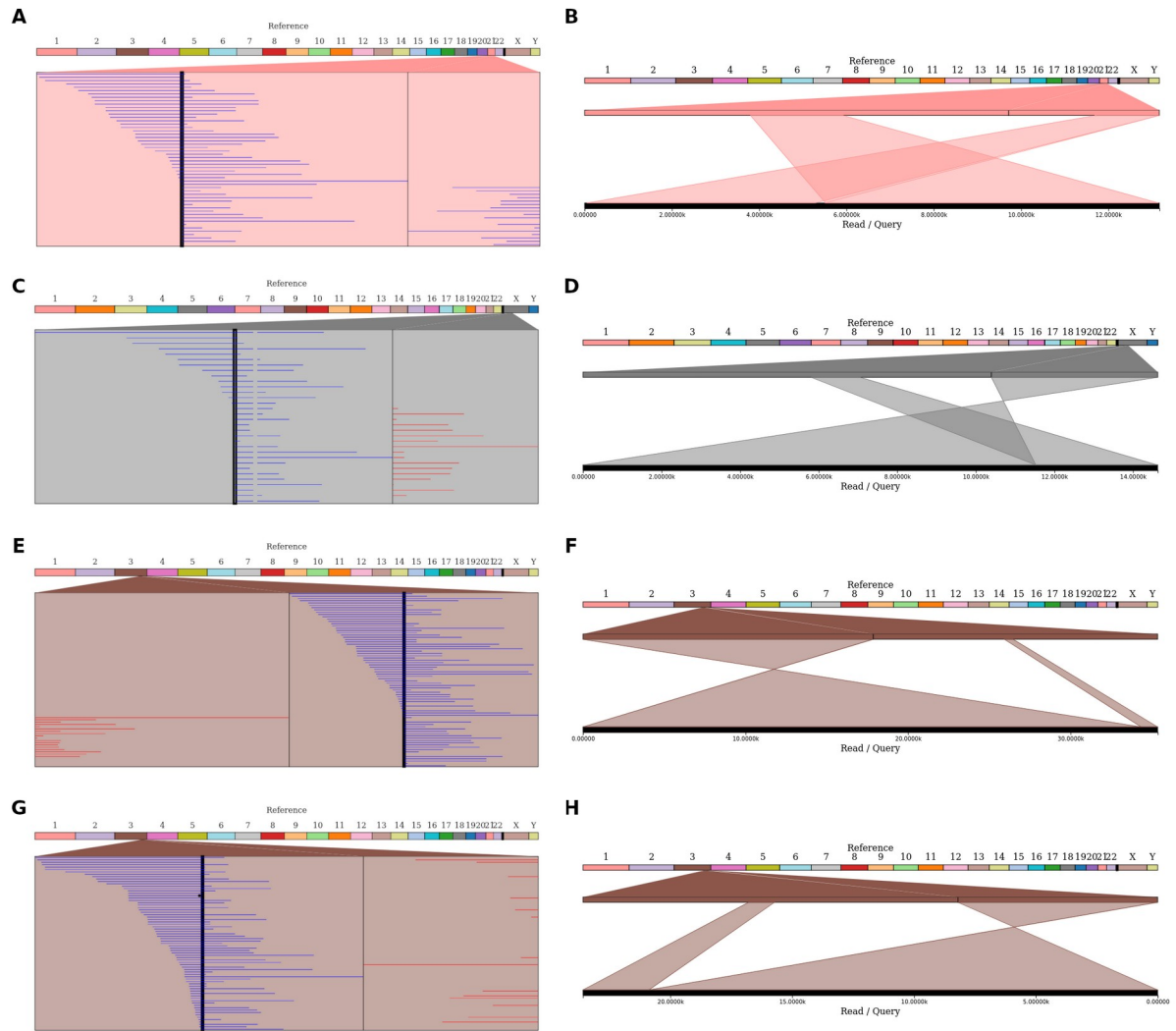
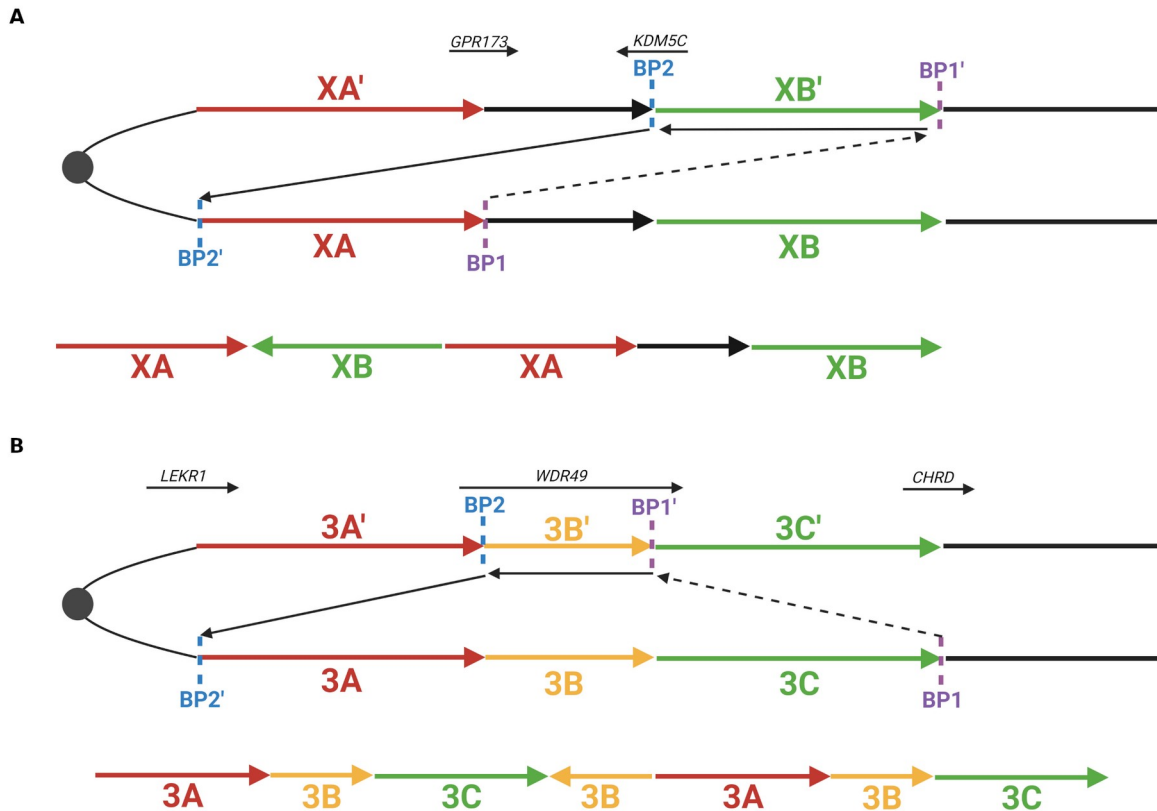


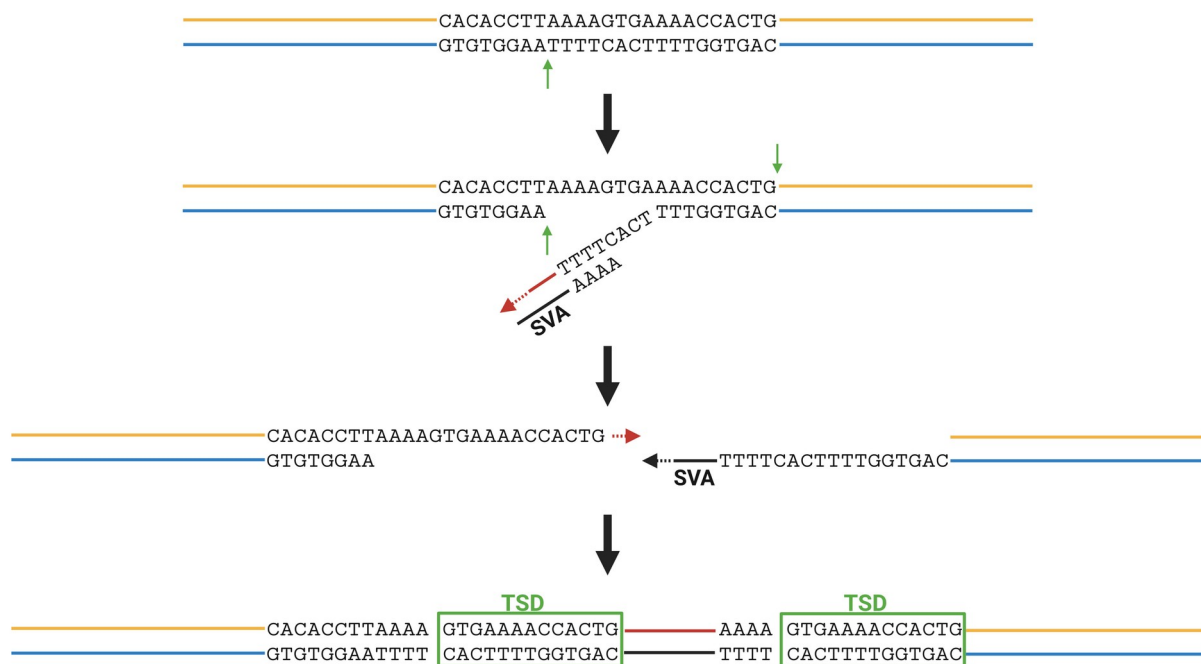
Figure Annexe 10 : Visualisation avec Ribbon Genome Viewer de l'ADN de UPN6 à UPN9 à partir de données de séquençage long-read. (A, C, E, G) Visualisations avec Ribbon des lectures couvrant les régions 21A (UPN6), XA (UPN7), 6A (UPN8) et 5A (UPN9), montrant une duplication en tandem. (B, D, F, H) Visualisations avec Ribbon d'une lecture unique couvrant les mêmes régions respectives chez UPN6 à UPN9.



Figure_Annexe 11 : Visualisation avec Ribbon Genome Viewer de l'ADN de UPN10 à UPN12 à partir de données de séquençage long-read. (A, C, E, G,). Visualisations avec Ribbon des lectures couvrant les régions 21A (UPN10, montrant une triPLICATION en tandem), XA-XB (UPN11 montrant une inversion), 3A-3B et 3B-3C (UPN12, montrant des inversions). **(B, D, F, H, J)** Visualisations avec Ribbon d'une lecture unique couvrant les mêmes régions respectives chez UPN10 à UPN12.



Figure_Annexe 12 : Modèle prédictif de l'émergence de variants complexes chez les patients 8 et 9. (A) Modèle prédictif de l'émergence de l'événement DUP/INV-NML-DUP chez UPN11. La formation de cet événement débiterait par un arrêt puis un effondrement de la fourche de réplication près du point de cassure BP1, situé à la fin de la région A. Une micro-homologie favorise alors l'invasion de brin au niveau de BP1', situé à la fin de B' sur le brin opposé, produisant la jonction 1. La réplication de l'ADN se poursuit ensuite dans la direction opposée jusqu'à ce qu'un second effondrement de la fourche se produise près de BP2, situé au début de B'. Une réparation par NHEJ (réparation non homologue) survient ensuite sur le brin original près de BP2', situé au début de A, créant la jonction 2. **(B)** Modèle prédictif de l'émergence de l'événement DUP-TRP/INV-DUP chez UPN12. La formation de cet événement commencerait par un arrêt puis un effondrement de la fourche de réplication près de BP1, situé à la fin de la région C. Une micro-homologie favorise l'invasion de brin au niveau de BP1', situé à l'extrémité de B' sur le brin opposé, produisant la jonction 1. La réplication de l'ADN se poursuit ensuite dans la direction opposée jusqu'à un second effondrement de la fourche près de BP2, situé au début de B', suivi d'une réparation par NHEJ sur le brin original au niveau de BP2', situé au début de A, créant la jonction 2.



Figure_Annexe 13 : **Modèle mécanistique de l'insertion d'un SVA.** (1) l'endonuclease LINE-1 clivant l'ADN cible au site riche en A/T (AATTTT), (2) l'appariement de la queue poly-A du SVA avec l'extrémité 3' de l'ADN, (3) la synthèse d'ADNc à partir de l'ARN SVA, et (4) la formation de la duplication du site cible (TSD) après insertion.

Table_Annexe 5: Synthèse des anomalies étudiées: type d'anomalies, mécanismes sous-jacent ainsi que la reclassification

Patient	Anomalie identifiée	Particularités du point de cassure	Mécanisme sous-jacent probable	Classification initiale	Reclassification
UPN 1	Duplication en Tandem Récurrente	LINEs (L1PA2) aux points de cassure	NAHR	Pathogène	Pathogène
UPN 2	Duplication en Tandem Récurrente	LINEs (L1PA2) aux points de cassure	NAHR	Pathogène	Pathogène
UPN 3	Inversion potentiellement Récurrente	Duplications Segmentaires aux points de cassure	NAHR	Aucun variant détecté	Pathogène
UPN 4	Triplcation en Tandem	Microhomologie de 3bp	MMBIR/FoSTeS	VSI	Probablement Pathogène
UPN 5	Chromoanasyntèse associé à une translocation	Multiples microhomologies + PATRR	MMBIR/FoSTeS + NHEJ	VSI	Probablement Pathogène
UPN 6	Duplication en Tandem et Délétion	Microhomologie de 3bp	MMBIR/FoSTeS ou NHEJ/MMEJ	DUP : VSI DEL : Pathogène	DUP : VSI DEL : Pathogène
UPN 7	Duplication en Tandem	Microhomologie de 2bp	MMBIR/FoSTeS ou NHEJ/MMEJ	Pathogène avec partie du phénotype non expliqué	Pathogène avec partie du phénotype non expliqué
UPN 8	Duplication en Tandem	Microhomologie de 3bp	MMBIR/FoSTeS ou NHEJ/MMEJ	VSI	VSI + hypothèse d'implication de SIM1 (nécessite des études fonctionnelles)
UPN 9	Duplication en Tandem	Microhomologie de 4bp	MMBIR/FoSTeS ou NHEJ/MMEJ	VSI	VSI + hypothèse d'implication de PURA (nécessite des études fonctionnelles)
UPN 10	Triplcation en Tandem	Insertion de 26bp	MMBIR/FoSTeS	VSI	Probablement Pathogène
UPN 11	DUP/INV-NML-DUP	Microhomologie de 2bp + extrémités franches	MMBIR/FoSTeS + NHEJ	VSI	Probablement Pathogène
UPN 12	DUP-TRP-DUP	Microhomologie de 3bp + extrémités franches	MMBIR/FoSTeS + NHEJ	Probablement Pathogène	Probablement Pathogène
UPN 13	Duplication en Tandem	Microhomologie de 7bp	MMBIR/FoSTeS ou NHEJ/MMEJ	Pathogène	Bénin
UPN 14	Duplication en Tandem	Insertion de 2bp	C-NHEJ	Pathogène	Bénin
UPN 15	Duplication Insertionnelle inversée	Microhomologie de 6bp à distance de 10bp	MMBIR/FoSTeS	Non détecté	Probablement Pathogène

Article Annexe 1

RESEARCH ARTICLE

Long-read sequencing of recurrent *FGF12* duplications in epilepsy: Insights into structural mechanisms and aberrant isoforms

Jade Fauqueux¹ | Laurence Chaton² | Pierre Cleuziou³ | Anne-Sophie Diependaële⁴ | Nathalie Bach⁵ | Nicolas Gruchy⁶ | Marion Gerard⁶ | Jean-Pascal Meneboo⁷ | Céline Villenet⁷ | Martin Figeac⁷ | Emilie Ait-Yahya⁸ | Caroline Thuillier⁹ | Elise Boudry⁹ | Adeline Trauffer³ | Sylvie Nguyen-The-Tich³ | Simon Boussion^{1,10} | Roseline Caumes^{1,10} | Jamal Ghoumid^{1,10} | Thomas Smol^{1,9} 

¹Univ. Lille, Unité de Recherche ULR7364 RADEME - Maladies RAres du Développement, Fédération Hospitalo-Universitaire FHU-G4 Génomique, Lille, France

²Centre Hospitalier Universitaire de Lille, Service de Neurophysiologie Clinique, Lille, France

³Centre Hospitalier Universitaire de Lille, Clinique de Pédiatrie, Service de Neuropédiatrie, Lille, France

⁴Centre Hospitalier Universitaire de Caen, Service de Neurologie Unité Explorations Fonctionnelles Neurologiques, Caen, France

⁵Centre Hospitalier Universitaire de Caen, Service de Pédiatrie, Caen, France

⁶Univ. Caen Normandie, Unité de Recherche UR7450 BioTARGen, Fédération Hospitalo-Universitaire FHU-G4 Génomique, Centre Hospitalier Universitaire de Caen, Service de Génétique, Caen, France

⁷Univ. Lille, Centre National de la Recherche Scientifique CNRS, Inserm, Centre Hospitalier Universitaire de Lille, Institut Pasteur de Lille, Plateformes Lilloises en Biologie & Santé, Lille, France

⁸Centre Hospitalier Universitaire de Lille, Cellule de Bioinformatique, Plateau Commun de Séquençage, Lille, France

⁹Centre Hospitalier Universitaire de Lille, Institut de Génétique Médicale, Lille, France

¹⁰Centre Hospitalier Universitaire de Lille, Clinique de Génétique, Lille, France

Correspondence

Thomas Smol, CHU Lille, Institut de Génétique Médicale, avenue Oscar Lambret, F-59000 Lille, France.
Email: thomas.smol@chu-lille.fr

Abstract

Objective: Fibroblast growth factor 12 (*FGF12*), a member of the fibroblast homologous factor family, plays a key role in the modulation of voltage-gated sodium (Nav) channels. Pathogenic variants in the *FGF12* gene leading to a gain-of-function mechanism and partial duplication encompassing the *FGF12* gene leading to a loss-of-function mechanism are associated with developmental and epileptic encephalopathy (DEE), characterized by developmental delay, intellectual disability, ataxia, and drug-resistant epilepsy. We report two patients with DEE harboring de novo recurrent intragenic duplications of *FGF12* identified by long-read sequencing (LRS).

