Identification de variants structuraux et non-codants par approche multiomique chez des patients atteints de troubles du neurodéveloppement

Kevin Riquin

COLLEGE ED BIOLOGIE DOCTORAL SANTE PAYS DE LA LOIRE

Nantes Université

THESE DE DOCTORAT

NANTES UNIVERSITE

ECOLE DOCTORALE N° 605
Biologie-Santé
Spécialité : Génétique, Génomique et Bioinformatique

Par
Kevin RIQUIN

Identification de variants structuraux et non-codants par approche multi-omique chez des patients atteints de troubles du neurodéveloppement

Thèse présentée et soutenue à NANTES, le 10/10/2023
Unité de recherche : L'institut du Thorax

Rapporteurs avant soutenance :

Caroline Schluth-Bolard	PU-PH Université de Strasbourg
Vincent Cantagrel	DR Université Paris Cité

Composition du Jury :

Président :	Vincent Cantagrel	DR Université Paris Cité
Examinateurs :	Caroline Schluth-Bolard	PU-PH Université de Strasbourg
	Pascale Saugier-Veber	MCU-PH Université de Rouen
Dir. de thèse :	Stéphane Bézieau	PU-PH Université de Nantes
Co-dir. de thèse :	Benjamin Cogné	PH Université de Nantes

Remerciements

Je remercie tout d'abord chaleureusement le Pr Caroline Schluth-Bolard, le Dr Vincent Cantagrel et le Dr Pascale Saugier-Veber d'avoir accepté de faire partie de mon jury. Merci pour votre temps et pour votre expertise.

Merci à Stéphane et Benjamin d'avoir encadré ma thèse pendant ces trois années.

Un grand merci à tous les membres de l'équipe pour leur sympathie et leur bonne humeur. Je me risque à une liste en espérant n'oublier personne. Merci Betty, Virginie, Laetitia, Loïc, Marine, Salam, Valéna, Léa, Wallid, Frédéric,Thomas, Bérénice.

Je remercie aussi ceux qui, longtemps avant cette thèse, ont contribué à me former et ont confirmé mon goût pour la recherche. Je pense en particulier à Sarah et Alexandre, qui ont accompagné mon premier vrai contact avec la recherche médicale. J'ai gardé de notre rencontre un excellent souvenir et un sens de la rigueur que j'espère avoir mis à profit dans ce travail de thèse.

Et bien sûr, parce que tout parait plus facile quand on est soutenu, merci à ma famille. Merci à eux d'avoir accepté mon manque de disponibilité parfois. Merci Marine pour ton soutien. Merci à Sarah qui m'a demandé quotidiennement des comptes sur l'avancée de ma thèse pour me maintenir motivé. Enfin, merci à Ezra qui, sans le vouloir, m'a permis de dormir moins pour travailler plus ☺

Table des matières

Liste des abréviations

ACPA : Analyse chromosomique par puce à ADN

AD : Autosomique dominant

AR : Autosomique récessif

ceRNA : Competing endogenous RNA

DI : Déficience intellectuelle

ET : Éléments transposable

FXS : Syndrome de l'X fragile

GWAS : Genome wide association studies

HERV : Retrovirus endogènes humain

HR : Recombinaison homologue

LCR : Répétitions à faible nombre de copies

LINE : Long interspersed elements

lncRNA : Long ARN non-codant

LTR : Long terminal repeats

miRNA : Micro-ARN

MMEJ : Microhomology-Mediated End-Joining

NAHR : Non-Allelic Homologous Recombination

NHEJ : Non-Homologous End-Joining

OGM : Optical Genome Mapping

ORF : Cadre de lecture (open reading frame)

RGD : Retard global de développement

SD : Duplications segmentaires

SE : Séquençage d'exome

SG : Séquençage de génome

STR : short tandem repeats

SV : Variants structuraux

SVA : SINE/variable number tandem repeat (VNTR)/Alu

TAD : Topologically associating domains

TDAH : Troubles déficit de l'attention avec ou sans hyperactivité

TND : Troubles du neurodéveloppement

TSA : Troubles du spectre de l'autisme

VNTR : Variable number of tandem repeats

XLR : Récessif lié à l'X

Liste des figures

Liste des tableaux

AVANT-PROPOS

Les troubles neurodéveloppementaux (TND) sont un ensemble de troubles hétérogènes dont certains, comme la déficience intellectuelle, ont fréquemment été associés à des causes monogéniques. Connaitre cette cause moléculaire est importante pour les familles. Elle permet, entre autres, de rompre l'isolement en constituant ou rejoignant des associations de patients, de prodiguer un conseil génétique, mais aussi, parfois, de susciter l'espoir d'une prise en charge thérapeutique.

Néanmoins, malgré les impressionnantes avancées des techniques de séquençage à haut débit, beaucoup de familles sont encore confrontées à l'impasse diagnostique. Le plafonnement à environ 50% du rendement diagnostique des TND par séquençage d'exome montre les limites d'une approche uniquement basée sur le génome codant. L'utilisation du séquençage de génome (SG) est sans aucun doute la solution d'avenir, mais s'intéresser au non-codant appelle de nouveaux challenges. Cette thèse propose d'explorer une approche intégrative du SG en le complétant d'une approche transcriptomique et de confirmations par optical genome mapping (OGM) puis d'exposer ses avantages pour la détection et l'interprétation des variants.

Dans l'introduction nous décrirons les principaux TND et les causes génétiques puis nous nous pencherons sur les techniques de diagnostic moléculaire que nous avons intégrées à notre approche. Enfin, nous nous intéresserons aux variants structuraux et aux régions non-codantes régulatrices du génome qui sont à portée d'étude grâce au SG.

Nous verrons ensuite les résultats de l'étude sur notre approche intégrative appliquée à une cohorte de 33 patients et une présentation des variants structuraux et non-codants identifiés comme cause de TND.

INTRODUCTION

Chapitre 1. Généralités sur les troubles du neurodéveloppement (TND)

1.1. Définitions

Les TND sont un ensemble de troubles dont la définition actuelle a été proposée en 2013 dans le DSM-5. Ils y sont définis comme « Un groupe d'affections apparaissant au cours de la période de développement et provoquant des déficits qui entraînent des altérations du fonctionnement ». Plus récemment, l'International Statistical Classification of Diseases and Related Health Problems (ICD) les a définis comme : « Des troubles comportementaux et cognitifs qui surviennent au cours de la période de développement et qui impliquent des difficultés significatives dans l'acquisition et l'exécution de fonctions intellectuelles, motrices, langagières ou sociales spécifiques ». Ils regroupent donc des troubles suivants, se manifestant avant l'âge de 18 ans, et dont les caractéristiques sont, sans équivoque, neurodéveloppementales :

- Les troubles déficit de l'attention avec ou sans hyperactivité (TDAH)
- La déficience intellectuelle (DI)
- Les troubles du spectre de l'autisme (TSA)
- Les troubles moteurs neurodéveloppementaux (trouble de la coordination, mouvements stéréotypés, tics)
- Les troubles spécifiques de l'apprentissage (lecture, expression écrite et déficit du calcul)
- Les troubles de la communication

Il existe un continuum entre ces différentes catégories. En effet, des études ont montré que 22% à 83% des enfants atteints de TSA présentent des symptômes qui répondent aux critères pour le TDAH, et inversement, 30% à 65% des enfants atteints de TDAH présentent des symptômes cliniquement significatifs de TSA [1]. Par ailleurs, près de 41% des personnes atteintes de DI présentent également un TSA, tandis que près de 71% des personnes atteintes de TSA présentent également une DI [2]. Des comorbidités neurologiques sont également signalées, l'épilepsie est 2,3 à 3 fois plus fréquente chez les enfants souffrants de TDAH [3,4]. Chez les patients présentant une DI, l'épilepsie est retrouvée avec une prévalence d'environ 22% mais elle peut-être plus élevée pour les troubles sévères [5]. La prévalence globale des TND est difficile à déterminer. Les rares études publiées n'incluent pas systématiquement toutes les catégories de troubles et leurs terminologies ayant évoluées, la définition n'est pas toujours homogène entre les études. Une prévalence d'environ 7% a néanmoins été rapportée en 2017 aux États-Unis pour les TND à l'exception des TDAH et troubles de l'apprentissage [6]. Plus spécifiquement, entre 2019 et 2020, chez les enfants de 3 à 17 ans aux États-Unis, les prévalences étaient d'environ 9% pour les TDAH, 2,5% pour les TSA et 1,4% pour la DI [7,8].

L'étiologie présumée des TND est complexe et, dans de nombreux cas, inconnue. On suppose tout de même qu'ils sont principalement dus à des facteurs génétiques ou environnementaux qui sont présents dès la naissance. Toutefois, le manque de stimulation appropriée ou de possibilités d'apprentissage adéquates peut également être un facteur contribuant aux TND et doit être systématiquement pris en compte dans leur évaluation. Certains TND peuvent également résulter d'une blessure, d'une maladie ou

d'une autre atteinte du système nerveux central, lorsqu'elle survient au cours de la période de développement (http://id.who.int/icd/entity/1516623224).

1.2. Déficience intellectuelle

Selon le DSM-5, la déficience intellectuelle est caractérisée par trois critères devant être retrouvés conjointement :

- Déficit des fonctions intellectuelles comme le raisonnement, la résolution de problèmes, la planification, l'abstraction, le jugement, l'apprentissage scolaire et l'apprentissage par l'expérience.

- Déficit des fonctions adaptatives qui se traduit par un échec dans l'accession aux normes habituelles de développement socioculturel permettant l'autonomie et la responsabilité sociale. Ces déficits limitent le fonctionnement dans un ou plusieurs champs d'activité de la vie quotidienne comme la communication, la participation sociale ou l'indépendance.

- Ces troubles débutent pendant la période développementale.

La DI est classée en plusieurs niveaux de sévérités : Légère, moyenne, grave et profonde. Elle peut être isolée ou syndromique, c'est-à-dire, associée à d'autres caractéristiques cliniques ou comorbidités. Elle est 1,5 fois plus fréquente chez les hommes que chez les femmes [9]. Le déficit des fonctions intellectuelles doit être confirmé par l'évaluation clinique et les tests d'intelligence individuels standardisés comme l'évaluation du quotient

intellectuel (QI). Il suit une distribution normale de moyenne 100 et d'écart-type 15. On considère qu'un individu ayant un QI environ 2 écarts-types en dessous de la moyenne, c'est-à-dire, un score inférieur à 70, présente un déficit (Figure 1).

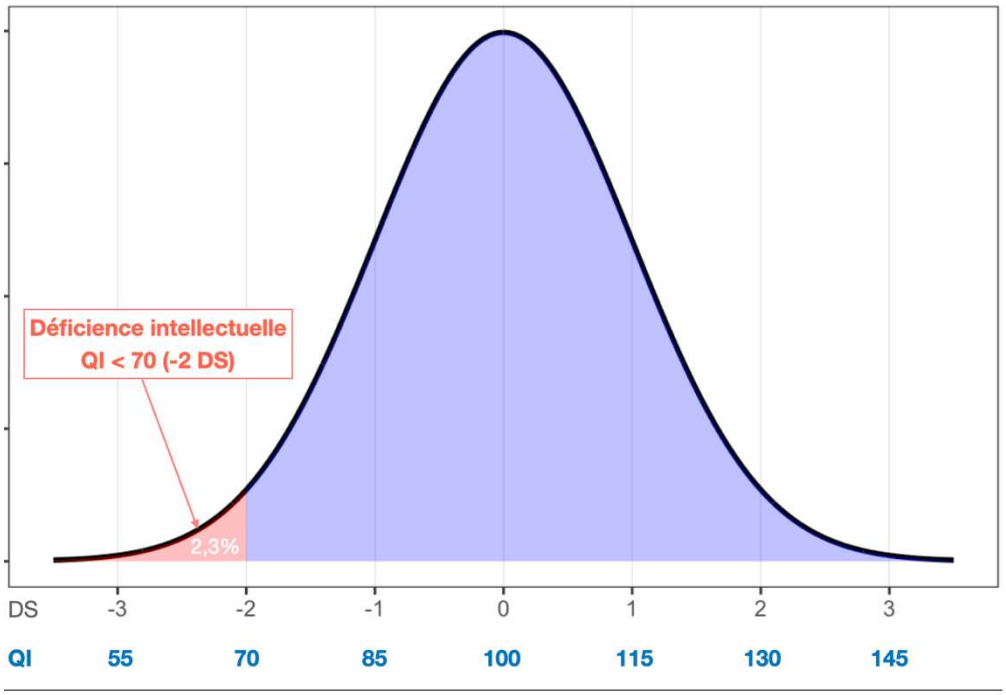


Figure 1 : Distribution du QI dans la population.

Le QI n'est qu'un estimateur approximatif de la DI et il ne permet pas d'évaluer précisément le déficit des fonctions adaptatives. Le sujet peut présenter des déficits du comportement adaptatif tellement sévères qu'ils dégradent de façon importante son fonctionnement quotidien, et ce, malgré un QI supérieur à 70. De fait, le DSM-5 recommande que les différents niveaux de sévérité ne soient plus définis sur la base du QI mais sur l'évaluation du fonctionnement adaptatif avec des tests comme le Vineland (VABS). De plus, il faut noter que les mesures du QI perdent de leur relevance pour les notes les plus basses. Toutefois, en pratique, tous les services n'ont pas nécessairement les ressources pour mener ces tests et il est courant que le diagnostic de DI soit posé par les généticiens cliniques, sur la seule base des fonctions intellectuelles et adaptatives

observées lors de la consultation. Enfin, lorsqu'on constate un retard dans les acquisitions chez l'enfant de moins de 5 ans ne pouvant subir les tests standardisés, on posera plutôt le diagnostic de retard global du développement (RGD).

1.3. Troubles du spectre de l'autisme (TSA)

Selon le DSM-5, les caractéristiques essentielles du trouble du spectre de l'autisme sont « des déficits persistants de la communication sociale réciproque et des interactions sociales ainsi qu'un mode restreint et répétitif des comportements, des intérêts et des activités. Ces symptômes sont présents depuis la petite enfance et limitent ou retentissent sur le fonctionnement de la vie quotidienne ». L'héritabilité estimée des TSA varie entre 37 % à plus de 90 % [10]. Actuellement, près de 15 % des TSA sont associés à une cause génétique. Cependant, même lorsqu'un TSA isolé est associé à un variant pathogène connu, il semble que sa pénétrance soit incomplète. Outre les principaux symptômes, environ 31 % des personnes atteintes de TSA présentent également une DI [11]. Cette comorbidité est par exemple observée pour les TND impliquant les gènes *CHD8* (TSA : ~95% ; DI : ~55%), *SHANK3* (TSA : ~55% ; DI : ~45%), *SCN2A* (TSA : ~65% ; DI : ~65%), *GRIN2B* (TSA : ~70% ; DI : ~70%), *SETD5* (TSA : ~40% ; DI : ~90%), *FOXP1* (TSA : ~40% ; DI : ~80%) [12] (Figure 2).

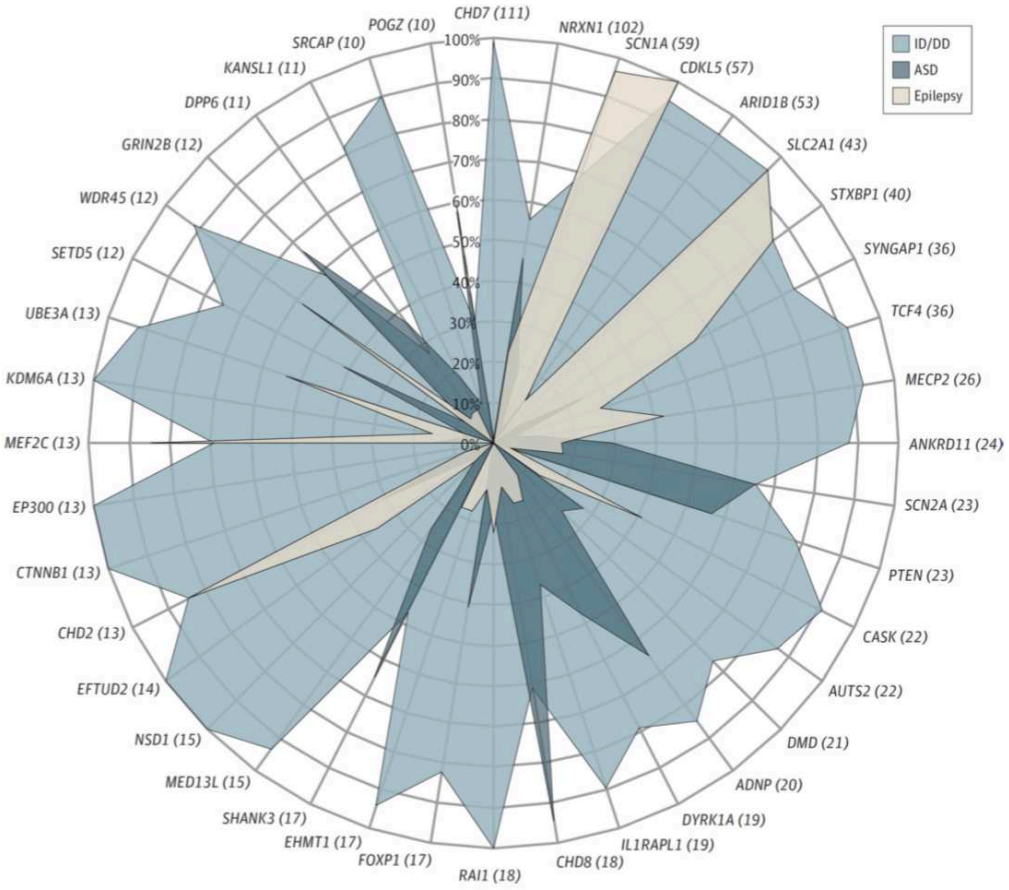


Figure 2 : Fréquence des DI/RGD, TSA et de l'épilepsie chez des patients avec variant perte de fonction. Tiré de Gonzalez-Mantilla et al. 2016.

1.4. Trouble déficit de l'attention avec ou sans hyperactivité (TDAH)

Le DSM-5 défini les TDAH comme : « Un mode persistant d'inattention et/ou d'hyperactivité-impulsivité qui interfère avec le fonctionnement ou le développement ». L'inattention et l'hyperactivité-impulsivité sont chacun composés de 9 symptômes dont au moins 6 sont nécessaires pour poser un diagnostic clinique (au moins 5 chez l'adulte et l'adolescent de 17 ans et plus). Ces symptômes ne doivent pas pouvoir être expliqués par une autre maladie mentale ou psychotique. Il s'agit donc essentiellement d'un diagnostic d'exclusion. La prévalence, variable selon le sexe est estimée à 5,8% environ

chez les enfants et 2,5% chez les adultes [13,14]. Des disparités géographiques sont observées sans qu'il soit possible de conclure si elles sont dues à des différences diagnostiques, méthodologiques ou culturelles (Figure 3).

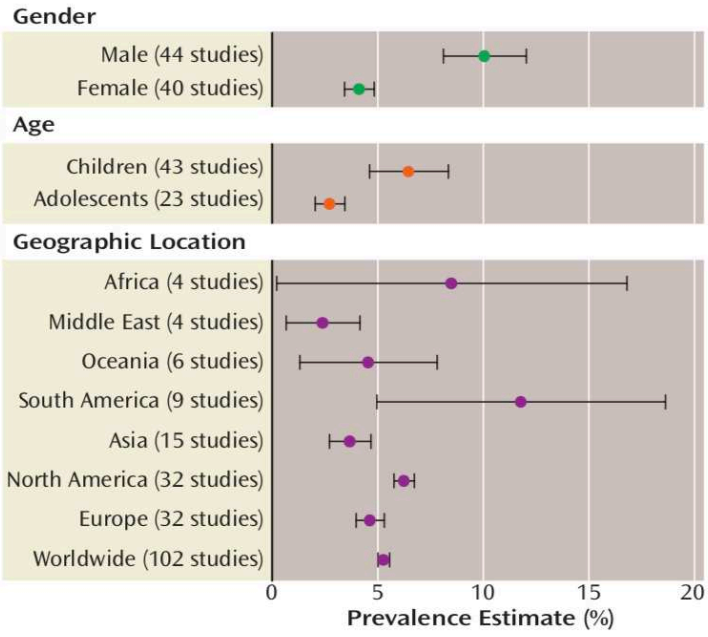


Figure 3 : Prévalence globale des TDAH selon les caractéristiques démographiques et géographiques. Tiré de Polanczyk *et al*, 2007

Les TDAH entrainent des conséquences scolaires, professionnelles et sociales pour les patients. Un moindre niveau scolaire chez les enfants et une plus forte probabilité de chômage chez les adultes sont rapportées [15,16].

L'héritabilité élevée, estimée entre 77–88%, suggère là aussi, une composante génétique. Toutefois, des analyses GWAS ont montré qu'une grande partie de l'héritabilité du TDAH est polygénique et due à de nombreux variants à effets faibles. Dans ces études, l'héritabilité due aux SNP était d'approximativement 22% seulement [17]. Les formes monogéniques sont rares et n'apparaissent généralement pas de façon isolée.

Néanmoins, le TDAH est fréquemment présent dans plusieurs maladies rares monogéniques comme le syndrome de Williams ou le syndrome de l'X fragile.

1.5. Étiologie des TND

Le développement du cerveau humain est un processus qui commence à la troisième semaine de gestation avec la différenciation des cellules progénitrices neurales et qui se poursuit au moins jusqu'à la fin de l'adolescence, voire tout au long de la vie. Ce processus étroitement régulé dépend de plusieurs étapes comme la prolifération des précurseurs neuronaux, la migration, la différenciation, la synaptogénèse, la myélinisation et l'élagage des synapses [18]. Le dysfonctionnement d'une ou plusieurs de ces étapes peut entraîner des troubles du neurodéveloppement plus ou moins sévères. Nous allons voir maintenant les principales étiologies génétiques ou environnementales des TND.

1.5.1. Les causes non-génétiques

Parmi les causes non génétiques de TND, nous retrouvons les expositions *in utero* à des toxiques comme les agents tératogènes ou l'alcool. Ce dernier est d'ailleurs la principale cause non génétique à l'échelle mondiale. La prévalence du syndrome d'alcoolisation fœtale (SAF) est variable d'une région du globe à une autre mais reste relativement élevée en Europe [19]. Les infections maternelles pendant la grossesse peuvent également porter gravement atteinte au neurodéveloppement du fœtus avec ou sans passage de la barrière placentaire. Des parasites comme *Toxoplasma gondii* ou de nombreux virus peuvent avoir un effet direct sur le cerveau fœtal et déclencher des réponses pro-inflammatoires, avec des conséquences potentiellement graves. Les agents pathogènes

viraux comprennent notamment le virus de la rubéole, le virus de l'herpès simplex et le parvovirus B19, le virus Zika (ZIKV) et le cytomégalovirus (CMV) [7]. Parmi les autres causes pré- ou périnatales augmentant le risque de TND, nous pouvons citer la prématurité, la gémellité (surtout monochoriale), les anoxies périnatales, les accidents neurologiques vasculaires ou traumatiques et certaines pathologies maternelles (par exemple la dysthyroïdie) [20].

1.5.2. Les causes génétiques

Les anomalies chromosomiques de nombre

Les anomalies chromosomiques de nombre (aneuploïdie) sont une cause importante de DI. Plus précisément, c'est la trisomie 21 (T21) ou syndrome de Down [21] qui en est la première avec une incidence d'environ 1 sur 1000 naissances [22]. Cette prévalence atteint même 1% lorsque l'âge maternel est supérieur à 40 ans. Seules deux autres trisomies complètes sont viables : La trisomie 13 (syndrome de Patau) et la trisomie 18 (syndrome d'Edwards). Le phénotype de ces dernières est beaucoup plus sévère et la mortalité est élevée puisque plus de 90% des enfants atteints décèdent dans leur première année. Toutefois, avec une prévalence inférieure à 1/6000, elles sont beaucoup moins fréquentes que la T21 [23,24]. On rapporte également de très rares cas de trisomies en mosaïque comme celles du chromosome 8 (syndrome de Warkany) ou chromosome 9 et qui sont, elles aussi, impliquées dans une forme de DI syndromique.

Les variants structuraux (SV)

Les SV tels que les variations du nombre de copies (CNV), les translocations ou les inversions, qu'ils soient à l'échelle chromosomique ou sub-chromosomique sont une cause importante de TND. Ils seront spécifiquement détaillés au chapitre 3.

Le syndrome de l'X fragile (FXS)

Le FXS (MIM : 300624), également connu sous le nom de syndrome de Martin-Bell, a été décrit pour la première fois en 1943 par Martin et Bell comme une forme de DI héréditaire liée à l'X [25]. La cause moléculaire a été identifiée en 1991 en 5'UTR du gène *FMR1* (Xq27.3) [26]. Il s'agit de la deuxième cause de DI après la trisomie 21 (2,4 % de toutes les DI), la première cause de DI héréditaire et la cause la plus répandue de DI chez les hommes. Plus de 99 % des personnes atteintes du FXS présentent une perte de fonction de *FMR1* causée par un nombre accru de triplets de nucléotides CGG dans la région 5′ UTR. La maladie se déclare chez des individus porteurs d'une répétition de 200 triplets ou plus [27]. Les répétitions de 55 à 199 triplets, appelées pré-mutations, sont instables et le nombre de répétition tend à augmenter à chaque génération jusqu'à atteindre le stade de mutation complète [28]. Au niveau de la clinique, le phénotype est variable et généralement moins sévère chez la femme tandis que chez l'homme, la DI est presque constante et on retrouve dans 30-50% des cas un TSA, et dans 12-23% des cas un TDAH. Le diagnostic du FXS par PCR de la région répétée et le statut de méthylation de l'allèle sont réalisés en première intention chez les enfants présentant une DI ou un RGD.

Les principaux gènes impliqués

Selon la base de données SysNDD [29], plus de 1500 gènes sont connus pour être impliqués dans des TND (associations confirmées au 30/06/2023). Toutefois, chacun de

ces gènes n'est responsable que d'une faible part des cas. La consultation des résultats d'une importante cohorte d'individus atteints de TND sévères, le DDD (Deciphering Developmental Disorders), nous permet toutefois de voir que certains gènes mutés sont surreprésentés. Ces gènes sont : *ARID1B* (Syndrome de Coffin-Siris), *SCN2A, ANKRD11* (Syndrome KBG), *SATB2* (Glass syndrome), *SYNGAP1, DYRK1A, MED13L, STXBP1, CTNNB1, KCNQ2, KMT2A* (Syndrome de Wiedemann-Steiner), *FOXP1, PACS1* (Syndrome de Schuurs-Hoeijmakers), *SMARCA2* (Syndrome de Nicolaides-Baraitser) et *WDR45* [30] (Figure 4).

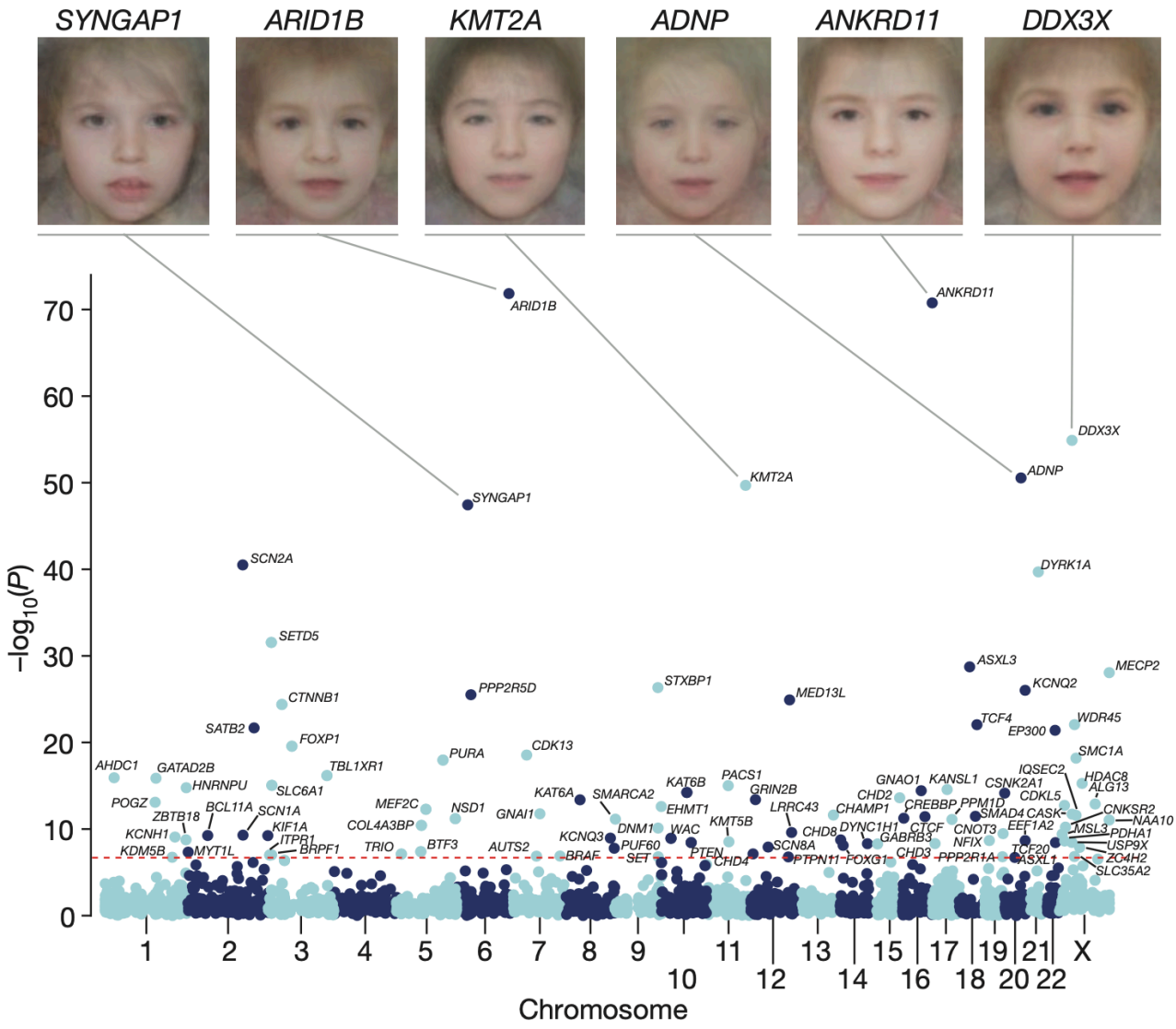


Figure 4 : Gènes surreprésentés dans la cohorte DDD. Tiré de Deciphering Developmental Disorders Study et al., 2017.