Methods: We applied LRS to the DNA and cDNA of patients with *FGF12* duplication to fully characterize the DNA's structural organization and its transcriptional consequences. Additionally, we reanalyzed electroencephalographic (EEG) data

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Epilepsia* published by Wiley Periodicals LLC on behalf of International League Against Epilepsy.

from patients at different timepoints to identify phenotypal specificities and refine the electroclinical spectrum.

Results: These duplications, spanning approximately 536 kbp, were mediated by nonallelic homologous recombination between L1PA2 elements (LINE-1 Primate-specific subfamily A, number 2). cDNA analysis revealed aberrant transcripts, one predicted to encode an elongated FGF12 protein and another leading to premature termination. Both patients shared overlapping clinical features, including postepilepsy onset regression, global developmental delay, and ataxia. EEG studies revealed a marked early encephalopathic pattern with disorganized and high-voltage slow background activity with multifocal spikes at onset evolving later into subcontinuous generalized spike and wave activation.

Significance: Our findings are consistent with previous reports linking structural variants to functional disruption, suggesting impaired Nav channel activity due to a shift in inactivation to hyperpolarized potential, leading to a loss-of-function effect. These findings underscore the utility of LRS for DNA and cDNA analysis in resolving structural variants and expanding the electroclinical spectrum of patients with *FGF12* duplications.

KEYWORDS

FGF12, intellectual disability, long-read sequencing cDNA analysis

1 | INTRODUCTION

Fibroblast growth factor 12 (FGF12), encoded by the *FGF12* gene and previously designated FHF1, is a member of the fibroblast growth homologous factor (FHF) family, which includes *FHF1* (*FGF12*), *FHF2* (*FGF13*), *FHF3* (*FGF11*), and *FHF4* (*FGF14*).¹ *FGF12* is predominantly expressed in the developing and mature nervous system, cardiac and skeletal muscle tissues, and olfactory epithelium.² The *FGF12* transcripts produce two major transcripts. The longer transcript, designated A, produces a protein with an N-terminal nuclear localization-like sequence (NLS) that allows its nuclear localization. In contrast, the shorter transcript, designated B, produces a protein without the NLS sequence, resulting in its cytoplasmic localization.^{1,3} FGF12 interacts with voltage-gated sodium (Nav) channels, including Nav1.2 (SCN2A), Nav1.5 (SCN5A), and Nav1.6 (SCN6A), and modulates the fast inactivation kinetics of these channels.^{2,4}

Recurrent pathogenic heterozygous missense variants in *FGF12* have been associated with autosomal dominant developmental and epileptic encephalopathies (Mendelian Inheritance in Man #617166), and have been reported in more than 20 publications.^{5–9} The associated phenotype includes developmental delay, intellectual disability, axial hypotonia, ataxia, and early onset drug-resistant epilepsy.¹⁰ Electrophysiological studies

Key points

- LRS identified aberrant *FGF12* transcripts, including an elongated protein and a truncated variant.
- *FGF12* duplications in DEE impair Nav channel function, leading to a loss-of-function effect and epilepsy-related regression.
- EEG analysis showed an early encephalopathic pattern, with high-voltage slow activity and multifocal spikes evolving into generalized spike-waves.

have demonstrated a gain-of-function effect of these recurrent variants, which causes depolarizing shifts in the voltage-dependent fast inactivation of Nav1.6.^{5,11} In addition to single nucleotide variants, partial duplications of the *FGF12* gene with a similar phenotype have been reported in six patients.^{12–14} The duplication is expected to produce a longer transcript associated with a loss-of-function mechanism.^{12,15}

We report two additional patients carrying a recurrent partial duplication of *FGF12*. Using DNA long-read sequencing (LRS), we fully resolved the underlying structural variant without prior assumptions, demonstrating that it corresponds to a direct tandem

duplication. Additionally, we examined aberrant *FGF12*-bearing transcripts using full-length cDNA LRS. We also provide a detailed description of the clinical and electroclinical phenotypes of the disease. This study highlights the central role of LRS in elucidating structural variants and offers new insights into their functional consequences.

2 | MATERIALS AND METHODS

2.1 | Ethics statement

Clinical details were collected, and informed consent was obtained for genetic studies. Blood samples were collected from the affected individuals and their parents, after informed consent was obtained. Analyses were performed on a diagnostic basis in accordance with the bioethical rules of French law. This study was approved by the Comité de Protection des Personnes ethics committee (reference no. 2023-A00473-42). Informed written consent was obtained from all participants or their legal guardians.

2.2 | Cell culture

Human lymphoblastoid cell lines (LCLs) were established by Epstein-Barr virus immortalization of the subject's blood lymphocytes and maintained in RPMI 1640 medium (Thermo Fisher Scientific) supplemented with 15% fetal bovine serum (Thermo Fisher Scientific) and 1% penicillin-streptomycin (Life Technologies).

2.3 | DNA and RNA extraction

DNA was extracted from patient lymphocytes using the QIAamp DNA Blood Kit (Qiagen) according to the manufacturer's instructions. RNA was extracted from the LCL of Patient 1 using the PureLink RNA Minikit (Invitrogen).

2.4 | Array comparative genomic hybridization

Copy number variations were detected by comparative genomic hybridization (CGH) array experiments using a 60-K oligonucleotide microarray (SurePrint G3 Human CGH Microarray Kit, 8×60 K, Agilent Technologies). Data were analyzed using CGH-Analytics software v2.7 and the ADM2 algorithm.

2.5 | Nanopore LRS

DNA libraries were prepared using the Ligation Sequencing Kit v14 (Oxford Nanopore Technologies) and subsequently sequenced following the standard protocol on a GridION sequencer (Oxford Nanopore Technologies). After sequencing, base calling was performed using the Guppy base caller (v6.4.6) in high-accuracy (HAC) mode. Reads were mapped to the human reference genome (GRCh38) using MiniMap2 (v2.24) and NGMLR (v0.2.7).^{16,17} Structural variations were detected using Sniffles2 (v2.0.7) and CuteSV (v2.1.1). Sniffles2 calling was tested with specific parameters to enhance detection of complex variants: "--long-dup-length 5000000" and "--minsupport 3" and CuteSV with "--max_size -1" and "--min_support 3". Visual analysis of structural variations in target regions were conducted with the Integrative Genomics Viewer (v2.18).^{18,19}

cDNA synthesis was then performed using the SuperScript III First-Strand Synthesis System Kit (Thermo Fisher Scientific), followed by real-time polymerase chain reaction (RT-PCR) using the NEBNext High-Fidelity Polymerase (New England Biolabs). Several primers were designed to amplify and analyze both wild-type (WT) and aberrant transcripts based on the proposed hypothesis. The goal was to amplify the junction of the tandem duplication in *FGF12* transcripts. For aberrant transcripts, primers targeting exons 4 to 3 were used, whereas for WT transcripts, primers targeting exons 3 to 6 were used; these primers also amplified aberrant transcripts (primers are available upon request; Figure S1). PCR amplified products were prepared as nanopore libraries and loaded onto a PromethION flow cell. Base calling was performed using the Guppy base caller (v6.4.6) in HAC mode. The resulting reads were aligned to the human reference genome GRCh38 using Minimap2 with splice-aware alignment settings to accommodate transcriptomic data. All chimeric alignments were mapped to the genome using the BLAST tool to determine their genomic origin, allowing us to distinguish between WT and aberrant transcripts.²⁰

2.6 | Breakpoint analysis

Sequencing primers were designed to amplify the specific breakpoints in the two patients. An initial PCR of approximately 6000bp was performed using Roche Expand Long Template Polymerase (Roche). This was followed by nested PCR using NEBNext High-Fidelity 2X Polymerase (New England Biolabs). PCR products were extracted from a 2% agarose gel, purified using a PCR clean-up kit (Macherey-Nagel), and sequenced by the Sanger method (primers are available upon request).

3 | RESULTS

3.1 | Clinical description and EEG features of Patient 1

Patient 1 was born to healthy unrelated parents after an uncomplicated pregnancy. Early psychomotor development was normal (walking at 14 months, first words at 12–13 months, good social interactions). No dysmorphic features could be seen. Epilepsy began abruptly at 18 months of age, with recurrent afebrile generalized motor status epilepticus (three status epilepticus incidents in 1 month) and frontotemporal seizures with rapid bilateralization. First electroencephalography (EEG), performed in the context of multiple daily seizures, showed disorganized, slowed (3.5 Hz on average, delta predominant 1.5 Hz intertwined with theta 4–5 Hz), and atypical very high-voltage background activity (150–400 μ V) with no anteroposterior gradient in wakefulness and rare focal spikes often diffuse over one hemisphere, either right or left (Figure 1A). Delta activity decreased in sleep, whereas theta activity increased, and bilateral sleep spindles were present (Figure 1B). EEG monitoring recorded bilateral tonic-clonic seizures, lasting 1–1.5 min, with a myoclonic jerk, tonic phase, and bilateral clonic phase (Figure 1C,D) and seizures with frontotemporal onset before bilateralization. A few weeks after the onset of epilepsy, the patient showed severe psychomotor regression, including complete loss of oral language, severe relational disorder (lack of eye contact, no smiling, laughing, or crying), pyramidal syndrome, loss of gait, and ataxia. Initial magnetic resonance imaging (MRI), cerebrospinal fluid analysis, and immunologic and metabolic tests were normal. The patient had no extraneurological involvement and a normal cardiac evaluation. Given the severity of the disease, he was treated with immunoadsorption at 20 months of age on the assumption of seronegative encephalitis. Because it resulted in a partial clinical improvement (regaining the ability to walk), this was followed with rituximab infusions for 18 months.

After the initial acute phase, epilepsy evolved with monthly bursts of bilateral tonic-clonic, tonic, tonic with focalization, and multiple types of complex focal seizures lasting 3–5 days. Figure 2A,B presents a 1-min focal seizure with right temporal onset registered at 3 years old. EEGs showed prolonged slow hypersynchrony on awakening lasting several minutes, with generalized delta rhythm and altered contact (Figure 2C), an increase in diffuse triphasic waves during wakefulness, and more frequent multifocal spikes. Of the 29 EEGs performed over 5 years, only one showed normal voltage, with subnormal background activity and rare paroxysmal activity during sleep, recorded after a 2-month seizure-free period. This

transient improvement was not related to any therapeutic modification. Evolution of brain MRI revealed mild cerebral and cerebellar atrophy (Figure S2). A positron emission tomography brain scan showed marked, symmetrical hypometabolism in the frontal, parietal, and temporal associative areas.

At the age of 5 years, myoclonic and atypical absence seizures appeared, along with a substantial increase in subcontinuous generalized spikes and waves on the EEG. The frequency of tonic and bilateral tonic-clonic seizures increased, occurring several times per day. Background activity was characterized by isolated generalized spikes or bursts of spike and waves at 2.5 Hz, along with atypical absence seizures and concomitant myoclonus (Figure 2D), intertwined with bi- and triphasic delta activity. His neurological condition deteriorated significantly, with complete loss of motor skills, lack of feeding autonomy, and a drastic reduction in daily wake time, with no responsiveness to stimulation.

Since epilepsy started, several antiepileptic therapies were tried. Ketogenic diet, vitamin B6, vitamin B8, and cannabidiol (10 mg/kg/day) were ineffective. Levetiracetam (55 mg/kg/day), sodium valproate (30 mg/kg/day), lamotrigine (poorly tolerated at 1 mg/kg/day), clonazepam (.15 mg/kg/day), clobazam (.15 mg/kg/day), topiramate (poorly tolerated at 2 mg/kg/day), and phenobarbital (5 mg/kg/day) showed transient and partial efficacy. Sodium channel blockers were introduced after genetic diagnosis at the age of 5 years 3 months. Carbamazepine (20 mg/kg/day) temporarily reduced seizure activity, without improvement of background abnormalities on EEG, and lost efficacy after a few months. Phenytoin successfully treated bilateral tonic-clonic status epilepticus, but long-term use resulted in increased ataxia and worsening myoclonus. At the last follow-up (6.5 years old), the patient was treated with phenobarbital and felbamate (Figure S3). He remained bedridden, poorly reactive, and experienced tonic and bilateral tonic-clonic seizures several times per day.

3.2 | Clinical description and EEG features of Patient 2

Patient 2 is the third child of a nonconsanguineous couple with no family history of epilepsy. He was born eutrophic at full term after an uncomplicated pregnancy. Epilepsy began at 5 months of age, with generalized epileptic seizures characterized by hypotonia and nonreactivity lasting approximately 2 min, followed by normal recovery. He rapidly presented recurrent seizures with head deviation and clonic movements of the upper limbs. Initial EEG showed a moderately slowed baseline rhythm, along

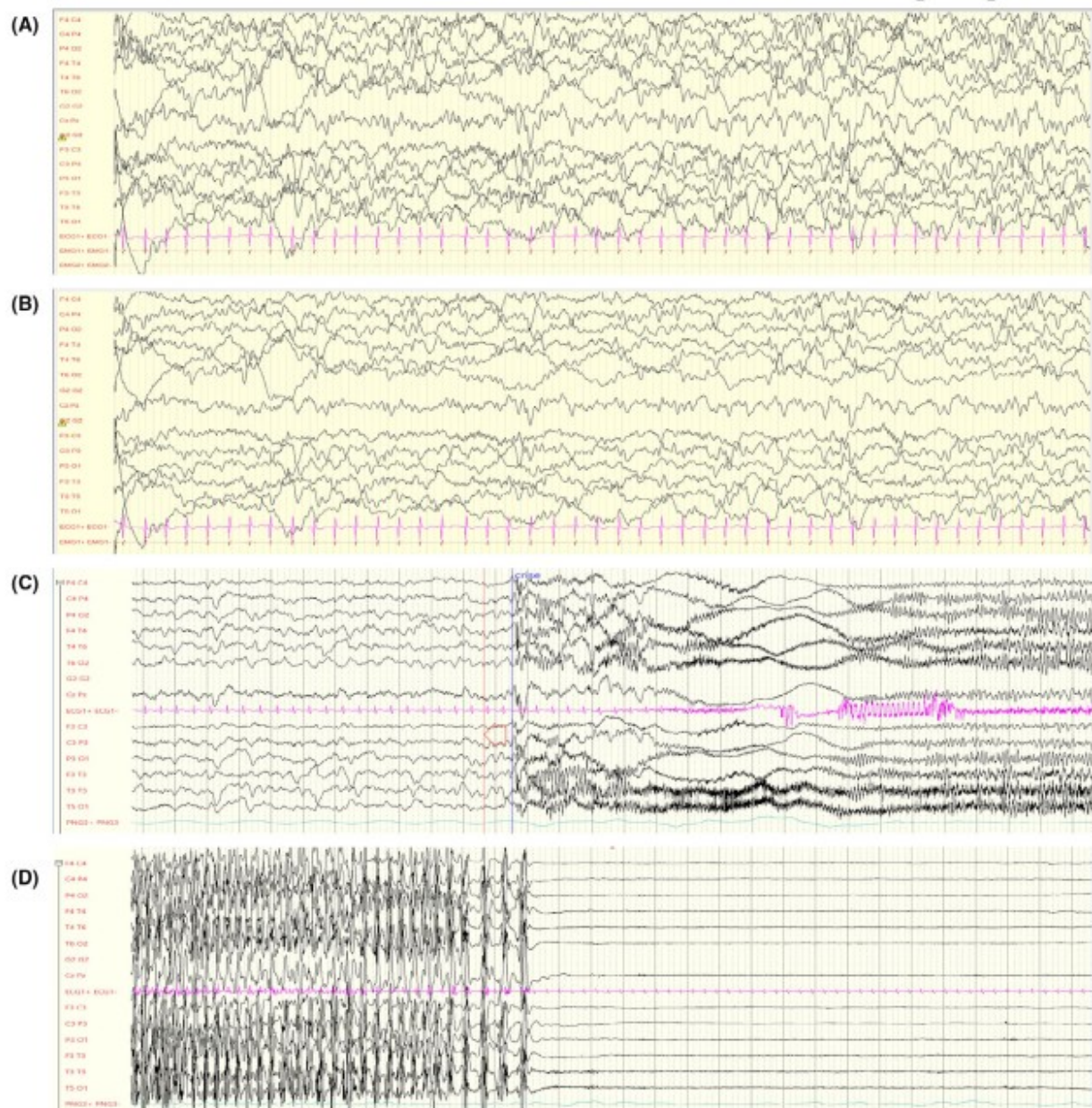


FIGURE 1 Patient 1 electroencephalographic (EEG) background at onset of epilepsy (acute phase 1) at 18 months in wakefulness (A) and sleep (B) and during seizures (C, D). (A) In wakefulness, abnormally high-voltage slow disorganized activity between 200 and 400 μ V, with rare multifocal spikes. Filter settings: .53–70 Hz, 50-Hz notch filter; sensitivity: 20 μ V/mm. (B) In sleep, with reduced delta activity at 1.5 Hz, and richer in theta at 4–5 Hz; sleep spindles were found bilaterally. Filter settings: .53–70 Hz, 50-Hz notch filter; sensitivity: 15 μ V/mm. (C, D) Bilateral tonic-clonic seizures followed by a suppressed EEG for 1 min after. Filter settings: .53–70 Hz, 50-Hz notch filter; sensitivity: 25 μ V/mm.

with slow waves and spikes in the temporal regions during sleep (Figure 3A). Despite the beginning of epilepsy, he reached early developmental motor milestones (sitting at 9 months, walking at 14 months). Language regression was noted at 15 months in the context of serous otitis, followed by an absence of further language development and

increasing social withdrawal. A diagnosis of severe autism spectrum disorder was confirmed at the age of 4 years. The patient achieved independent walking, with no signs of pyramidal syndrome or ataxia. He is well integrated into a specialized medical-educational facility. Brain MRI performed at 6 months and at 2 years of age was normal.

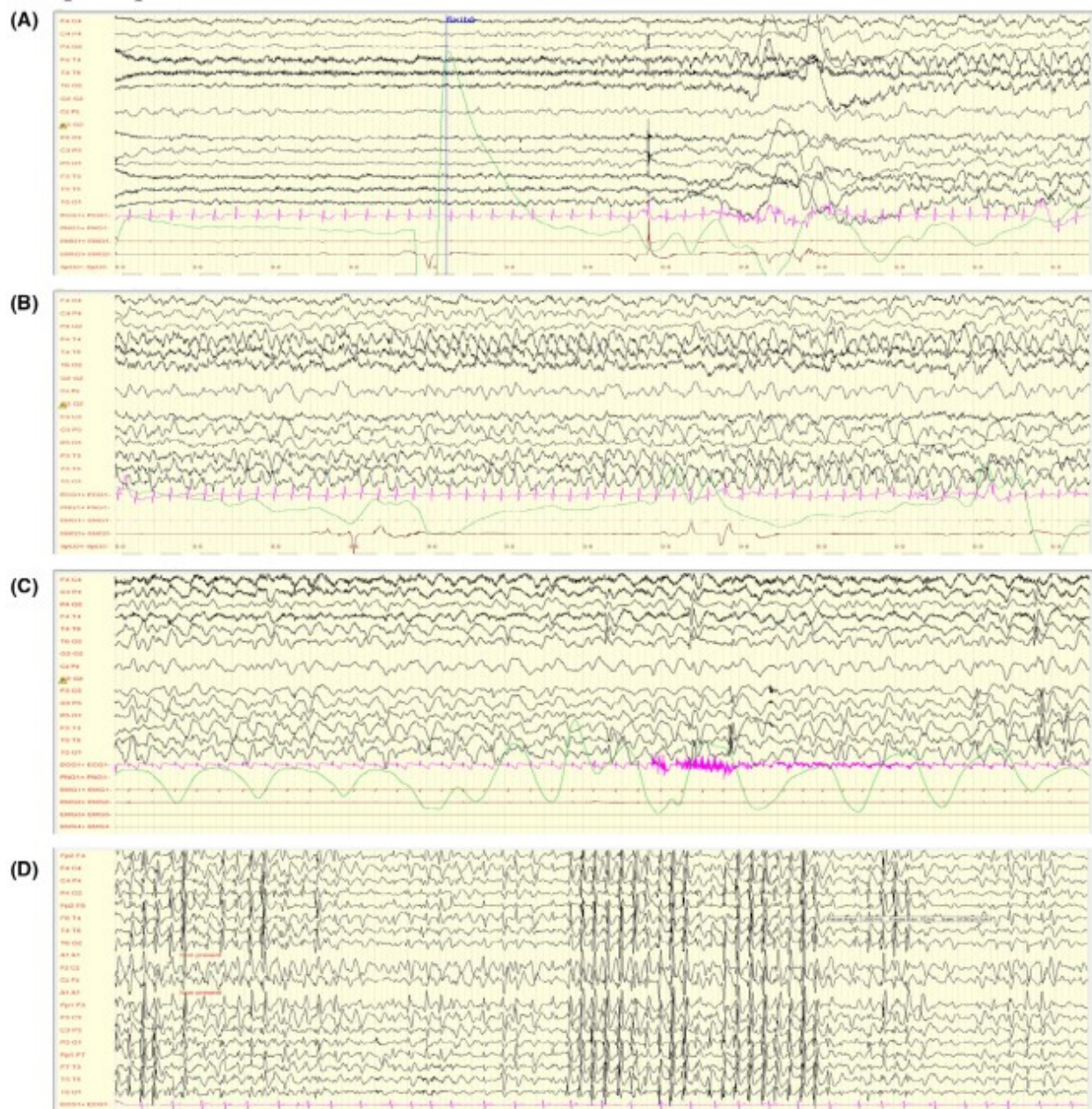


FIGURE 2 Patient 1 electroencephalographic (EEG) background at age 3 years during seizures (“honeymoon” phase 2 with intermittent seizures; A, B) and 1 month later (C), and at last follow-up at 6 years old (phase 3 aggravation; D). (A, B) Bilateral tonic-clonic seizures followed by a suppressed EEG for 1 min after. Filter settings: .53–70 Hz, 50-Hz notch filter; sensitivity: 25 μ V/mm. (C) Slow hypersynchrony on awakening lasting several minutes, generalized delta rhythmic, with altered contact during this period on several successive EEGs and an increase in diffuse triphasic waves. Filter settings: .53–70 Hz, 50-Hz notch filter; sensitivity: 10 μ V/mm. (D) Background of diffuse encephalopathy rich in bi- and triphasic delta activity, with isolated generalized spikes or bursts of spikes-waves at 2.5 Hz, with atypical absence and concomitant myoclonus. Filter settings: .53–70 Hz, 50-Hz notch filter; sensitivity: 10 μ V/mm.

He had 1–2 seizures per month during his first years of life. Administered antiepileptic treatments were poorly effective (sodium valproate, levetiracetam, oxcarbazepine, lamotrigine, and topiramate). At approximately 3 years old, EEG revealed global slowing

with abnormally high-voltage and disorganized delta activity (Figure 3B).

At 6 years old, he experienced a significant increase in seizure frequency, consisting primarily of bilateral tonic-clonic episodes that could be prolonged, along

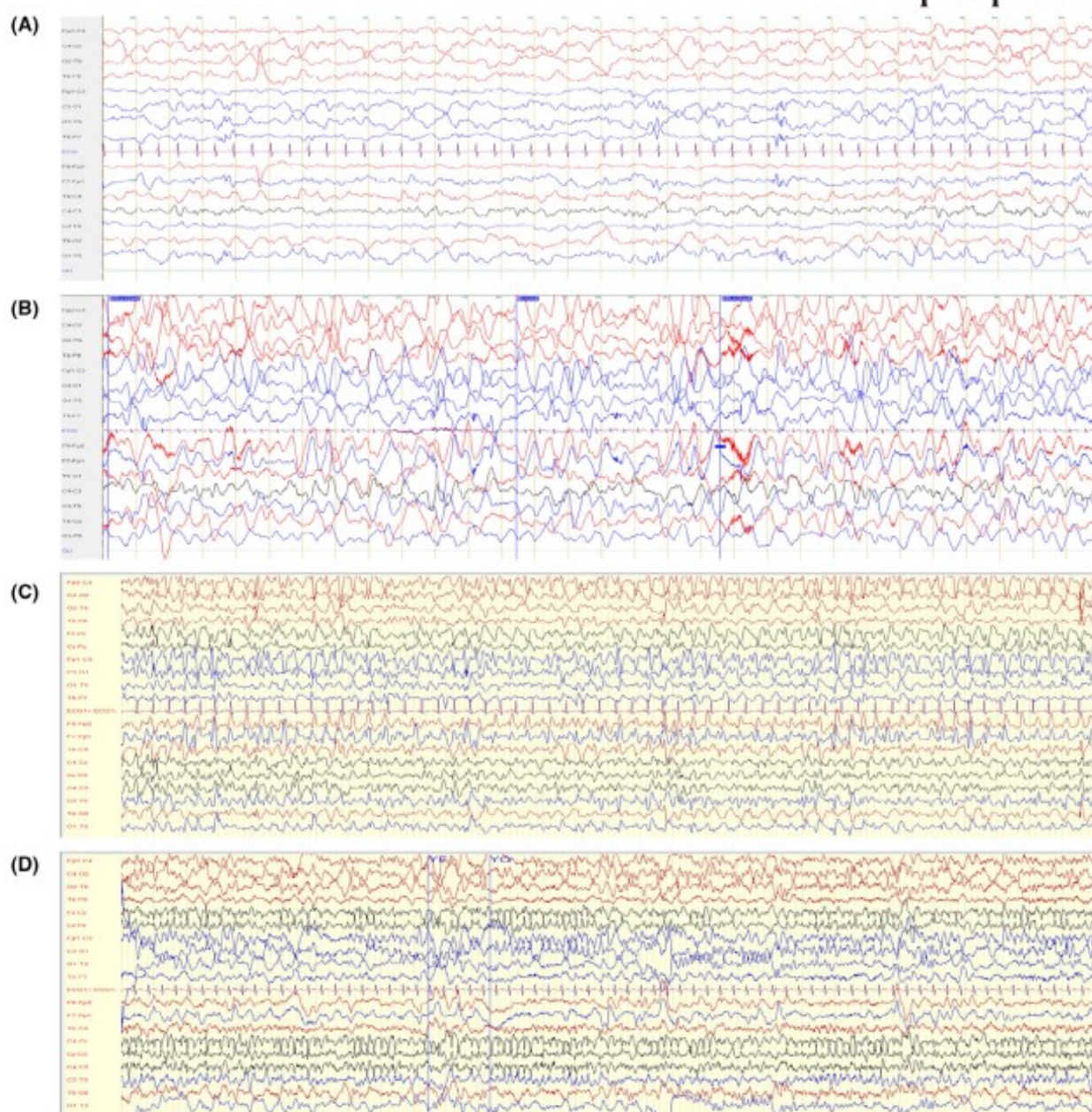


FIGURE 3 Electroencephalographic (EEG) background of Patient 2. (A) Acute phase 1 at 6 months old during sleep, showing slow waves and spikes in left temporal and occipital regions. Filter settings: .53–70 Hz, 50-Hz notch filter; sensitivity: 20 μ V/mm. (B) At 4 years old, showing slow disorganized background with high-voltage delta waves. Filter settings: .53–70 Hz, 50-Hz notch filter; sensitivity: 15 μ V/mm. (C) At 6 years old, showing slow disorganized and high-voltage background with bi- and triphasic waves and continuous spikes-waves in the frontal regions. Filter settings: .53–70 Hz, 50-Hz notch filter; sensitivity: 20 μ V/mm. (D) At 6 years old, showing slow disorganized background with sharp theta rhythmic discharge on central vertex regions. Filter settings: .53–70 Hz, 50-Hz notch filter; sensitivity: 10 μ V/mm.

with multiple types of focal seizures. Background activity on EEG deteriorated and showed a predominant delta encephalopathy pattern, with frequent bi- and triphasic delta waves in addition to bilateral frontal spike-wave discharges (Figure 3C). Further treatment modifications did not provide significant improvement (lacosamide,

clonazepam, phenytoin, perampanel, topiramate, cannabidiol, ketogenic diet) or even worsened seizure activity (clobazam; Figure S3). Follow-up EEG continued to show persistently disorganized activity, with asymptomatic seizures localized to the central right and midline regions (Figure 3D). His level of consciousness altered

progressively. On last follow-up at the age of 6.5 years, the patient demonstrated profound intellectual disability, limited autonomy in daily activities, and speech impairments, speaking only a few words.

At the age of 7 years, the patient experienced a 2-month episode of refractory status epilepticus, with a suspected but unconfirmed diagnosis of limbic encephalitis. Management included admission to the intensive care unit and plasma exchange therapy. The myoclonic status epilepticus resolved following oral administration of phenytoin and zonisamide, with a gradual and complete return to the patient's baseline status: independent walking, presence of stereotypies, absence of spoken language, and emotional expression through smiling and vocalizations. No signs of pyramidal syndrome were observed.

3.3 | Characterization of the structural rearrangements by LRS

Both patients had a de novo gain of minimal size of 536 kbp within the *FGF12* gene at the 3q28q29 locus as previously detected by array CGH analysis: arr[GRCh38] 3q28q29 (192 182 695_192 719 315)×3. Targeted LRS was performed using the adaptive sampling method, focusing on the region chr3:163 000 000–195 000 000, which includes *FGF12* among more than 120 genes. After sequencing enrichment, the targeted region achieved 22-fold coverage for both patients, whereas the entire genome had 2–3-fold coverage. Sequencing generated 12.6 GB and 6.6 GB of data, corresponding to 14.9 million and 11.37 million reads for Patient 1 and Patient 2, respectively. The average supporting read size over the copy-gain variant containing *FGF12* was 15 508 bp for Patient 1 and 7948 bp for Patient 2.

The precise coordinates of the identified duplications were as follows: NC_000003.12:g.192153910_192743885dup in Patient 1 and NC_000003.12:g.192155125_192745081dup in Patient 2 (Figure 4A). Furthermore, Structural variant calling using CuteSV detected the duplication in both patients, further confirming these events. The breakpoints were located within the L1PA2-A and L1PA2-B (LINE-1 Primate-specific subfamily A, number 2) LINE elements (Long Interspersed Nuclear Element) at chr3:192 149 316–192 155 328 and chr3:192 739 214–192 745 242, respectively. These LINE elements share 98.14% sequence homology. Chimeric read alignments indicated a direct tandem duplication for both cases, with the 5' breakpoint located in the last intron, common to both isoforms. The 3' breakpoint was located downstream of the 5' untranslated region in an intergenic region (Figure 4B–4C). The involvement of L1PA2-A and L1PA2-B LINE elements in both

patients suggests a nonallelic homologous recombination (NAHR) mechanism (Figure 4B). This mechanism and the occurrence of duplications are further supported by Sanger sequencing analysis of the breakpoints (Figure S4), which revealed homologous sequences of 73 bp in Patient 1 and 53 bp in Patient 2.

3.4 | LRS of *FGF12* major isoforms A and B

cDNA analysis was performed on *FGF12* isoform A and isoform B. Both WT transcripts share the terminal exons, referred to as exons 3 to 6. They differ in their first exons; WT transcript A contains exon 1A, whereas WT transcript B contains exon 1B and exon 2. Exons 1A, 1B, and 2 together form the initial transcript exons (ITEs).

RNA was extracted from human LCLs derived from Patient 1 and reverse transcribed in cDNA fragments. PCR primers were designed to amplify potential aberrant transcripts, and the resulting amplicon library was sequenced using LRS technology. By analyzing the sequences of the last exons, we identified two different types of aberrant isoforms in addition to the WT isoforms (Figure 5A). The first, aberrant isoform A, involved a fusion of the coding sequences of exon 5 and exon 3, skipping exon 6, in the sequence ITE-3-4-5-3-4-5-6 (Figure 5B). The second, aberrant isoform B, was characterized by the insertion of 123 nucleotides from the intergenic region chr3:192 738 861–192 738 983, named "INS," after exon 5, followed by exon 3, in the sequence ITE-3-4-5-Ins-3-4-5-6 (Figure 5B). The WT proteins yielded from isoform A and isoform B consist of 243 and 181 amino acids (aa), respectively. The aberrant isoform A protein based on the isoform B is predicted to be 320 aa, with no premature termination due to the in-frame fusion of exon 5 and exon 3 (Figure 5C). In contrast, the aberrant isoform B protein is predicted to be truncated with a premature stop codon (Figure 5C).

4 | DISCUSSION

Resurgent sodium currents are essential for regulating the excitability of central and peripheral neurons. In this study, we applied LRS to investigate the structural organization of recurrent intragenic microduplications in *FGF12* and their effects on transcription in two patients with developmental and epileptic encephalopathy (DEE), both harboring an overlapping phenotype and specific EEG features.

Recurrent *FGF12* microduplications with breakpoints located within the L1PA2-A and L1PA2-B LINE elements have been previously identified in patients with DEE

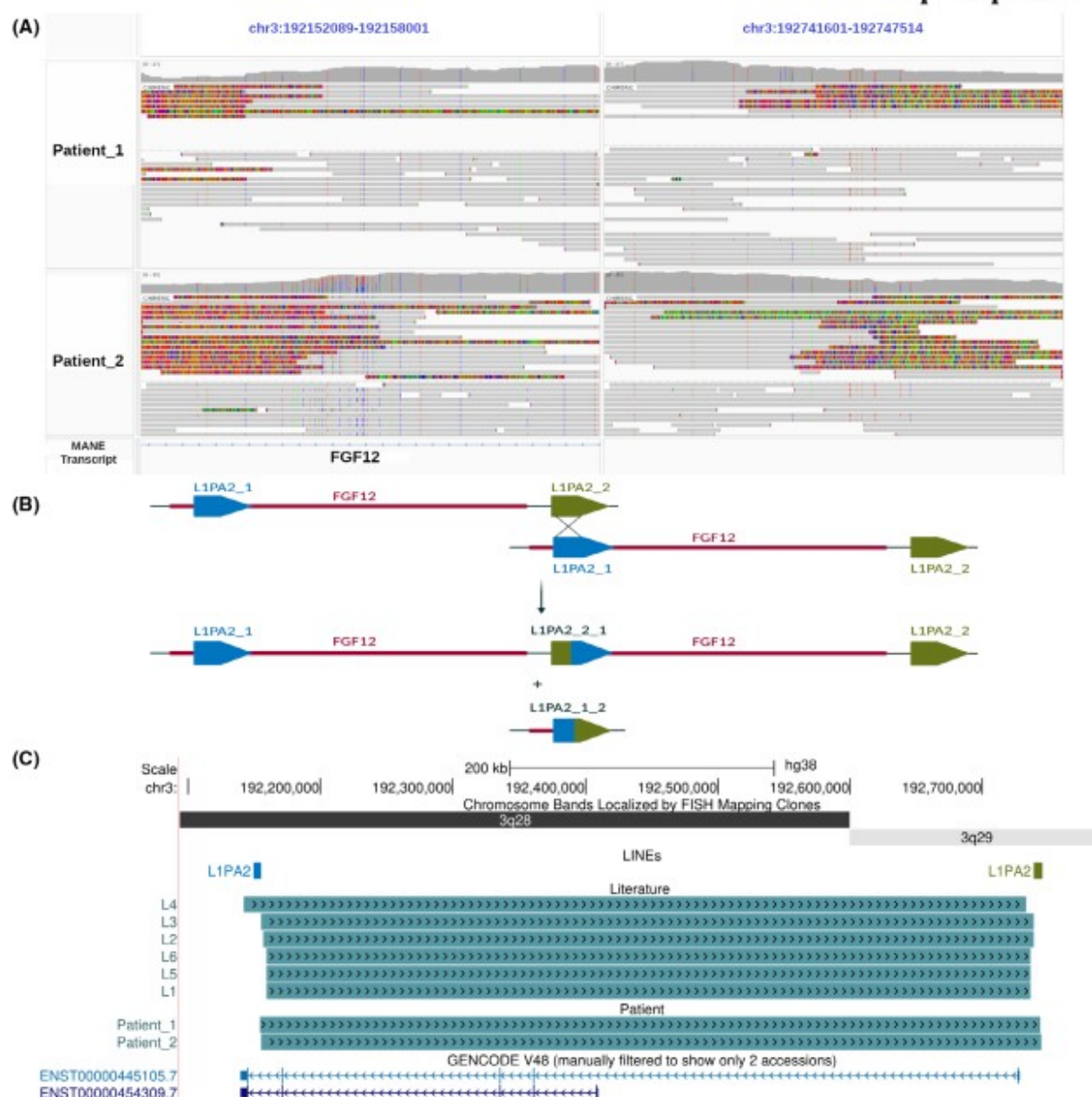


FIGURE 4 Study of DNA rearrangement encompassing *FGF12* gene. (A) Visualization of long-read sequencing reads overlapping the *FGF12* duplication breakpoints in Patients 1 and 2, displayed using the Integrated Genome Viewer. (B) Schematic representation of the repair mechanism through nonallelic homologous recombination, precisely localized on L1PA2 LINE elements (LINE-1 Primate specific subfamily A, number 2; Long Interspersed Nuclear Element) at the *FGF12* duplication breakpoint. (C) Overview of duplications identified in patients from the literature (L1–L6) and in patients from our study (Patient 1 and Patient 2), along with the localization of L1PA2 elements in the UCSC Genome Browser (University of California, Santa Cruz; chr3:192080000–193020000). (FISH = Fluorescence In Situ Hybridization; MANE = Matched Annotation from NCBI and EMBL-EBI).