L'analyse basée sur l'ontologie des gènes associés aux TND de la base SysNDD montre leur grande hétérogénéité ainsi que la diversité des voies biologiques impliquées. Nous pouvons tout de même constater que certaines d'entre elles, comme celles impliquées dans le métabolisme, les transports, le développement du système nerveux et les mécanismes transcriptionnels sont très représentées [29]. Une revue détaillée de toutes les voies dépasserait le cadre de cette introduction, aussi, nous nous limiterons donc aux plus représentées et à celles concernant les gènes les plus fréquemment en cause dans la DI.

Troubles du métabolisme

Plusieurs troubles métaboliques comptent la déficience intellectuelle parmi leurs symptômes. La cause de cette DI, rarement isolée, peut trouver son origine dans trois mécanismes pathogènes survenant au cours d'étapes critiques du neurodéveloppement : une accumulation d'un composé toxique, l'absence d'un substrat indispensable ou un déficit énergétique [31]. Les voies concernées sont, entre autres, le métabolisme des acides aminés, des acides gras, des lipides, les désordres mitochondriaux ou du transport peroxysomal. Un exemple de trouble du métabolisme des acides aminés est la phénylcétonurie (PCU), une condition liée à un déficit en phénylalanine hydroxylase (PAH). La PAH permet normalement la transformation de la phénylalanine en tyrosine. Son déficit entraîne une accumulation de phénylalanine dans les organes, dont le cerveau, pour lequel elle est toxique. Du coté des maladies peroxysomales, nous pouvons évoquer le syndrome de Zellweger, une maladie à transmission autosomique récessive causée par des variants dans un des 12 gènes de transport des molécules dans le peroxysome.

Cohésinopathies et chromatinopathies

Parfois, la dérégulation d'une voie s'accompagne d'un phénotype évocateur, comme pour le syndrome de Cornélia de Lange (CdL) (MIM : 122470, 300590, 300882, 610759 and 614701). Cette pathologie due à une anomalie dans la formation du complexe cohésine (« **cohésinopathie** ») est causée dans sa forme classique par des variants dans les gènes *NIPBL*, *SMC1A*, *SMC3*, *RAD21* et *HDAC8* codant des sous-unités ou des régulateurs du complexe. De façon intéressante, le séquençage de patients présentant un phénotype de type CdL a permis de découvrir des variants dans des gènes codant des facteurs clés associés à la chromatine mais qui n'avaient jusqu'alors aucun lien connu avec le complexe cohésine et qui sont, pour certains, déjà associés à des pathologies de présentations différentes. Ces mutations affectent par exemple des gènes tels que *KMT2A*, *EP300*, *AFF4*, *ANKRD11*, *SETD5* ou *BRD4* [32] (Figure 5).

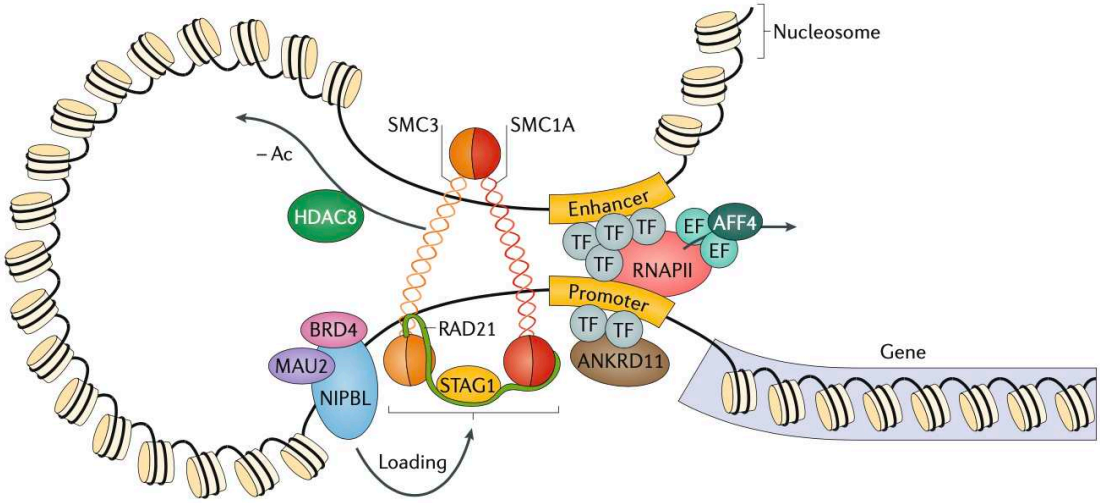


Figure 5 : Le complexe cohésine et les gènes impliqués. Tiré de Kline *et al*, 2018

Cette découverte a conduit certains auteurs à inscrire les cohésinopathies dans l'ensemble plus large, les « **chromatinopathies** ». Il s'agit d'un ensemble diversifié de maladies liées à des gènes jouant un rôle crucial dans la modification des marques

épigénétiques et, par conséquent, dans la régulation de l'expression génique. Parmi ces affections, on retrouve des syndromes impliquant des lysines méthyltransférase ou déméthylases tels que le syndrome de Kabuki (*KMT2D* et *KDM6A*) [33] et le syndrome de Wiedemann-Steiner (*KMT2A*) [34]. On retrouve également des gènes de la famille SWI/SNF impliqués dans le remodelage ATP-dépendant de la chromatine comme ceux impliquées dans le syndrome de Coffin-Siris (*ARID1A*, *ARID1B*, *ARID2*, *SMARCA4*, *SMARCC2*, *SMARCB1*, *SMARCE1*).

Anomalies de la neurogénèse

La neurogenèse repose sur les cellules progénitrices qui se divisent dans la zone sous-ventriculaire [35]. Les défauts d'amplification ou d'apoptose de ces progéniteurs entraînent des altérations de la taille du cerveau, et sont donc, souvent, associés à une microcéphalie primaire, au retard de développement et à l'épilepsie [36]. Plusieurs gènes ont été associés à une microcéphalie primaire dont : *MCPH1*, *ASPM*, *CDK5RAP2*, *CENPJ*, *STIL*, *WDR62*, *CEP152* et *CEP63*. Il est à noter que tous ces gènes associés à la microcéphalie codent des protéines en lien avec le centrosome. Outre leur rôle dans la prolifération cellulaire et donc l'amplification des cellules progénitrices neurales, les centrosome sont également impliqués dans la coordination des microtubules et dans d'autres processus cellulaires tels que la migration [37]. Par ailleurs, des gènes impliqués dans la prolifération cellulaire peuvent également être responsables de phénotypes de macrocéphalie comme le gène *PTEN*, associé à la maladie de Lhermitte-Duclos (MIM : 158350).

Anomalies de la migration, différentiation et maturation des neurones

Comme nous venons de l'évoquer, anomalies de la neurogénèse et anomalies de la migration peuvent générer des phénotypes comparables et mettre en cause les mêmes gènes. De fait, malgré la diversité des étiologies, les déficits structurels du cerveau, tels que la microcéphalie ou la lissencéphalie sont aussi souvent attribuables à des variants dans des gènes contrôlant l'architecture du cytosquelette. En effet, la migration des précurseurs neuronaux permet, elle aussi, l'expansion du cerveau. Une migration anormale des précurseurs contribue, là encore, en grande partie aux phénotypes de lissencéphalie. Le principal gène responsable de la lissencéphalie, *PAFAH1B1* (LIS1), est un régulateur essentiel de la dynéine cytoplasmique, moteur des microtubules, et effecteur clé du transport neuronal au cours du développement cérébral. Parmi les autres gènes en cause, nous pouvons également citer *TUBA1A* codant une protéine constituant les microtubules [38].

Anomalies de la connectivité neuronale

La perturbation de l'assemblage et de la connectivité des circuits neuronaux impliquant des gènes régulant le guidage des axones et la fonction synaptique est en cause dans beaucoup de TND. Entre autres exemples, des variants dans les gènes *SHANK3*, *MECP2*, *FMR1*, *CACNA1C*, et *UBE3A* ont montré qu'une partie de leur pathogénicité était due à un mécanisme de « **synaptopathie** » [39].

Chapitre 2. Le diagnostic moléculaire des TND

Comme nous l'avons vu, identifier la cause d'un TND et poser un diagnostic moléculaire peut présenter certaines difficultés. Le nombre croissant de gènes en cause et la grande

hétérogénéité phénotypique compliquent la tâche des biologistes. Au cours des années, le diagnostic des TND a connu plusieurs avancées techniques avec des rendements croissants. Toutefois, il est toujours délicat de comparer ces rendements entre les différentes études. En effet, leurs méthodologies, notamment en ce qui concerne les critères d'inclusion des patients, les analyses précédemment effectuées, les critères de classification des variants peuvent fortement varier (Figure 6).

	Cohort characteristics	Additional information
a)	Down syndrome-like phenotype	-
b)	Developmental disorder	-
c)	Fragile X-like phenotype male patients	-
d)	Non-syndromic intellectual disability	Karyotype
e)	Syndromic intellectual disability	Karyotype
f)	Non-syndromic intellectual disability	FISH
g)	Syndromic intellectual disability	FISH
h)	X-linked intellectual disability	-
i)	Non-syndromic intellectual disability	-
j)	Consanguineous families	-
k)	Syndromic intellectual disability	-
l)	Intellectual disability	-
m)	Consanguineous families	CMA + TNGS
n)	Sporadic intellectual disability	-
o)	Intellectual disability and/or developmental disorder	-
p)	Non-syndromic neurodevelopmental disorder	-
q)	Syndromic neurodevelopmental disorder	-
r)	Severe intellectual disability	Negative CMA and ES result
s)	Severe intellectual disability	Cumulative diagnostic
t)	Intellectual disability	Low-coverage GS + medical ES

Figure 6 : Rendement diagnostique de plusieurs techniques appliquées à la DI. Tiré Maia *et al*. 2012.

Les méthodes cytogénétiques conventionnelles (caryotype et l'hybridation in situ en fluorescence) permettent d'identifier les anomalies chromosomiques de structure et les aneuploïdies, qui sont responsables d'environ 15 % des cas de DI. Les recherches d'expansion de triplet pour l'X fragile et les techniques par puces (ACPA et SNP array) ont

permis d'améliorer le diagnostic mais leur rendement ne dépasse pas les 20% [40]. L'arrivée du séquençage à haut débit (SHD), tout d'abord selon une approche ciblée sur un panel de gènes puis sur l'exome et enfin le génome a considérablement changé le diagnostic des TND. Dans ce chapitre, nous aurons un aperçu du SHD, de son histoire, et de son utilité en diagnostic des TND. Nous verrons dans un second temps l'optical génome mapping (OGM) et le RNA-Seq clinique qui sont également des techniques applicables seules ou en complément du séquençage de génome (SG).

2.1. Le séquençage à haut débit (SHD)

2.1.1. Histoire du séquençage

Au moment où j'écris ces lignes, 70 ans se sont écoulés depuis la découverte de la structure de l'ADN par Watson et Crick sur la base des travaux de Rosalind Franklin et Maurice Wilkins. Pendant ces sept décennies, notre connaissance du génome humain a été l'objet d'une évolution impressionnante. Le plus notable dans l'histoire du séquençage est sans doute la fulgurante accélération des développements techniques durant les 20 dernières années. Ces nouveaux outils et méthodes ont permis de générer un important volume de données qui nourrit aujourd'hui la génétique fondamentale et la médecine. Nous pouvons dire que peu de domaines des sciences biologiques peuvent se réjouir d'avoir connu une telle dynamique (Figure 7).

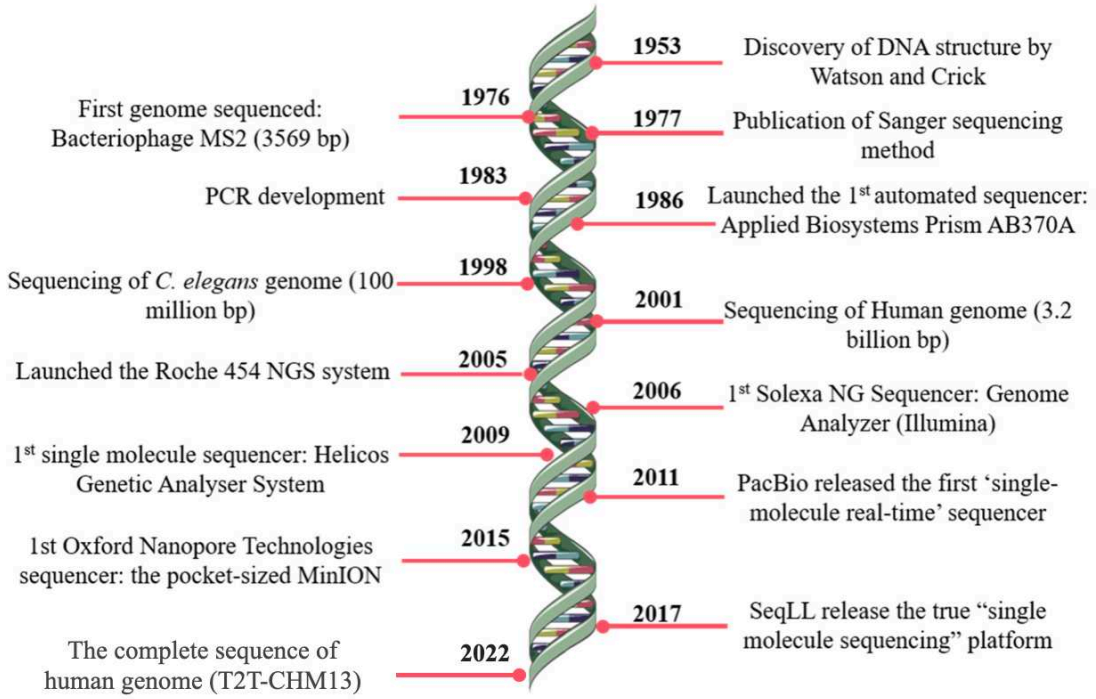


Figure 7 : Frise chronologique du séquençage d'ADN. Adapté de Pereira *et al*. 2020.

Pourtant, générer les premières séquences était un processus long et fastidieux. En 1965, Robert Holley et ses collègues ont séquencé le tout premier ARN de transfert (Alanine tRNA) de *Saccharomyces cerevisiae*. Déterminer ces 77 nucléotides a été un travail colossal qui a employé 5 personnes pendant 3 ans (soit un débit d'environ 2 bases par mois) et a nécessité comme matière première 140 kg de levure [41]. Toutefois, il faudra encore attendre 11 ans pour que le premier génome d'un organisme soit séquencé en totalité, le petit génome ARN (seulement 3569 pb) du bactériophage MS2 publié en 1976 [42].

En 1977 Sanger *et al.* ont proposé la première méthode de séquençage par synthèse enzymatique. Cette méthode repose sur l'ajout de désoxynucleotides (dNTP) en excès et de didésoxynucléotides (ddNTP) radio-marqués en quantité limitée. L'incorporation de ces ddNTP entrainent un arrêt de l'élongation et donc un marquage du dernier nucléotide

du fragment. Les fragments sont ensuite séparés par taille et analysés par électrophorèse sur gel pour déterminer la séquence. Dans ses débuts, cette technique ne permettait de séquencer que des ADN monocaténaires comme le génome du bactériophage ΦX174 [43]. Cette méthode a continué de s'améliorer avec la possibilité de séquencer des génomes bicaténaires puis avec l'introduction du marquage fluorescent et de l'électrophorèse capillaire. Elle a, pendant de nombreuses années, été largement répandue en tant que "séquençage de première génération" et est toujours utilisée aujourd'hui pour les courtes séquences et les confirmations de génotypes. Cette avancée technologique majeure a permis de progressivement séquencer des génomes de plus en plus longs jusqu'à proposer une première ébauche de génome humain en 2001. En 2003, une première version complète est annoncée et clos le « Human genome project » initié en 1988. Ce génome entier a mobilisé des centaines de chercheurs partout dans le monde pendant 13 ans. Le coût total de ce projet est estimé à 3 milliards de dollars [44] ce qui rendait, en l'état, l'idée d'un séquençage de génome humain à grande échelle totalement hors d'atteinte. En 2005, l'arrivée des techniques dites à « haut débit » ou de « seconde génération » ont nourri l'espoir d'un génome à moins de 1000 dollars. La chute des coûts a en effet été spectaculaire dans les 5 années qui ont suivi (Figure 8). Cet objectif, qui pouvait sembler ambitieux à l'époque, est aujourd'hui atteint et ouvre maintenant la voie à la médecine personnalisée et un diagnostic par SG en routine.

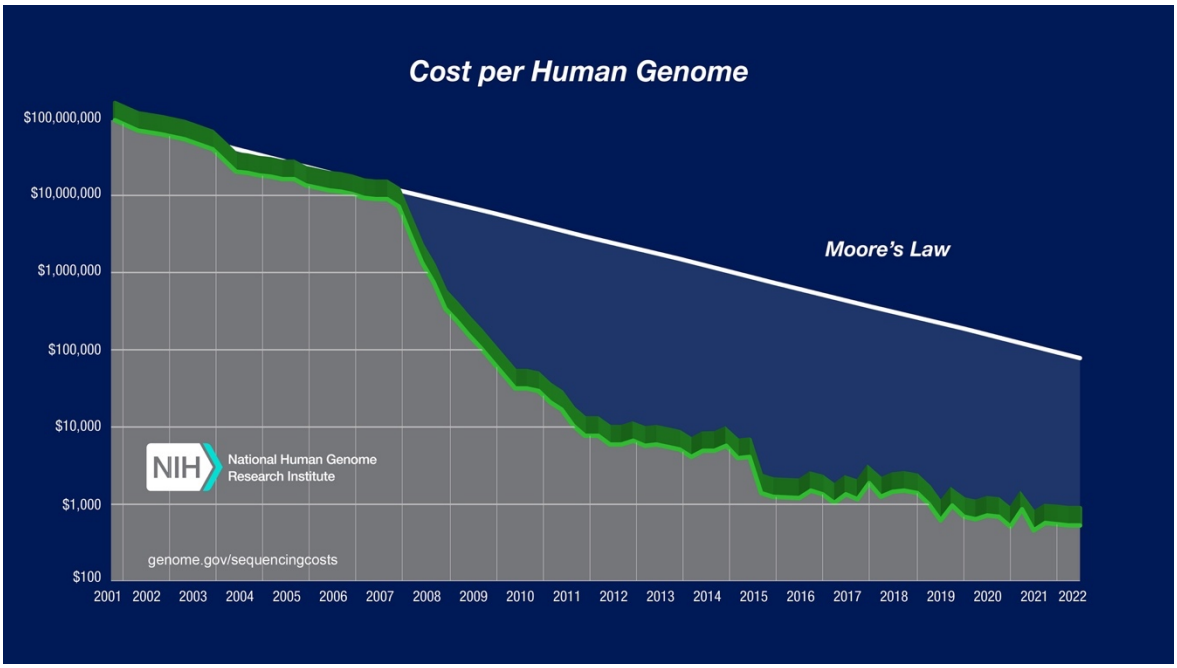


Figure 8 : Evolution du coût de séquençage d'un génome humain. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: **www.genome.gov/sequencingcostsdata**. Accessed 07/07/2023.

2.1.2. Le séquençage d'exome et génome

La chute du coût du SHD a permis de le mettre progressivement en application en diagnostic. Ainsi, le nombre de gènes connus pour être impliqués dans la DI a connu une augmentation soutenue après 2008, tout particulièrement pour les maladies à transmissions autosomiques dominantes (Figure 9). Cela est, en partie, dû au fait que les approches familiales principalement utilisées jusque-là n'étaient pas adaptées à l'identification de ces variants *de novo* à effet fort. En effet, la transmission d'un variant pathogène d'un parent à ses enfants est considéré comme très improbable (toutefois, des cas d'enfants sévèrement atteints ayant hérité du variant d'un parent *pauci* symptomatique sont parfois observés). Grâce à la mise en œuvre d'approches SHD en trio le rôle des formes autosomiques dominantes a été fortement reconsidéré. En effet,

dans les populations Occidentales chez qui la consanguinité est rare, la DI est principalement sporadique et, au global, les mutations *de novo* représenteraient jusqu'à 42 % des DI sévères [30].

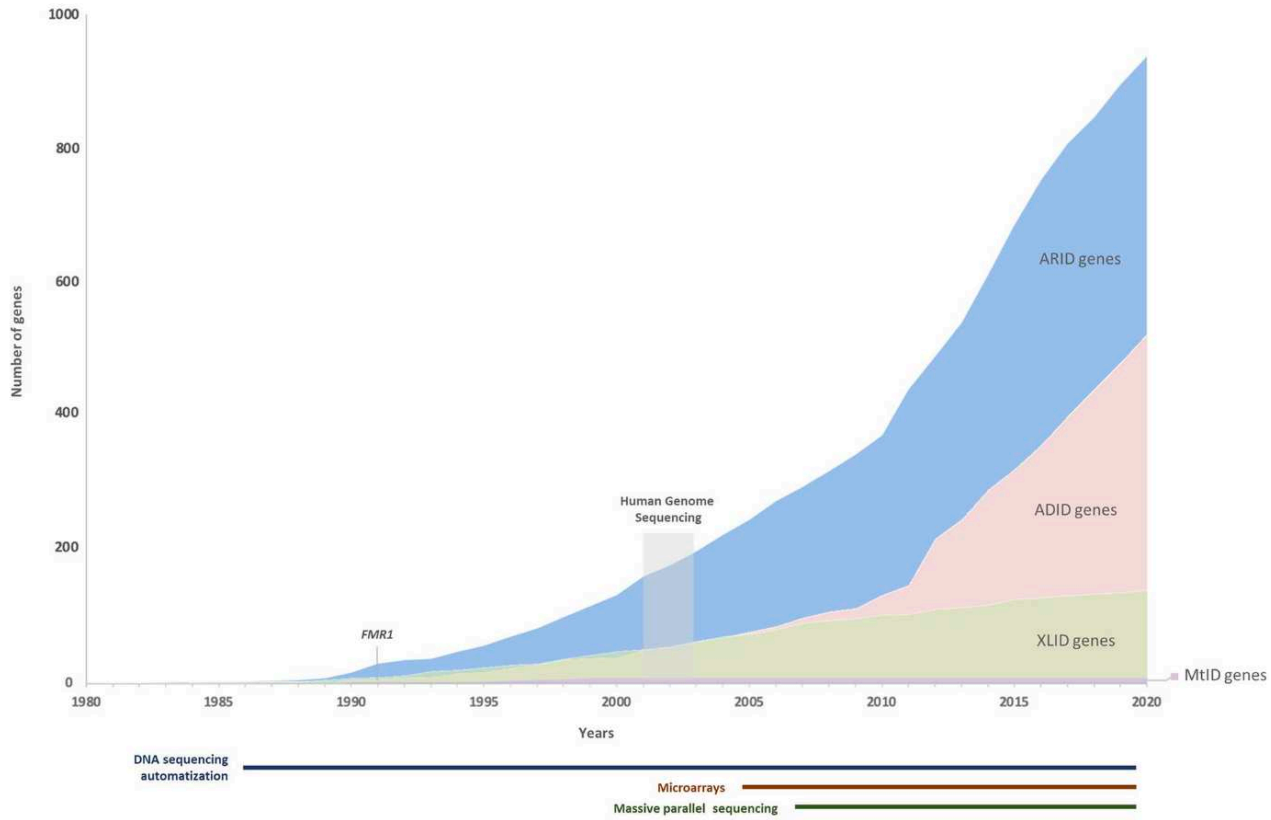


Figure 9 : Évolution du nombre de gènes connus liés à la DI. Tiré de Maia et *al.* 2021

Le séquençage à visée diagnostique repose essentiellement sur trois principales approches, le séquençage ciblé (panel de gènes), le séquençage d'exome (SE), le séquençage de génome (SG). Avant de présenter ces techniques, nous tenons à préciser ici notre choix terminologique. Les termes anglais *whole exome sequencing* (WES) ou *whole genome sequencing* (WGS) sont historiquement utilisés dans la littérature mais certains auteurs préfèrent ne plus utiliser le mot *whole*. Nous avons également choisi ce positionnement.

La technologie de séquençage la plus utilisée aujourd'hui en diagnostic est celle du fabriquant Illumina (San Diego, CA, USA). Elle repose sur le séquençage par synthèse. L'ADN est tout d'abord fragmenté puis des adaptateurs dont la séquence est connue sont ajoutés à chaque extrémité des fragments. Ces adaptateurs serviront pour l'amplification et le séquençage. L'ADN est ensuite chargé sur une flow-cell sur laquelle les adaptateurs s'hybrident avec des séquences complémentaires. Chaque fragment est amplifié pour générer des clusters de plusieurs milliers de copies. Enfin, les *primers*, l'ADN polymérase et des nucléotides marqués par des fluorophores sont ajoutés (une couleur par type de base). Lors de l'élongation, l'incorporation d'un nucléotide s'accompagne d'une émission de fluorescence enregistrée par une caméra puis cette information est convertie en base nucléotidique. Cette technique permet un séquençage en parallèle massif de milliers de fragments et donc un débit très élevé. En revanche, une des limitations de la technologie Illumina est la petite taille des fragments séquencés qui est de 300 pb maximum (*short read*). En effet, le *short read* ne permet pas une détection optimale de tous les types de variants structuraux, nous reviendrons plus en détails sur ce point. La distinction entre panel, SE et SG porte sur la technique de préparation des échantillons. Les panels de gènes et le SE reposent sur la capture de séquences cibles. Le principe, décrit par Gnirke *et al*. en 2009, fait appel à un ensemble de sondes d'ARN biotinylés qui, après hybridation avec leurs cibles, sont extraites à l'aide de billes magnétiques recouvertes de streptavidine [46]. Notons que ces protocoles de préparation incluent des étapes de PCR qui génèrent un biais préjudiciable pour la détection des variations du nombre de copies. Aujourd'hui, le SE est l'approche couramment utilisée en première intention en association avec les puces d'hybridation génomique comparative (ACPA). Son coût est abordable et offre un rendement diagnostique variant entre 21% et 66% [47,48]. Cet intervalle très large s'explique par la

difficulté à comparer les différentes études. Comme nous l'avons déjà évoqué, les designs de ces travaux varient fortement et le rendement diagnostique ne sera pas le même pour des TND isolés ou syndromiques, pour des approches différentes (solo, trio, familiales) mais aussi selon le type d'investigations préalables déjà effectuées.

Study	Country	N	Previous Investigation	Cohort Phenotype	DY
Studies with singleton approach					
Al-Kasbi et al. (2022)	Oman	188	K, CMA, TGP	GDD/ID	27%
Levchenko et al. (2022)	Russia	133	K or CMA	Non-specific GDD/ID	27%
Chen et al. (2021)	Taiwan	49	CMA	Unexplained moderate–severe ID	51%
Nouri et al. (2021)	Iran	61	K	Unexplained ID/DD	66%
Valentino et al. (2021)	Italy	84	CMA	ID, without ASD	39%
Hu et al. (2019)	Iran	404	NA	Unexplained ID in consanguineous family	54%
Xiao et al. (2018)	China	33	CMA	Unexplained ID/DD	57%
Studies with trio or familial approach					
Guo et al. (2021)	China	21	NA	ID	42%
Hiraide et al. (2021)	Japan	101	NA	Unexplained ID/DD	54%
McSherry et al. (2021)	Turkey	21	NA	Clinical suspicion of non-syndromic ARID	48%
Taskiran et al. (2021)	Turkey	59	CMA	ID, born to consanguineous parents	49%
Xiang et al. (2021)	China	17	NA	Unexplained ID	59%
Harripaul et al. (2018)	Pakistan and Iran	192	CMA	Unexplained ID in consanguineous family	46%
Snoeijen-Schouwenaars et al. (2018)	Netherland	100	Single-gene testing	Unexplained Epilepsy and ID	25%
Zhao et al. (2018)	Sweden	28	NA	ID/DD with dysmorphic features/congenital anomalies	21%

Tableau 1 : Rendements du SE pour la DI/RGD dans quelques études récentes. Tiré de Ko and Chen 2023. Abbreviations: AR = autosomal recessive; CMA = chromosome microarray; DD = developmental delay; DY = diagnostic yield; GDD = global developmental delay; ID = intellectual disability; K = karyotyping; N = number of cases with ID/GDD; NA = not available; TGP = targeted gene panel.

Avec l'arrivée du génome à 1000 dollars, la question de l'implémentation du SG en première intention en remplacement de l'ACPA et du SE s'est rapidement posée. Initialement, le SG à même montré qu'il pouvait surpasser le SE pour la détection de variants dans les régions exoniques grâce à une meilleure couverture. De plus, contrairement au SE l'absence de biais d'amplification permet une meilleure sensibilité

pour la détection de CNV [49,50]. Mais, c'est sa capacité théorique à détecter des variants structuraux équilibrés ou dans des régions non-codantes qui le rend particulièrement intéressant.

2.1.3. Workflow d'analyse et d'interprétation des résultats

Alignement des *reads* de séquençage :

Cette partie de l'analyse des données est dans l'ensemble très strandardisée, que ce soit pour le SE ou SG. Le guide des bonnes pratiques du GATK (Genome Analysis ToolKit) donne un aperçu du workflow le plus souvent utilisé tant en recherche qu'en diagnostic (https://software.broadinstitute.org/gatk/best-practices/). La toute première étape est le contrôle qualité des données de séquençage brutes générées sous forme de fichier FASTQ et le retrait des séquences des adaptateurs Illumina. Les FASTQ ne contiennent que des *reads* bruts ainsi que le score qualité de l'identification de chaque base. L'objectif de l'étape suivante, appelée « **alignement** » est donc de déterminer les coordonnées de chacun de ces *reads* sur le génome humain. Il s'agit ici du processus le plus exigeant en ressources informatiques et donc le plus long. Une recherche directe de motifs exacts dans le génome de référence présenterait une complexité algorithmique telle que les ressources informatiques disponibles dans les laboratoires de diagnostic ne suffiraient pas pour une analyse dans un temps raisonnable. Les algorithmes d'alignement reposent donc sur des transformées telles que la transformée de Burrows-Wheeler (BWT). La BWT d'une chaîne S est une permutation réversible de S qui permet de rechercher un motif P dans S en temps linéaire par rapport à la longueur de P, indépendamment de la longueur de S. En d'autres termes, la BWT transforme la chaîne de caractère du génome de référence de manière à faciliter la recherche efficace de

motifs correspondant aux *reads* de séquençage, en un temps qui dépend uniquement de la longueur de ces *reads*. Les algorithmes d'alignement bwa [51] et bowtie [52] sont les deux principaux basés sur le BWT. Le fichier de sortie de l'alignement est sous un format standard appelé BAM (Binary alignment map) sous sa forme binaire, SAM (Sequence alignment map) sous sa forme lisible par un humain ou bien CRAM (Compressed reference-oriented alignment map) pour sa forme compressée.

Le *calling* des variants :

Le *calling* des variants se fait à partir des fichiers BAM par comparaison avec le génome de référence. Le fichier de sortie est en général un fichier au format VCF (Variant call format). Il existe un standard pour le format VCF (https://samtools.github.io/hts-specs/VCFv4.3.pdf), mais il n'est pas rare que des algorithmes dévient un peu de cette norme ce qui peut compliquer le traitement des fichiers par la suite. Si l'algorithme de *calling* des SNV/indels HaplotypeCaller (GATK) est assez consensuel, il n'en va pas de même pour les expansions de répétions, insertions d'éléments mobiles et plus généralement de tous les variants structuraux. Ces outils sont nombreux, et reposent sur des méthodes de détection différentes et offrent des performances très variables. Il est donc indispensable de combiner ces outils afin de maximiser les chances d'identifier un éventuel variant causal. Augmenter le nombre d'algorithme s'accompagne inévitablement d'une augmentation du nombre de faux positifs. Il faut donc également optimiser la filtration des variants lors de l'étape d'interprétation. Nous évoquerons les différents outils de *calling* des SV selon leur type au chapitre 3.

Temps d'analyse et parallélisation :

Beaucoup des processus informatiques à l'œuvre pour l'alignement ou le *calling* de variants peuvent être parallélisés ce qui permet de diviser le temps nécessaire. Pendant plusieurs années, la solution pour augmenter la parallélisation était d'augmenter le nombre de CPU. Toutefois, cela augmente aussi le coût des équipements informatiques. Il est donc nécessaire de trouver un compromis. L'analyse bioinformatique complète d'un génome humain à une couverture de 40X peut ainsi prendre près de 24h, soit beaucoup plus de temps que le séquençage en lui-même. Aujourd'hui, de nouvelles solutions commencent à apparaitre. Par exemple, la société Illumina a développé la puce « Dragen Bio-IT » qui, grâce à une architecture dédiée à l'analyse génomique, permet une analyse complète d'un génome humain en environ 30 minutes. Un autre exemple, qui ne nécessite pas d'avoir à investir dans une solution propriétaire, est l'utilisation des capacités de parallélisation des cartes graphiques (GPU). Les GPU sont conçus pour traiter un nombre important d'opérations en parallèle (au prix toutefois d'une moindre rapidité séquentielle). Certains algorithmes permettant d'utiliser les GPU pour l'analyse de génome annoncent un temps d'alignement divisé par dix [53]. Ces méthodes n'en sont qu'à leurs débuts mais prendront probablement une place importante au cours des prochaines années.

Annotation, filtration et interprétation des résultats

La variabilité génétique au sein des population est telle qu'un SE identifie plusieurs dizaines de milliers de variants et le SG plusieurs millions. Sans une annotation efficace et une filtration drastique, aucune interprétation biologique n'est possible. L'annotation est la première étape permettant d'inférer un rôle transcriptionnel à ces variants en fonction de leur localisation (exon, intron, site d'épissage, intergénique …) et du type d'évènement

(faux sens, non-sens, frameshift). Plusieurs outils ou agrégateurs d'outils existent comme ANNOVAR [54], VEP [55] ou encore AnnotSV pour les variants structuraux [56]. L'annotation dépend de la base de transcrits de gènes utilisée (NCBI, refSeq, ENSEMBL, UCSC ou GENCODE). Ainsi, dans le cas d'une analyse de génome, le choix de la version la plus récente de ces bases est indispensable pour bénéficier des informations les plus à jour sur les régions non-codantes dont la connaissance évolue très rapidement. La filtration est, quant à elle, une étape critique du workflow d'analyse. Il s'agit d'arriver à l'obtention d'une courte liste exploitable tout en évitant une filtration trop strigente qui ferait prendre le risque d'éliminer un variant causal. Au cours des dernières années, cette tâche a été facilitée par la mise à disposition de la communauté scientifique d'importantes bases de données de variants. Les premières bases de données publiques générées à partir de données de séquençage ont été publiées par le projet 1000 génomes [57] suivi par l'Exome Sequencing Project [58]. Le besoin d'une base de données de population de référence plus importante et plus diversifiée a ensuite conduit à la création de l'ensemble de données Exome Aggregation Consortium (ExAC) en 2014 [59]. Puis, avec l'ajout de données de SG, cette base a été renommée Genome Aggregation Database (gnomAD) [60]. Cette dernière est aujourd'hui une des plus utilisées pour la filtration de variants en fonction de leur fréquence dans la population [61]. En complément de ces outils, plusieurs algorithmes de prédiction de pathogénicité peuvent apporter une aide précieuse pour l'interprétation des variants candidats. Au moins 28 outils sont rapportés dans les études comparatives. Ils sont basés sur des algorithmes différents avec des performance variables et reposent parfois sur l'intégration de prédictions d'autres outils. Les méthodes reposent notamment sur la conservation de séquence inter-espèce (SIFT [62], FATHMM [63]) ; l'altération de la séquence ou de la structure protéique (PolyPhen-2 [64], SNPeffect [65]) ; l'analyse par *machine learning*

supervisé (REVEL [66], CADD [67], BayesDel [68], VEST3 [69]) ; l'analyse des altérations des sites d'épissage (SpliceAI [70]). Paradoxalement, les outils VEST3, REVEL, FATHMM et BayesDel qui se sont révélés les plus performants dans plusieurs analyse comparatives restent beaucoup moins cités dans la littérature que les outils PolyPhen-2, SIFT et CADD [71].

2.1.4. Séquençage long read

Le séquençage de type *short read* a toutefois montré ses limites en diagnostic et plus précisément ses performances modestes dans la détection de variants structuraux. En effet, les *reads* de moins de 300 bases, sont trop courts pour détecter plus de 70 % des SV du génome, les variations structurelles de taille intermédiaire de moins de 2 kb étant particulièrement sous-représentées. La principale cause est la difficulté à aligner ces *reads* avec le génome de référence dans les larges régions répétées (voir chapitre 3). Les technologies dites *long read* peuvent, quant à elles, générer des séquences continues d'une longueur allant de 10 kb à plusieurs mégabases qui peuvent facilement couvrir la totalité d'une région répétitive [72,73]. Deux technologies principales sont aujourd'hui les plus utilisées le PacBio SMRT sequencing et l'Oxford Nanopore Technologies sequencing (ONT) (Figure 10).

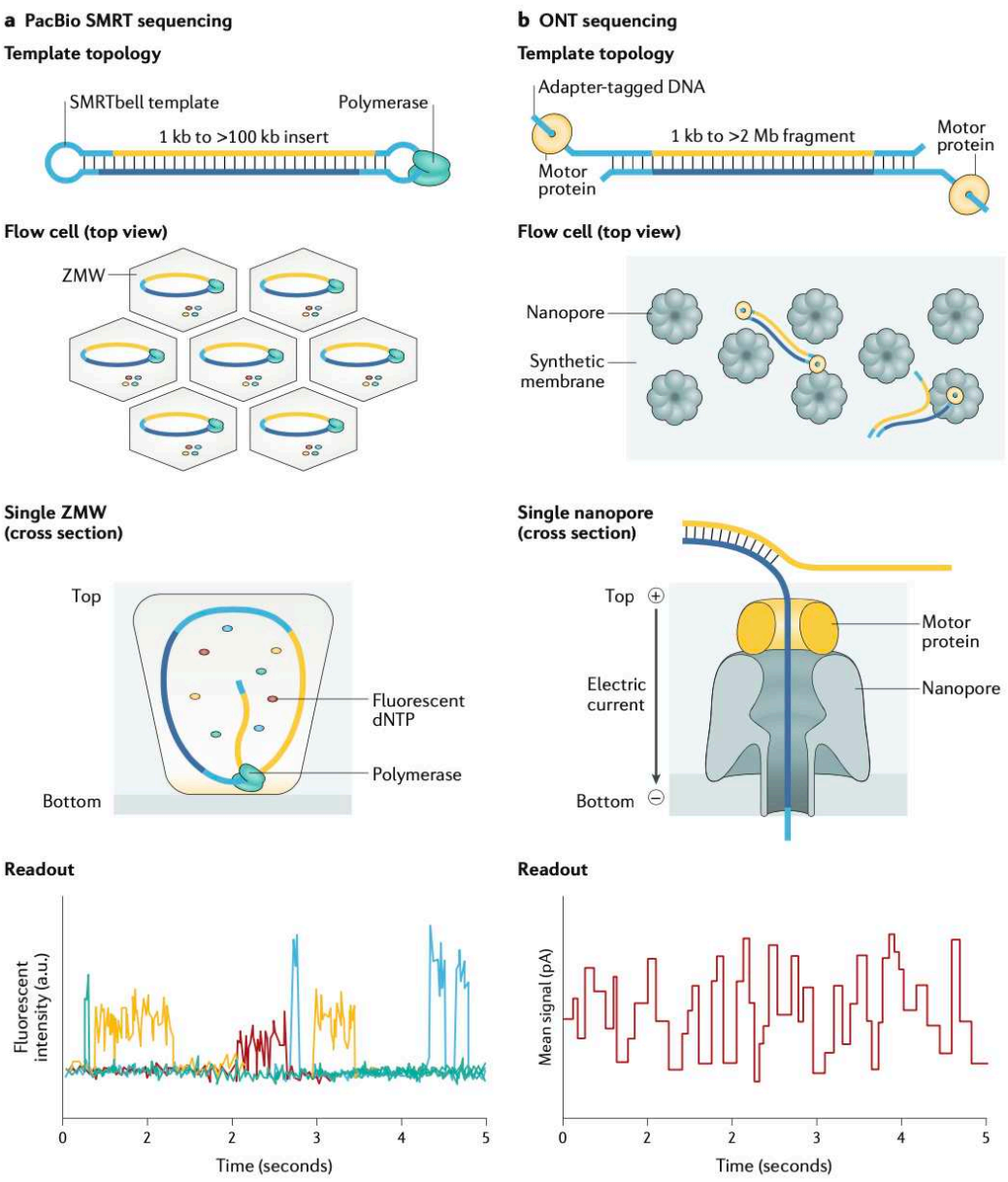


Figure 10 : Techniques SG long read PacBio SMRT et Oxford Nanopore Technologies. Tiré de Logsdon et al. 2020

Le PacBio SMRT repose sur la ligature de deux adaptateurs en épingle à cheveux à des longs fragments d'ADN double brin. La bibliothèque est composée d'un ensemble de ces ADN circulaires appelés SMRTbell. Chaque SMRTbell est associé à une ADN polymérase puis déposé dans un puit (un puit par SMRTbell). Ce modèle circularisé est séquencé par synthèse de manière répétée pour générer des *reads* HiFi avec une précision supérieure à 99,9 % [74]. La technologie ONT repose sur des nanopores protéiques intégrés dans une

membrane électriquement isolante. Lorsque des brins d'ADN individuels passent à travers les pores, ils provoquent des modifications du courant ionique, qui sont ensuite analysées pour caractériser chaque base nucléotidique en temps réel [75]. Il est difficile de comparer les performances des deux technologies. Les études publiées sur ce sujet montrent les points forts et limites de chaque approche [76], mais les améliorations régulièrement apportées par les fabricants peuvent rendre rapidement ces informations obsolètes.

2.2. L'optical genome mapping

La cartographie optique de l'ADN est apparu ces dernières années comme une technique intéressante pour l'étude des variants structuraux, et plus particulièrement les événements équilibrés [77,78]. Cette méthode est basée sur la reconnaissance par l'enzyme DLE-1 d'un motif de 6pb (14 à 17 motifs par 100kb) et par le marquage de ces derniers. Les molécules isolées peuvent atteindre une taille maximale de l'ordre de la mégabase et la densité en sondes permet de viser une résolution de 500 pb pour les CNV [79]. Les molécules d'ADN de haut poids moléculaire marquées sont libérées de leur état d'enchevêtrement (thermodynamiquement favorable) et étendues pour permettre une lecture séquentielle de leurs marqueurs fluorescents. Cette extension peut être réalisée à l'aide de diverses méthodes comme un passage sur surface de polymère chargée ou de verre silanisé ou bien par passage à l'intérieur d'un nanocanal [77]. Par exemple, avec cette dernière méthode, utilisée notamment par le fabriquant Bionano (San Diego, CA, USA) un champ électrique est appliqué aux longues molécules d'ADN pour les faire passer dans un réseau de nanocanaux, ce qui contraint les molécules se linéariser [80] (Figure 11). La technologie OGM a été initialement développée pour permettre d'atteindre

une résolution et un débit supérieurs aux techniques cytogénétiques classiques. À ce titre, elle présente un grand intérêt pour les laboratoires d'hématologie et oncogénétique. La question de son utilisation en génétique constitutionnelle se pose également et l'approche optimale reste à définir. Il faut ici souligner que la principale limite de cette technique est sa résolution plus faible que celle du séquençage. Comme nous l'avons vu, les approches de SG *long read* permettent déjà de dépasser les problèmes d'alignement rencontrés avec le SG *short read* dans les régions répétées et ainsi améliorer la détection des SV tout en conservant une résolution d'une paire de bases. Toutefois, avec un faible coût et une simplicité de mise en œuvre et d'analyse des données, l'OGM est une solution plus facilement envisageable en pratique à court terme. Par ailleurs, il semblerait que certains SV détectés en OGM et tout particulièrement les duplications et inversions, ne soient pas toujours détectés en SG *long read* [81].

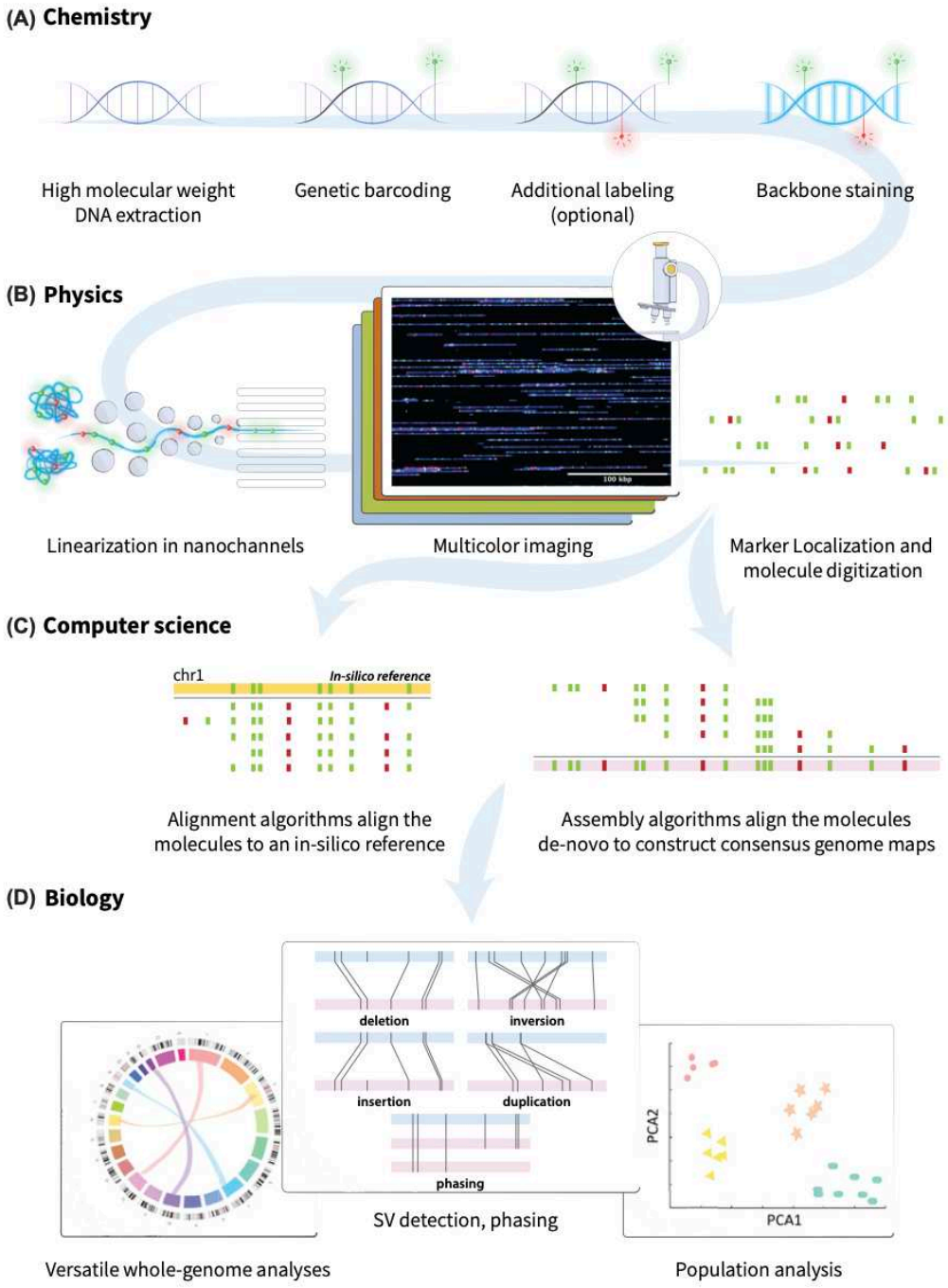


Figure 11 : Principe de l'optical genome mapping. Tiré de Jeffet et al. 2021

2.3. Le RNA-Seq clinique

Alors que le SE devenait peu à peu une routine et que le SG apparaissait comme une solution prometteuse, le problème du plafonnement du rendement diagnostique a amené plusieurs équipes à se tourner vers le RNA-Seq. Cette approche complémentaire a très tôt montré son utilité dans plusieurs pathologies tissus spécifiques comme des maladies musculaires ou mitochondriales rares [82–85]. L'accès direct à des tissus en lien avec la pathologie étudiée (biopsie de cellules musculaires ou de fibroblastes) rendait cette approche appropriée et donnait l'espoir d'obtenir de bons rendements diagnostiques. Toutefois, en pratique, le fait d'avoir recours à un geste invasif pour obtenir l'ARN peut potentiellement limiter l'adhésion des patients et des familles. De plus, un tissu relevant n'est pas toujours accessible pour toutes les maladies rares. En 2019, Frésard *et al*, ont montré qu'il était possible d'utiliser le sang total, et cela pour un ensemble hétérogène de maladies rares [86]. Les gains de rendement rapportés dans toutes ces études sont, sans trop de surprise, très variables et vont de 7,5 % à 36 % en fonction du tissu échantillonné et du phénotype clinique (Tableau 2). L'apport ne se limite d'ailleurs pas seulement à l'identification de variants chez des patients pour lesquels les précédents séquençages d'ADN étaient non-concluants. L'information supplémentaire apportée par le RNA-Seq peut également servir à résoudre ou prioriser des variants de signification incertaine (VUS) [87,88]. Si l'utilisation du sang total présente un avantage pratique évident, la question de ses performances diagnostiques fait débat. Dans une étude récente, Murdock *et al.* (2021) ont comparé le rendement pour différentes maladies mendéliennes à partir de sang total et de fibroblastes de peau et en sont arrivés à la conclusion que les fibroblastes présentaient une expression plus élevée et plus homogène (Figure 12) des gènes cliniquement pertinents que le sang total. Ils ont également constaté que l'événement causal était manqué dans le sang dans la moitié des cas [89].

Les principaux événements recherchés par RNA-Seq clinique sont les anomalies d'expression, les anomalies d'épissage et l'expression monoallélique.

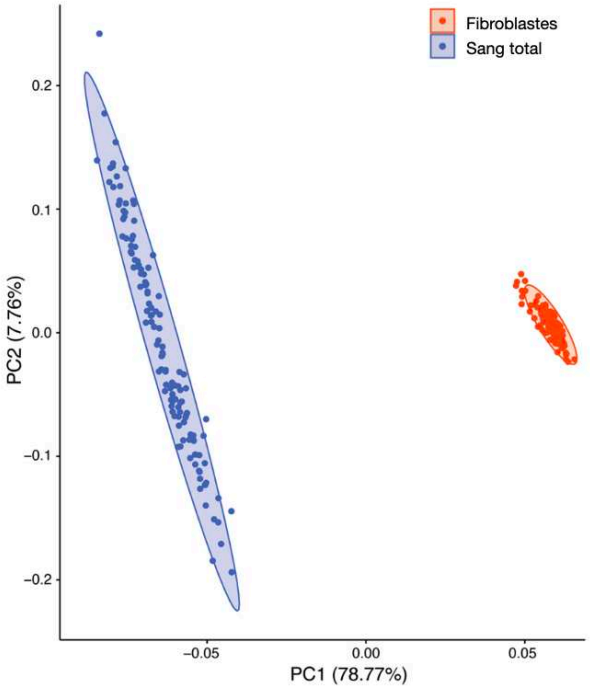


Figure 12 : Analyse en composantes principales de l'expression des gènes (TPM) dans le sang total et les fibroblastes cutanés. Tiré de Murdock et *al.* 2021

Etude	Tissu	Pathologie	Effectif	Rendement
Cummings et al., 2017	Muscle	Dystrophie musculaire congénitale liée au collagène de type VI	50	25%
Kremer et al., 2017	Fibroblastes	Maladies mitochondriales	48	10%
Frésard et al., 2019	Sang total	Diverses maladies	94	7,5%
Lee et al., 2020	Sang total, fibroblastes, muscle, moelle	Diverses maladies	234	18%
Murdock et al., 2020	Sang total, fibroblastes	Diverses maladies	115	17%
Yépez et al., 2022	Fibroblastes	Maladies mitochondriales	303	16%
Colin et al., 2022	Sang total	TND	30	8,7%

Tableau 2 : Principales études sur le diagnostic par RNA-Seq clinique

2.3.1. Les anomalies d'épissage

Les variants des "site canoniques d'épissage" (CSS) situés à moins de 2 pb d'une jonction exon-intron sont bien connus pour être des candidats perte de fonction. La contribution des variants d'épissage à des sites non canoniques n'est pas négligeable non plus puisque qu'il été rapporté sur une cohorte de patients présentant un trouble du développement que jusqu'à 27 % des variants d'épissage *de novo* pathogènes se trouvent à des positions non canoniques [90]. Plusieurs études ont développé le concept d'une région "proche de l'épissage", généralement des dizaines de paires de bases autour d'une jonction exon-intron, qui contient de nombreux motifs d'épissage conservés. Cependant, il n'existe pas de standard pour leur interprétation et leur annotation n'est pas complète [90–92]. En outre, les variants de point de branchement

putatifs et les variants introniques profonds (situés à plus de 100 pb d'une jonction exon-intron), peuvent également perturber l'épissage, et leur contribution globale aux maladies rares n'est pas connue [93]. Enfin il existe des sites de régulation de l'épissage (SRE *Splicing Regulatory Elements*) situé tant dans les exons que les introns. Ils peuvent être activateurs comme les *Exonic Splicing Enhancers* (ESE) et les *Intronic Splicing Enhancers* (ISE) ou inhibiteurs comme les *Exonic Splicing Silencers* (ESS) et les *Intronic Splicing Silencers* (ISS). Ces séquences de 6 à 8 bases, peu conservées, fixent des protéines activatrices ou inhibitrices de l'épissage. Ces sites peuvent, eux aussi, être le siège de variants pathogènes [94] (Figure 13).

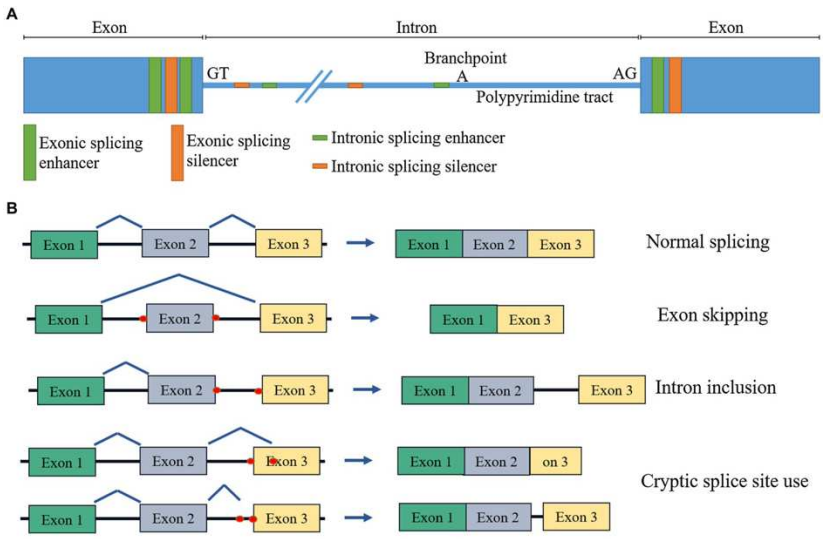


Figure 13 : Sites d'épissages et éléments régulateurs d'épissage. Tiré de Lord et Baralle 2021.