through array CGH and confirmed via Sanger sequencing (Table 1).^{12–14,21} LRS enabled the precise identification of the complete structural rearrangement as a direct tandem duplication between the two interspersed paralogous repeats, L1PA2-A and L1PA2-B, as previously suspected.^{12,14} Unlike whole-genome short-read sequencing,

LRS provides an unbiased and comprehensive approach to investigate the two-dimensional organization of the genome, thereby supporting the formulation of physiopathological hypotheses.^{22,23} This structural change is likely to be driven by the NAHR mechanism, which exploits sequence homology between the directly oriented

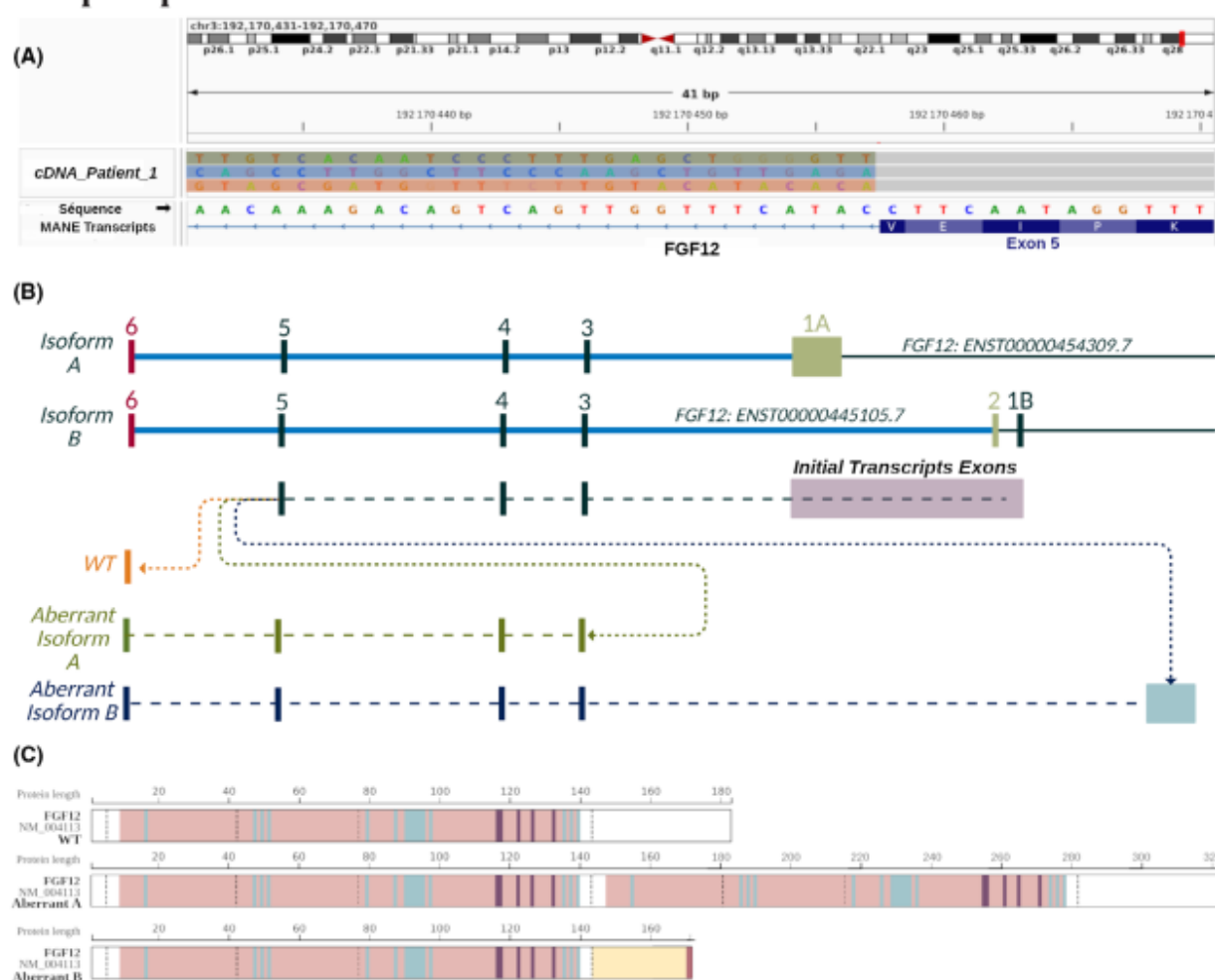


FIGURE 5 Study of the impact on *FGF12* RNA transcripts. (A) Visualization of long-read aberrant transcripts from Patient 1 of exon 5, displayed using the Integrated Genome Viewer. (B) Modeling of the sequences of the wild-type (WT) transcript isoforms (A and B) and aberrant transcripts based on the visualization of sequencing of transcripts. (C) Modeling of the sequence of the WT and aberrant proteins (yellow: intergenic sequence; red: stop codon). (MANE = Matched Annotation from NCBI and EMB-EBI).

L1PA2-A and L1PA2-B repeats to mediate recombination. This process can result in either deletions or duplications, as observed in other genomic instability hotspots flanked by repeat elements.²⁴ Similarly, the recurrent *FGF12* duplication exhibits highly conserved breakpoints, with homologous regions differing by only a few base pairs, further supporting an NAHR-mediated mechanism.^{12–14,21}

As member of the FHF subfamily, *FGF12* has at least two transcription initiation sites that generate two major isoforms. Transcript A (ENST00000454309.7; NM_021032) encodes FGF12 isoform A, a protein with a longer N-terminal sequence compared to FGF12 isoform B, encoded by transcript B (ENST0000044105.7; NM_004113). Structurally, FHF proteins share a β -trefoil fold, with a conserved C-terminal region that covers a hydrophobic surface on the β -trefoil fold.^{23,25,26} The β -trefoil fold of FGF12 contains unique motifs that facilitate interactions

with the cytoplasmic tails of Nav channels, thereby modulating their activity. These interactions enhance neuronal excitability by shifting the voltage dependence of rapid inactivation of the Nav channels.^{2,3,26} FGF12 contains key surface residues essential for Nav channel interactions. The FGF12 isoform A possesses distinct properties due to a conserved motif of 15–20 residues at the N-terminus, which acts as an open-channel blocking particle. This motif facilitates both rapid and prolonged inactivation by maintaining the channel in a long-term inactivated state.^{2,26} A pathogenic missense variant, p.(Arg144His) in isoform A, has been identified as a cause of DEE. This substitution disrupts the interaction FGF12–Nav channel interaction, causing a depolarizing shift in channel inactivation and leading to a gain-of-function effect.^{3,5,15}

The recurrent *FGF12* duplications identified in this study included the central motif and key residues essential

TABLE 1 Clinical table of newly reported and previously published patients with the *FGF12* recurrent duplication of 500 kb and *FGF12* missense variant.

Characteristic	Patient 1	Patient 2	Shi et al. 2017	Verheye et al. 2020	Abraham et al. 2024	Willemssen et al. 2020, Patient 1	Willemssen et al. 2020, Patient 2	Willemssen et al. 2020, Patient 3	Duplication synthesis	Missense variant synthesis, Trivisano et al., n = 16
Patient age at last consultation	6 years	6 years	8 years	7 years	22 years	10 years	5 years	30 years	6–22 years	1 month to 33 years
Genomic variant	3q28q29 (chr3:192 153 910–192 743 885) (hg38)	3q28q29 (chr3:192 154 622–192 744 509) (hg38)	3q28q29 (chr3:192 159 189–192 736 886) (hg38)	3q28q29 (chr3:192 156 263–192 739 147) (hg38)	3q28q29 (chr3:192 155 384–192 738 956) (hg38)	3q28q29 (chr3:192 142 300–192 733 325) (hg38)	3q28q29 (chr3:192 159 179–192 736 896) (hg38)	3q28q29 (chr3:192 159 179–192 736 896) (hg38)		
Inheritance	De novo	De novo	De novo	De novo	De novo	De novo	Inherited from Patient 3	Mother of Patient 2, de novo		
Age at diagnosis	2 years	5 months	NA	NA	16 years	NA	NA	NA		
Neurological impairment										
Intellectual impairment	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes (100%)	Yes (87.5%)
Early psychomotor development	Normal	Normal	Normal	Normal	Normal	Normal	DD	DD	Normal (4/8), DD (2/8)	
Sitting age	NA	9 months	NA	NA	NA	NA	NA	NA		
Walking age	15 months	15 months	NA	NA	NA	NA	NA	NA		
Age of first words	12 months	2 words at 18 months (dad, mom)	NA	NA	NA	NA	NA	NA		
Current language level: oral [words, sentences], written [syllabic, fluent reading]	Absence of language	Few words	Few words	Single words	No speech (before her death at age 22 years)	No speech	Reduced vocabulary, drooling	Speech difficulties	Abnormal speech: 100%	
Autonomy in daily life: cleanliness, eating, dressing	Lack of autonomy	Lack of autonomy	Lack of autonomy, ataxic walking	Loss of practical hand skills, independent walking	Loss of the ability to speak and eat, wheelchair dependent	Wheelchair dependent, independent walking at 5 y.o.	Walking independently at 5 y.o.	Care supervisor for her daily activities, limited writing and reading, special education	Reduced autonomy: 100%	
IQ	Profound ID	NA	Profound ID	NA	NA	Severe ID	NA	Mild to moderate	Mild to Profound	Mild to severe
ASD or autistic traits	Yes	Yes	NA	NA	NA	No	Yes, level 3	NA	Yes (23%), no (7%), NA (70%)	Yes (43.7%), no (25%), other (18.7%)
Neurological signs	Global hypotonia	NA	Aphasia, ataxia	Hand apraxia	NA	Ataxia	Uncoordinated gait	Unsteadiness		
Behavioral issues	Poor eye contact	Short attention span	NA	Inappropriate laughing and crying	NA	NA	NA	NA		
Sleep disturbance	Yes	No	Yes	Yes	NA	NA	NA	NA		

(Continues)

TABLE 1 (Continued)

Characteristic	Patient 1	Patient 2	Shi et al. 2017	Verheyen et al. 2020	Abraham et al. 2024	Willemssen et al. 2020, Patient 1	Willemssen et al. 2020, Patient 2	Willemssen et al. 2020, Patient 3	Duplication synthesis	Misense variant synthesis, Trivisano et al., n = 16
Treatments already used	VPA, LTG, CZP	LCM, EZD	PB, EZD, VPA, Kbr	VPA, PHT, CBZ, plasmapheresis	PHT, VPA, GBP	NA	VPA	NA		PB (13/16), VPA (8/16), PHT (11/16), GVG (5/16), TPM (6/16), CZP (1/16), PN (3/16), LEV (9/16), KD (3/16), PER (3/16), VNS (3/16), PRG (1/16), CBZ (7/16), LTG (2/16), RUF (1/16), CLB (2/16), ESM (1/16), ZNS (1/16), Kbr (1/16), RTG (1/16), OXC (1/16), LCM (1/16), CLZ (1/16), PLP steroids (1/16), folate acid (2/16), biotin (1/16)
Effective treatment			PHT	PHT, LTG, LEV, TPM, CLB	PHT, VPA, GBP	VPA, EZD, LTG, TPM	VPA	VPA, CBZ	PHT (3/8), VPA (4/8), LTG (2/8), LEV (1/8), EZD (2/8), TPM (2/8), CBZ (1/8), GBP (1/8)	PHT (8/16), VNS (1/16), CBZ (2/16), RUF (1/16), LTG (1/16), PB (2/13), VPA (2/16), ASM (2/16)
Aggravating treatment	No	NA	NA	NA	NA	NA	No	NA	Not reported	
Regression of acquisitions	Yes at 20 months	Yes	Yes	Yes	Yes, 7 y.o.	Yes, 3 y.o.	No	No	Yes (6/8 75%)	
EEG, first seizure background trace	Disorganized background activity	Diffuse delta hypersynchrony of sleep onset, ampic, then polymorphic delta rhythm overlaid by sleep spindles, slow Rolandic spikes	NA	NA	NA	NA	At 13 months: normal EEG	NA		Interictal: slow (56%), normal (31%), no description (23%)

(Continues)

TABLE 1 (Continued)

Characteristic	Patient 1	Patient 2	Shi et al. 2017	Verheyen et al. 2020	Abraham et al. 2024	Willemssen et al. 2020, Patient 1	Willemssen et al. 2020, Patient 2	Willemssen et al. 2020, Patient 3	Duplication synthesis	Misense variant synthesis, Trivisano et al., n = 16
EEG, first seizure paroxysmal abnormalities	Rare focal spikes of variable focalization, often diffuse over one hemisphere	Left temporoparietal paroxysmal wave spikes	NA	NA	NA	NA	At 13 months: normal EEG	NA		Interictal: multifocal SW (62%), diffuse SW (6%), focal SW (6%), suppression burst (6%)
EEG, follow-up background trace	Background of diffuse encephalopathy rich in bi- and triphasic delta activity		Severe suppression of background activity	Background suppression when NCSF; or sleep background, diffuse slowing	At 8 y.o. normal EEG	At 3 y.o.: slowed background activity	NA	NA	Abnormal background (4/5)	Slow (64%), normal (14%), no description (22%)
EEG, follow-up paroxysmal anomalies	Isolated generalized spikes or bursts at 2.5 Hz with atypical absence and concomitant myoclonus		Paroxysmal bursts	Periodic bursts of slow wave activity (NCSF); or primary generalized and bioccipital epileptic discharges	At 8 y.o. normal EEG	At 3 y.o.: multifocal epileptic discharges	At 2 y.o.: generalized epileptic activity	NA	Paroxysmal abnormality (5/6)	Multifocal SW (64%), focal SW (21%), generalized and focal paroxysmal in temporal regions (15%)
MRI										
MRI abnormalities: description and picture if possible	At 4 y.o.: corticocortical atrophy	2018: normal MRI 2024: nodular lesion of the same signal as the 5 mm cortex of the white matter of the left cerebellar peduncle; probable gray matter heterotopia	At 8 y.o.: mild cerebellar and cerebellar atrophy	NA	At 18 y.o.: T2 hyperintensity in left posterior periventricular white matter consistent with a perivascular space	Bilateral delayed myelination in the parieto-occipital region	Mild prominence of the subarachnoid space in the frontal regions bilaterally, otherwise normal	Normal		Cerebellar atrophy (19%), normal (38%), mild cerebellar atrophy (6%), temporal sclerosis/cerebellar folia (6%), cerebellar atrophy (19%), mild Chiari I (6%), mild enlargement of lateral ventricle (6%)
Pyramidal signs: spasticity, hyperreflexia	Progressive spasticity of the 4 limbs	No	No	NA	NA	NA	NA	NA		
Other										
Height, SD	At 6 y.o.: -1 SD	-0.5 SD	At birth: -1.2 SD	NA	NA	NA	NA	NA		
Weight, SD	At 6 y.o.: M	-0.5 SD	At birth: -0.5 SD	NA	NA	NA	NA	NA		
Head circumference, SD	At 6 y.o.: -1 SD	NA	At birth: -0.4 ET	NA	NA	NA	NA	NA		
Appetite disorders: food selection, small appetite	No	No	Dysphagia when seizures occurred	NA	NA	Progressive feeding difficulties, reflux	NA	NA		

for interactions with Nav channels, specifically Lys9, Ile11, Arg52, Tyr93, and Val95 (as designated in FGF12 isoform A nomenclature).¹⁵ These duplications were associated with two aberrant transcripts, each originating from transcription starts of either transcript A or transcript B. The first aberrant transcript, derived from WT transcript A, is composed of eight exons, corresponding to the following sequence: exon 1A-3-4-5-3-4-5-6, with no predicted occurrence of a premature stop codon. The second aberrant transcript, derived from WT transcript B, is composed of seven exons and includes an intergenic region. It corresponds to the following sequence: exon 1B-2-3-4-5-INSsequence-3-4-5-6 (Figure 5). Previous studies using RT-PCR and Sanger sequencing on blood and hair follicles from a patient with *FGF12* duplication suggested the presence of a long aberrant transcript containing all exons in-frame. This hypothesis is further supported by our study, with the identification by LRS on cDNA of the presence of the same extended transcript, corresponding to aberrant transcript A.^{14,15} The aberrant transcript could potentially be fully translated, yielding a hypothetical longer protein with two central β -trefoil folded regions, as exons 5 and 3 remain in-frame, preventing a frameshift. Although we have not analyzed the protein sequence, we suspect that the aberrant structural conformation may lead to protein instability and degradation. Our results also validated the existence of a second aberrant transcript characterized by an out-of-frame insertion of intergenic content, leading to a premature stop codon.¹⁴ It is also important to note that *FGF12* is affected by both recurrent and nonrecurrent structural variants. For example, a patient with biallelic structural variants—including a deletion of exon 6 and a duplication of exons 3–4 of *FGF12*—was recently reported.²⁷ In that study, LRS enabled the resolution of small, complex biallelic structural variants. The patient exhibited focal motor seizures progressing to bilateral tonic-clonic seizures and status epilepticus, with apparent treatment responsiveness. These findings illustrate the heterogeneity of *FGF12*-related rearrangements and emphasize the utility of LRS in characterizing clinically relevant, complex structural variants.