Il existe aujourd'hui de puissants outils de prédiction de l'effet d'un variant sur l'épissage, comme SpliceAI [70]. Ils peuvent permettre de réduire drastiquement le nombre de candidats avec un variant dans une région proche de l'épissage mais avec un risque de faux négatif encore difficile à quantifier précisément. De plus, les variants prédits comme

impactant l'épissage devront, quoi qu'il en soit, être validés expérimentalement (validation par Minigene par exemple). Le RNA-Seq est un moyen de s'affranchir de cette limitation. La plupart des pipelines d'analyse RNA-Seq clinique incluent un module de détection des anomalies d'épissage qui permet de détecter un épissage statistiquement aberrant au sein d'une cohorte d'échantillons. Il est ainsi possible de détecter et valider en une seule étape. Il existe plusieurs outils comme LeafCutter [95] ou FRASER [96] mais c'est ce dernier qui semble le plus utilisé aujourd'hui. FRASER, repose sur 3 étapes d'analyse. 1°/ une carte des sites d'épissage est générée à partir des *split reads* supportant les jonctions exon-exon ainsi que les *reads* entiers chevauchant les sites d'épissage. À partir de cette information, des métriques d'épissage qui quantifient les accepteurs alternatifs (ψ5), les donneurs alternatifs (ψ3) et les efficacités d'épissage au niveau des donneurs (θ5) et des accepteurs (θ3) sont calculées. 2°/ un modèle statistique contrôle les covariations des échantillons. 3°/ Les valeurs aberrantes sont détectées comme des points de données qui s'écartent de manière significative du modèle ajusté (Figure 14).

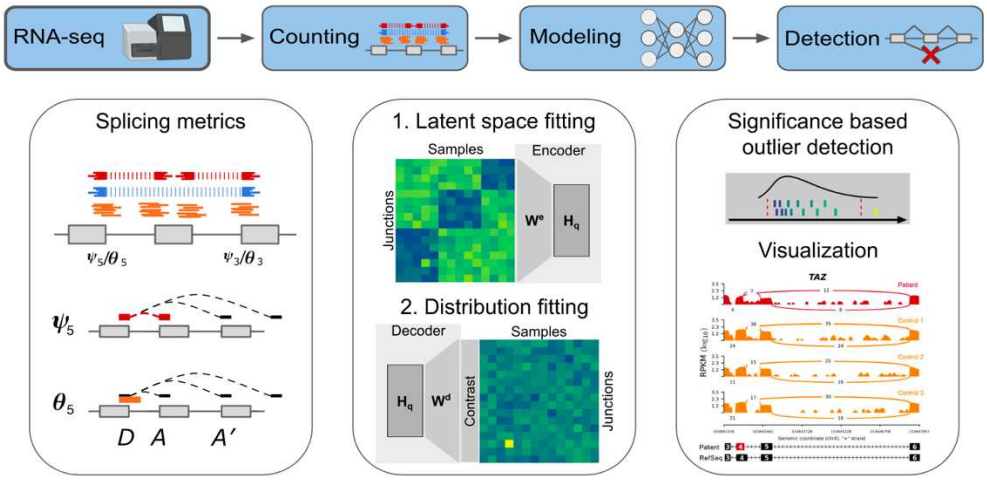


Figure 14 : Aperçu du workflow d'analyse d'épissage aberrant FRASER. Tiré de Mertes *et al.* 2021

2.3.2. Les anomalies d'expression

Les variants localisés dans des régions codantes et susceptibles d'entrainer une perte d'expression (non-sens, décalage du cadre de lecture) sont en règle générale plutôt simples à interpréter. Pourtant, tous ces événements, comme les SV équilibrés, ne sont pas forcément détectés en SE ou SG. De plus, il peut être difficile de relier un variant non-codant dans une région régulatrice à son gène cible seulement sur la base des données de SG. Les algorithmes couramment utilisés en RNA-Seq tels que DESeq2 [97] et edgeR [98] ont été conçus pour comparer l'expression des groupes d'échantillons sur un mode « cas *versus* contrôles » avec un grand nombre de réplicats. Cette configuration n'est pas vraiment adaptée aux maladies rares pour lesquelles il n'est souvent pas possible de constituer ces groupes de réplicats. C'est donc vers une approche par recherche *d'outliers* au sein d'une population hétérogène de patients que les outils actuels se sont tournés. Les premières méthodes proposées n'intégraient pas toujours une analyse statistique solide et une prise en compte des covariables ce qui limitait leur portée [83]. Aujourd'hui, l'outils le plus cité, OUTRIDER, s'appuie sur un autoencodeur pour représenter le nombre de *reads* attendu en fonction des variations dues à des facteurs techniques, environnementaux ou génétiques communs. Le nombre de *reads* issus des données RNA-seq sont considérés comme étant distribués selon une loi binomiale négative avec une dispersion spécifique à chaque gène. Pour détecter les valeurs aberrantes, l'algorithme identifie les nombres de *reads* qui s'éloignent significativement de cette distribution [99].

2.3.3. Les expressions monoalléliques

Lorsqu'un gène candidat est associé à un mode d'hérédité récessif, les variants présents uniquement sur un seul allèle ne sont généralement pas prioritaires après séquençage d'ADN. Pourtant, si l'expression de l'allèle non muté est inhibée par un mécanisme indépendant (cis-régulation, modifications épigenétique) ce mode d'hérédité peut tout de même être compatible avec le phénotype observé. De fait, la détection d'une expression monoallélique (MAE) d'un variant rare pathogène a été rapportée pour plusieurs cas [82,83,100–102]. Une MAE peut également être mise en évidence dans des gènes haploinsuffisants en complément de recherche d'aberration d'expression. Les outils existants sont ANEVADOT [100] ou le module MAE du pipeline Drop [103]. Le principe général repose sur le comptage des *reads* alignés sur chaque allèle aux positions génomiques des variants hétérozygotes. Ces outils nécessitent donc d'avoir réalisé *a minima* un SE au préalable.

2.3.4. Le calling de variants

Le taux de mutation germinale *de novo* par génération est évalué à environ $1{,}8.10^{-8}$ soit environ 40 mutations par génération. Une étude du projet 1000 génomes a identifié 35 et 49 mutations germinales *de novo* dans deux familles mais aussi 952 et 643 mutations *de novo* non germinales (somatique ou culture cellulaire) et plus de 1000 faux positifs [104]. Indirectement, ces données mettent en lumière l'intérêt que peut présenter le *calling* de variants (SNP/indels) à partir des données de RNA-Seq tout particulièrement si le tissu diffère de celui utilisé pour le séquençage d'ADN. En effet, il est ainsi possible d'exclure

par comparaison des variations somatiques ou apparues lors des cultures cellulaires le cas échéant mais aussi de mettre à jour un possible mosaïcisme.

Chapitre 3. Les variants structuraux

Comme nous l'avons évoqué précédemment, le SG *short read* remplace progressivement le séquençage d'exome pour le diagnostic moléculaire des maladies génétiques rares. Pour suivre cette évolution technologique, il est très vraisemblable que le SG *long read* le remplacera à son tour quand les coûts auront diminué suffisamment pour rendre son utilisation en routine abordable. Avec ces nouvelles techniques, la détection des variants structuraux prend un nouveau souffle. La sensibilité s'améliore, des types de variants difficilement détectables le deviennent, et l'identification d'événements complexes est maintenant à notre portée. La définition des SV a évolué au cours du temps et au gré de l'amélioration des techniques. Elle est tout d'abord passée de larges remaniements visibles sur caryotype à des remaniements impliquant tout fragments de taille supérieure à 1kb. Cette dernière frontière, qui était principalement fixée par les limites des techniques d'identification sur puce, laissait de nombreux variants de tailles inférieures inaccessibles. Aujourd'hui, grâce au SHD, cette limitation est surmontée et une définition qui fait consensus considère les SV comme des réarrangements génomiques impliquant des régions de plus de 50 paires de bases [105]. Ces événements peuvent affecter le phénotype de plusieurs façons, notamment en altérant le dosage d'un gène ou par rupture ou fusion de gènes. Le taux de SV *de novo* par génération a été estimé entre 0,160 et 0,29 événements par génome dont 0,13 événements cliniquement significatifs [106,107]. Cette estimation ayant été déterminée par SG *short read,* il est probable que ce taux soit une hypothèse basse. Nous allons voir dans ce chapitre les différents types

de SV, leurs implications en pathologie humaine, les régions à risque de remaniements structuraux ainsi que les mécanismes d'apparition.

3.1. Les types de SV

3.1.1. Variations du nombre de copies (CNV)

Certains SV, appelés variations du nombre de copies, entrainent un gain (duplication) ou une perte (délétion) d'un segment chromosomique et sont donc dit non-équilibrés. Le tout premier CNV impliqué dans une pathologie autosomique dominante, une duplication responsable de la maladie de Charcot-Marie-Tooth de type 1A a été rapportée en 1991 par Lupski *et al* [108]. Toutefois, on estime que des CNV couvrent 12% du génome dans la population générale [109]. Ainsi, de nombreux gènes sont concernés ce qui peut rendre difficile l'interprétation des variants détectés en diagnostic. Aujourd'hui, les CNV d'une taille supérieure à 1Mb sont aisément détectables en ACPA et le SE offre de bonnes performances pour des événements plus petits. Néanmoins, la capacité du SE à détecter des délétions ou duplications de quelques exons seulement peut être limitée en fonction du kit de capture utilisé et de la couverture des exons concernés. Le SG permet, quant à lui, de dépasser cette difficulté en permettant, dans beaucoup de cas, l'identification précise des points de cassure même introniques ou intergéniques.

3.1.2. Translocations

La translocation est un réarrangement anormal des chromosomes pouvant être réciproque, quand elle est due à un échange de segments entre l'extrémité de deux chromosomes non homologues, ou robertsonienne quand deux chromosomes fusionnent pour former un chromosome unique. Elle peut également impliquer un transfert de segment chromosomique vers une autre région du même chromosome ou d'un autre, avec ou sans échange. On parle d'un événement équilibré lorsqu'il n'y a ni perte ni gain de matériel génétique. Environ 6% des translocations sont associées à un phénotype anormal [110]. Les translocations robertsonniennes, quant à elles, ne causent pas de phénotype visible mais sont en revanche impliquées dans l'infertilité. Les mécanismes expliquant la pathogénicité de ces événements équilibrés sont la rupture d'un gène, un effet de position (éloignement d'une région régulatrice par exemple), ou une disomie uniparentale pour les gènes soumis à empreinte. Le caryotype et l'hybridation in situ en fluorescence (FISH) sont les approches traditionnelles pour détecter les translocations. Le caryotype ne peut toutefois identifier les points de cassure qu'à une échelle chromosomique. Il est donc impossible de déterminer précisément quels gènes sont potentiellement touchés. Même si la FISH permet d'atteindre dans certains cas des résolutions de l'ordre du gène, le point de cassure exact reste inconnu. Il est donc notamment difficile de savoir si la translocation est réellement équilibrée ou si il y a une perte ou un gain de matériel au niveau des cassures [111]. En effet, des analyses par ACPA chez des patients présentant des translocations apparemment équilibrées ont montré que la présence de CNV, ou d'événements complexes, associés aux translocations est fréquente (40-60%) [112–114].

3.1.3. Inversions

Les inversions sont des réarrangements génomiques dans lequel un segment d'un chromosome est inversé par rapport à sa position d'origine. Il est depuis longtemps possible de détecter les inversions de grandes régions chromosomiques en utilisant les mêmes techniques cytogénétiques que pour les translocations notamment les caryotypes à bandes G. Cependant, comme nous l'avons dit précédemment, la résolution de ces derniers est très limitée et ne s'applique qu'à l'identification de variants de plusieurs mégabases. De plus, même les inversions de grandes tailles peuvent échapper à la détection si le segment inversé n'entraîne qu'une faible différence de marquage. Les techniques moléculaires permettant de détecter efficacement les inversions de plus petites tailles ont longtemps fait défaut et n'ont pas rencontré le même succès que pour les autres types de SV [115]. Cela s'explique en partie par le fait que, contrairement aux CNV, elles ne peuvent pas être détectées comme des changements dans la profondeur de couverture. De plus, comme nous le verrons par la suite, les points de cassure sont souvent dans des régions répétées ce qui rend leur détection particulièrement incertaine en SG *short read*.

Tout comme les translocations, les inversions peuvent avoir un effet pathogène direct sur l'expression d'un gène par rupture de ce dernier ou en l'éloignant de régions régulatrices. Un effet pathogène indirect a également été observé. En effet, à la suite d'une inversion, de nouvelles régions répétées complexes peuvent se former au niveau des points de cassure. Ces nouvelles régions d'homologie peuvent elles-mêmes devenir sujettes à la formation de CNV récurrents. Cela a été rapporté pour plusieurs syndromes (Tableau 3) notamment pour le syndrome de Williams-Beuren (MIM : 194050) et le syndrome de Koolen de Vries (MIM : 610443) dont les parents de patients porteurs de la délétion

causale sont plus fréquemment eux-mêmes porteurs d'une inversion polymorphique fréquente dans la population [116–119].

Bande	Taille (Mb)	Pathologie/Réarrangement	Référence
5q35.2-q35.3	1,9	Syndrome de Sotos (microdélétion)	[120]
7q11.23	1,5	Syndrome de Williams-Beuren (microdélétion)	[116]
15q11-q13	4	Syndrome d'Angelman (délétion)	[121]
15q13.3	2	Microdélétion 15q13.3	[122]
17q21.31	0,9	Syndrome de microdélétion 17q21.31	[123]

Tableau 3 : Exemples de CNV récurrents associés à des inversions. Adapté de Feuk, 2010.

3.1.4. Insertions

Les insertions de grandes tailles (> 50pb) peuvent relever de plusieurs mécanismes comme les insertions d'éléments mobiles (voir partie 3.2.2) ou les duplications non en tandem (voir partie 3.1.5 traitant des SV complexes). Les effets pathogènes de ces insertions seront détaillés dans les parties concernées.

3.1.5. Les variants structuraux complexes

Une partie des SV n'est parfois pas classable dans une seule de ces catégories. On parle dans ce cas de SV complexes qui, par définition, comprennent plus de deux points de cassure [124]. Les SV complexes peuvent être séparés en quatre grandes classes, les

duplications en tandem avec délétions imbriquées, les duplications non en tandem, les événements de délétion-inversion-délétion et les événements de duplication-inversion-duplication [125]. Dans une étude de 2016 sur 235 individus (dont 71 présentant un TSA) Brandler *et al* ont rapporté en moyenne 251 SV complexes par individu montrant ainsi que ces événements sont courants dans la population. Par ailleurs, ils ont observé que les duplications non en tandem étaient la forme la plus courante de SV complexe. Ces insertions se sont produites dans des orientations directes et inversées avec la même probabilité, et 22 % étaient interchromosomiques. Enfin, la majorité (73 %) présentait des délétions au niveau du site d'insertion [125].

3.2. Les régions répétées : Hotspots d'apparition des SV

Plusieurs hypothèses mécanistiques peuvent expliquer l'apparition des SV. Une étude de 2010 sur 1054 larges SV issus de 17 génomes a toutefois montré que, dans au moins la moitié des cas, le mécanisme repose sur une homologie de séquences [126]. Ainsi, même si des points de cassure de réarrangements structuraux ont été identifiés dans l'ensemble du génome, ils prédominent dans des régions de faible complexité génomique, telles que les régions répétées. Cela suggère donc que les SV ne sont pas uniquement des événements aléatoires, mais résultent d'une prédisposition aux réarrangements médiés par des *hotspots* de recombinaison [127]. Plus de la moitié du génome humain est composée de régions répétées (Figure 15). Comme nous l'avons vu précédemment, les régions avec de fortes homologies de séquence peuvent présenter un challenge pour le SG *short read* en rendant l'alignement des *reads* à ces *loci* difficile sinon impossible. Cette limitation est d'autant plus critique que, comme nous venons de

l'évoquer, les variants structuraux sont principalement médiés par ces mêmes homologies.

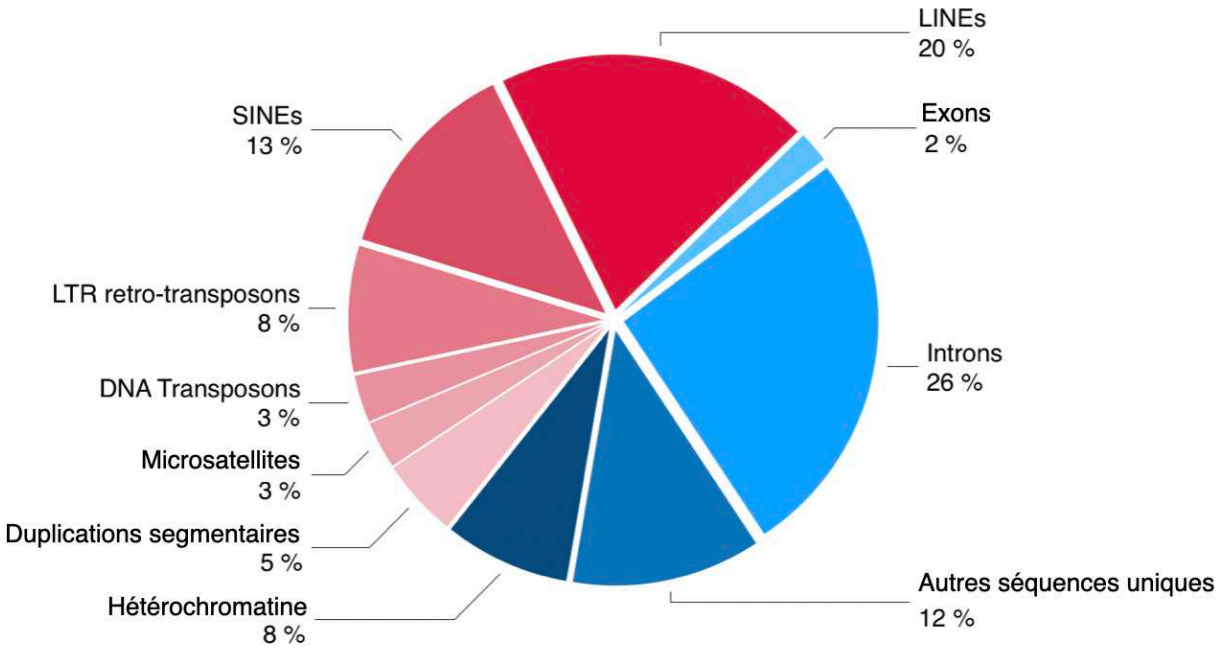


Figure 15 : Composition du génome humain. Plus de la moitié est composée de régions répétées.

3.2.1. Répétitions en tandem

L'**ADN satellite** est constitué d'un ensemble de séquences répétitives principalement présent dans l'hétérochromatine. Plusieurs types ont été identifiés en fonction de la longueur de leurs monomères (β-satellites : ~ 68pb ; α-satellites : ~171pb ; satellites III : 5pb). Les satellites III et α-satellites sont présents sur tous les chromosomes et principalement dans les régions centromériques. Les α-satellites ont d'ailleurs montré qu'ils pouvaient se lier avec la protéine centromérique CENP-B et pourraient jouer un rôle dans le *silencing* de l'hétérochromatine. Les β-satellites, quant à eux, sont principalement

présents sur les bras courts des chromosomes acrocentriques [128,129]. **Les minisatellites**, aussi appelés VNTR (*variable number of tandem repeats*), sont des répétitions dispersées sur l'ensemble du génome. Leur définition sur la base de la taille varie entre les auteurs (10-60pb ou 6-100pb). Leur polymorphisme dans la population, mais avec tout de même un certain degré de proximité entre individus apparenté, en a fait une source d'information utile en médecine légale [129]. Ces éléments sont instables et sont donc sources de variations génétiques. Les répétitions en tandem les plus connues en pathologie humaine restent toutefois les **short tandem repeats (STRs)**, aussi appelées **Microsatellites**. Ces STR composées de motifs de 2 à 6 pb représentent ~3% du génome [130]. Leur nature hautement répétitive est susceptible de provoquer un glissement réplicatif (*replication slippage*) et donc une variation de la longueur de cette région au cours des divisions cellulaires. Leur taux de mutation est donc particulièrement élevé [131]. Nous avons déjà évoqué l'expansion de triplets CGG dans le gène *FMR1* responsable du syndrome de l'X fragile mais plus de 40 autres troubles, principalement neurodégénératifs ont été rapportés en lien avec des expansions de STR. Nous pouvons par exemple citer des expansions de triplets CAG dans les gènes *ATXN1, ATXN2, ATXN3, ATXN7, CACNA1A, PPP2R2B* ou *TBP* tous responsables d'ataxies cérébelleuses transmises sur un mode autosomique dominant (AD). Parmi les autres maladies transmises sur un mode AD, nous pouvons également évoquer la maladie de Huntington (MIM : 143100) due à une expansion de CAG dans le gène *HTT* et pour les transmissions récessives liées à l'X (XLR) l'expansion CAG dans le gène *AR* responsable de la maladie de Kennedy (MIM : 313200). Outre l'X fragile, d'autres variations du nombre de motifs STR sont impliquées spécifiquement dans les TND, comme par exemple la présence d'un nombre de répétitions du triplet CCG supérieur à 200 dans le gène *FMR2* responsable d'une DI (MIM : 309548) [132].

Actuellement, la détection des expansions de STR est encore souvent réalisée par PCR de façon ciblée, mais le SG *short read* permet également leur identification. Plusieurs algorithmes ont été développés récemment pour une détection à partir des données de SG : ExpansionHunter [133], exSTRa [134], STRetch [135], et TREDPARSE [136]. Ils ont tous démontré une bonne capacité à détecter les expansions de STR lorsque la taille de l'allèle étendu est supérieure à la longueur des *short reads*. Tous ces outils se focalisent toutefois sur des *loci* et motifs connus alors que des études récentes ont montré que beaucoup de STR pathogènes ont une structure plus complexe. Un nouvel outil, ExpansionHunter de novo, a été développé pour dépasser cette limitation. Il ne nécessite pas de connaissance préalable des coordonnées génomiques, il détecte les changements de motifs nucléotidiques et enfin il est applicable à la fois aux motifs courts et longs [137].

3.2.2. Éléments répétés transposables

A la fin des années 1940, la généticienne Barbara McClintock, découvre lors de ses études sur le maïs que certains gènes peuvent être mobiles et que ces déplacements peuvent entrainer la perte d'expression d'autres gènes au site d'insertion [138]. Ces séquences d'ADN appelées « gènes sauteurs » puis éléments transposables ou transposons ont, quelques années après, été observés dans d'autres organismes, y compris l'homme, chez qui elles représentent près de 45% du génome (Figure 15). Cette découverte vaudra à Barbara McClintock le prix Nobel de médecine en 1983. Les éléments transposables (ET) peuvent être classés en deux grandes classes selon leur mode de transposition. La classe I regroupe les éléments faisant intervenir une étape de

rétrotranscription d'un intermédiaire ARN et qui peut donc s'assimiler à un mécanisme de copier-coller. Nous voyons ici que le nombre de ces ET de classe I peut donc croitre au gré des événements de transposition. La classe II, quant à elle, ne transpose que des séquences ADN et donc, dans la majorité des cas, nous pouvons l'assimiler à un mécanisme de couper-coller [139]. Ces transposons à ADN, n'ont jamais montré de preuve d'activité chez l'humain et des estimations bio-informatiques ont suggéré qu'aucune nouvelle transposition n'aurait eu lieu chez les primates depuis environ 37 millions d'années [140]. Les rétrotransposons (classe I) sont de plusieurs types dont certains ont gardé une activité leur conférant un potentiel pathogène propre. Ce pouvoir pathogène inhérent à la rétrotransposition est, en l'état actuel des connaissances, considéré comme relativement marginal dans le cadre des maladies génétiques constitutionnelles [141]. Mais, leur nature répétée et leur fréquence dans le génome en fait des éléments particulièrement sujets au risque de remaniements structuraux. Nous allons brièvement décrire les caractéristiques de quelques-uns de ces rétrotransposons (Figure 16).

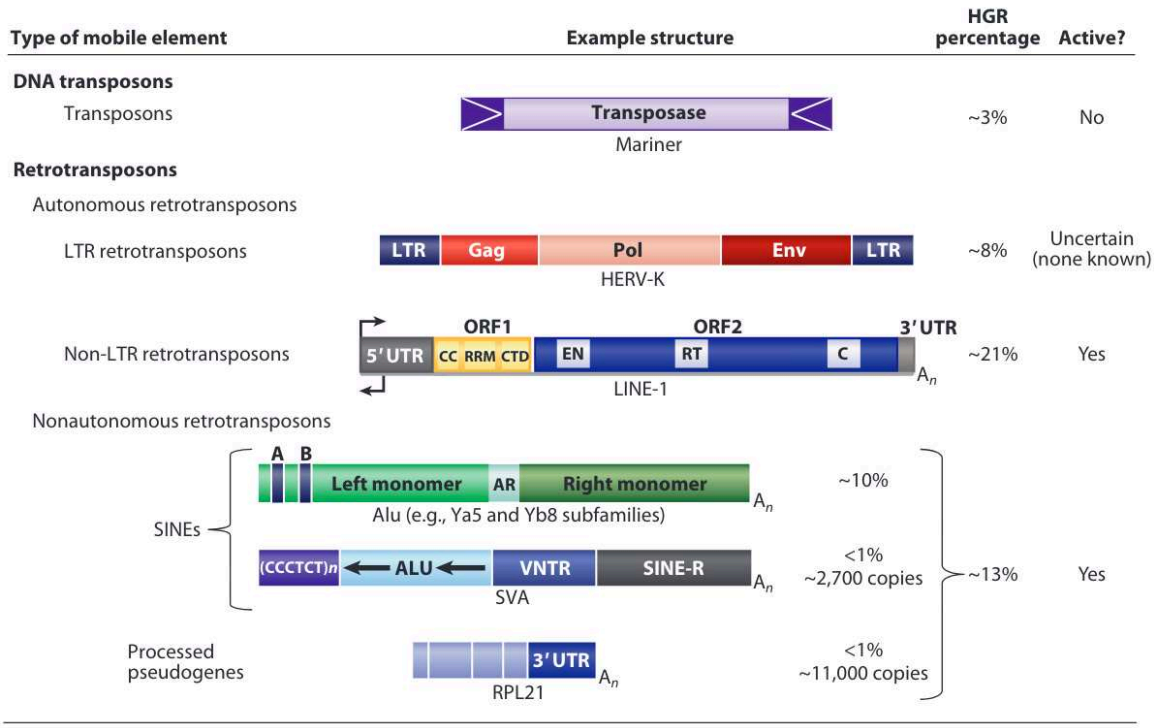


Figure 16 : Type d'éléments transposables dans le génome humain. Adapté de Beck et al. 2011

Les rétrotransposons LTR :

Ces ET, tels que les retrovirus endogènes humains (HERV), sont le résultat d'infections ancestrales par des retrovirus qui se sont intégrés au génome dans la lignée germinale et couvrent maintenant environ 8% du génome [130]. Du fait de leur origine virale, ils ont une organisation génomique comparable à celle des retrovirus mais ont perdu la capacité d'exprimer une protéine d'enveloppe fonctionnelle. Ils sont donc maintenant cantonnés à un cycle intracellulaire et à une transmission verticale mendélienne. Bien que l'on retrouve chez plusieurs membres du type HERV-K un ORF intact, ils sont généralement considérés comme inactifs. Cette absence de rétrotranscription n'exclue toutefois pas une activité transcriptionnelle dont les conséquences pathologiques notamment en cancérologie et autoimmunité restent débattues [142].

Les *long interspersed elements* (LINE):

Trois sous-types de rétrotransposons appartiennent à la catégorie des LINE : Les LINE-1 (L1), LINE-2 (L2) et LINE-3 (L3). Les L2 et L3 sont des séquences fossiles ayant fortement dérivé et n'ayant plus d'activité. Les L1, en revanche, sont les seuls rétrotransposons autonomes actifs connus chez l'humain et ils sont également ceux qui couvrent la plus grande part du génome. En effet, bien qu'il soient moins nombreux que les Alu dont nous parlerons ensuite, leur taille moyenne de 6kb leur permet de constituer environ 20% du génome humain [130,143]. Malgré cette fréquence élevée, le nombre de L1 réellement actifs est très limité et ne dépasserait pas 80 à 100 éléments [144]. Les L1 actifs abritent en 5'UTR un promoteur de l'ARN polymérase. Deux cadres de lectures sont traduits et codent des protéines indispensables à la rétrotransposition s'associant préférentiellement à leur propre ARN messager codant (cis-préférence). L'ORF1 code une protéine d'environ ~40 kDa (ORF1p) et l'ORF2 une endonucléase et rétrotranscriptase de ~150 kDa (ORF2p).