Pathogenic missense variants have been linked to a gain-of-function effect on Nav channels, whereas duplications have been linked to loss of function. Despite these opposite mechanisms, both alterations give rise to an identical DEE phenotype, differing only by the age at epilepsy onset. Gain-of-function variants lower the firing threshold of Nav channels, leading to neuronal hyperexcitability and early onset seizures.^{5,15} In contrast, loss-of-function duplications do not trigger hyperexcitability and are associated with a slightly later seizure onset. The putative aberrant elongated variant protein may shift the voltage dependence of Nav channel

inactivation toward more hyperpolarized potentials, impairing channel activity, further supporting a loss-of-function mechanism.¹⁵ In the hippocampus, mossy fibers carry signals from granule cells in the dentate gyrus to the CA3 region. These fibers target excitatory pyramidal cells and GABAergic inhibitory interneurons. These interneurons, which release the neurotransmitter GABA, play a key role in suppressing neuronal activity, thereby regulating excitation and maintaining network balance.^{28–30} Nav channels, particularly Nav1.2, are highly expressed in the dentate gyrus granule cells, which project to the CA3 region. Interestingly, mossy fibers form connections with specific inhibitory interneurons known as basket cells approximately 50 times more frequently than with excitatory pyramidal cells. Loss of function in these Nav channels reduces the efficiency of inhibitory signaling, thereby increasing hyperexcitability in the CA3 region.^{28–30} This altered excitability may explain the later onset of epileptic seizures (average onset at 19 months for duplicated variant against 19 days for missense variants), a timing that corresponds with the expression of these cells in a more mature brain and is observed in patients with *FGF12* duplication.^{5,12–15} Thus, as previously suspected, patients with *FGF12* duplication presented with a later onset of seizures and showed an apparent stagnation or regression in development after seizure onset.¹²

Most of patients with *FGF12* pathogenic missense variants presented with drug-resistant epilepsy and EEG showing mainly multifocal nonspecific abnormalities. When EEG data were available, half presented with slow background activity and interictal focal or multifocal spikes; discontinuous patterns of suppression burst were rarely seen. Ictal EEG was reported as focal (with rapid spread to bilateral regions) or generalized with low-voltage fast activity, followed by widespread spikes and postictal suppression.^{9,10} Conversely, in patients with *FGF12* duplications, we have observed a more complex EEG pattern, with some common specificities (Table 1). First, at the onset of epilepsy, both patients presented a marked alteration in electrogenesis compatible with severe encephalopathy, suggesting an inflammatory etiology. Both patients showed a very slow and high-voltage activity, as described in encephalitis.³¹ One of our patients, like others in the literature, was treated with immunomodulatory drugs.²¹ In treated patients, a three-phase evolution was observed. The first phase corresponded to a storm of seizures with cognitive regression (Figure 6). Phase 2 was a “honeymoon phase,” with partial response to therapy, corresponding to a seizure-free period, but with significant improvement in cognitive performance. Finally, in phase 3, we observed a worsening of the epilepsy with nonconvulsive status

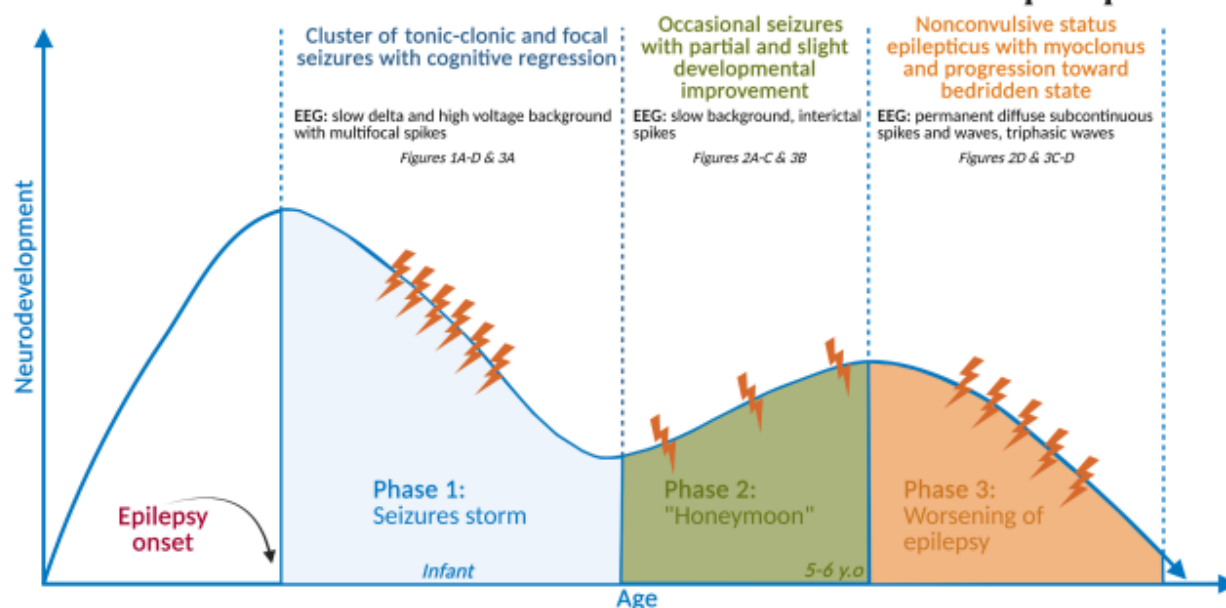


FIGURE 6 Three-phase evolution of epilepsy and neurodevelopment. Following initial development, patients experience a seizure storm (phase 1), a transient improvement (phase 2), and a later decline with nonconvulsive status epilepticus and motor deterioration (phase 3). EEG, electroencephalography; y.o., years old.

epilepticus, obtundation, myoclonus (accompanied by diffuse subcontinuous spikes and waves with abundant triphasic waves on EEG), and progression toward bedridden state.^{12,14,21} Both patients reported in this study had bilateral tonic-clonic seizures and focal temporal seizures at epilepsy onset. This temporal localization is consistent with the hypothesis of disruption in the balance between excitatory and inhibitory signaling within these regions.²⁸ The kinetics of seizure onset and progression in *FGF12* duplications may suggest a disease course similar to Nav1.2 (SCN2A) loss of function, with a later onset compared to Nav1.2 (SCN2A) gain of function.²⁸ The parallel with Nav1.2 alterations could be extended to treatment approaches. Like patients with pathogenic missense variants, those with duplications exhibited significant resistance to antiepileptic treatments despite various therapeutic approaches. Although phenytoin demonstrated partial efficacy in *FGF12* duplications, its overall impact was limited and failed to produce substantial clinical improvement.^{10,12,14}

We have provided the first comprehensive characterization of *FGF12* duplications using LRS approaches on DNA combined with cDNA analysis, to reveal the complete structural rearrangement caused by NAHR mechanisms and their impact on transcripts. This resulted in two distinct aberrant isoforms: an elongated isoform A and a frameshift isoform B. The electroclinical consequences of *FGF12* duplications appear to be more specific than previously suspected compared to

point pathogenic missense variants, with clinical regression observed after seizure onset and distinctive EEG patterns at onset and during evolution/progression, suggesting that these duplications could be suspected based on EEG findings.

AUTHOR CONTRIBUTIONS

Jade Fauqueux and Thomas Smol wrote the manuscript. Jade Fauqueux and Caroline Thuillier conducted molecular genetic studies, assisted by Céline Villenet. Jean-Pascal Meneboo, Emilie Ait-Yahya, and Martin Figeac carried out the bioinformatic analysis for long-read genome sequencing. Jade Fauqueux handled the bioinformatics analysis for long-read amplicon sequencing. Jade Fauqueux and Thomas Smol interpreted the LRS data. Elise Boudry and Nicolas Gruchy performed the array CGH analysis. Roseline Caumes, Laurence Chaton, Pierre Cleuziou, Adeline Trauffer, Sylvie Nguyen-The-Tich, Nicolas Gruchy, Marion Gerard, Simon Boussion, Nathalie Bach, and Anne-Sophie Diependaële collected and evaluated clinical and genetic data. Thomas Smol and Jamal Ghoumid revised the manuscript. All authors contributed to discussions of the results, provided comments on the manuscript, and approved the final version.

ACKNOWLEDGMENTS

We thank the patients and their families for participating in this study. We thank Delphine Ceraso, Alexis Leurent,

Heidi Tampere, and Pauline Grave for their technical assistance with LRS, primer design, and Sanger experiments.

FUNDING INFORMATION

This work was supported by the University Hospital of Lille, France (Budget Projet Innovation).

CONFLICT OF INTEREST STATEMENT

None of the authors has any conflict of interest to disclose.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2013. This study was approved by the Comité de Protection des Personnes ethics committee (reference #2023-A00473-42).

PATIENT CONSENT STATEMENT

All study participants signed written informed consent prior to their inclusion in the study.

ORCID

Thomas Smol  <https://orcid.org/0000-0002-0119-5896>

REFERENCES

- Olsen SK, Garbi M, Zampieri N, Eliseenkova AV, Ornitz DM, Goldfarb M, et al. Fibroblast growth factor (FGF) homologous factors share structural but not functional homology with FGFs. *J Biol Chem*. 2003;278:34226–36.
- Goldfarb M. Fibroblast growth factor homologous factors: evolution, structure, and function. *Cytokine Growth Factor Rev*. 2005;16:215–20.
- Biadun M, Karelus R, Krowarsch D, Opalinski L, Zakrzewska M. FGF12: biology and function. *Differentiation*. 2024;139:100740.
- Hanada Y, Nakamura Y, Ozono Y, Ishida Y, Takimoto Y, Taniguchi M, et al. Fibroblast growth factor 12 is expressed in spiral and vestibular ganglia and necessary for auditory and equilibrium function. *Sci Rep*. 2018;8:11491.
- Siekierska A, Isrie M, Liu Y, Scheldeman C, Vanthillo N, Lagae L, et al. Gain-of-function FHF1 mutation causes early-onset epileptic encephalopathy with cerebellar atrophy. *Neurology*. 2016;86:2162–70.
- Guella I, Huh L, McKenzie MB, Toyota EB, Bebin EM, Thompson ML, et al. De novo FGF12 mutation in 2 patients with neonatal-onset epilepsy. *Neurol Genet*. 2016;2:e120.
- Villeneuve N, Abidi A, Cacciagli P, Mignon-Ravix C, Chabrol B, Villard L, et al. Heterogeneity of FHF1 related phenotype: novel case with early onset severe attacks of apnea, partial mitochondrial respiratory chain complex II deficiency, neonatal onset seizures without neurodegeneration. *Eur J Paediatr Neurol*. 2017;21:783–6.
- Takeguchi R, Haginoya K, Uchiyama Y, Fujita A, Nagura M, Takeshita E, et al. Two Japanese cases of epileptic encephalopathy associated with an FGF12 mutation. *Brain Dev*. 2018;40:728–32.
- Velišková J, Marra C, Liu Y, Shekhar A, Park DS, Iatckova V, et al. Early onset epilepsy and sudden unexpected death in epilepsy with cardiac arrhythmia in mice carrying the early infantile epileptic encephalopathy 47 gain-of-function FHF1(FGF12) missense mutation. *Epilepsia*. 2021;62:1546–58.
- Trivisano M, Ferretti A, Bebin E, Huh L, Lesca G, Siekierska A, et al. Defining the phenotype of FHF1 developmental and epileptic encephalopathy. *Epilepsia*. 2020;61:e71–e78.
- Paprocka J, Jezela-Stanek A, Koppola A, Rydzanicz M, Kosińska J, Stawiński P, et al. FGF12p.Gly112Ser variant as a cause of phenytoin/phenobarbital responsive epilepsy. *Clin Genet*. 2019;96:274–5.
- Willemsen MH, Goel H, Verhoeven JS, Braakman HMH, de Leeuw N, Freeth A, et al. Epilepsy phenotype in individuals with chromosomal duplication encompassing FGF12. *Epilepsia Open*. 2020;5:301–6.
- Abraham A, Ramsey K, Belnap N, Szelinger S, Jepsen W, Balak C, et al. FGF12 copy number variant associated with epileptic encephalopathy. *Clin Genet*. 2024;106:114–5.
- Shi R-M, Kobayashi T, Kikuchi A, Sato R, Uematsu M, An K, et al. Phenytoin-responsive epileptic encephalopathy with a tandem duplication involving FGF12. *Neurol Genet*. 2017;3:e133.
- Seiffert S, Pendziwiat M, Bierhals T, Goel H, Schwarz N, van der Ven A, et al. Modulating effects of FGF12 variants on Nav1.2 and Nav1.6 being associated with developmental and epileptic encephalopathy and autism spectrum disorder: a case series. *EBioMedicine*. 2022;83:104234.
- Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma Oxf Engl*. 2018;34:3094–100.
- Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet*. 2018;19:329–46.
- Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol*. 2024;42:1571–80.
- Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol*. 2020;21:189.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.
- Verheyen S, Speicher MR, Ramler B, Plecko B. Childhood-onset epileptic encephalopathy due to FGF12 exon 1–4 tandem duplication. *Neurol Genet*. 2020;6:e494.
- Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, et al. Targeted long-read sequencing identifies missing disease-causing variation. *Am J Hum Genet*. 2021;108:1436–49.
- Mantere T, Kersten S, Hoischen A. Long-read sequencing emerging in medical genetics. *Front Genet*. 2019;10:426.

24. Szafranski P, Kośmider E, Liu Q, Karolak JA, Currie L, Parkash S, et al. LINE- and Alu-containing genomic instability hotspot at 16q24.1 associated with recurrent and nonrecurrent CNV deletions causative for ACDMPV. *Hum Mutat*. 2018;39:1916–25.
25. Xiao Y, Theile JW, Zybura A, Pan Y, Lin Z, Cummins TR. A-type FHF's mediate resurgent currents through TTX-resistant voltage-gated sodium channels. *elife*. 2022;11:e77558.
26. Goldfarb M. Fibroblast growth factor homologous factors: canonical and non-canonical mechanisms of action. *J Physiol*. 2024;602:4097–110.
27. Ohori S, Miyauchi A, Osaka H, Lourenco CM, Arakaki N, Sengoku T, et al. Biallelic structural variations within *FGF12* detected by long-read sequencing in epilepsy. *Life Sci Alliance*. 2023;6:e202302025.
28. Wolff M, Johannesen KM, Hedrich UBS, Masnada S, Rubboli G, Gardella E, et al. Genetic and phenotypic heterogeneity suggest therapeutic implications in *SCN2A*-related disorders. *Brain J Neurol*. 2017;140:1316–36.
29. Kaplan MR, Cho MH, Ullian EM, Isom LL, Levinson SR, Barres BA. Differential control of clustering of the sodium channels *Na(v)1.2* and *Na(v)1.6* at developing CNS nodes of Ranvier. *Neuron*. 2001;30:105–19.
30. Acsády L, Kamondi A, Sik A, Freund T, Buzsáki G. GABAergic cells are the major postsynaptic targets of mossy fibers in the rat hippocampus. *J Neurosci*. 1998;18:3386–403.
31. Galovic M, Schmitz B, Tettenborn B. C15EEG in inflammatory disorders, cerebrovascular diseases, trauma, and migraine. In: Schomer DL, da Lopes Silva FH, Schomer DL, da Lopes Silva FH, editors. *Niedermeyers Electroencephalography basic Principles and clinical applications*. 7th ed. New York, NY: Oxford University Press; 2017. <https://doi.org/10.1093/med/9780190228484.003.0015>

SUPPORTING INFORMATION

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

How to cite this article: Fauqueux J, Chaton L, Cleuziou P, Diependaële A-S, Bach N, Gruchy N, et al. Long-read sequencing of recurrent *FGF12* duplications in epilepsy: Insights into structural mechanisms and aberrant isoforms. *Epilepsia*. 2025;00:1–19. <https://doi.org/10.1111/epi.18609>

Article Annexe 2



Research article

Combining long-read DNA and RNA sequencing to enhance molecular understanding of structural variations leading to copy gains

Jade Fauqueux^a, Jean-Pascal Meneboo^b, Roseline Caumes^{a,c}, Luc Thomès^a,
Emilie Ait Yahya^d, Caroline Thuillier^e, Elise Boudry^e, Celine Villenet^b,
Martin Figeac^b, Jamal Ghomid^{a,c}, Thomas Smol^{a,e,*}

^a ULR 7364—RADEME, Univ. Lille, FHU-G4 Génomique, Lille F-59000, France

^b Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41-UAR 2014-PLBS, FHU-G4 Génomique, Lille F-59000, France

^c CHU Lille, Clinique de Génétique, Lille, Lille F-59000, France

^d CHU Lille, Cellule de Bioinformatique, Plateau Commun de Séquençage, CHU Lille, Lille F-59000, France

^e CHU Lille, Institut de Génétique Médicale, Lille, Lille F-59000, France

ARTICLE INFO

Keywords:

Long Read Sequencing

Nanopore

Cytogenomics

RERE

ZMYM2

ABSTRACT

Structural variants (SVs) significantly contribute to human disease, but their complexity often makes accurate characterization difficult with conventional methods. Advances in long-read sequencing (LRS) offer potential by spanning kilobases and directly resolving SVs. In this study, we examined two individuals with unresolved SVs. LRS on both DNA and cDNA provided single-base resolution of all breakpoint junctions, revealing detailed rearrangement structures and underlying mechanisms. Transcriptomic analyses identified abnormal fusion transcripts and clarified their functional consequences, including haploinsufficiency and potential dominant-negative effects. In one case, a triplication affecting the *ZMYM2* gene was precisely mapped, revealing a truncated variant that may escape nonsense-mediated decay. In the second case, a highly complex reciprocal translocation involving *RERE* and *FHAD1* disrupted *RERE* expression, with Hi-C data showing minimal impact on enhancer-promoter interactions. Due to their complexity, these SVs were not fully resolved by standard methods. By integrating LRS with transcriptomic and chromosomal conformation analyses, we provided a comprehensive understanding of SV formation and its pathogenic impact. Our findings emphasize the need for advanced genomic approaches to resolve complex SVs, enhance diagnostic accuracy, and inform clinical management.