Les *short interspersed elements* (SINE):

Les SINE regroupent deux types d'éléments, les Alu et les SVA, tous deux propres à la lignée des primates. Les séquences **Alu** ont une taille moyenne de 300 pb ce qui en fait des éléments beaucoup plus petits que les LINE. Mais leur nombre élevé, estimé à ~1,1 millions chez l'humain [145], fait qu'ils constituent tout de même environ 10% du génome [130]. Ils sont également classés en sous-familles sur la base de leur dérive génétique. Les éléments les plus jeunes, les AluY, sont aussi les plus actifs et les AluYa5 et AluYa8 sont ceux les plus souvent impliqués dans les insertions à caractère pathogène [146] . Les Alu actifs sont constitués de deux monomères séparés par une séquence riche en

adénosine (Figure 16). Le monomère de gauche contient le promoteur interne de l'ARN polymérase III pour sa transcription. À la différence des L1, les Alu ne sont pas autonomes. Ils ne codent pas d'endonucléase ni de rétrotranscriptase et sont dépendant de l'ORFp2 de L1 pour leur mobilité [147]. Tous les Alu n'ont pas le même niveau d'activité et les éléments plus anciens sont généralement inactifs. Plus précisément, le niveau d'activité est inversement corrélé au nombre de changements nucléotidiques par rapport à la séquence consensus. Il a ainsi été rapporté que les Alu présentant plus de 10% de divergence avec le consensus étaient éteints [148]. Les seconds types de SINE, appelés **SVA** (SINE/variable number tandem repeat (VNTR)/Alu) sont des éléments composites contenant : Une répétition hexamérique de longueur variable (CCCTCT)n, une séquence inversée de type Alu, une région à nombre variable de répétitions en tandem, une séquence dérivée de l'extrémité 3′ d'un élément HERV-K10 (SINE-R) et une queue poly(A) [149]. Ils sont apparus dans la lignée des primates il y a environ 25 millions d'années. Comme pour les Alu, ils ont été classés en sous-familles. Les sous-familles SVA_A, SVA_B, SVA_C et SVA_D sont apparues avant la divergence entre l'homme, le chimpanzé et le gorille, tandis que les sous-familles SVA_E et SVA_F, plus récentes sont, elles, limitées à la lignée humaine. On estime leur nombre à moins de 3000 copies chez l'humain [150]. Comme les Alu, ils ne sont pas autonomes et dépendent de L1 pour leur rétrotransposition.

Rétrotransposons et pathologies humaines

La mobilisation des rétrotransposons est une source de modifications pathogènes du génome. Le premier cas de maladie directement causée par une insertion d'éléments mobiles (MEI) a été rapporté en 1988 par Kazazian et al. Il s'agissait d'une insertion *de novo* d'une séquence L1 chez un patient atteint d'hémophilie A (MIM : 306700) [151]. Les

MEI *de novo* se produisent avec des fréquences variables selon le type d'éléments. On les estime à 1/40 naissances pour les Alu, 1/63 naissances pour les L1 et 1/63 naissances pour SVA [152]. À la lecture de ces fréquences nous pouvons dire que ces mobilisations ne sont pas des événements anecdotiques. Néanmoins, l'impact direct des MEI sur le phénotype est à relativiser car la part de variants pathogènes causés par des insertions *de novo* d'éléments mobiles n'est estimée qu'à 0,3% [141]. Il est intéressant de noter que la fréquence des MEI *de novo* constitutionnels est vraisemblablement limitée par des mécanismes de défense de l'hôte. En effet, il existe des voies d'inactivation des ET (piRNA, méthylation de l'ADN, modification des histones) qui sont particulièrement actives dans la lignée germinale. Le défaut de ces voies et la réactivation des ET dans les cellules germinales est d'ailleurs, associé à une infertilité [153]. Dans une revue exhaustive parue en 2016, Hancks et Kazazian ont recensé 119 MEI pathogènes impliquant des L1, Alu et SVA. Cette vue d'ensemble a notamment permis de mettre en avant les variations du consensus du site de clivage par l'endonucléase de L1. En effet, ils ont noté que le site de clivage 5′-TTTT/AA-3′ n'est pas absolu et peut donc être mieux défini comme 5′YYYY/RR-3′ (où Y est pyrimidine et R une purine). On note parmi tous ces cas des événements récurrents et, plus précisément, des sites de récurrence. Ces hotspots sont la cible de L1 mais peuvent être l'objet d'une insertion de n'importe quel type de rétrotransposons et même, pour un cas, d'un processed-pseudogene [154]. Le gène NF1 impliqué dans la neurofibromatose de type 1 (MIM : 162200) est un exemple de hotspot de MEI. Plusieurs insertions pathogènes ont été rapportées chez des patients non apparentés [155,156] et en particulier six insertions différentes identifiées dans une région relativement petite de 1,5 kb comprenant trois sites d'intégration rencontrés deux fois chacun. Toutefois, malgré cette susceptibilité apparente, les MEI ne représenteraient qu'environ 0,4% de toutes les mutations touchant *NF1* [157].

Les insertions *de novo* de rétrotransposons ne représentent qu'une part des événements structuraux pathogènes. Leur homologie de séquence et leur fréquence dans le génome les rend potentiellement sujets à des recombinaisons pouvant conduire à l'apparition d'autres types de SV. Cela est particulièrement le cas des Alu dont l'homologie entre éléments est de 71 % en moyenne [158]. Par exemple, 45,5 % (20/43) des CNV décrits en 17p13.3 étaient médiées par des paires Alu, 56 % (9/16) dans le locus *FOXF1*, 68 % (39/57) dans le locus *SPAST*, 88 % (29/33) dans le locus *VHL* et 100 % (45/45) dans le locus *EPCAM* [159]. Par ailleurs, des simulations ont évalué à 82,8 % la part du génome humain à risque d'instabilité médiée par les LINE [160]. En revanche, au vu du faible nombre de copies de SVA dans le génome (moins de 1%) la probabilité de SV médiés par ces derniers est faible. Des cas ont toutefois été observés ponctuellement comme deux délétions intragéniques dans *NF2* responsables d'une Neurofibromatose de type 2 (MIM :) rapportés par Legoix et al [161].

3.2.3. Les duplications segmentaires ou low copy repeats (LCR)

Les duplications segmentaires (LCR), sont définies comme des segments d'ADN de plus de 1 kb présentant une homologie supérieure à 90% [162]. Elles peuvent être inter-chromosomique, entre chromosomes non-homologues, et intra-chromosomiques, voire sur une même bande (aussi appelées répétitions à faible nombre de copies). Pendant les deux dernières décennies, leur proportion dans le génome était estimée entre 3,5% et 5,2%. Ces estimations fluctuaient pour des raisons de divergence de méthodes, notamment sur le choix de la longueur des segments et du génome de référence [163–165]. De plus, les stratégies d'assemblage basées sur le SG *short read* montrent des

limites dans la différenciation entre séquences dupliquées hautement similaires et vrais chevauchements. Ceci laissait penser que la proportion de LCR était sous-estimée et que les régions non couvertes des génomes de référence pouvaient être enrichies en LCR. Plus récemment, avec la publication du génome de référence télomère à télomère sans gap (T2T-CHM13), ces estimations ont pu être revues avec plus de précision. La proportion de LCR a ainsi été portée à environ 7% [166]. Leur implication en pathologie humaine est bien documentée. Plusieurs remaniements à l'échelle chromosomique liées à des LCR et impliqués dans des TND ont été identifiés par analyses cytologiques (Tableau 4).

Syndrome	Réarrangement	Localisation	Taille (Mb)	Références
Syndrome de Smith Magenis	Délétion	7p11.2	5	[167]
Syndrome de Prader–Willi	Délétion	15q11–15q13	4	[168,169]
Syndrome d'Angelman	Délétion	15q11–15q13	4	[168,169]
Syndrome de Williams	Délétion	7 q11.23	1,6	[170,171]
Syndrome de délétion 22q11.2	Délétion	22q11.2	3	[172–174]

Tableau 4 : Exemples de syndromes liés à des SV médies par des LCR. Adapté de Emanuel & Shaikh, 2001.

3.3. Mécanismes d'apparition des SV

La plupart des variants structuraux résultent de mécanismes de recombinaison ou de réparation qui se produisent après une rupture de double brin (DSB) ou de mécanismes de réplication après une rupture ou un blocage de la fourche de réplication. La majorité

des DSB de l'ADN, qu'elles surviennent au cours d'un processus pathologiques ou physiologiques, sont réparées par un processus impliquant des jonctions non homologues (NHEJ). Ces réparations peuvent survenir à n'importe quelle étape du cycle cellulaire contrairement à celles faisant intervenir des régions homologues (HR) qui se limitent, elles, aux phases S/G2. Le processus qui détermine si la cellule utilisera la HR ou la NHEJ à la suite d'une DSB est imparfaitement connu. Une des hypothèses est que pendant la phase S, la chromatide soeur est physiquement très proche, fournissant ainsi un donneur d'homologie pour la HR. En dehors des phases S/G2, la NHEJ est donc privilégiée [175].

Trois mécanismes sont à l'origine de la majeure partie des SV de la lignée germinale : Les processus médiés par la micro-homologie impliquant de courtes séquences de 2 à 20 pb (MMEJ aussi appelé Alt-NHEJ) pour 28%, la recombinaison homologue non allélique (NAHR) pour 22% et la rétrotransposition d'éléments LINE-1 pour 19% [126]. Il existe néanmoins des disparités entre type de SV. Dans une étude récente, Porubsky *et al.* ont apporté des éléments d'observation à partir de SG *long read* et OGM qui tendent à confirmer que le mécanisme principal d'apparition des inversions est différent de celui des insertions et CNV. La plupart des inversions équilibrées résolues par séquençage (72 %) présentaient des régions inversées répétées flanquantes d'une longueur d'au moins 200 pb, suggérant un mécanisme d'apparition basé sur la NAHR ce qui dépasse les estimations pour les insertions et les délétions supérieures à 50 pb (15 %-25 %). Sur les inversions candidates NAHR, 77 % présentaient des LCR inversées flanquantes, tandis que les 23% restantes présentaient des séquences d'éléments mobiles (L1 et Alu) [176]. Ces nouveaux résultats contrastent avec les études plus anciennes qui privilégiaient des mécanismes basés sur le NHEJ [177] . Nous pouvons diviser les SV en deux groupes

selon leur récurrence. Les réarrangements récurrents sont identiques en taille et en séquence à un *locus* donné chez des individus non apparentés. La NAHR est généralement le mécanisme responsable de ces réarrangements dont les points de cassure sont dans des régions de forte homologie (LCR ou transposons). Les réarrangements non récurrents, quant à eux, sont des réarrangements dont les séquences et les tailles à un *locus* donné sont uniques chez des individus non apparentés (Figure 17). Pour autant, les réarrangements non récurrents ne surviennent probablement pas de façon totalement aléatoire. Des organisations génomiques particulières, telles que des séquences répétées susceptibles de former des structures non-B d'ADN sont fréquemment observées en association avec des points de cassure de SV non récurrents. Le mécanisme est imparfaitement connu mais une des hypothèses est qu'ils rendent certaines régions susceptibles de former des structures secondaires qui peuvent conduire à l'effondrement des fourches de réplications, à la formation de cassures doubles brins ou à une perturbation de la progression de l'ADN polymérase [178].

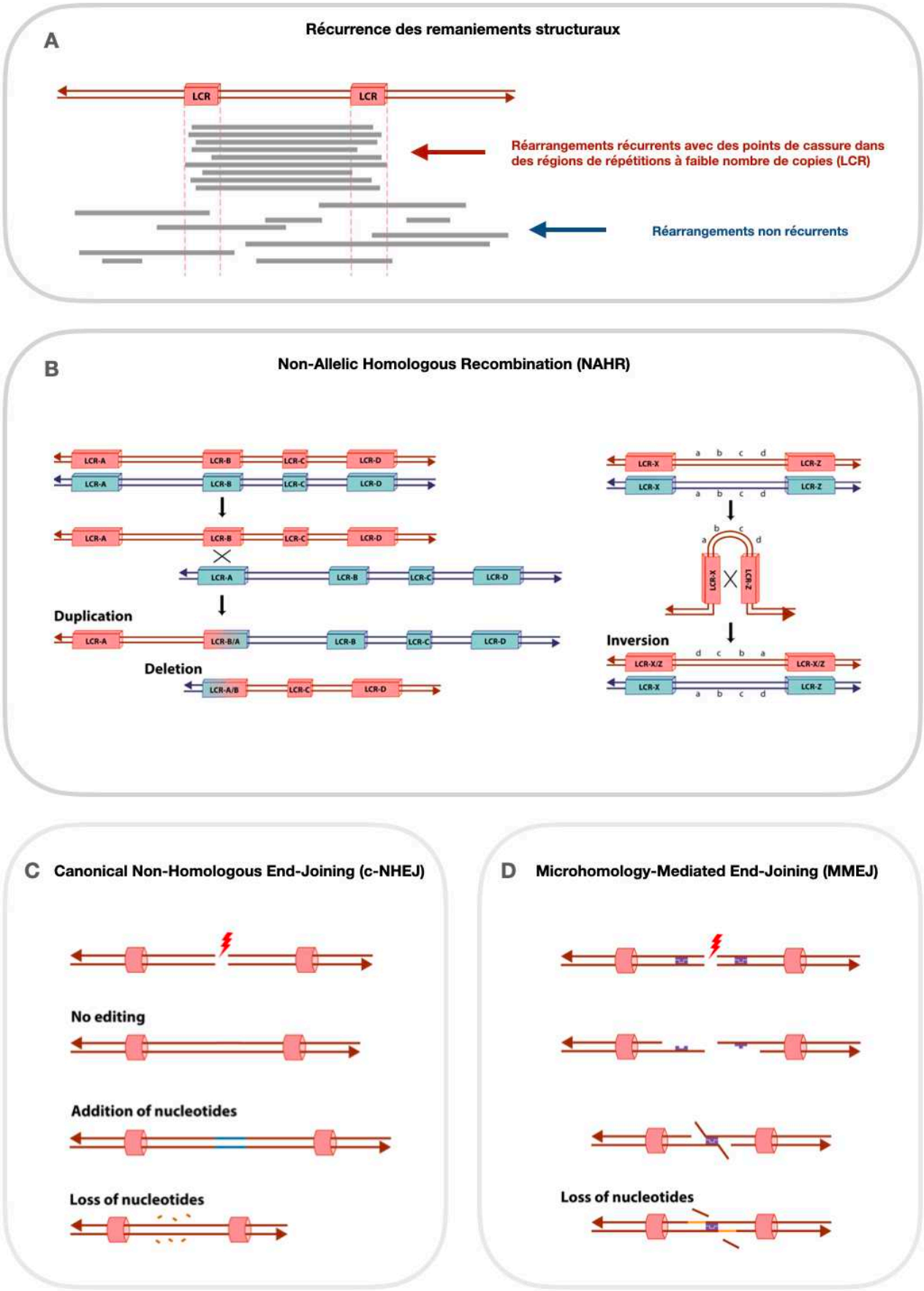


Figure 17 : Principaux mécanismes d'apparition des variants structuraux. Adapté de Burssed *et al*, 2022.

3.3.1. Canonical non-homologous end joining (c-NEHJ)

Dans la voie de réparation des DSB par c-NEHJ, lorsqu'une cassure double brin est détectée un hétérodimère des protéines Ku70 et Ku80 se lie à chaque extrémité, ce qui permet de recruter les nucléases, polymérases et ligases nécessaires à la réparation. Le c-NHEJ est un processus imprécis, et des gains ou pertes de nucléotides peuvent être observées aux points de jonction. Par ailleurs, pour augmenter l'efficacité de la ligature des extrémités, des nucléotides peuvent être ajoutés afin de créer une microhomologie. Ces pertes ou gains de nucléotides, en l'absence de microhomologies préexistante, sont d'ailleurs considérés comme une signature en faveur des mécanismes d'apparition de SV médiés par des NEHJ [179].

3.3.2. Microhomology-mediated end-joining (MMEJ)

Il a été observé qu'en l'absence des gènes impliqués dans la voie c-NHEJ, un processus de *end-joining* alternatif le remplaçait. Cette voie est appelée *alternative nonhomologous end-joining* (Alt-NHEJ) ou *microhomology-mediated end-joining* (MMEJ). Chez beaucoup d'auteurs, ces termes sont considérés comme interchangeables, mais d'autres introduisent une distinction. En effet, selon eux, cette voie est caractérisée avant tout par son indépendance aux protéines de la c-NHEJ (Ku-indépendante), et la présence d'une microhomologie, quoi que fréquente, n'est pas systématique [180]. Par ailleurs, la voie c-NHEJ peut être utilisée même en présence d'une microhomologie, bien que celle-ci ne soit pas nécessaire au processus. Cette voie, tout comme la c-NEHJ est susceptible de commettre des erreurs, tout particulièrement des délétions [181]. Lorsqu'une microhomologie est présente, elle est d'une longueur de 2 à 10 pb. Au-delà, ce sont les mécanismes de HR qui deviennent prépondérants (Figure 18). Les processus de

réparations HR se font sans retrait de nucléotides au site de cassure, ils sont donc intrinsèquement moins source d'erreurs au site de réparation. Toutefois, comme nous le verrons avec le NAHR, ils peuvent être sources de recombinaisons pathogènes générant des SV.

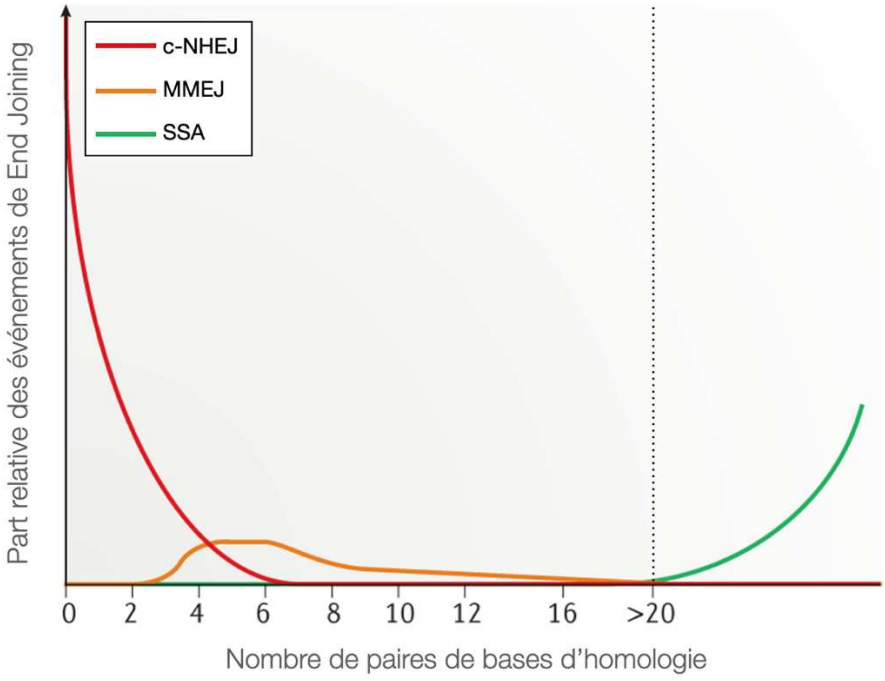


Figure 18 : NEHJ. Mécanisme et part relative en sein des événements de end joining. Adapté de Chang *et al*, 2017.

3.3.3. Non-allelic homologous recombination (NAHR)

La plupart des SV récurrents sont formés par un mécanisme impliquant les LCR appelé NAHR. Comme nous l'avons vu, les LCR sont des régions instables du fait de leurs tailles et de leurs degrés d'homologie qui les rendent susceptibles aux réarrangements génomiques. La NAHR peut survenir lorsque des copies non-alléliques de LCR s'alignent par erreur en raison de leur haut niveau d'identité. Ce cross-over inégal conduit aux

remaniements qui génèrent les SV. Les NAHR entre LCR, intrachromosomique ou interchromatidienne, de même orientation, causent des duplications et des délétions. Les inversions, quant à elles, sont formées lorsque les LCR se trouvent dans l'orientation opposée. Il en va de même pour la NAHR entre LCR intrachromatides, mais, dans ce cas, aucune duplication ne semble se former. Enfin, les NAHR entre des LCR interchromosomiques peuvent être à l'origine de translocations [127,179,182,183].

3.3.4. Mécanismes basés sur la réplication

Des mécanismes basés sur la réplication ont été proposés pour expliquer les réarrangements complexes notamment ceux avec de multiples points de cassure, des insertions de segments d'ADN et une microhomologie. Ils sont dus à un changement de matrice d'ADN simple brin au cours de la réplication survenant par exemple après un blocage de la fourche de réplication [178]. Le blocage de la fourche de réplication peut conduire au mécanisme appelé *Fork Stalling and Template Switching* (FoSTeS). Au cours de la réplication, la fourche de réplication s'arrête à un endroit. Le brin retardé se détache de la matrice originale et, en raison de la présence d'une microhomologie, passe à une autre matrice au niveau d'une autre fourche de réplication active et redémarre la synthèse de l'ADN. Dans la nouvelle fourche, l'ADN est copié et le brin naissant peut à nouveau se désengager et envahir d'autres fourches de réplication ou retourner à sa fourche d'origine, tout cela grâce à la présence de microhomologie. L'effondrement de la fourche de réplication, quant à lui, est dû à une rupture de la fourche au cours de la réplication. Une cassure double brin à extrémité unique se forme et la résection de l'extrémité 5′ génère un *overhang* 3' libre qui peut envahir une autre région génomique [182]. Deux mécanismes peuvent être à l'œuvre. Le premier, la réplication induite par la rupture (BIR),

nécessite de longs tronçons de séquences d'ADN homologues. Le second, la réplication induite par la rupture médiée par la microhomologie (MMBIR), lui, nécessite la présence d'une microhomologie. Comme dans le cas du FoSTeS, de multiples dissociations et invasions peuvent avoir lieu jusqu'à ce que la fin d'un chromosome soit atteinte. Pour tous ces mécanismes basés sur la réplication, le passage à une fourche de réplication située en aval, entraîne une délétion, tandis que le passage à une fourche située en amont entraîne une duplication. Le passage à un chromosome différent entraîne des translocations. Des segments inversés peuvent également être formés [179].

3.4. Outils bio-informatiques de détection des SV

Comme nous l'avions déjà souligné, la méthode de *calling* des SV est moins consensuelle que celle pour les SNP/indels. Plus précisément, il n'existe pas aujourd'hui d'outil polyvalent capable de détecter tous les types de SV. Il est donc nécessaire d'en combiner plusieurs pour obtenir la meilleure performance. Les algorithmes actuels reposent principalement sur la détection de trois signaux : La discordance entre paires de *reads*, les *reads* scindés (*split reads*) et les variations de profondeur de lecture (*read depth*) [105]. Les paires de *read* sont obtenues par séquençage d'un même fragment d'ADN suivant les deux orientations. La distance séparant les deux paires est donc connue. Les paires discordantes peuvent donc être des paires dont la distance est incohérente avec celle attendue, ce qui peut suggérer un CNV, sur ces chromosomes différents (translocation) ou avec une orientation incohérente (inversion). Les mêmes inférences peuvent être faites à partir des *split reads* qui sont des *reads* fragmentés en deux segments ou plus. Les outils basés sur ces deux sources d'information sont, entre autres : MANTA [184], DELLY [185] et LUMPY [186]. Ces algorithmes présentent

l'avantage de pouvoir déterminer assez précisément les coordonnées des points de cassure (principalement grâce aux *split read*). En revanche, en contrepartie de cette capacité, ils se montrent peu performants dans la détection d'événements dont les points de cassure sont dans de grandes régions répétées. Les méthodes reposant sur les variations de *read depth* permettent de contourner ce problème pour la détection de CNV. Ce type d'algorithme suppose une distribution aléatoire de la *read depth* et cherche des écarts à cette distribution. Le principe est que les régions dupliquées présentent une *read depth* significativement plus élevée et les délétions une *read depth* plus faible. Toutefois, cela se fait au prix d'une grande incertitude sur les coordonnées des points de cassure [105]. Les outils basés sur la *read depth* sont, entre autres : CNVnator [187], Canvas [188] ou GATK4. Des outils spécifiques ont par ailleurs été développés pour les insertions d'éléments mobiles dont la détection est particulièrement délicate du fait de l'homologie entre ces éléments. Ces algorithmes, tels que Mobster [189] et MELT [190], ajoutent à certains des signaux précédemment évoqués la consultation d'une base d'éléments transposables pour caractériser plus précisément les MEI.

Chapitre 4. Le génome non-codant

Bien qu'il fût admis depuis longtemps qu'une majeure partie du génome ne code pas de protéine, l'estimation de la proportion exacte a fluctué au gré des capacités techniques du moment. Mais, en 2003, le projet international de séquençage du génome humain (Human Genome Project, officiellement lancé en 1990) a permis de la quantifier avec précision en montrant que les séquences exoniques ne représentaient qu'environ 1,5 à 2%. La question des régions répétées et plus généralement du génome non-codant ne peut donc pas être dissociée du SG. Il apparait comme une évidence qu'un des

avantages du SG sur le SE est la détection de variants dans des séquences uniques non-codantes comme les micro-ARN, long ARN non-codants ou autres régions régulatrices.

Par le passé, l'intérêt de ces régions non-codantes a fait débat parmi les généticiens, certains les considérant comme des vestiges moléculaires de l'évolution sans aucune fonction. D'ailleurs, en 1972, Susumu Ohno inventa le terme de « junk DNA » (ADN poubelle) pour décrire l'ensemble de ces régions [191]. En 2012, le projet ENCODE (Encyclopedia of DNA Elements) a permis de cartographier systématiquement les régions de transcription, d'association de facteurs de transcription, de structure de la chromatine et de modification des histones. Étonnamment, une importante partie du génome humain, 80,4%, est couverte par au moins un élément identifié par ENCODE. Il a ainsi été montré que 62% du génome est en réalité transcrit [192]. Cette découverte met définitivement un terme au concept de « l'ADN poubelle » et, dans le même temps, apporte de nouvelles questions pour la compréhension des maladies génétiques. Afin d'avoir une vision plus précise du génome non-codant, nous allons maintenant détailler certains de ces éléments répétés ou uniques.

4.1. Les pseudogènes

Les pseudogènes ont été définis à l'origine comme des gènes aberrants présentant une grande similarité de séquence avec les gènes fonctionnels, mais ayant perdu leur capacité de codage, principalement en raison de la présence de codons stop prématurés ou d'un décalage du cadre de lecture [193]. Toutefois, la question d'un rôle fonctionnel de ces gènes a par la suite été soulevée, notamment en étudiant la conservation de ces séquences chez les primates non humains. Des taux de mutation et des contenus en GC

différents des régions inter-géniques environnantes ont apporté la preuve d'une pression de sélection compatible avec une fonction biologique [194].

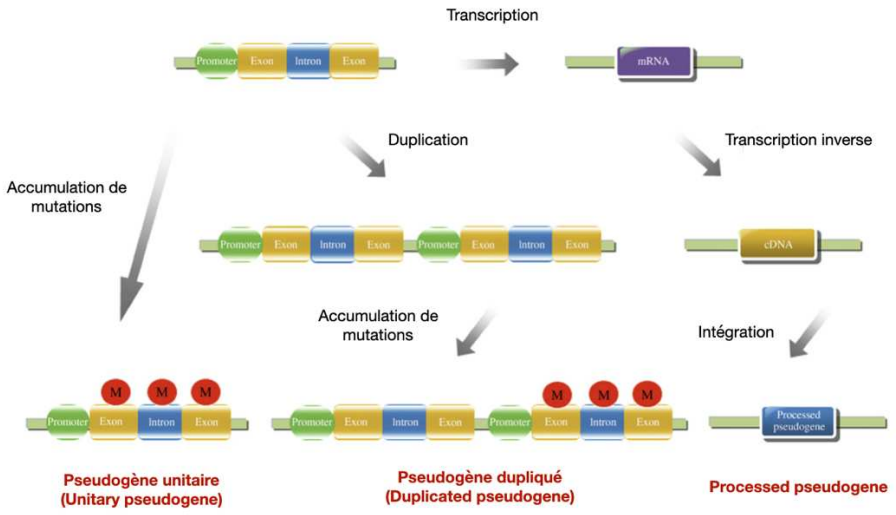


Figure 19 : Les types de pseudogènes et leurs mécanismes d'apparition. Adapté de Li et al, 2013

La version 34 du Genecode Project [195] référence un peu plus de 15 000 pseudogènes dont plus de 10 000 *processed pseudogenes*. Des études protéomiques ont rapporté qu'au moins 140 pseudogènes étaient traduits en peptides. Si l'hypothèse d'une action au niveau protéique reste à démontrer, plusieurs pseudogènes ont déjà une fonction régulatrice avérée.