1. Introduction

Structural variants (SVs) are large-scale genomic rearrangements that affect DNA segments of 50 base pairs (bp) or more. SVs are broadly classified into two main types based on their effect on DNA dosage: unbalanced and balanced variants. Unbalanced variants, also known as copy number variants (CNVs), include duplications and deletions that result in DNA gains and losses, disrupting genetic dosage or the regulatory environment. In contrast, balanced variants, such as inversions and balanced translocations, do not change the overall DNA content but reorganize the genomic structure. These rearrangements can affect gene expression or regulation, particularly when breakpoints disrupt functionally critical genes [1–3].

Accurate characterization of SVs is essential for precise diagnosis. Their detection typically involves conventional and molecular

cytogenetic techniques, including karyotyping, fluorescence *in situ* hybridization (FISH), comparative genomics hybridization arrays (CGH), and short read sequencing (SRS) [1]. However, these approaches face significant limitations, particularly in resolving complex genomic rearrangements (CGRs) and SVs within highly repetitive genomic regions. Such challenges complicate the precise characterization of some duplications, which is critical for assessing their potential effects on nearby genes or regulatory elements [4]. As a result, some SVs identified by SRS or CGH may remain classified as variants of uncertain significance (VUS), presenting significant challenges for clinical interpretation and decision-making. [5].

In this context, the use of long read sequencing (LRS) technologies to characterize complex events has significantly advanced the understanding of SVs [6,7]. Reads generated by LRS technologies range from 10 to 150 kbp, enabling the capture of full-length or near full-length SVs.

* Corresponding author at: CHU Lille, Institut de Génétique Médicale, Lille, Lille F-59000, France.

E-mail address: thomas.smol@univ-lille.fr (T. Smol).

<https://doi.org/10.1016/j.csbj.2025.04.031>

Received 26 February 2025; Received in revised form 22 April 2025; Accepted 23 April 2025

Available online 24 April 2025

2001-0370/© 2025 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

By reducing sequencing bias and resolving the organization of complex regions at both the DNA and cDNA levels, LRS strategies are considered a key component of cytogenomics [8].

In this study, we employed LRS on both DNA and cDNA to fully characterize two unresolved structural cytogenetic cases: a two-copy gains of the *ZMYM2* gene and a copy gain of the *RERE* gene. The *ZMYM2* gene encodes a zinc finger transcriptional regulator involved in promoting and maintaining cell identity [9]. Monoallelic *ZMYM2* loss-of-function variants cause an autosomal dominant craniofacial syndrome with renal and cardiac anomalies (NECRC, MIM #619522). Additionally, some evidence suggests a potential dominant-negative mechanism through escape from nonsense-mediated decay (NMD) [10]. The *RERE* gene encodes a nuclear receptor coregulator that regulates the retinoic acid signal [11]. Pathogenic variants within *RERE* are associated with an autosomal dominant neurodevelopmental disorder characterized by brain, ocular, and cardiac anomalies (MIM #605226). Most pathogenic variants in *RERE* lead to loss of function, although severe phenotypes have also been reported with specific missense variants [11,12].

Our goal was to integrate a comprehensive DNA and cDNA approach using LRS to refine the classification of structural variants previously designated as variants of uncertain significance.

2. Materials and methods

Peripheral blood samples were collected from the affected individuals and their parents after obtaining informed consent. This study was approved by the *Comité de Protection des Personnes* (CPP) ethics committee, reference #2023-A00473-42. Informed written consent was obtained from all participants or their legal guardians.

Clinical descriptions Patient 1 is a 2-year-old boy born to unrelated parents. The patient was delivered at 39 weeks' gestation after a pregnancy with suspected congenital heart defect that was not confirmed after birth. At 15 months of age, his weight was 9.7 kg (-1 SD), his height was 75 cm (-2 SD), and his cranial circumference was 44 cm (-3 SD). Clinical features include axial hypotonia that began at 2 months of age, feeding difficulties, moderate developmental delay, postnatal microcephaly and cryptorchidism. CGH analysis identified a two-copy gain at 13q12.11 encompassing the *PSPCL1*, *ZMYM5*, and *ZMYM2* genes for Patient 1.

Patient 2 is an 11-year-old girl born to unrelated parents after an uncomplicated full-term pregnancy. At 10 years of age, the patient weighed 41 kg (+1 SD), was 1.46 m tall (+2 SD), and had a cranial circumference of 55 cm. She presented with speech delay, with her first words appearing around 3 years of age. She showed global delay and moderate intellectual disability. A copy gain at 1p36.23 encompassing the *RERE* and *ENO1* genes was initially identified by CGH for Patient 2.

2.1. CGH-array

Patient peripheral blood samples were collected in EDTA tubes. Genomic DNA was extracted using the CHEMAGIC-STAR robotic extraction system (Hamilton, Reno, US-NV). Experiments were performed on genomic DNA. Agilent Human Genome CGH 60 K oligonucleotide arrays (Agilent, Santa Clara, CA, USA) were used according to the manufacturer's protocols.

2.2. Nanopore long-read DNA sequencing

DNA libraries were prepared using the Ligation Sequencing Kit SQK-LSK114 (Oxford Nanopore Technologies ONT, Oxford, UK). Sequencing was then performed according to the standard protocol on an ONT PromethION sequencer, using a PromethION flow cell (R10.4.1) connected to a GridION sequencer as the computing device. After sequencing, basecalling was performed using the Guppy basecaller (version 6.4.6) in High Accuracy (HAC) mode.

The resulting reads were mapped to the human reference genome GRCh38 using MiniMap2 (version 2.24) [13]. These mapped data were then used for downstream analysis. Sniffles2 (version 2.0.7) and CuteSV (version 2.1.1) were used to detect structural variation [14,15]. Sniffles2 calling was optimized with specific parameters to enhance detection of complex variants: `-long-dup-length 5000000` and `-minsupport 3`, while CuteSV was used with the parameters `-max size -1` and `-minsupport 3`. Visual inspection of structural variants within target regions was performed using tools the Integrative Genomics Viewer (IGV version 2.10). Breakpoints were confirmed by Sanger sequencing using specific primers, which are available in Additional Table S3.

2.3. qPCR targeting analysis

Quantitative PCR was performed using genomics DNA extracted from the patient and the PowerUp SYBR Green kit (ThermoFisher Scientific, Waltham, MA, USA) with primers specifically designed to quantify copy number within the variant locus. Copy number was normalized to that of a control sample processed under the same conditions as the patient sample. Relative primer levels were calculated using the $2^{-\Delta\Delta Ct}$ method with normalization to a reference gene, and statistical analysis was performed using appropriate software. All experiments were performed in triplicate.

2.4. Optical genome mapping (OGM)

Ultra-high molecular weight DNA was isolated from patient whole blood using the Bionano Prep SP protocol for frozen human blood (Bionano Genomics, San Diego, CA, USA). Briefly, DNA was labeled with the DLE-1 enzyme at the CTTAAG motif and the backbone was stained using the Direct Label and Stain Kit according to the Bionano Prep DLS protocol (Bionano Genomics). The labeled DNA was quantified using the Promega Quantifluor HS assay and then loaded onto a Saphyr chip for linearization and imaging on the Saphyr instrument.

2.5. Cell culture

Human lymphoblastoid cell lines (LCL) were established by Epstein-Barr virus immortalization of the subject's blood lymphocytes and maintained in RPMI 1640 medium (ThermoFisher Scientific) supplemented with 15 % fetal bovine serum (FBS) (ThermoFisher Scientific) and 1 % penicillin-streptomycin (Life Technologies).

2.6. Nanopore long-read cDNA sequencing

RNA was extracted from LCL using the PureLink RNA Minikit (Invitrogen, Carlsbad, CA, USA). For the ONT sequencing experiment, the Direct cDNA sequencing protocol (SQK-LSK114 kit) was followed according to the manufacturer's instructions. Briefly, for each patient, 2–4 µg of total RNA was reverse transcribed using patient-specific primers, followed by strand switching to produce full-length cDNA. The prepared cDNA libraries were then loaded onto separate PromethION flow cells, one for each patient, for sequencing. Base calling was performed using the Guppy basecaller (version 6.4.6) in High Accuracy (HAC) mode. The resulting reads were aligned to the human reference genome GRCh38 using Minimap2 with splice-aware alignment settings to accommodate transcriptomic data. RT-qPCR was performed to examine the distribution of the transcripts of candidate gene in both patients.

2.7. Conventional cytogenetic analysis

Karyotyping was performed using RH G-banding on metaphase spreads prepared from peripheral blood of patient 2 according to standard procedures. FISH analysis was performed on metaphase and interphase nuclei from blood lymphocytes using fosmid clones

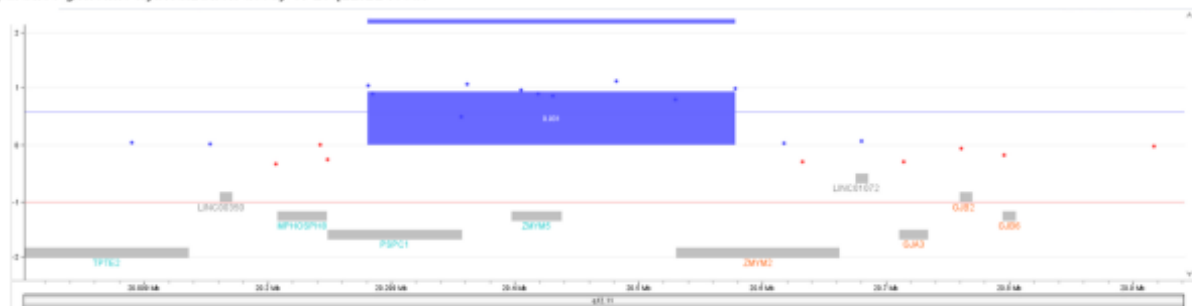
G248P82408E3, G248P86196F8, G248P82014D9, and G248P89339F11 for patient 2.

2.8. Hi-C sequencing

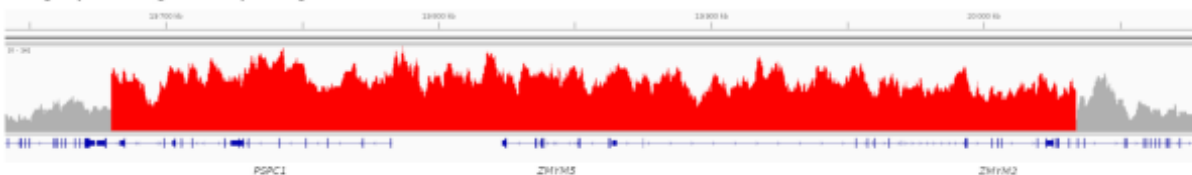
Hi-C sequencing was performed for patient 2 on LCLs using the Arima HiC kit for mammalian cell lines (Arima Genomics, Carlsbad, CA, USA). Briefly, 10 million cells were cross-linked with 2 % formaldehyde, lysed, and digested with enzymes provided in the Arima kit. The resulting fragments were labeled with biotin, ligated, and cross-linking was reversed. Next, the biotin-labeled DNA fragments were sheared to a size of approximately 400 bp using a Covaris S220 instrument (Covaris

LLC, Woburn, MA, USA) and AFA tubes. Biotin-labeled fragments were then selectively enriched with beads from the kit. DNA was further processed through repair steps according to the Arima protocol using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, Ipswich, MA, USA). Sequencing libraries were prepared by ligation of adapters and indexes using the NEBNext Multiplex Oligos for Illumina kit (New England Biolabs). PCR purification and size selection were performed using Agencourt AMPure XP beads (Beckman Coulter, A63881). The libraries were subjected to deep sequencing (~300 million reads) in a 150 bp paired end run on a NextSeq550 sequencer (Illumina, San Diego, CA, USA). Paired reads obtained from the sequencing were aligned using the Burrows-Wheeler Aligner (BWA-MEM 0.7.17) [16]. Parameters were

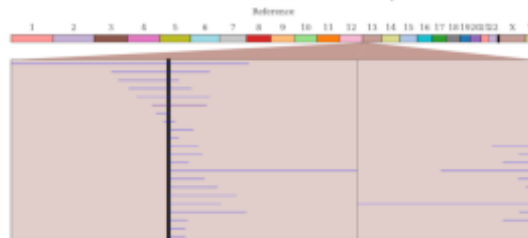
A Comparative genomic hybridization assay of 13q12.11 locus



B Sequencing depth of Long-Read Sequencing data



C Ribbon Visualization of chimeric reads observed in the 13q12.11 locus



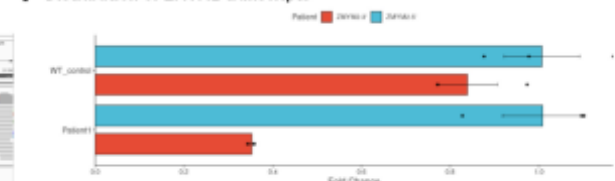
E IGV visualization of DNA and cDNA reads



D Subway Plot



F Distribution of ZMYM2 transcripts



G Plot of total RERE transcripts

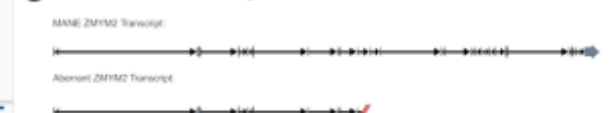


Fig. 1. Molecular analysis of the two-copy gain in Patient 1. (A) Comparative genomic hybridization assay (CGH) analysis reveals a 0.3 Mb gain encompassing the entire *PSPC1* and *ZMYM5* genes, along with a partial duplication of the *ZMYM2* gene. (B) Coverage plot from DNA long-read nanopore sequencing shows a two-copy gain encompassing the entire *PSPC1* and *ZMYM5* genes and exons 1–10 of *ZMYM2*, with precise breakpoint identification. (C) Ribbon Genome Viewer visualization of long-read nanopore sequencing data identifies a tandem triplication event [43]. (D) Schematic representation of wild-type and triplicated alleles, with boxed regions highlighting copy gain loci. (E) Integrated Genome Viewer (IGV) visualization of DNA and cDNA long-read nanopore sequencing data showing genome breakpoints and their effect on *ZMYM2* transcripts. (F) RT-qPCR analysis quantifying total transcripts (blue bars, targeting the 5' region of *ZMYM2*) and wild-type transcripts (red bars, targeting the 3' region of *ZMYM2*). (G) Schematic representation of MANE (Matched Annotation from NCBI and EMBL-EBI) and aberrant *ZMYM2* transcripts.

set in accordance to the documentation of HiCExplorer 3.7.3 [17]. Briefly, mates of the Hi-C sequencing read pairs were aligned separately with parameters setting a high gap extension penalty (-E50) and the penalty for 5' and 3' clipping to zero (-L0). As proposed in the documentation, we also set the matching score to 1 (-A1) and the mismatch penalty to 4 (-B4). Alignments were converted from SAM to BAM files and sorted using Sambamba [18]. For further data exploration, Hi-C contact matrices were produced at various resolutions using HiCExplorer. Resolutions of 50 kb and 1 Mb were selected for the visualization of local contact or TADs and inter-chromosomal contacts

respectively.

3. Results

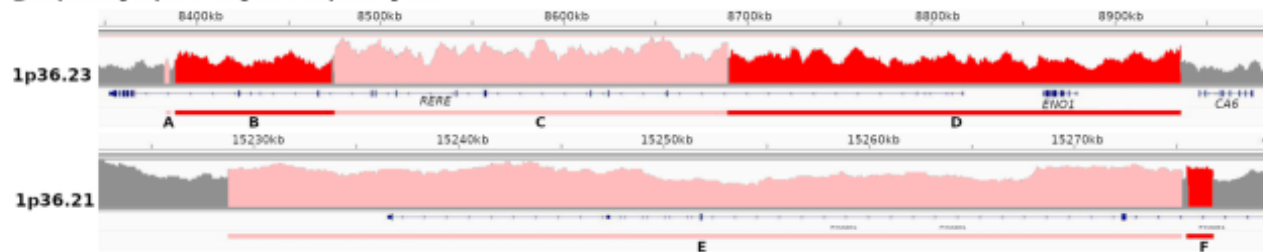
3.1. LRS enables base-level characterization of a triplication involving the ZMYM2 gene

A two-copy gain at 13q12.11 encompassing the *PSPC1*, *ZMYM5*, and *ZMYM2* genes was initially identified by CGH for Patient 1 (Fig. 1A). This finding was subsequently confirmed by targeted qPCR and

A Comparative genomic hybridization assay of 1p36.23 locus



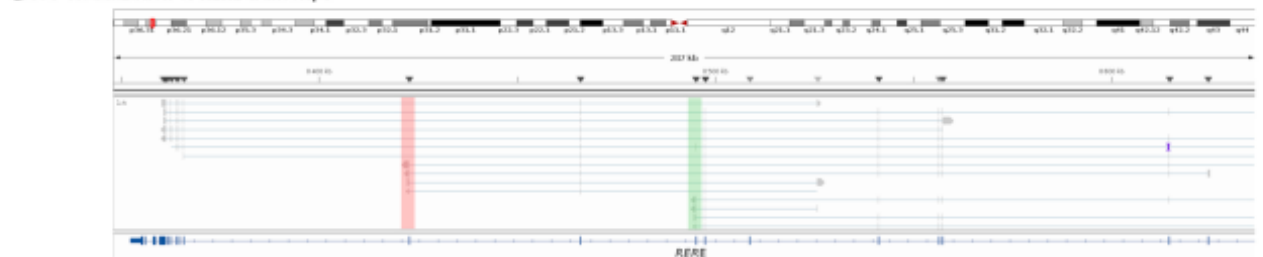
B Sequencing depth of Long-Read Sequencing data



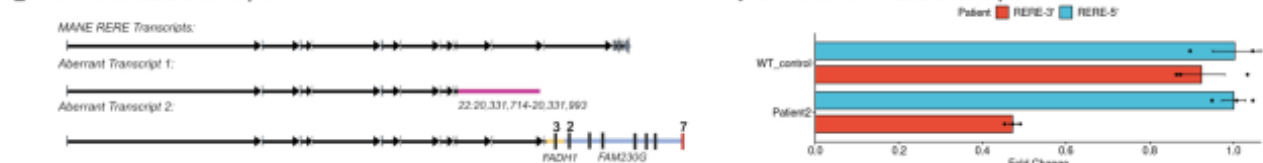
C Subway Plot



D IGV Visualization of RERE transcript



E Plot of total RERE transcripts



F Distribution of RERE transcripts

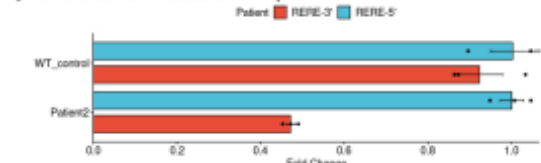


Fig. 2. Molecular Analysis of the Highly Complex RERE Rearrangement in Patient 2. (A) Comparative genomic hybridization assay (CGH) analysis identifies a 0.5 Mb gain encompassing the entire *ENO1* gene and a partial duplication of the *RERE* gene. (B) Coverage plot of DNA long-read nanopore sequencing reveals regions with five-copy and two-copy gains, labeled A to F. (C) Subway plot depicting the complete complex rearrangement observed in Patient 2. Gray structures represent wild-type chromosomes, yellow denotes the derivative chromosome 1, and red denotes the derivative chromosome 22. (D) IGV visualization of cDNA long-read Nanopore sequencing data, showing full-length wild-type and aberrant *RERE* transcripts terminating at exon 10 and exon 12. Green bar indicate exon 10 of the *RERE* gene and Red bar indicate exon 12 of the *RERE* gene. (E) Schematic map of the MANE and Aberrant *RERE* transcripts. (F) RT-qPCR analysis quantifying total transcripts (blue bars, targeting the 5' region of *RERE*) and wild-type transcripts (red bars, targeting the 3' region of *RERE*).

considered to be *de novo* (Additional Figure S1A).

LRS using Nanopore technology was performed to characterize the rearrangement. All sequencing data are reported in Additional Table S4. Visual read depth analysis and the structural variants callers confirmed the presence of the copy gains (Fig. 1B). Chimeric reads identified the breakpoints at chr13:19,680,196–20,033,950, with each copy gain spanning 353,755 bp and arranged in a direct tandem configuration (Fig. 1C). This duplication encompassed the entire *PSPC1* and *ZMYM5* genes, while the *ZMYM2* gene was disrupted by truncation at intron 10 (Fig. 1D–E). This specific copy number gain has not been previously reported in the gnomAD CNV or DGV population databases. While *ZMYM2* is known to be associated with disease, alterations in *PSPC1* and/or *ZMYM5* have not been reported in pathology. In addition, predictive sensitivity scores for copy number gains or losses did not suggest significant pathogenic potential for *ZMYM5* or *PSPC1* [19].

Chimeric sequence analysis revealed junction sequences both upstream and downstream of the copy gains, with a 3-bp microhomology at the breakpoints that was validated by Sanger sequencing (Additional Figure S1B). Furthermore, the absence of similar repeat elements in the region and at the breakpoint suggests that a non-allelic homologous recombination (NAHR) mechanism is unlikely (Additional Tables S1–S2).

The detailed structural characterization provided by LRS confirmed the tandem-direct nature of the rearrangement. This analysis revealed two wild-type alleles of *ZMYM2*, in addition to two truncated alleles containing exons 1–10. Given the predictive sensitivity scores from the pTriplo tool and the absence of previously reported cases with copy number gains affecting these genes, no further investigation into the impact on *PSPC1* and *ZMYM5* was deemed necessary.

cDNA analysis using LRS Nanopore sequencing confirmed the presence of both full-length *ZMYM2* transcripts and aberrant truncated transcripts terminating near exon 10 with polyadenylation signals (Fig. 1E). RT-qPCR was performed to quantify the distribution of full length versus aberrant transcripts and revealed a 50:50 ratio of full length to truncated transcripts in the patient. This finding suggests a frameshift consequence of the triplication SV involving *ZMYM2* (Fig. 1F–G).

3.2. LRS unravels a highly complex structural variant involving the *RERE* gene

CGH analysis identified a copy gain at 1p36.23 spanning 499 kbp and encompassing the *RERE* and *ENO1* genes in Patient 2 (Fig. 2A). Further examination revealed three consecutive probes within this region with a log ratio of approximately 1, suggesting the presence of a second copy gain within the first, resulting in a DUP-TRP-DUP configuration. The rearrangement was subsequently confirmed by targeted qPCR and considered to be *de novo*. This specific copy number gain has not been previously reported in the gnomAD CNV or DGV population databases [20,21].

LRS using Nanopore technology was performed to further characterize the rearrangement. All sequencing data are reported in Additional Table S4. Depth of coverage analysis confirmed the unbalanced DUP-TRP-DUP, with segment sizes of 87,189 bp, 213,819 bp, and 246,772 bp, named 'B', 'C' and 'D', respectively, corresponding to the CGH result (Additional Table S1). Additionally, three smaller CNVs were identified: a copy gain upstream of the 'B' segment, named 'A', with a size of 1871 bp; and two additional copy gains located 7 Mb downstream at the 1p36.21 locus, with segment sizes of 46,495 bp and 1034 bp, named 'E' and 'F', respectively (Fig. 2B, Additional Figures S2–S3). These variants were validated by qPCR with fold-change values of 1.5, consistent with duplications, and 2.0, consistent with triplications (Additional Figure S4A). For clarity, we will use only the A–B–C–D–E–F designation for each genomic segment (Fig. 2C, Additional Figure S3). Detailed analysis of chimeric read patterns identified multiple organizations involving these different segments, notably an interaction

between locus 1p36.23 and locus 1p36.21, and between these two loci and the 22q11.2 region (Fig. 2C). This suggested a more complex rearrangement pattern.

In this configuration, two potential hypotheses of rearrangement were possible: one involving an insertion in chromosome 22 and another involving a reciprocal translocation. Complementary cytogenetic analyses were then performed to differentiate between a translocation from an insertion. Karyotyping and chromosome painting confirmed the presence of 22pter on 1pter and 22qter on 1qter, supporting the hypothesis of a reciprocal translocation (Supplementary Figure S4B–C). OGM and FISH were also performed with probes targeting the 'B', 'C', 'D', and 'E' regions. OGM identified eight breakpoints, while 'A' and 'F' were not detected due to their small size, and confirmed the translocation t(1;22) involving a highly repeated region in the 22q11.2 locus (Supplementary Figure S4D–F). Thus, the complete SV, initially identified as a duplication of the *RERE* gene, was found to involve derivative chromosomes 1 and 22 resulting from a t(1;22) translocation, with two unbalanced complex regions at both breakpoints that include genomic segments from the *RERE* region.

The rearrangement configuration indicated the presence of one wild-type allele of the *RERE* gene and two aberrant alleles arising from the der(1) and der(22) chromosomes. LRS transcript analysis revealed the presence of a full-length wild-type transcript along with two distinct aberrant transcript types (Fig. 2D–E). The first aberrant transcript aligned with the *RERE* gene up to exon 10 and continued on chromosome 22, corresponding to der(1). The second aberrant transcript extended to exon 12 of the *RERE* gene, then aligned with exon 3 of the *FHAD1* gene, followed by exons 2–7 of the long non-coding RNA *FAM230G*, corresponding to der(22). Notably, the terminal sequence of exon 12 in *RERE* and the start sequence of exon 3 in *FHAD1* are phased, suggesting the possible formation of a fusion transcript involving *RERE*, *FHAD1*, and *FAM230G* (Fig. 2E). RT-qPCR was performed to quantify the distribution of full length versus aberrant transcripts and revealed a 50:50 ratio of full length to truncated transcripts in the patient. (Fig. 2F). This finding suggests a frameshift equivalent of the CGR involving *RERE* from both der(1) and der(22).

Hi-C sequencing was then used to investigate the effect of the characterized CGR on the 3D organization of chromosomes 1 and 22. Inter-chromosomal contacts between chromosomes 1 and 22 were confirmed, consistent with the translocation hypothesis (Fig. 3A). In addition, intra-chromosomal interactions were observed on chromosome 1 between regions 1p36.23 and 1p36.21 (Fig. 3B). Given the disorganization of the 1p36 region, we investigated the potential for novel enhancer-gene interactions. While the translocation appeared to alter chromatin architecture, its effect on enhancer-promoter associations was minimal or negligible (Fig. 3C). No significant novel enhancer-promoter interactions were observed, suggesting limited regulatory disruption (data not shown).

4. Discussion

The accurate characterization of SVs, especially complex events involving copy gains, remains one of the major challenges in cytogenomics. The advent of new sequencing technologies, particularly long-read sequencing for both DNA and cDNA, has greatly improved the ability to unravel the organization and consequences of SVs [6–8,22]. In this study, the LRS approaches facilitated the characterization of two CGRs, enabling the precise identification of breakpoint locations and a clear definition of the nature of the SVs.

LRS provided a comprehensive analysis for patient 1, precisely characterizing the CNV as a tandem triplication with a breakpoint within intron 10 of the *ZMYM2* gene. This approach localized the triplication and its breakpoints at the nucleotide level, with a 3-bp microhomology at the two breakpoints (Additional Table S2, Additional Figure S1B). This microhomology, in the context of copy gains, provides valuable insight into potential mechanisms and strongly suggests the

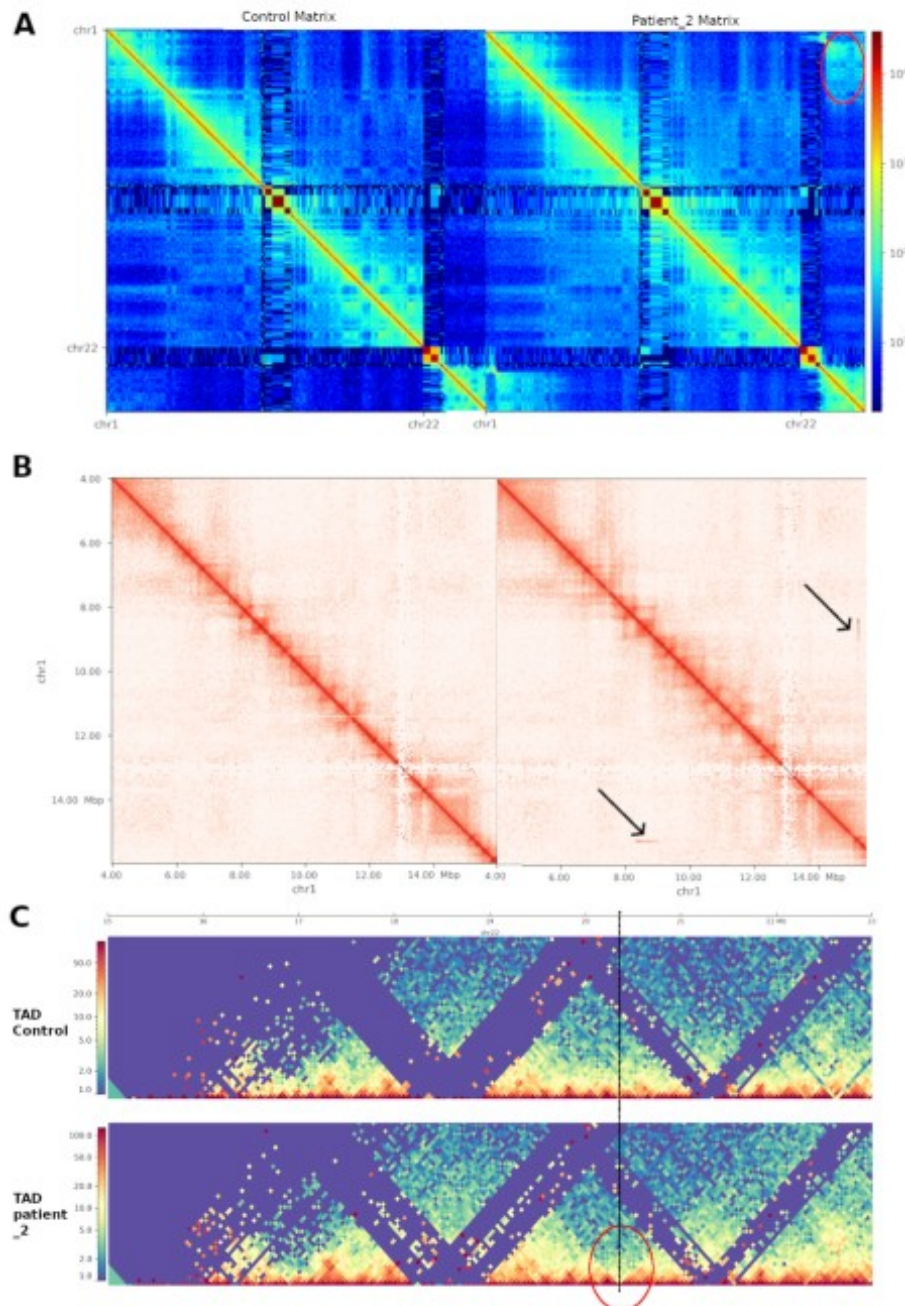


Fig. 3. 3D Genome Analysis Targeting the Structural Rearrangement of Patient 2. **(A)** Hi-C cis contact maps for chromosomes 1 and 22 comparing a wild-type (WT) control and patient 2. Red frames highlight ectopic interchromosomal contacts observed between chromosomes 1 and 22 in patient 2. **(B)** Hi-C cis contact map for chromosome 1 in WT control and patient 2. Black arrows indicate ectopic intra-chromosomal contacts within chromosome 1 in patient 2. **(C)** Topologically associated domain (TAD) map of chromosome 22 (chr22:15,000,000–23,000,000; hg38) derived from Hi-C data for WT control and patient 2. Alterations in the TAD structures in patient 2 are evident compared to the WT control.

involvement of replication-based processes such as microhomology-mediated break-induced replication (MMBIR) or fork stalling and template switching (FoSTeS). These mechanisms likely utilized the 3 bp microhomology as a repair template during the rearrangement event [2, 23,24]. Such processes can repeatedly collapse during replication, resulting in structurally more complex variants such as the observed triplication (Fig. 1D). The direct tandem triplication of *ZMYM2* is different from copy gains mediated by the non-allelic homologous recombination (NAHR) mechanism [1], such as type I triplications,

which lacks flanking duplications or inversions, or type II triplications, which typically contain inversions and flanking duplications [1,25]. The NAHR mechanism does not apply in this case due to the absence of repeat elements in the region (Additional Tables S1–S2).

In previous reports, microcephaly was documented in 4 of 14 patients, neurodevelopmental disorders in 9 of 14, hypotonia in 3 of 14, feeding difficulties in 2 of 14 and cryptorchidism in 1 patient. Patient 1, who presented with developmental delay, cryptorchidism and microcephaly, was reminiscent of this condition [10]. LRS on cDNA from

Patient 1 provided insight into the effect of the *ZMYM2* rearrangement on transcript profiles. Notably, we observed a 50:50 distribution of wild-type and truncated transcripts in the LCL sample, which is predicted to result in the production of an aberrant *ZMYM2* protein. We hypothesize that the truncated *ZMYM2* transcripts may escape nonsense-mediated decay (NMD). As suggested by Connaughton *et al.*, similar truncating pathogenic variants were identified in a family, that displayed a comparable phenotype. Using Bioluminescence Resonance Energy Transfer (BRET) assays, which measure protein-protein interactions by transferring energy from a luciferase to a fluorophore, the authors demonstrated that some truncating pathogenic variants could disrupt the *ZMYM2* interactome. If these variants allow the transcript to escape nonsense-mediated decay (NMD), they may impact DNA-binding transcription factors, co-repressors, and proteins involved in chromatin regulation and organization [10]. These alterations may have downstream effects on key factors involved in urinary tract development and may contribute to the observed urinary tract malformations in the patient described in this study [26]. In particular, truncated *ZMYM2* variants have been shown to partially interact with FOXPI, a transcription factor implicated in both neurodevelopmental and renal diseases [27]. In line with this hypothesis, these findings suggest that the truncated *ZMYM2* proteins could disrupt the native interactome by competing with wild-type proteins, thereby contributing to the patient's phenotype.

LRS of the genome of Patient 2 revealed a CGR with multiple copy number gains, interspersed inversions, and a reciprocal translocation between chromosomes 1 and 22, with breakpoints within the *RETF* and *FHADI* genes (Fig. 2C). LRS facilitated the precise identification of all copy number gains, their arrangements, and the exact locations of the breakpoints, including base-pair resolution sequencing of the breakpoints themselves [6,28,29]. Despite the challenging nature of the breakpoint locations, which often included repetitive elements, regions of microhomology were identified between interacting genomic regions (Additional Table S1–S2).

We propose a two-step mechanism underlying the rearrangement. In the first step, repair by mechanisms such as MMBIR/FOSTeS likely induced chromoanagenesis on chromosome 1, leading to the observed CNV. Microhomologies ranging from 3 to 8 base pairs may have served as repair templates. The microhomology between segments 'C' and 'F' was not directly observed, possibly due to the loss of the corresponding chimeric read, suggesting that this interaction was disrupted by subsequent repair events [2,23,24].