4.2. Les petits ARN

Petits ARN nucléaires

Les snRNA constituent un petit groupe de transcrits non-codants, non polyadénylés, très abondants et présents dans le nucléoplasme. Les snRNA peuvent être divisés en deux

classes, Sm et Lsm, sur la base de caractéristiques de séquence communes et de cofacteurs protéiques. La classe Sm des snRNA comprend U1, U2, U4, U4atac, U5, U7, U11 et U12, tandis que la classe Lsm est composée de U6 et U6atac. À l'exception de la snRNP U7, qui intervient dans le traitement du pré-ARNm par les histones, les autres snRNP riches en uridine sont des éléments des splicéosomes et catalysent l'élimination des introns du pré-ARNm [196]. Deux splicéosomes ont été observés et diffèrent dans leur composition et dans la nature des sites d'épissage reconnus. Ils ont été appelés splicéosome « majeur » et « mineur » du fait des différences de fréquences des site reconnus. En effet, chez l'humain, seulement 700 à 800 gènes contiendraient des introns ciblés par le splicéosome mineur. Le splicéosome majeur (U2-dependant) contient les snRNA U1, U2, U4, U5, et U6 tandis que le splicéosome mineur (U12-dependent) contient les snRNA U11, U12, U4atac, et U6atac. Les introns de type U2 sont de type GT-AG, alors que les introns de type U12 appartiennent pour la plupart aux sous-types AT-AC ou GT-AG (Figure 20). Quelques exemples d'introns de type U12 avec d'autres combinaisons de résidus terminaux ont été rapportés expérimentalement mais avec une efficacité réduite [197].

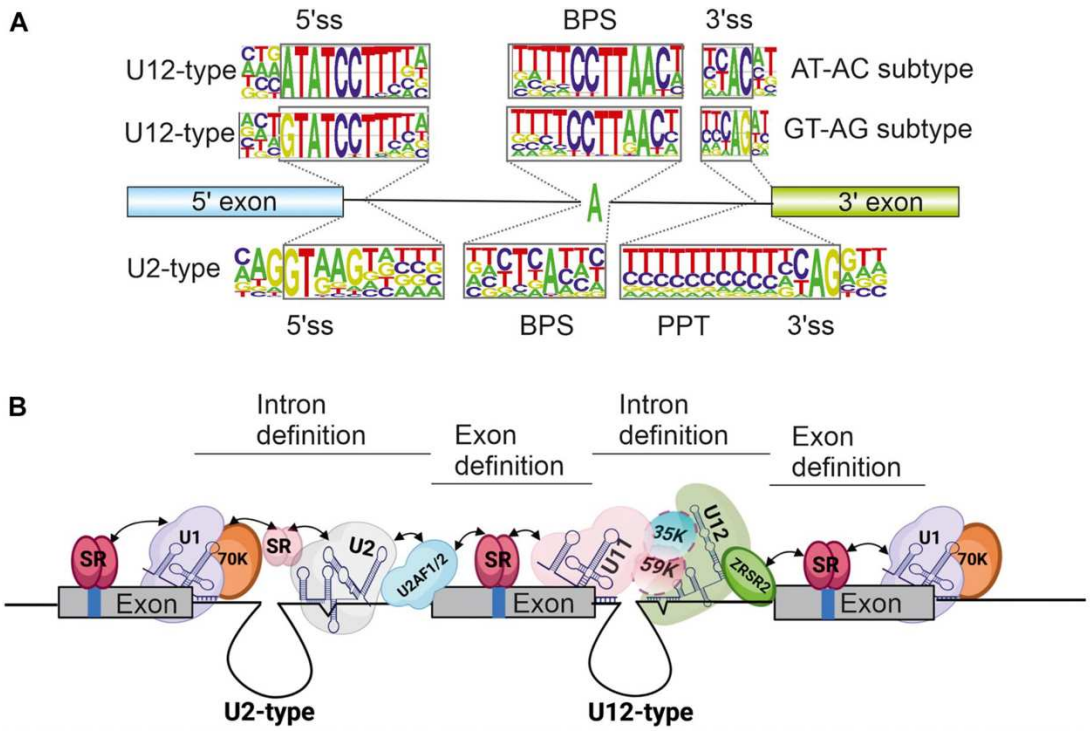


Figure 20 : Séquences consensus des sites d'épissage des splicéosomes mineurs et majeurs. Tiré de Akinyi et Frilander, 2021

À ce jour, quatre pathologies sont associées à des variants dans deux snRNA constituant le splicéosome mineur. L'ataxie spinocérébelleuse 33 (MIM : 620208) causée par des mutations bialléliques de RNU12 [198] ainsi que le syndrome de Lowry-Wood (MIM : 226960) [199], le syndrome de Roifman (MIM : 616651) [200] et le syndrome de Taybi-Linder (MIM : 210710) [201] tous trois causés par des variants bialléliques du gène RNU4ATAC.

Petits ARN nucléolaires

Les petits ARN nucléolaires (snoRNA) sont de petits ARN non codants d'une longueur de 60 à 300 nt largement présents dans les nucléoles des cellules eucaryotes. Les gènes exprimant les snoRNA sont situés dans des régions introniques de gènes codants ou dans des régions non-codantes. Ils peuvent être classés en trois groupes : les snoRNA à boîte H/ACA, les snoRNA à boîte C/D et les petits ARN cajal (scaRNA). Les deux premiers

types participent à la maturation des ARN ribosomiques [202]. À ce jour, peu de snoRNA ont été associés à des TND. Ont toutefois été rapportés des variants pathogènes bialléliques dans *SNORD118* responsables d'une microangiopathie cérébrale avec calcifications et kystes (MIM : 614561) [203] et des délétions de *SNORD116-1* identifiées chez des enfants présentant un phénotype cliniques assimilé au syndrome de Prader-Willi (MIM : 176270) [204]. *SNORD116-1* pourrait donc jouer un rôle dans la physiopathologie du syndrome de Prader-Willi (MIM : 176270).

ARN intéragissant avec Piwi

Les piRNA (PIWI-interacting RNAs) sont une classe de petits ARN régulateurs portant des terminaisons 3′ modifiées en 2′-O-méthyle et donc le rôle est de guider un type de protéine argonaute spécifique de type PIWI. Ils sont spécifiques aux animaux mais la conservation des séquences est faible et les voies de biogénénèse peuvent être différentes. Ils sont, en partie, originaires d'un cluster riche en piRNA mais d'autres sources génomiques sont également rapportées comme une biogénèse à partir du *processing* d'ARN de transfert, les *tRNA-derived fragments* (tRFs) [205]. Un des rôles le mieux compris des piRNA, et cela pour de nombreuses espèces est le *silencing* des rétrotransposons dans la lignée germinale. Toutefois, ce n'est probablement pas leur seule fonction. En effet, de nombreux piRNA correspondent à des séquences génomiques uniques non liées à des éléments transposables. Des preuves de plus en plus nombreuses suggèrent donc qu'ils réguleraient également l'expression des ARNm de l'hôte [206]. À ce jour, aucune association entre des variants dans des gènes précurseurs de piRNA et des TND n'ont été rapportés.

Les microARN :

Les microARN (miARN) sont de petits ARN non-codants d'une longueur d'environ 22 nucléotides. Ils jouent le rôle de molécules guides dans le *silencing* des ARNm. Ils ciblent la plupart des transcrits codant des protéines, et sont donc impliqués dans presque tous les processus du développement. Leurs réseaux d'action sont complexes, un miRNA peut avoir plusieurs ARNm cibles et un ARNm peut être régulé par plusieurs miRNA [207,208]. Plusieurs voies de biogenèse ont été rapportées. Environ la moitié de tous les miRNA connus sont intra-géniques et principalement issus des introns (mirtrons) de gènes codant des protéines. Dans ce cas, il est généralement transcrit par l'ARN polymérase II en même temps que le gène hôte et fait l'objet d'une première maturation en précurseurs de miRNA (pré-miRNA) par le splicéosome puis la debranching enzyme 1 (DBR1). Les miRNA inter-géniques sont, quant à eux, transcrits par l'ARN polymérase III, indépendamment d'un gène hôte et régulés par leurs propres promoteurs. Dans cette voie, la maturation des miRNA primaires (pri-miRNA) en pré-miRNA est effectuée par un complexe composé de la protéine DGCR8 et de la ribonucléase Drosha. Les deux voies se rejoignent au moment de l'exportation des pré-miRNA dans le cytoplasme, médiée par l'Exportine 5. Les pré-miRNA sont clivés par la ribonucléase Dicer pour produire un duplexe miRNA mature. Un des brins (5p ou 3p) est ensuite chargé dans une protéine Argonaute (AGO1-4 chez l'humain) pour former le miRNA-induced silencing complex (miRISC) [209] (Figure 21).

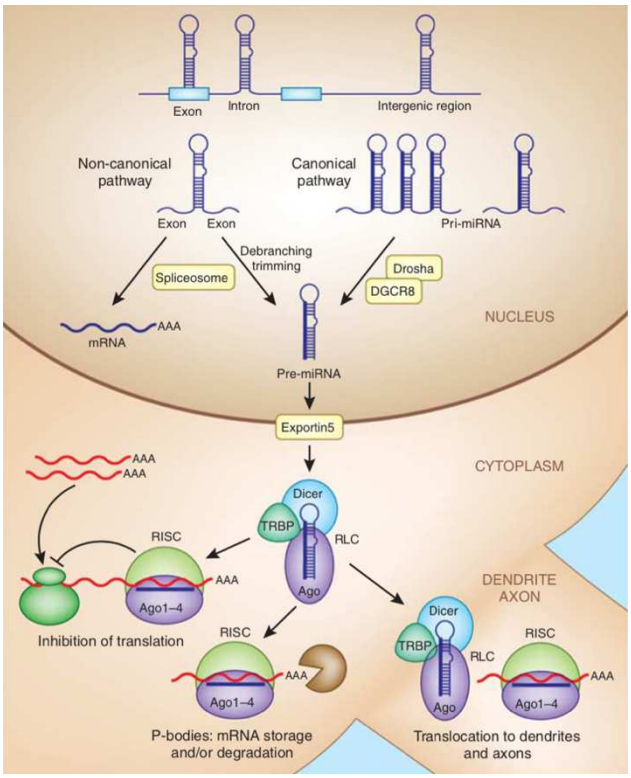


Figure 21 : Aperçu schématique de la biogénèse des miRNA dans le cerveau. Tiré de O'Carroll & Schaefer, 2013.

L'action régulatrice du complexe miRISC repose sur son interaction avec des séquences complémentaires en 3'UTR de l'ARNm cible. La reconnaissance de la cible repose sur une région cruciale appelée « graine » impliquant les nucléotides 2 à 8 du miRNA. Dans les faits, cette graine peut impliquer les nucléotides 2-8, 2-7 et 2-6 [210]. C'est le degré de complémentarité globale (graine et régions supplémentaires) du miRNA avec son ARNm cible qui détermine le mode de régulation post-transcriptionnel. Un faible degré de complémentarité entraine une inhibition de la traduction tandis que les degrés de complémentarité élevés conduisent à la dégradation de l'ARNm cible par clivage enzymatique [211]. Bien que les miRNA soient généralement associés à une inhibition de leurs gènes cibles, quelques études ont montré un effet activateur pour certains d'entre eux. Des cas d'activation de la traduction ont été observés par association de miRNA aux

protéines AGO2 et Fragile-x-mental retardation related protein 1 (FRX1). Il a ainsi été montré que Let-7 et miR369-3 pouvaient avoir un rôle inhibiteur ou bien activateur de la traduction selon l'étape du cycle cellulaire [212]. Toutefois, dans l'état actuel des connaissances, le rôle activateur semble relativement rare et c'est l'inhibition qui reste la règle.

Un autre élément s'ajoute à la complexité des réseaux de régulation des miRNA : Ils sont eux-mêmes soumis à une régulation par des ARN endogènes concurrents (*competing endogenous RNA* : ceRNA). Ces ARN, appartenant au groupe des longs ARN non-codants, peuvent entraver l'activité des miRNA en les séquestrant (éponges à miRNA) ce qui augmenterait l'expression de leurs gènes cibles (Figure 22).

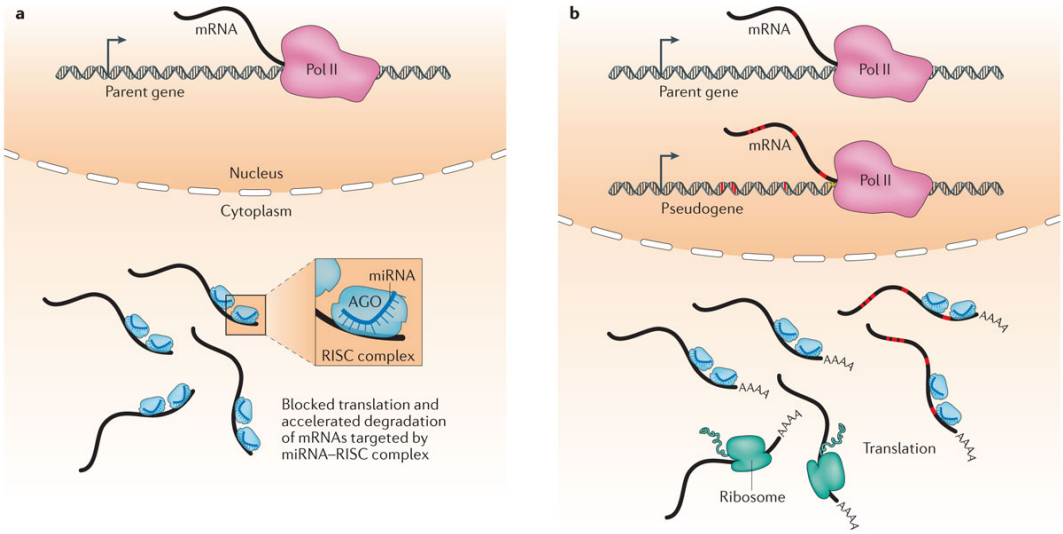


Figure 22 : Régulation des miRNA par des ARN endogènes concurrents. Tiré de Thomson & Dinger, 2016).

Deux classes de ceRNA sont rapportés dans les études comme agissant en tant que ceRNA fonctionnels : les transcrits dérivés de pseudogènes et les ARN circulaires (circRNA). L'effet de séquestration est aujourd'hui bien établi mais l'impact fonctionnel sur l'expression des gènes cibles reste toutefois discuté par certains auteurs [213].

Le premier miRNA à avoir été associé à une maladie mendélienne est le miR-96. Deux variants, miR-96(n.13G>A) et miR-96(n.14C>A) ségrégant avec une surdité non syndromique (MIM: 611606) ont été rapportés dans deux familles par Mencia et al. en 2009 [214]. Ces mutations situées dans la graine de miR-96 ont un fort impact sur sa biogénèse mais aussi sur les capacités de ciblages de l'ARNm. Par la suite, une autre mutation miR-96(n.57T>C) a été associée à cette maladie. Cette dernière est localisée en dehors de la séquence du miR-96 mature et le mécanisme pathogène repose alors sur une réduction de la stabilité de l'épingle à cheveux du pré-miRNA et donc une baisse du niveau d'expression du miR-96 [215]. Un même variant pathogène dans le miR-184 (n.57C>T) a été identifié dans trois familles américaine, irlandaise et espagnole non apparentées atteintes du syndrome EDICT. Un syndrome autosomique dominant caractérisé par une dystrophie endothéliale, une hypoplasie de l'iris, une cataracte congénitale et un amincissement du stroma cornéen [216–218]. En 2015, Conte *et al.* ont associé une dystrophie rétinienne et le colobome de l'iris avec ou sans cataracte congénitale (MIM 616722) à une mutation hétérozygote du miR-204 (n.37C>T) [219]. Plus récemment, en 2019, une nouvelle dysplasie spondylo-épiphysaire (MIM: 618618) a été identifiée dans trois familles non apparentées dans le cadre d'un projet de diagnostic moléculaire des troubles squelettiques congénitaux ultrarésistants. Une même substitution hétérozygote située au niveau du premier nucléotide de la graine du miARN miR-140-5p hautement conservée a été identifiée chez deux patients

(MIR140:NR_029681.1:n.24A>G). Le mécanisme pathogène serait probablement dû à un gain de fonction. En effet, cette graine mutée, n'est pas partagée avec d'autres miRNA connus et l'analyse informatique a suggéré que le complexe miR-140(:n.24A>G) AGO2 entrerait en compétition avec la RNA-binding protéine YBX1 pour les mêmes sites de liaison à l'ARN d'où un impact négatif sur la stabilisation des ARNm et la régulation de l'épissage des ARN cibles de YBX1 [220].

4.3. Les longs ARN non-codants (lncRNA)

Les long ARN non-codants (lncRNA) sont des ARN d'une longueur supérieure à 200 nt et qui sont supposés ne pas coder de protéine d'une taille supérieure à 100 acides-aminés. Chez l'humain, on en dénombre entre 30 000 et 60 000 [221] et, contrairement aux gènes codants, la majorité d'entre eux n'est pas conservée entre espèces [222]. L'absence de conservation pourrait suggérer, à première vue, que ces RNA n'ont pas de rôle fonctionnel. Toutefois, cette donnée est à relativiser. En effet, si leurs séquences ne sont souvent que peu ou pas conservées, leurs structures le sont parfois [223]. De plus, les preuves d'un rôle fonctionnel de lncRNA dont la séquence n'est pas conservée ont d'ores et déjà été apportées [224,225]. Leur fonction ne reposerait donc pas uniquement sur la nature de leur séquence mais aussi sur leur organisation structurelle. Tous les lncRNA n'ont pas une organisation moléculaire identique. Environ 50 % d'entre eux possèdent une queue polyA, 98 % sont épissés [226] et beaucoup possèdent également des coiffes m7G. Ces caractéristiques, semblables à celles des ARNm, permettent à une catégorie de lncRNA de quitter le noyau et rejoindre le cytoplasme où ils peuvent interagir avec les ARNm pour réguler leur traduction. Une des classifications des lncRNA est basée sur leur localisation génomique (Figure 23). Les lncRNA intergéniques (lincRNA) sont des lncRNA

dont les gènes ne se chevauchent avec aucun autre gène et qui sont éloignés de plus de 1 kb des gènes voisins. De nombreux lncRNA chevauchent d'autres gènes et sont orientés en sens ou anti-sens par rapport à eux. Ils peuvent chevaucher des régions exoniques ou introniques. Les lncRNA anti-sens exoniques peuvent naturellement s'hybrider avec une région des ARNm du gène qu'ils chevauchent. Enfin, les lncRNA bidirectionnels sont transcrits à partir du même promoteur qu'un autre gène, mais dans la direction opposée. Les gènes intergéniques et anti-sens sont les types les plus courants chez l'homme [216]. D'autres classifications, basées sur la fonction ont été proposées mais nécessitent une validation expérimentale [227].

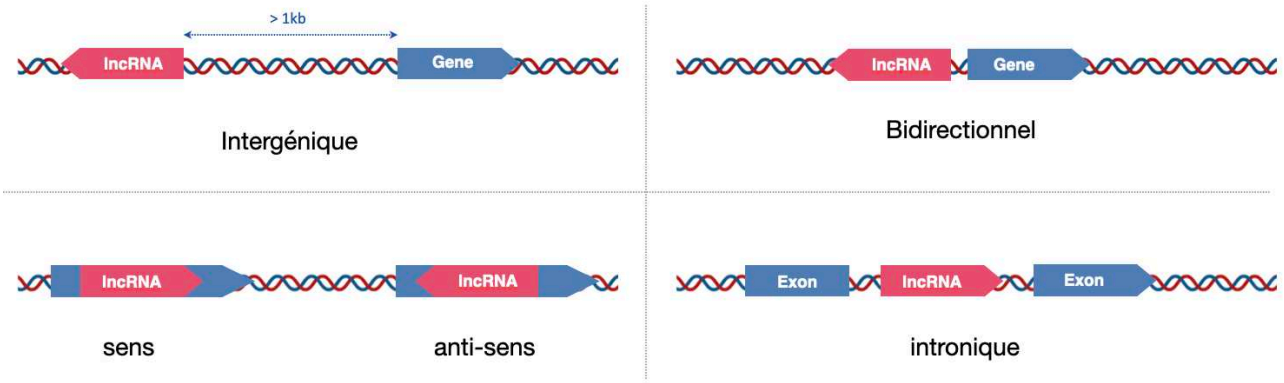


Figure 23 : Différents types de lncRNA

Mécanismes de régulation par les lncRNA nucléaires

Plusieurs mécanismes de régulation transcriptionnelle par des lncRNA nucléaires ont été rapportés dans la littérature. Ils peuvent impliquer des **interactions ARN-ADN**, comme avec le lncRNA PAPAS. Ce lncRNA est transcrit en anti-sens d'un gène précurseur d'ARN ribosomique avec lequel il s'hybride sous forme d'un triplex ARN-ADN puis recrute le remodeleur de chromatine CHD4/NuRD pour inhiber l'expression de son gène hôte [228]. Des mécanismes d'**interactions ARN-ARN** ont également été rapportés expérimentalement comme par exemple pour le lncRNA MEG3. Dans une région de

MEG3 conservée au cours de l'évolution, deux motifs distaux interagissent par complémentarité pour former des structures de pseudo-nœuds alternatives et mutuellement exclusives ("*kissing loops*"). Cette nouvelle structure permet d'interagir avec des facteurs de transcription comme p53. Les mutations qui perturbent la formation des *kissing loops* empêchent la stimulation de p53 dépendante de MEG3 in vivo et perturbent le repliement de MEG3 in vitro [229]. Enfin, des **interactions ARN-protéines** ont été décrites, comme les modifications tridimensionnelles favorisant les interactions *enhancer-promoteur* au sein des *topologically associating domains* (voir partie 4.4) médiées par le lncRNA HOTTIP, ou la séquestration de protéines inhibitrices de la transcription par le lncRNA Jpx (Figure 24). Il est important de noter que ces catégories d'interactions ne sont pas mutuellement exclusives.

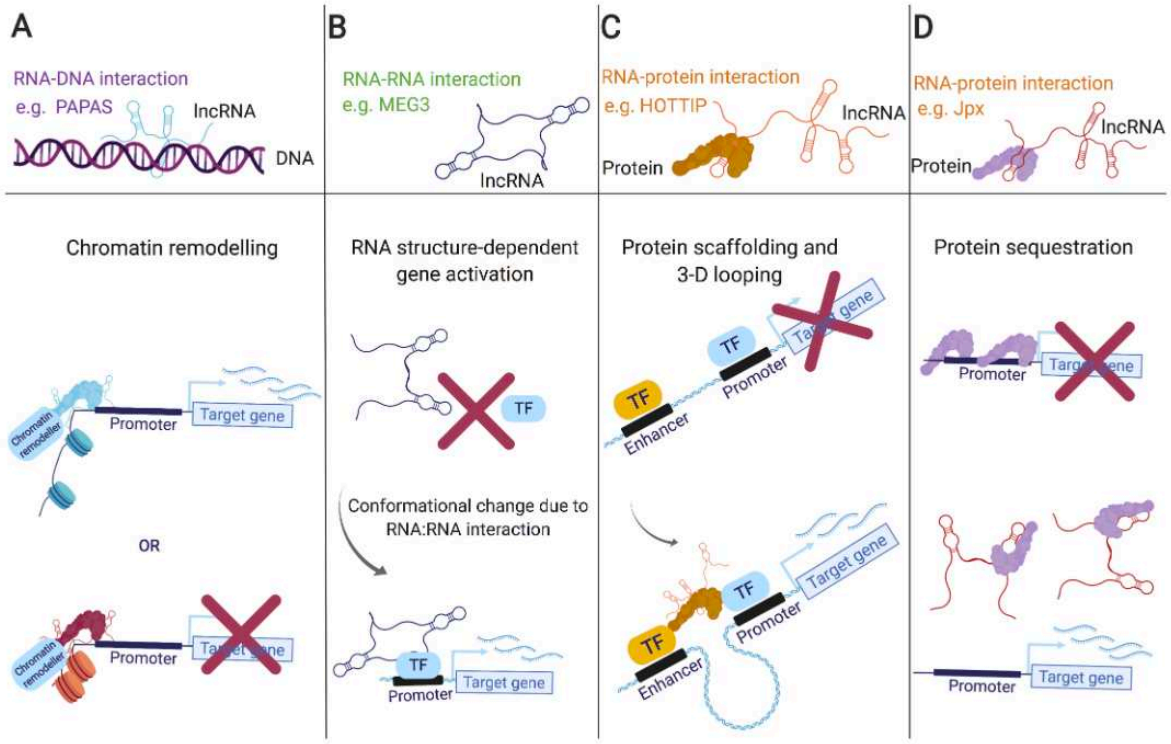


Figure 24 : Mécanismes de régulation par les lncRNA nucléaires. Tiré de Tsagakis et al, 2020.

Mécanismes de régulation par les lncRNA cytoplasmiques

Les lncRNA aptes à rejoindre le cytoplasme peuvent assurer une régulation post-transcriptionnelle suivant plusieurs mécanismes. Ils peuvent notamment former des complexes ribonucléoprotéiques comme c'est le cas, par exemple, avec *LncMoyD*. *LncMyoD* est localisé à proximité du gène *MyoD* par lequel il est directement activé durant la différenciation des myoblastes. L'inactivation de *LncMyoD* inhibe fortement la différenciation musculaire terminale. Il se lie directement à la protéine 2 de liaison à l'ARN de l'IGF2 (IMP2) et régule négativement la traduction des gènes de prolifération, tels que *N-Ras* et c-*Myc,* médiée par IMP2. Bien que la séquence d'ARN de LncMyoD ne soit pas bien conservée entre l'homme et la souris, son locus, la structure de son gène et sa fonction sont préservés [230]. Un autre mécanisme régulateur a déjà été évoqué brièvement dans la partie 4.2 : Les *competing endogenous RNA* (ceRNA)*.* Cette catégorie de lncRNA présente des sites d'hybridation avec des miRNA eux-même régulateurs post-transcriptionnels. L'hybridation des ceRNA avec un *pool* de miRNA entraine une déplétion de ces derniers et donc une surexpression de leurs cibles. Ce type de lncRNA est parfois appelé « éponge à miRNA ». Ce mécanisme a, par exemple, été observé avec linc-ROR qui lève la répression de SOX9 en hybridant directement plusieurs miARN, dont miR-15b, miR-33a, miR-129, miR-145 et miR-206 [231].

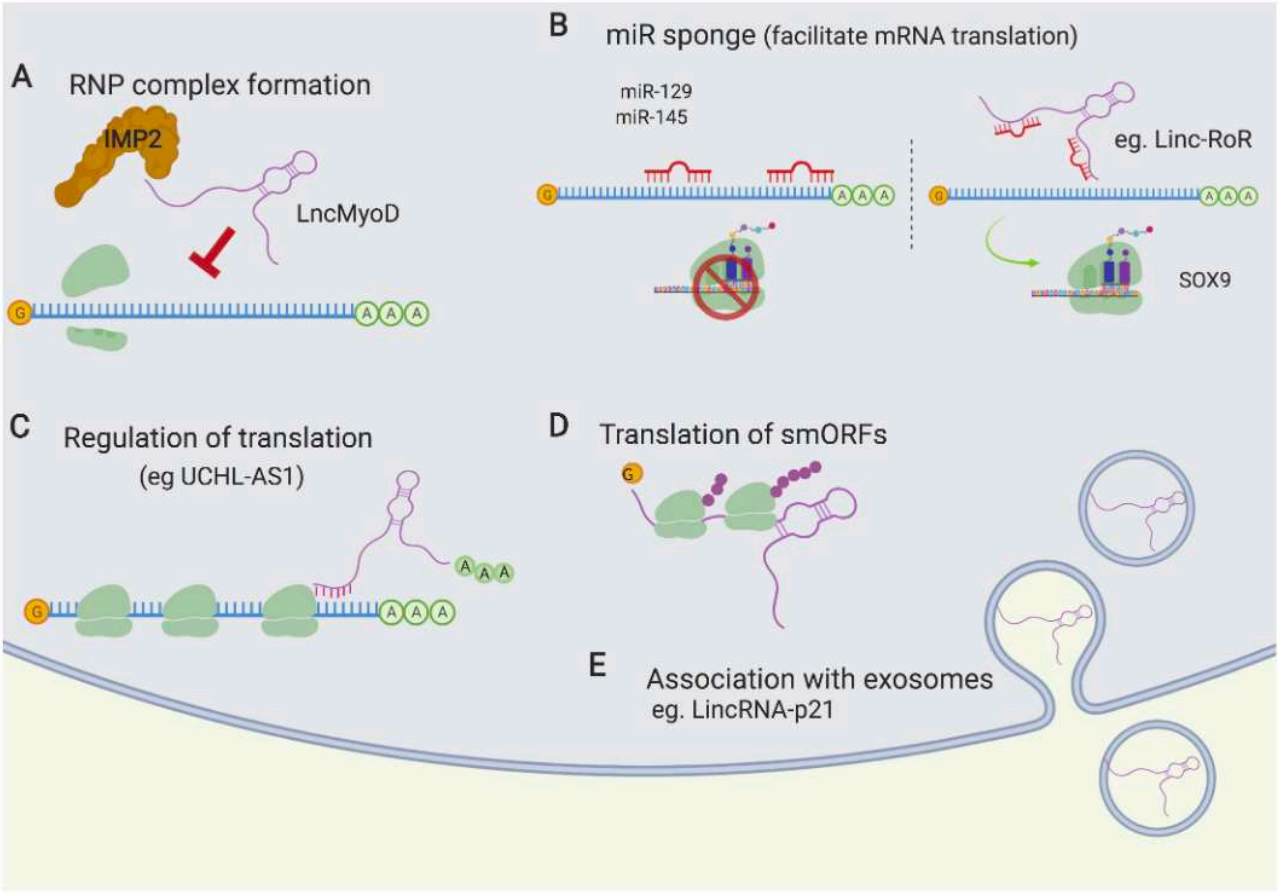


Figure 25 : Mécanismes d'action des lncRNA cytoplasmiques. Tiré de Tsagakis et al, 2020.

Enfin, les lncRNA porteurs d'une coiffe 5' et polyadénylés peuvent interagir avec les ribosomes et être traduits. Des études chez l'homme, la souris, la drosophile et la levure ont détecté des lncRNA dans des complexes liés aux ribosomes, révélant que des petits ORF codant des peptides de moins de 100 acides-aminés (smORF), présents dans certains lncRNA, sont en fait traduits [232]. La fonction de ces peptides n'est, à ce jour, pas clairement élucidée. Un des lnRNA les plus connus est XIST (*X-inactive specific transcript*). Cet ARN non-codant, d'environ ~19,2 kb, est exprimé spécifiquement chez la femme. Il est porteur d'une coiffe en 5', qu'une queue poly-A et subi un épissage alternatif. Les études ont montré qu'il est un acteur majeur du développement précoce en assurant la compensation de dose par inactivation aléatoire d'un des deux chromosomes

X chez la femme [233,234]. Il a été démontré, dans un modèle murin, que des mutations de Xist empêchent l'inactivation de l'X et sont létales au stade embryonnaire [235]. Le mécanisme de cette inactivation reposerait sur le recrutement par XIST de partenaires, formant plusieurs complexes lncRNA-protéines et recouvrant un chromosome X pour l'inactiver. XIST recrute SPEN et les complexes hnRNPK, PRC1, PRC2, pour déposer des marques chromatiniennes répressives afin d'établir et de maintenir le *silencing* [236].

Beaucoup de publications rapportent des liens entre lncRNA et régulation de processus physiologiques ou physiopathologiques. Ce lien est assez richement documenté pour plusieurs maladies complexes, notamment les cancers (Chung et al., 2011; Sanchez Calle et al., 2018). En ce qui concerne le neurodéveloppement, là-aussi, un ensemble d'études sur des modèles murins a montré un rôle de plusieurs lncRNA conservés [237,238] et des dérégulations sont suspectées dans des manifestations pathologiques [239] (Tableau 5). Malgré cela, les preuves d'une implication directe d'un lncRNA dans des TND d'origines monogéniques restent limitées. Ang *et al.* ont rapporté en 2019 l'identification d'un patient dont le génome présente une translocation équilibrée perturbant le locus lnc-NR2F1 sans aucune autre variation génétique pathogène détectable et qui présente des troubles neurodéveloppementaux. Toutefois, le père du patient est porteur de la même translocation et ne souffre que de dyslexie et de bégaiement. Les études sur modèle murin confirment un rôle neurodéveloppemental de lnc-NR2F1 mais le fait que le variant soit hérité d'un parent ne présentant pas le même phénotype et l'absence d'autres patients identifiés incite à la prudence [240].

lncRNAs Directly Involved in ID				
lncRNA	Observed Anomaly	Associated Phenotype Observed		Mechanism/Evidence
lnc-NR2F1	[t(5:12)]	Developmental and speech delay		Disruption of *lnc-NR2F1*, which controls neuronal migration and other NDD-genes
LINC00299	2p25.1 disruption	Speech delay, ID, bipolar disorder, epilepsy, Angelman-like syndrome		*LINC00299* increased levels in patients
lncRNAs regulating genes involved in ID				
lncRNA	Regulated gene	Possible association with disease	OMIM disease	Mechanism/Evidence
BC200	*FMR1*	Fragile-X syndrome	#300624	Repression of local translation, interacting with FMRP
GAS5	*GSTM3*	Down syndrome	#190685	Downregulation of GAS5 in DS patients Upregulation in Klinefelter syndrome patients
NRON	*NFAT*	Down syndrome	#190685	Reduced NFAT causes DS-like phenotype; *NRON* modulates NFAT activity.
EVF2	*MECP2*	Rett syndrome X-linked intellectual developmental disorder	#312750 #300055 #300260	*EVF2* associates with MECP2 at regulatory elements in interneurons, controlling *DLX5*, *DLX6* and *GAD1* expression
AK081227 ***AK087060***	*MECP2*	Rett syndrome X-linked intellectual developmental disorder	#312750 #300055 #300260	Upregulated levels in *Mecp2*-null mice; *AK081227* downregulates *Gabrr2*
ZEB2-NAT	*ZEB2*	Mowat–Wilson syndrome	#235730	*ZEB2-NAT* controls *ZEB2* by retaining the first intron of *ZEB2* pre-mRNA
RMST	*SOX2*	Microphthalmia and optic nerve hypoplasia and abnormalities of the central nervous system	#206900	*RMST* physically associates with SOX2 regulating neurogenesis pathways
Sox2ot	*SOX2*	Microphthalmia and optic nerve hypoplasia and abnormalities of the central nervous system	#206900	*Sox2ot* represses *SOX2* RNA levels
SYNGAP1-AS	*SYNGAP1*	Mental Retardation Autosomal Dominant 5	#612621	*SYNGAP1-AS* is upregulated in post-mortem brains

Tableau 5 : Exemples lncRNA impliqués dans la DI. Tiré de Liaci *et al.* 2022.

Une disruption du lncRNA LINC00299 a aussi été mise en cause chez plusieurs patients présentant une DI mais avec des phénotypes très variables [241,242]. Une des études rapporte notamment le cas d'une mère et son fils porteurs d'une translocation dont le point de cassure est situé dans le gène *LINC00299*. La mère présente une DI modérée mais le phénotype de son fils est, lui, beaucoup plus sévère. Toutefois, cette différence peut en partie s'expliquer par la présence d'un chromosome surnuméraire dérivé chez ce dernier [242] .

4.4. Les topologically associating domains (TAD)

Comme nous l'avons déjà brièvement évoqué, l'effet délétère des SV ne se limite pas à la disruption de gènes ou une modification de la séquence codante. Un effet dit « positionnel » est maintenant bien documenté. Un des principaux mécanismes positionnels fait intervenir des modifications tridimensionnelles de la chromatine, médiées

par des déstructurations des *topologically associating domains* (TAD). Les TAD sont définis comme des régions génomiques présentant une fréquence élevée d'auto-interaction, alors que peu ou pas d'interactions sont observées avec les TAD voisins [243]. En raison de leur structure isolante, les TAD modulent les connexions entre les éléments régulateurs, tels que les promoteurs et les enhancers, et jouent donc un rôle essentiel dans la régulation transcriptionnelle. Ces structures sont très conservées d'une espèce à l'autre et d'un type de cellule à l'autre [244]. Le modèle actuellement proposé pour expliquer la formation des TAD repose sur un mécanisme d'extrusion de boucle. Brièvement, les frontières des TAD sont matérialisées par des séquences consensus auxquelles se lient les protéines à doigt de zinc CCCTC-binding factor (CTCF). Une boucle d'ADN est extrudée au travers d'un anneau formé par un complexe de cohésine jusqu'à être bloquée par les protéines CTCF. Des SV peuvent modifier les frontières de TAD en les supprimant, en les déplaçant, ou en en créant de nouvelles (Figure 26). Cela conduit à des expressions ectopiques de certains gènes avec de possibles conséquences développementales.

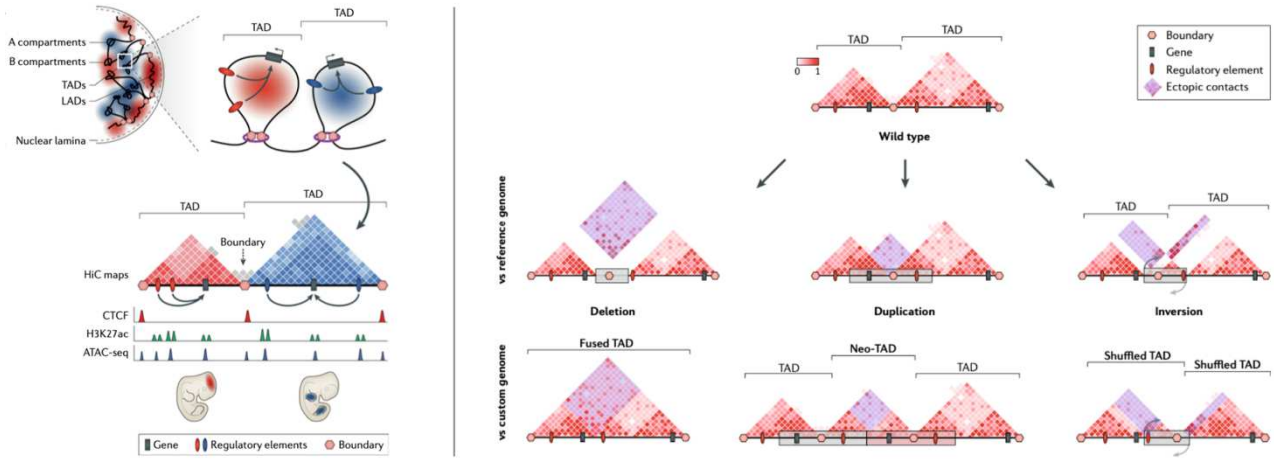


Figure 26 : Effet des variants structuraux sur les TAD. Tiré de Spielmann *et al*, 2018).

Lupianez *et al.* ont montré en 2015 que des malformations distinctes des membres humains sont causées par des délétions, des inversions ou des duplications altérant la structure du locus *WNT6/IHH/EPHA4/PAX3*. Ces SV conduisent à l'apparition d'interactions ectopiques entre des promoteurs de gènes impliqués dans le développement des membres et un groupe d'enhancers. Cela conduit à diverses malformations de membre (brachydactylie, polydactylie, F-syndrome) en fonction du type de SV (Figure 27). Ce remaniement ne se produit que si le variant perturbe un domaine frontière associé au CTCF.

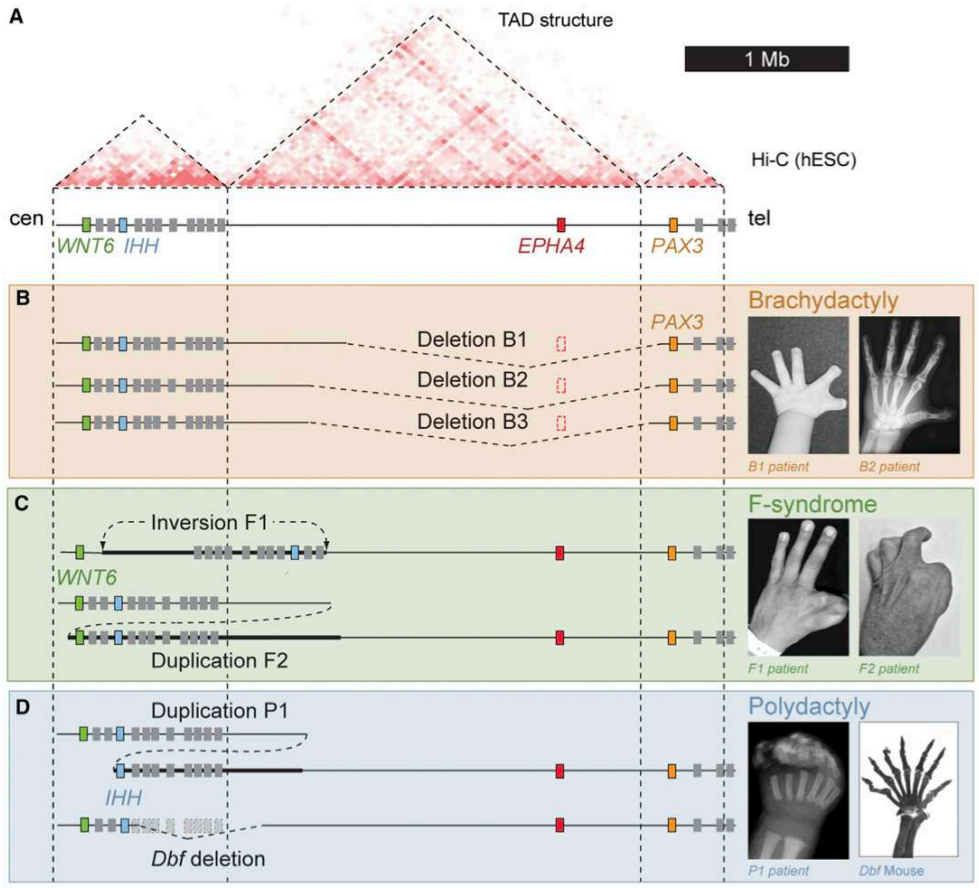


Figure 27 : Différentes variations structurelles pathogènes des membres chez l'homme et la souris sont liées au TAD EPHA4. Tiré de Lupianez et al, 2015.

OBJECTIFS DE LA THÈSE

Aujourd'hui, le séquençage de génome prend une place de plus en plus importante en diagnostic. Le SG *short read* est le plus répandu et ses limites sont bien documentées. Bien que le SG permette de détecter plus de variants structuraux que le SE, certains types de SV restent difficiles à identifier. Par ailleurs, la capacité du SG à découvrir des variants dans les régions non-codantes soulève la question de leur difficile interprétation. L'hypothèse initiale de mon travail de thèse est qu'il est possible d'améliorer significativement le rendement diagnostique du SG en intégrant un nombre restreint de données complémentaires telles que le RNA-Seq et les techniques long fragments. Bien évidemment, au-delà de l'apport à court terme que représente l'amélioration du rendement, nous pensions également que cette approche peut être bénéfique pour l'identification de nouveaux gènes candidats en recherche.

Notre première étape était donc de réaliser une analyse SG *short read* d'une cohorte de 33 patients. Pour cela, j'ai utilisé des outils standards de détection des SNV/indels, puis, j'ai développé notre propre pipeline d'analyse des SV à partir d'un ensemble d'outils complémentaires. Dans un second temps, nous avons proposé une analyse RNA-Seq aux patients pour lesquels le SG n'était pas concluant. Enfin, nous avons réalisé une analyse OGM chez les patients pour lesquels un SV identifié en SG nécessitait plus d'investigations pour mieux comprendre son organisation structurale.

En parallèle, j'ai cherché à identifier un tissu alternatif pour la réalisation des analyses RNA-Seq. En effet, les fibroblastes qui sont actuellement souvent utilisés, et que nous avons-nous-même choisi pour cette étude, présentent certaines contraintes. Nous avons donc étudié la culture de cellules souches urinaires comme type cellulaire d'intérêt pour éventuellement remplacer les prélèvements sanguins ou les biopsies cutanées.

Ces résultats ont été publiés dans *journal of medical genetics* le 26 juillet 2023.

METHODE

Recrutement des patients et critères d'inclusions.

33 patients suivis dans divers centres hospitaliers du Grand Ouest ont été sélectionnés pour cette étude. Tous présentaient une déficience intellectuelle (HP:0001249) ou un retard global de développement (HP: 0001263) isolé ou syndromique. La sévérité allait de légère à grave. L'étiologie génétique était fortement suspectée chez ces individus pour lesquels aucune cause environnementale n'a été documentée. Les individus dont un parent était symptomatique étaient exclus. Tous avaient déjà bénéficié d'un SE en solo ou trio et d'une ACPA non-conclusifs.

Séquençage de génome

Le SG des 33 trios a été réalisée au Centre National de Recherche en Génomique Humaine (CNRGH, Evry, France). La génération de librairies a été réalisée à partir d'ADN extrait d'échantillons sanguins avec le kit TruSeq®DNA PCR-free (Illumina, San Diego, CA, USA). Les génomes ont été séquencés sur un HiSeq X5 (Illumina, San Diego, CA, USA) à une profondeur moyenne minimale de 30X. Les séquences d'ADN ont été alignées sur le génome de référence du génome humain GRCh37 à l'aide de bwa 0.7.12. Les SNV et les petites insertions/délétions (INDEL) ont été identifiés conformément aux recommandations de GATK (v3.4).

Pipeline d'analyse des SV

Il n'existe pas de pipeline standard, validé, pour une analyse exhaustive des variants de structure. J'ai donc choisi de développer un outil permettant de détecter le maximum

d'événements tout en fournissant une sortie interprétable par un humain en un temps raisonnable. Le pipeline que j'ai développé visait à proposer le meilleur compromis entre sensibilité et interprétabilité. Cet outil devait permettre de 1°/ lancer en parallèle plusieurs algorithmes afin d'optimiser le temps d'analyse et ainsi avoir un résultat pour les 33 trio en moins d'une semaine. 2°/ comparer et filtrer les sorties de chacun des algorithmes. 3°/ annoter les sorties pour faciliter l'interprétation.

Une des difficultés majeures était de déterminer avec précision si un variant était hérité ou *de novo*. En effet, les coordonnées des points de cassures d'un SV hérité retourné par les algorithmes ne sont pas toujours rigoureusement les mêmes. Il a donc été nécessaire de mettre en place un système de calcul et de comparaison tenant compte de l'incertitude sur les points de cassures (une estimation de cette incertitude peut être extraite des VCF). Étant donné que les SV *de novo* sont attendus comme rares, j'ai choisi des modes de calculs qui tendent à favoriser leur conservation. En effet, l'étape d'interprétation humaine se focalise prioritairement sur les événements *de novo.* Un faux positif peut être facilement éliminé lors de l'interprétation tandis qu'un variants considéré hérité à tort risque fortement de ne pas être considéré et vérifié. Chaque type de SV a nécessité un mode de calcul propre.

Les variations du nombre de copies (CNV), les variants structurels (SV) et les courtes répétitions en tandem (STR) ont donc été détectés à l'aide de ce pipeline personnalisé (https://gitlab.univ-nantes.fr/kriquin/sv-genome) qui combine des algorithmes avec des méthodes de *calling* complémentaires et des outils d'annotation. Il est basé sur quatre algorithmes : GATK, MANTA, Delly et Expansion Hunter *de novo*.

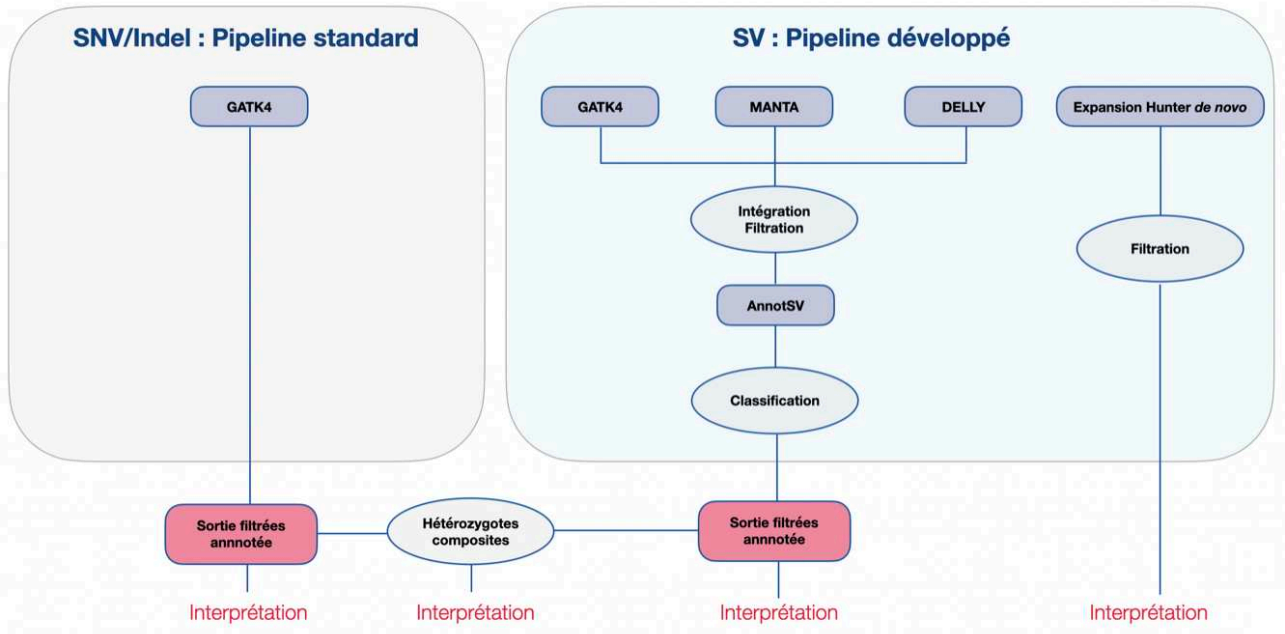


Figure 28 : Pipeline d'analyse des données de séquençage de génome.

CNV

Les CNV sont détectés à l'aide d'une méthode basée sur les variations de profondeur de lecture. J'ai choisi l'algorithme GATK 4.1.4.1 (Broad Institute) configuré en mode cohorte avec des fenêtres glissantes de 1000 pb.

Les premiers essais de cet algorithme ont retourné un nombre important de faux positifs. Ces artéfacts étaient essentiellement dus à des variations de profondeur de lecture dans des régions avec une faible qualité de mapping. J'ai donc ajouté une étape de filtration permettant d'éliminer les intervalles se trouvant dans des régions de duplications segmentaires ou pseudo-autosomiques. J'ai également tenté de rajouter un niveau de filtration en identifiant les coordonnées de tous les intervalles dont le MAPQ moyen était inférieur à 20 et qui correspondaient donc vraisemblablement à des régions dans lesquelles l'alignement est difficile. Afin de ne pas être trop stringent, cette liste de régions à exclure a été faite à partir d'un nombre variable d'échantillons. J'ai pu constater

qu'en utilisant plus de 20 échantillons pour cette détection, le nombre de régions à exclure augmentait de façon modérée ce qui suggère que la plupart des intervalles problématiques pouvaient être déterminés à partir de seulement 20-30 échantillons. Au-delà, le risque de filtrer à tort une région d'intérêt me semblait trop élevé (Figure 29).

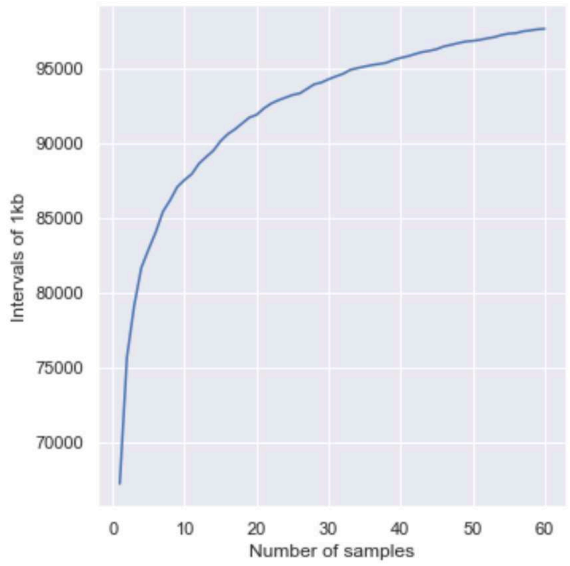


Figure 29 : Nombre d'intervalles exclus en fonction du nombre d'échantillon.

Afin de confirmer la pertinence de ce filtre, un *calling* de CNV a été effectué sur la cohorte entière en excluant ces intervalles *low-MAPQ*. Bien que cela ait quelque peu diminué le nombre de faux positifs, le gain s'est révélé assez marginal et n'a pas considérablement facilité l'analyse. Ce filtre n'a donc pas été retenu pour l'intégration au pipeline. Par ailleurs, les filtres sur les duplications segmentaires et régions pseudo-autosomique se sont, quant à eux, révélés particulièrement efficaces pour réduire le bruit de fond.

Détection des CNV par discordance de paires et *split reads*:

Les SV ont été détectés par MANTA 1.6.0 (Illumina Inc.) 13 et DELLY 0.8.7 (European Molecular Biology Laboratory). Ces deux algorithmes ont été intégrés à mon pipeline car ils sont relativement rapides et mes essais préliminaires ont montré qu'ils pouvaient être

complémentaires pour la détection des petits événements, tout particulièrement les insertions de moins de 100 pb.

Détection des STR:

Les courtes répétitions en tandem ont été génotypées par Expansion Hunter *de novo* 0.9.0 (Illumina Inc.). Contrairement à une version plus ancienne de l'algorithme appelée simplement « Expansion Hunter », cette version est capable de détecter naïvement les STR et ne nécessite pas de liste de coordonnées de STR connus.

Annotation et priorisation :

Tous les SV ont été annotés à l'aide de AnnotSV 2.2 et un tag de priorisation (LOW, MEDIUM, HIGH) a été attribué sur la base du mode de transmission, de la fréquence allélique du variant dans les bases de données et de la localisation. Les variants *de novo*, homozygotes, hétérozygotes composites et hémizygotes ont été sélectionnés et leurs aperçus IGV en trio ont été extraits automatiquement pour une vérification manuelle. Les variants identifiés ont été classés selon la norme ACMG pour l'interprétation des variants de séquence. Pour les variants homozygotes, le degré de consanguinité a été estimé à l'aide d'Automap 1.2. Tous les SV ont été confirmés par caryotype, qPCR ou optical genome mapping.

Insertions d'éléments mobiles *de novo*

Les insertions d'éléments mobiles *de novo* ont été recherchées chez les individus négatifs après SG à l'aide de MELT 2.2.2.

RNA-Seq

Le séquençage de l'ARN a été effectué sur les fibroblastes de peau de neuf individus et d'un parent en deux *batchs* distincts. La bibliothèque a été générée à l'aide du kit NEBNext UltraTM II Directional RNA Library Prep (New England Biolabs, Ipswich, MA, USA). L'ARN a été séquencé sur NextSeq 500 (Illumina, San Diego, CA, USA) avec environ 60-70 millions de *reads* par échantillon. Les *reads* ont été alignées à l'aide de STAR 2.6.1d en mode *two-passes* (GRCh37 release75, Gencode34) et les duplicats marquées avec Picard (Broad Institute). L'analyse des données RNA-Seq a été réalisée à l'aide de Drop 1.1.4. L'expression aberrante a été détectée à l'aide de OUTRIDER avec une matrice externe de 127 échantillons (https://doi.org/10.5281/zenodo.7510845). Afin d'évaluer le nombre minimal d'échantillons dans une analyse d'expression aberrante, j'ai renouvelé l'analyse avec 100, 60, 40 et 30 au total. Il était possible de retrouver l'ensemble de nos événements d'intérêt avec un minimum de 60 échantillons. Toutefois, j'ai pu constater que l'ajout d'un nombre important d'échantillons en matrice externe ne nuisait pas à la correction de l'effet batch, au contraire. La Figure 30-A montre un effet batch important avec 279 échantillons à comparer avec l'effet pour 55 échantillons en Figure 30-C. La correction de l'effet batch avec les 279 échantillons est meilleure (Figure 30-B) qu'avec 55 échantillons (Figure 30-D). Dans la mesure du possible, il est donc préférable de choisir un grand nombre d'échantillons même s'ils sont issus de plusieurs *batchs* et/ou de matrices externes. Toutefois, il est important de noter que cela peut rallonger considérablement le temps d'analyse. C'est pourquoi j'ai choisi de faire un compromis en utilisant un échantillonnage de 127 individus pour la matrice externe.