In a second step, two double-strand breaks (DSBs) on chromosomes 1 and 22 appear to have triggered the formation of the reciprocal translocation [30]. At the chromosome 22 breakpoint, a region containing multiple copies of a type A translocation breakpoint sequence—characterized by AT-rich, highly variable segments—was identified (Additional Figure S5A) [31]. This region tends to form cruciform secondary structures, which increase genomic instability and make the region more susceptible to DSBs (Additional Figure S5B) [32]. Similarly, the DSB on chromosome 1 may have been driven by an inversion near region 'B' which could have induced secondary structure formation, further promoting genomic instability and leading to a break [33].

Once these DSBs occurred, cellular repair mechanisms were engaged to restore genomic integrity. Based on the sequences identified at the breakpoints, non-homologous end joining (NHEJ) appears to be the most likely mechanism, facilitating the fusion of 1pTER/22pTER and 1qTER/22qTER (Fig. 2C, Additional Figure S5C). This repair process explains the reciprocal translocation and provides a plausible model for the observed complex rearrangement. This proposed two-step mechanism could have occurred during different developmental windows—premeiotic, meiotic, or postzygotic. While these stages are considered critical for the emergence of such genomic alterations, pinpointing the exact timing remains challenging.

Transcriptomic analysis and structural studies provided comprehensive insights into how the observed complex rearrangement affects

RETF gene expression. Monoallelic pathogenic variants of *RETF* are associated with a neurodevelopmental disorder characterized by brain, eye and heart abnormalities (MIM #605226) [12]. Pathogenic variants in *RETF* are predominantly loss-of-function. However, severe phenotypes have also been associated with specific missense variants within the atrophin-1 domain, suggesting a dominant-negative mechanism may contribute to more severe clinical presentations [11,12]. The cumulative data—showing one wild-type transcript alongside two aberrant truncating *RETF* transcripts (Fig. 2D–E)—strongly support haploinsufficiency as the most likely pathogenic mechanism.

RT-qPCR analysis, which revealed a 50:50 ratio of wild-type to aberrant transcripts, highlights several intriguing findings. The first aberrant transcript, comprising exons 1–10 of the *RETF* gene and aligned to chromosome 22, appears to be degraded via the NMD pathway [34]. The second aberrant transcript, spanning exons 1–12 of the *RETF* gene, exon 3 of the *FHADI* gene, and exons 2–7 of the *FAM230G* lncRNA, may have functional relevance. This fusion transcript suggests the possibility of producing a chimeric protein that may contribute to the observed phenotype. Notably, fusion transcripts involving lncRNAs, as observed in cancer studies, have been shown to generate functional proteins. This raises the possibility that the second aberrant transcript exerts a biological effect that contributes to the phenotype. The observed 50:50 ratio between the second aberrant transcript and the WT transcript further supports its potential functional relevance [35]. Consistent with the literature, patients with *RETF* variants leading to haploinsufficiency often present with milder and less specific phenotypes, aligning with the clinical features observed in Patient 2 [11,12]. Surprisingly, the complex SV involving multiple CNVs, inversions, and a translocation appears to have minimal impact on the regulatory environment and may only lead to *RETF* dysregulation. Although Hi-C data suggested a possible reorganization of topologically associated domains (Fig. 3C), no evidence of enhancer hijacking events was observed in this case [36–38].

In both reported cases, LRS significantly enhanced our understanding of CNVs and CGRs. Although both SVs occurred *de novo* in the patients, they could potentially be transmitted to future generations. The detailed characterization of these complex rearrangements is therefore essential for genetic counseling, enabling a better understanding of potential reproductive risks. This technology allowed us to identify the likely molecular diagnosis in these patients, who experienced a “diagnostic odyssey”. Nevertheless, analysis of LRS data in the context of medical molecular diagnosis remains challenging, and many bioinformatics issues need to be addressed. Current structural variant callers often fail to detect all breakpoints and to capture the full complexity of such rearrangements [39]. The two variant calling software tools used, CuteSV and Sniffles2, accurately both identified only the duplication in Patient 1 and a single inversion in Patient 2 [14,15]. Additionally, CuteSV identified an inversion between regions 'A' and 'B', and a deletion between regions 'A' and 'E'. The detection of this deletion is not surprising, as it reflects the contact observed between these two regions. However, the variant is inaccurately named: it is not a true deletion, but rather an insertion of a duplicated segment, which the caller did not take into account. This discrepancy highlights the limitations of current variant calling algorithms, especially when analyzing complex rearrangements. Factors such as structural variant complexity, sequencing quality, and read depth play a key role in variant detection and characterization. For example, low coverage or regions of high sequence complexity can hinder breakpoint resolution, leading to underrepresentation of specific rearrangements and incomplete variant profiles [39]. The development of improved algorithms that effectively integrate depth, read mapping consistency, and sequence context will be key to unlocking the full potential of long-read sequencing in genomic diagnostics. These advances could improve breakpoint detection and enable more accurate characterization of complex rearrangements. However, manual curation and complementary experimental validation remain essential components of comprehensive genomic analysis to

ensure accurate interpretation and clinical relevance of identified structural variants.

LRS has demonstrated its versatility by enabling both genome analysis and transcriptome characterization. Unlike traditional methods, it enables direct sequencing of full-length transcripts without prior PCR amplification, preserving native RNA features and minimizing amplification-induced bias [40,41]. This capability was essential for the accurate identification of truncated and fusion transcripts in both patients, providing critical insight into the functional impact of structural variants on gene expression. SRS techniques lack the ability to directly span large genomic regions, relying instead on statistical inference to piece together sequences. As a result, they often struggle to accurately resolve intricate fusion events and other complex structural variations [42]. While OGM is a powerful tool for detecting large structural variants and provides an intuitive visualization of genome-wide rearrangements, it has resolution limitations that may hinder its ability to capture finer details of complex genomic rearrangements (CGRs). Specifically, OGM has a resolution threshold of approximately 5 kb, which can result in missed detection of smaller events, as observed in our study with regions 'A' and 'F'. In a separate study comparing OGM to long-read sequencing (LRS) using Nanopore technology, OGM detected 74 % of the junctions in complex cases, whereas LRS alone identified 100 % of these junctions [7].

Our findings emphasize the importance of integrating long-read sequencing of DNA and cDNA into routine clinical diagnostics for rare diseases. These approaches provide essential insights into the characteristics of structural variants, facilitate hypotheses about their formation mechanisms and pathogenic effects, and ultimately improve the classification of VUS. By enhancing diagnostic accuracy, this comprehensive strategy supports more personalized and effective clinical management for affected individuals.

CRedit authorship contribution statement

Thuillier Caroline: Validation, Methodology, Investigation. **Yahya Emilie:** Software, Methodology, Investigation. **Meneboo Jean-Pascal:** Software, Methodology, Investigation, Formal analysis. **Fauqueux Jade:** Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis. **Thomès Luc:** Software, Methodology, Investigation. **Caumes Roseline:** Investigation, Data curation. **Figeac Martin:** Software, Methodology, Investigation. **Villenet Celine:** Formal analysis. **Smol Thomas:** Writing – review & editing, Writing – original draft, Validation, Resources, Project administration, Funding acquisition, Conceptualization. **Ghoumid Jamal:** Writing – review & editing, Writing – original draft, Validation, Resources, Project administration, Conceptualization. **Boudry Elise:** Formal analysis, Data curation.

Declaration of Competing Interest

All authors declare that they have no conflicts of interest.

Acknowledgments

We thank the patients and their families for participating in this study. This work was supported by the University Hospital of Lille, France (Grant No. BPI22-07-1531). We thank Delphine Ceraso, Alexis Leurent, Heidi Tampere, and Pauline Grave for their technical assistance with LRS, primer design, and RT-PCR experiments. We thank Julien Merlin for his technical assistance with FISH analysis.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.04.031.

Data Availability

All data sets used or analyzed in this study are available from the corresponding author upon request.

References

- [1] Weckselblatt B, Rudd MK. Human structural variation: mechanisms of chromosome rearrangements. *Trends Genet* 2015;31:587–99. <https://doi.org/10.1016/j.tig.2015.05.010>.
- [2] Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* 2016;17:224–38. <https://doi.org/10.1038/nrg.2015.25>.
- [3] Schuy J, Grochowski CM, Carvalho CMB, Lindstrand A. Complex genomic rearrangements: an underestimated cause of rare diseases. *Trends Genet TIG* 2022;38:1134–46. <https://doi.org/10.1016/j.tig.2022.06.003>.
- [4] Newman S, Hermetz KE, Weckselblatt B, Rudd MK. Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. *Am J Hum Genet* 2015;96:208–20. <https://doi.org/10.1016/j.ajhg.2014.12.017>.
- [5] Boycott KM, Rath A, Chong JX, Hartley T, Alkuray F, Baynam G, et al. International cooperation to enable the diagnosis of all rare genetic diseases. *Am J Hum Genet* 2017;100:695–705. <https://doi.org/10.1016/j.ajhg.2017.04.003>.
- [6] Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, et al. Targeted long-read sequencing identifies missing disease-causing variation. *Am J Hum Genet* 2021;108:1436–49. <https://doi.org/10.1016/j.ajhg.2021.06.006>.
- [7] De Clercq G, Vantomme L, Dewaele B, Callewaert B, Vanakker O, Janssens S, et al. Full characterization of unresolved structural variation through long-read sequencing and optical genome mapping. *Sci Rep* 2024;14:29142. <https://doi.org/10.1038/s41598-024-80068-z>.
- [8] Damaraju N, Miller AL, Miller DE. Long-read DNA and RNA sequencing to streamline clinical genetic testing and reduce barriers to comprehensive genetic testing. *J Appl Lab Med* 2024;9:138–50. <https://doi.org/10.1093/jalm/fjad107>.
- [9] Owen DJ, Aguilar-Martinez E, Ji Z, Li Y, Sharrocks AD. ZMYM2 controls human transposable element transcription through distinct co-regulatory complexes. *eLife* 2023;12:RP86669. <https://doi.org/10.7554/eLife.86669>.
- [10] Connaughton DM, Dai R, Owen DJ, Marquez J, Mann N, Graham-Paquin AI, et al. Mutations of the transcriptional corepressor ZMYM2 cause syndromic urinary tract malformations. *Am J Hum Genet* 2020;107:727–42. <https://doi.org/10.1016/j.ajhg.2020.08.013>.
- [11] Pregeau B, Kim BJ, Hernández-García A, Jordan VK, Cho MT, Schmur RE, et al. De Novo mutations of RERE cause a genetic syndrome with features that overlap those associated with proximal 1p36 deletions. *Am J Hum Genet* 2016;98:963–70. <https://doi.org/10.1016/j.ajhg.2016.03.002>.
- [12] Jordan VK, Pregeau B, Ge X, Giordano J, Wapner RJ, Balci TB, et al. Genotype-phenotype correlations in individuals with pathogenic RERE variants. *Hum Mutat* 2018;39:666–75. <https://doi.org/10.1002/humu.23400>.
- [13] Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma Oxf Engl* 2018;34:3094–100. <https://doi.org/10.1093/bioinformatics/bty191>.
- [14] Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* 2024. <https://doi.org/10.1038/s41587-023-02024-y>.
- [15] Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol* 2020;21:189. <https://doi.org/10.1186/s13059-020-02107-y>.
- [16] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM 2013. <https://doi.org/10.48550/arXiv.1303.3997>.
- [17] Ramirez F, Bhardwaj V, Arrigoni L, Lam KC, Grünig BA, Villaveces J, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun* 2018;9:189. <https://doi.org/10.1038/s41467-017-02525-w>.
- [18] Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prijs P, Sambamba: fast processing of NGS alignment formats. *Bioinforma Oxf Engl* 2015;31:2032–4. <https://doi.org/10.1093/bioinformatics/btv098>.
- [19] Riggs ER, Andersen EF, Cherry AM, Kantarci S, Kearney H, Patel A, et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet Med J Am Coll Med Genet* 2020;22:245–57. <https://doi.org/10.1038/s41436-019-0686-8>.
- [20] Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* 2023. <https://doi.org/10.1038/s41586-023-06045-0>.
- [21] 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. <https://doi.org/10.1093/nar/gkt958>.
- [22] Fujimoto A, Wong JH, Yoshii Y, Akiyama S, Tanaka A, Yagi H, et al. Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. *Genome Med* 2021;13:65. <https://doi.org/10.1186/s13073-021-00883-1>.
- [23] Bahrambeigi V, Song X, Sperle K, Beck CR, Hijazi H, Grochowski CM, et al. Distinct patterns of complex rearrangements and a mutational signature of microhomology are frequently observed in PLP1 copy number gain structural variants. *Genome Med* 2019;1:80. <https://doi.org/10.1186/s13073-019-0676-0>.

- [24] Currall BB, Talkowski Chiang C, Morton ME. OC. Mechanisms for structural variation in the human genome. *Curr Genet Med Rep* 2013;1:81–90. <https://doi.org/10.1007/s40142-013-0012-8>.
- [25] Shimajima K, Mano T, Kashiwagi M, Tanabe T, Sugawara M, Okamoto N, et al. Pelizaeus-Merzbacher disease caused by a duplication-inverted triplication-duplication in chromosomal segments including the PLP1 region. *Eur J Med Genet* 2012;55:400–3. <https://doi.org/10.1016/j.ejmg.2012.02.013>.
- [26] Estruch SB, Graham SA, Quevedo M, Vito A, Dekkers DHW, Deriziotis P, et al. Proteomic analysis of FOXF proteins reveals interactions between cortical transcription factors associated with neurodevelopmental disorders. *Hum Mol Genet* 2018. <https://doi.org/10.1093/hmg/ddy230>.
- [27] Wu S-T, Feng Y, Song R, Qi Y, Li L, Lu D, et al. Foxp1 is required for renal intercalated cell differentiation and acid-base regulation. *J Am Soc Nephrol JASN* 2024;35:533–48. <https://doi.org/10.1681/ASN.000000000000319>.
- [28] Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Presard L, et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med J Am Coll Med Genet* 2018;20:159–63. <https://doi.org/10.1038/gim.2017.86>.
- [29] Nakamichi K, Huey J, Sangermano R, Place EM, Bujakowska KM, Marra M, et al. Targeted long-read sequencing enriches disease-relevant genomic regions of interest to provide complete Mendelian disease diagnostics. *JCI Insight* 2024;9:e183902. <https://doi.org/10.1172/jci.insight.183902>.
- [30] Lieber MR, Gu J, Lu H, Shimazaki N, Tsai AG. Nonhomologous DNA end joining (NHEJ) and chromosomal translocations in humans. *Subcell Biochem* 2010;50:279–96. https://doi.org/10.1007/978-90-481-3471-7_14.
- [31] Delias N. A family of long intergenic non-coding RNA genes in human chromosomal region 22q11.2 carry a DNA translocation breakpoint/AT-rich sequence. *PLoS One* 2018;13:e0195702. <https://doi.org/10.1371/journal.pone.0195702>.
- [32] Correll-Tash S, Lilley B, Salmons H, Mlynarski E, Franconi CP, McNamara M, et al. Double strand breaks (DSBs) as indicators of genomic instability in PATRR-mediated translocations. *Hum Mol Genet* 2021;29:3872–81. <https://doi.org/10.1093/hmg/ddaa251>.
- [33] Ait Saada A, Guo W, Costa AB, Yang J, Wang J, Lobachev KS. Widely spaced and divergent inverted repeats become a potent source of chromosomal rearrangements in long single-stranded DNA regions. *Nucleic Acids Res* 2023;51:3722–34. <https://doi.org/10.1093/nar/gkad153>.
- [34] Hug N, Longman D, Cáceres JF. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res* 2016;44:1483–95. <https://doi.org/10.1093/nar/gkw010>.
- [35] Sánchez-Marín D, Silva-Cáceres MB, Porras-Reyes FI, García-Román R, Campos-Parrá AD. Breaking paradigms: Long non-coding RNAs forming gene fusions with potential implications in cancer. *Genes Dis* 2024;11:101136. <https://doi.org/10.1016/j.gendis.2023.101136>.
- [36] Wang X, Xu J, Zhang B, Hou Y, Song F, Lyu H, et al. Genome-wide detection of enhancer-hijacking events from chromatin interaction data in re-arranged genomes. *Nat Methods* 2021;18:661–8. <https://doi.org/10.1038/s41592-021-01164-w>.
- [37] Melo US, Jatzlau J, Prada-Medina CA, Flex E, Hartmann S, Ali S, et al. Enhancer hijacking at the ARHGAP36 locus is associated with connective tissue to bone transformation. *Nat Commun* 2023;14:2034. <https://doi.org/10.1038/s41467-023-37585-8>.
- [38] Ramani V, Shendure J, Duan Z. Understanding spatial genome organization: methods and insights. *Genom Proteom Bioinforma* 2016;14:7–20. <https://doi.org/10.1016/j.gpb.2016.01.002>.
- [39] Helal AA, Saad BT, Saad MT, Mosaad GS, Aboshanab KM. Benchmarking long-read aligners and SV callers for structural variation detection in Oxford nanopore sequencing data. *Sci Rep* 2024;14:6160. <https://doi.org/10.1038/s41598-024-56604-2>.
- [40] Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet* 2020;21:597–614. <https://doi.org/10.1038/s41576-020-0236-x>.
- [41] Vollmers C, Penland L, Kanbar JN, Quake SR. Novel exons and splice variants in the human antibody heavy chain identified by single cell and single molecule sequencing. *PLoS One* 2015;10:e0117050. <https://doi.org/10.1371/journal.pone.0117050>.
- [42] Nattestad M, Goodwin S, Ng K, Baslan T, Sedlazeck FJ, Rescheneder P, et al. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res* 2018;28:1126–35. <https://doi.org/10.1101/gr.231100.117>.
- [43] Nattestad M, Aboukhalil R, Chin C-S, Schatz MC. Ribbon: intuitive visualization for complex genomic variation. *Bioinforma Oxf Engl* 2021;37:413–5. <https://doi.org/10.1093/bioinformatics/btaa680>.