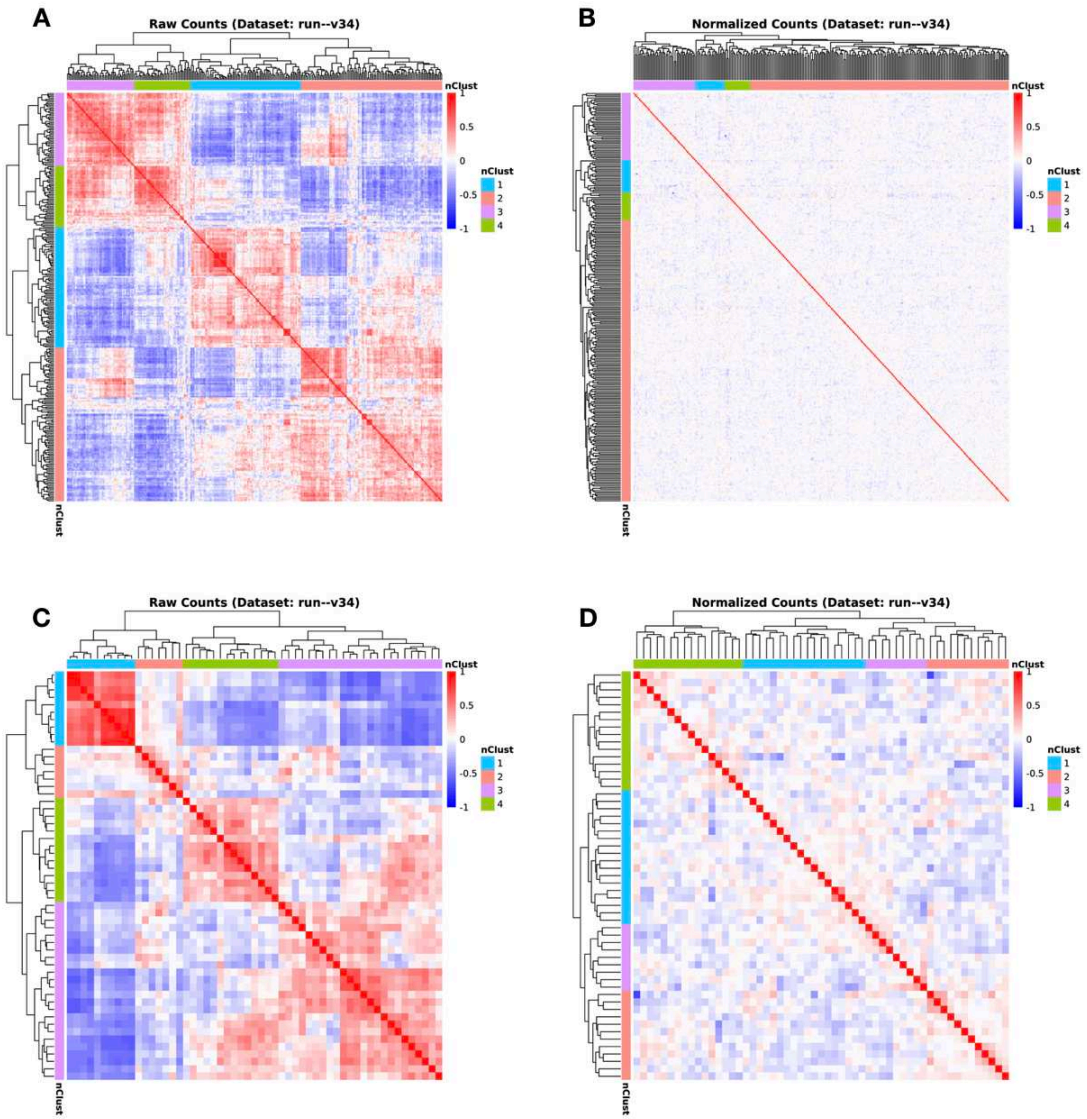


Figure 30 : Correction de l'effet batch avec différents nombres d'échantillons. A) Avec 279 échantillons avant correction. B) Avec 279 échantillons après correction. C) Avec 55 échantillons avant correction. C) Avec 55 échantillons après correction.

L'épissage aberrant a été détecté à l'aide de FRASER sans matrice externe. Les événements d'épissage ont été filtrés comme suit : valeur p ≤ 0,01, padjGene < 1, DeltaPsi ≤ -0,10 ou DeltaPsi ≥ 0,10. L'expression monoallélique a été détectée en utilisant le module MAE de Drop et filtrée comme suit : padj ≤ 0,05, cohort_freq ≤ 0,1, AFmax ≤ 0,01. Nous avons quantifié l'expression des gènes dans le sang total, les fibroblastes et les cellules souches dérivées de l'urine provenant de donneurs sans TND. En utilisant Kallisto 0.46.2 (index Kallisto : Ensembl 99 ; GRCh38), j'ai calculé le niveau

d'expression des gènes de chaque échantillon (en TPM). Les gènes ayant un TPM > 10 ont été considérés comme suffisamment exprimés.

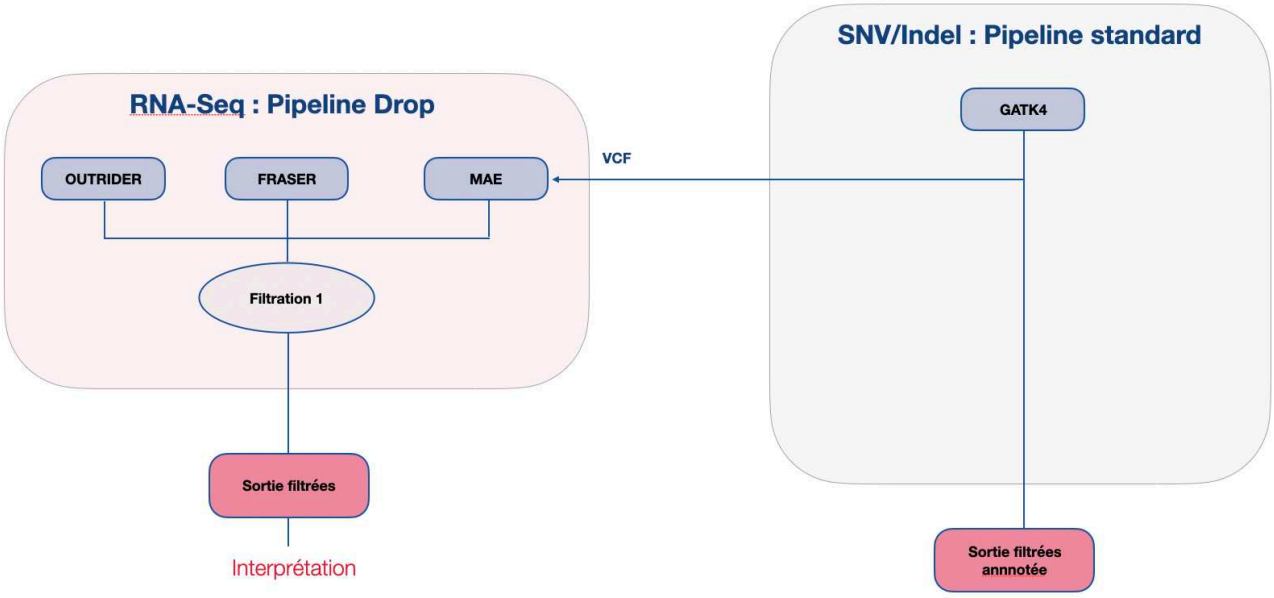


Figure 31 : Pipeline d'analyse des données de RNA-Seq.

Rendement diagnostique global

Le rendement diagnostique global a été déterminé comme la proportion de sujets de la cohorte ayant reçu un diagnostic moléculaire définitif. Le rendement diagnostique unique du SG a été calculé comme le pourcentage de sujets de la cohorte ayant reçu un diagnostic moléculaire concluant qui n'aurait pas pu être obtenu uniquement par la réanalyse des données de SE.

RÉSULTATS

Article : Integrating RNA-Seq into genome sequencing workflow enhance the analysis of structural variants causing neurodevelopmental disorders.

Integrating RNA-Seq into genome sequencing workflow enhance the analysis of structural variants causing neurodevelopmental disorders.

Kevin Riquin[1], Bertrand Isidor[1,2], Sandra Mercier[1,2], Mathilde Nizon[1,2], Estelle Colin[3,4], Dominique Bonneau[3,4], Laurent Pasquier[5], Sylvie Odent[5,6], Xavier Le Guillou Horn [7,8], Gwenael le Guyader[7], Annick Toutain[9,10], Vincent Meyer[11], Jean-François Deleuze[11], Olivier Pichon[2], Martine Doco-Fenzy[1,2], Stéphane Bézieau[1,2], Benjamin Cogné[1,2]

1 Nantes Université, CHU de Nantes, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France
2 Nantes Université, CHU de Nantes, Service de Génétique médicale, F-44000 Nantes, France
3 CHU Angers, Service de Génétique médicale, 49933 Angers Cedex 9, France
4 UMR CNRS 6214-INSERM 1083, Université d'Angers, 49933 Angers Cedex 9, France
5 Service de Génétique Clinique, ERN ITHACA, CHU Rennes, Rennes, France
6 Institut de Génétique et Développement de Rennes, IGDR UMR 6290 CNRS, INSERM, IGDR Univ Rennes, Rennes, France
7 Service de génétique médicale, CHU de Poitiers, Poitiers France
8 LabCom I3M-Dactim mis / LMA CNRS 7348, Université de Poitiers, Poitiers, France
9 UF de Génétique Médicale, Centre Hospitalier Universitaire, 37044 Tours, France
10 UMR 1253, iBrain, Université de Tours, INSERM, 37032 Tours, France
11 Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine (CNRGH), 91057, Evry, France

Abstract

Background: Molecular diagnosis of neurodevelopmental disorders (NDD) is mainly based on exome sequencing (ES), with a diagnostic yield ranging from 31% for isolated and 53% for syndromic NDD. As sequencing costs decrease, genome sequencing (GS) is gradually replacing ES for genome-wide molecular testing. As many variants detected by GS only are in deep intronic or non-coding regions, the interpretation of their impact may be difficult. Here, we showed that integrating RNA-Seq into the genome sequencing workflow can enhance the analysis of the molecular causes of NDD, especially structural variants, by providing valuable complementary information such as aberrant splicing, aberrant expression, and monoallelic expression.

Methods: We performed trio-GS on a cohort of 33 individuals with NDD for whom ES was inconclusive. RNA-Seq on skin fibroblasts was then performed in nine individuals for whom GS was inconclusive and optical genome mapping (OGM) was performed in two individuals with a structural variant (SV) of unknown significance.

Results: We identified pathogenic or likely pathogenic variants in 16 individuals (48%) and six variants of uncertain significance. RNA-seq contributed to the interpretation in three individuals, and OGM helped to characterize two SVs.

Conclusion: Our study confirmed that GS significantly improves the diagnostic performance of NDDs. However, most variants detectable by GS alone are structural or located in non-coding regions, which can pose challenges for interpretation. Integration of RNA-Seq data overcame this limitation by confirming the impact of variants at the transcriptional or regulatory level. This result paves the way for new routinely applicable diagnostic protocols.

Key messages:

WHAT IS ALREADY KNOWN ON THIS TOPIC:

Genome sequencing has shown that it can outperform the standard diagnostic procedures for neurodevelopmental disorders. Routine implementation in diagnostic laboratories poses new challenges, particularly in the detection and interpretation of structural or non-coding variants.

WHAT THIS STUDY ADDS:

The application of complementary techniques, such as RNA-Seq and optical genome mapping, on a case-by-case basis, may help enhance the diagnostic yield of genome sequencing for complex or difficult-to-interpret cases.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY:

Clinical RNA-Seq is a growing subject of research. However, its application in routine diagnostics has not yet been standardized. Our study provides novel insights into its gradual introduction, in association with genome sequencing.

Introduction

Neurodevelopmental disorders (NDD) are a group of behavioral and cognitive disorders that occur during the developmental period and involve significant difficulties in the acquisition and performance of specific intellectual, motor, language, or social functions [1]. Among this disorders, intellectual disability have frequently been associated with monogenic causes. Owing to cumulative efforts to improve molecular diagnosis, more than 1400 causative genes have been identified [2,3]. Although next-generation sequencing has significantly improved the diagnosis, most individuals still face many years of diagnostic odyssey.

Currently, molecular diagnosis is mainly based on exome sequencing (ES), which is a cost-effective technique with a diagnostic yield ranging from 31% for isolated

NDD, and 53% for syndromic NDD [4]. Thus, the molecular cause is still not identified in most of the cases. Possible reasons for this high proportion of unexplained cases are the inability of ES to detect variants in deep intronic and non-coding regions or difficulties in detecting structural variants (SVs). Over the past decade, studies have shown that genome sequencing (GS) outperforms the standard ES + array-CGH procedure with GS-unique diagnostic yields ranging from 19% to 25% [5–8].

Nevertheless, many variants detected in deep intronic or non-coding regions raise interpretation issues. Transcriptomic analysis can provide valuable complementary information to interpret noncoding variants, including aberrant splicing, aberrant expression, and monoallelic expression. To date, RNA-Seq studies have been particularly successful for neuromuscular disorders [9] for which appropriate tissue is accessible, with a diagnostic rate of up to 35%. Recent studies have shown an increased diagnostic rate of up to 17% after applying RNA-Seq to GS-negative individuals with diverse disorders [10]. To our knowledge, this approach has rarely been specifically applied to a cohort of individuals with neurodevelopmental disorders[11,12].

Here, we performed trio GS on a cohort of 33 individuals with NDD for whom previous genetic analyses were negative, including array-CGH and ES. RNA-Seq was then performed in nine probands with inconclusive GS. We also characterized two SVs by optical genome mapping (OGM). Finally, we report interesting cases.

Methods

Recruitment, Inclusion Criteria

33 families from the University Hospitals of Western France (HUGO network) were recruited based on suspected neurodevelopmental disorders. Phenotypic data were collected as Human Phenotype Ontology (HPO) terms. All probands presented with syndromic or nonsyndromic moderate-to-severe intellectual disability or global developmental delay. Parents were asymptomatic. Individuals were excluded if a non-genetic etiology was suspected or if a parent was symptomatic. The detailed clinical phenotypes are listed in Sup.Table 2. Informed consent for genetic analysis was obtained from the legal guardians according to the French law on bioethics and following the Helsinki Declaration. The study was approved by the ethics committee of Nantes University Hospital (number CCTIRS:14.556). Solo- or trio-ES and array-CGH have already been performed and were negative for each proband. RNA-seq was proposed to any individual for whom a variant of uncertain significance or no variant was identified after the GS.

Genome sequencing variant calling and annotation

Genome sequencing was performed on 33 trio at the French National Research Center for Human Genomics (CNRGH, Evry, France). Library generation was performed using DNA extracted from blood samples with TruSeq®DNA PCR-free kit (Illumina, San Diego, CA, USA). The genomes were sequenced on a HiSeq X5 (Illumina, San Diego, CA, USA) at a minimum mean depth of 30X. DNA sequences

were mapped to the reference human genome sequence (GRCh37) using bwa 0.7.12, and single nucleotide variants (SNVs) and small insertions/deletions (INDELs) were identified following GATK's best practices (v3.4). Copy number variations (CNVs), structural variants (SV), and short tandem repeats (STRs) were detected using a custom pipeline (https://gitlab.univ-nantes.fr/kriquin/sv-genome). Briefly, this pipeline combines algorithms with complementary calling methods and annotation tools. It is based on four algorithms: CNVs are detected using the read depth method GATK 4.1.4.1 (Broad Institute) with a sliding window of 1000 bp. SV are called by MANTA 1.6.0 (Illumina Inc.) [13] and DELLY 0.8.7 (European Molecular Biology Laboratory) [14]. Short tandem repeats were genotyped by Expansion Hunter de novo 0.9.0 (Illumina Inc.) [15] and *de novo* mobile element insertions were called in individuals negative after GS using MELT 2.2.2 [16]. All structural variants were annotated using AnnotSV 2.2 [17] and a prioritization tag was assigned based on inheritance status, allelic frequency of the variant in databases and localization. Identified variants were classified using the ACMG standard for the interpretation of sequence variants [18]. The degree of consanguinity was estimated using Automap 1.2 [19]. All SVs were confirmed by karyotyping, quantitative PCR, or optical genome mapping (OGM).

Diagnostic yield calculations

The overall diagnostic yield was determined as the proportion of probands in the cohort with a definitive molecular diagnosis. The GS-unique diagnostic yield was calculated as the percentage of probands in the cohort who received a conclusive

molecular diagnosis that would have been unattainable solely through the reanalysis of ES data.

Optical genome mapping (OGM)

High-molecular-weight DNA was extracted from frozen whole peripheral blood EDTA according to the manufacturer's instructions (Bionano Genomics, San Diego, CA, USA). Labeled DNA was loaded onto a Saphyr chip for linearization and imaging on the Saphyr instrument. After checking QC metrics, the *de novo* assembly and variant annotation pipeline were run using Bionano Solve v.7.1 software at 80X coverage. Each optical genome map was compared with a reference genome map. SVs were visualized using Bionano Access software V7.1 and compared to an optical genome map dataset of 300 human population control samples from apparently healthy individuals (provided by Bionano Genomics) to filter out common SVs and potential artifacts.

Prediction tools

The impact of missense mutations was predicted using Mobidetails [20]. The effect of the mutation on protein stability was predicted using I-Mutant 3.0 [21] and DUET [22]. The parameters used were: T=25°C and pH = 7. The PDB accession number of the structure of the APP protein used was 4PWQ [23].

RNA-Sequencing

RNA-Sequencing was performed on skin fibroblasts of nine individuals and one parent in two separate batches. The library was generated using NEBNext Ultra™ II Directional RNA Library Prep (New England Biolabs, Ipswich, MA, USA). RNA was sequenced using NextSeq 500 (Illumina, San Diego, CA, USA) with approximately 60-70 million reads per sample. Reads were aligned using STAR 2.6.1d [24] in two pass-mode (GRCh37 release75, Gencode34) and duplicate marked with Picard (Broad Institute). Analysis of RNA-Seq data was performed using Drop 1.1.4 [25]. Aberrant expression was detected using OUTRIDER [26] with external matrix of 127 samples (https://doi.org/10.5281/zenodo.7510845). Aberrant splicing was detected using FRASER [27] without external matrix. Splicing events were filtered as follows: pvalue ≤ 0.01, padjGene < 1, DeltaPsi ≤ -0.10 or DeltaPsi ≥ 0.10. Monoallelic expression was detected using the MAE module of Drop and filtered as follows: padj ≤ 0.05, cohort_freq ≤ 0.1, AFmax ≤ 0.01 [28,29]. We quantified the gene expression in whole blood, cultured fibroblasts, and cultured urine-derived stem cells (sup. Method) from donors without NDD. Using Kallisto 0.46.2 [30] (Kallisto index: Ensembl 99; GRCh38), we generated TPM values for the expressed genes in each sample. Genes with TPM > 10 were considered well expressed [31].

Results

Cohort

Our cohort was composed of 33 individuals, 19 of whom were female (57%). The median age was 12 years (range, 2–31 years). All individuals have syndromic or nonsyndromic intellectual disability (ID) (HP:0001249) or global developmental delay (GDD) (HP:0001263). Clinical data are available in Sup.Table 2. All individuals had a previous negative array-CGH and ES (Trio-ES 26/33, Solo-ES for 7/33). GS with a minimum mean depth of 30X was performed for each parent–child trio. RNA-seq was then proposed to all individual for whom a variant of uncertain significance or no variant was identified by GS. Nine families accepted.

Diagnostic yield

We report a diagnosis in 16 individuals (16/33; 48%) (Table 1; Sup.Table 1; Figure 1). Seven *de novo* variants and one inherited variant were identified in genes associated with an autosomal dominant neurodevelopmental disorder (8/33; 24%). Five cases (5/33; 15%) were resolved through the identification of biallelic variants in genes associated with autosomal recessive inheritance. In addition, we identified a hemizygous pathogenic variant inherited from the mother in an X-linked gene in one male participant, and a *de novo* variant in an X-linked gene in two female participants. No candidate variants involving short tandem repeats or mobile element insertion were identified.

SNV/INDELs:

SNVs or INDELs supported a diagnosis in ten individuals (10/16; 62%). Of these, seven were identified thanks to the reanalysis effect. They were detected by ES but not classified as pathogenic at the time of analysis. Better support for pathogenicity, a recent publication of the genes involved, or additional investigations allowed us to revise this classification. Among these variants, we reconsidered a maternally inherited nonsense variant in *MN1*. Indeed, the variant was identified as de novo in another affected individual [32] with very similar dysmorphic facial features. After analysis in maternal grand-parents, the variant was found *de novo* in the mother. Although the variant appeared heterozygous in the blood, we suspected mosaicism in the healthy mother. Additionally, a low-level mosaic frameshift variant in *NIPBL* was identified through targeted NGS analysis of saliva following inconclusive GS and ES. Notably, two cases were resolved by SNVs in exons of *H3-3A* that were not covered by the kit at the time of ES. These cases may have been resolved by renewing ES.

SV:

Pathogenic or likely pathogenic SVs were identified in six individuals (6/16; 37%). One balanced event was identified, a *de novo* inversion with a breakpoint in FOXP1, as previously reported [33]. The five others were CNVs that were too small to be detected by ES or Array-CGH. We also found 4 *de novo* SV of uncertain significance. Two were *de novo* balanced events with breakpoints in the intergenic

regions and no genes directly involved. One was a complex event involving the duplication of *FGF18* associated with an inverted insertion.

Contribution of complementary techniques

OGM:

OGM helped in the characterization of two SVs. In the first case (P25), we suspected an inversion of 2.2Mb following a 500kb duplication at the *TENM2*/*FGF18* locus. The OGM showed an inverted insertion, rather than an inversion/duplication. In the second case (P24), GS detected a duplication in *PUM1* with uncertainty in the breakpoint coordinates, and the OGM characterized this as an intragenic tandem duplication.

RNA-Seq:

We additionally collected skin biopsies from 10 individuals (nine probands and one parent) with inconclusive GS and performed RNA-Seq on cultured fibroblasts. We identified two cases with an abnormal splicing event caused by an intragenic CNV: skipping of two exons in *CBX3* (P23) and an abnormal splicing event in *PUM1* (P24). We also detected one sample with significant underexpression of the long non-coding RNA (lncRNA) *CHASERR*, which could be linked to *de novo* CNV. RNA-seq contributed to the diagnosis or interpretation of the variant in three individuals (3/9; 33%) (Sup Table 3).

To evaluate the relevance of fibroblasts as a source of biological material for clinical RNA-seq analysis, we quantified and compared gene expression levels among blood samples, fibroblasts, and an alternative cell type known as Urine Derived Stem Cells (USCs) (sup. Method). We observed that approximately 60% of NDD-associated genes were expressed in fibroblasts with TPM>10 and 37% were expressed in blood. In addition, *CBX3*, *PUM1* and *CHASERR* were expressed at a level that theoretically allowed their detection in blood samples (https://kriquin-univ-nantes.shinyapps.io/expression-rna-seq). Interestingly, USCs obtained by non-invasive means expressed approximately 63% of the NDD-associated genes. This could make them an interesting alternative to the fibroblasts.

Case/Sex	Type	Gene	Variant	Transmission	Interpretation	ACMG	Reason not reported by ES	Technique applied
P2/F	SNV indel	*NIPBL*	NC_000005.9:g.37059021_37059022del (Mosaicism 11%)	*De novo* [MOS]	Solved [MIM :122470]	5 (PVS1, PS2, PM2)	Mosacism Variant not found in blood	GS Panel
P4/M	SNV indel	*H3-3A*	NC_000001.10 :g.226259121G>C NM_002107.7 :c.352G>C :p.(Val118Leu)	*De novo* [HET]	Solved [MIM:619720]	5 (PP2, PS2, PM2, PP3, PM1)	Exon not covered	GS
P5/M	SNV indel	*MN1*	NC_000022.10 :g.28192754C>A NM_002430.3 :c.3778G>T :p.(Glu1260*)	Inherited from mother [MOS]	Solved [MIM :618774]	5 (PVS1,PS1,PM2)	Nonconcordant segregation	GS
P6/F	SNV indel	*H3-3A*	NC_000001.10:g.226259146A>G NM_002107.7:c.377A>G:p.(Gln126Arg)	*De novo* [HET]	Solved [MIM :619720]	5 (PS2,PM1,PM2, PP2,PP3)	Exon not covered	GS
P7/F	SNV indel	*TFE3*	NC_000023.10:g.48895930A>G NM_006521.6:c.572T>C:p.(Leu191Pro)	*De novo* [HET]	Solved [MIM :301066]	4 (PS2,PM1,PM2, PP3)	Gene now linked to ID	GS
P9/F	SNV indel	*BAP1*	NC_000003.11:g.52442077C>G NM_004656.4:c.272G>C:p.(Cys91Ser)	*De novo* [HET]	Solved [MIM :619762]	5 (PS2,PM1,PM2, PM5,PP3)	More support for pathogenicity	GS
P13/M	SNV indel	*HACE1*	NC_000006.11:g.105232328_105232331del NM_020771.4:c.1439_1442del:p.(Val480Alafs*7) - NC_000006.11:g.105297081_105297084del NM_020771.4:c.259_262del:p.(Lys87Glufs*27)	Autosomal recessive [COMP HET]	Solved [MIM:616756]	5 (PVS1,PM2,PM3) - 5 (PVS1,PM2,PM3)	Gene now linked to ID	GS
P15/F	SNV indel	*INTS11*	NC_000001.10:g.1248089C>T NM_017871.6:c.1295-9G>A:p.? - NC_000001.10:g.1248275T>C NM_017871.6:c.1186A>G:p.(Lys396Glu)	Autosomal recessive [COMP HET]	Solved [34]	4 (PM2,PM3,PP3, PP4) - 4 (PS3,PM2,PP3)	More support for pathogenicity	GS
P17/F	SNV indel	*TMEM147*	NC_000019.9:g.36036812_36036830del NM_032635.4:c.100_118del:p.(Lys34Serfs*33) - NC_000019.9:g.36038077C>G NM_032635.4:c.486C>G:p.(Tyr162*)	Autosomal recessive [COMP HET]	Solved [MIM:613585]	5 (PVS1,PM2,PM3) - 5 (PVS1,PM2,PM3)	More support for pathogenicity	GS

P28/M	SNV indel	*FITM2*	NC_000020.10:g.42935539A>C NM_001080472.4:c.515T>G:p.(Val172Gly) - NC_000020.10:g.42935585A>C NM_001080472.4:c.469T>G:p.(Phe157Val)	Autosomal recessive [COMP HET]	Solved [MIM :618635]	4 (PM1,PM2,PP3 PP4) - 4 (PM1,PM2,PP3, PP4)	Gene now linked to ID	GS
P1/F	SV	*FOXP1*	NC_000003.11:g.68954396_71064931inv NM_001244813.1:c.570-127_*2054002inv	*De novo* [HET]	Solved	5 (PVS1,PS2,PM2)	Copy neutral inversion not detected by ES	GS
P10/F	SV	*RSPRY1*	NC_000016.9:g.57230856_57252612del NM_133368.1:c.-155-7560_901+1665del	Autosomal recessive [HOM]	Solved [MIM :616723]	5 (PVS1,PM2,PM3)	Not detected	GS
P14/F	SV	*STEEP1*	NC_000023.10:g.118673832_118675628del NM_022101.3:c.514-245_607-80del	*De novo* [HET]	Solved [MIM : 301013]	4 (PVS1,PS2)	Not detected	GS
P24/F	SV	*PUM1*	NC_000001.10:g.3140901_31422000dup NM_001020658.1:c.2856+974_3435+489dup	*De novo* [HET]	Solved	4 (PS2,PM1,PM2)	Not detected	GS RNA-seq OGM
P27/M	SV	*SLC6A8*	NC_000023.10:g.152957273_152958417del NM_001142805.1:c.645-157_778-79del	Inherited from mother [HEM]	Solved [MIM :300352]	5 (PVS1,PS3,PM2)	Not detected	GS
P32/M	SV	*CHASERR*	NC_000015.9:g.93422237_93430600del	*De novo* [HET]	Solved	N/A	Non coding gene	GS RNA-seq

Table 1: Variants identified in this study. ACMG classe and evidence categories are determined following ACMG/AMP Standards and Guidelines. HOM: Homozygous; HEM: Hemizygous; HET: Heterozygous; COMP HET: Compound heterozygous; MOS: Mosaicism. NC_000001.10:g.31409001_31422000dup

Case reports

Individual P19 - Homozygous VUS in *APP*

This case involved a male presenting with severe ID, lack of language, abnormal behavior, autism, and epilepsy. Parents are related (their fathers are half-brothers). A homozygous variant inherited from healthy heterozygous parents was identified in the *APP* gene NM_000484.4:c.(440A>G) p.(His147Arg). This variant is located on chromosome 21 in a 7.08 Mb region with 99.96% homozygosity (Sup.figure 1). No other homozygous candidate variants have been identified in this region.

APP is a ubiquitous cell-surface receptor capable of dimerization. The most studied isoform is the 695 amino acid protein which is expressed at the surface of neurons. It has been observed that APP dimerization at the neuron membrane participates in neurite growth, neuronal adhesion, axonogenesis, and synaptogenesis. Mutations have been identified in the Aβ domain and are associated with early-onset Alzheimer's disease, but few damaging variants outside this domain have been documented. However, it is interesting to note that a homozygous nonsense variant NM_000484.3:c.1075C>T p.(Arg359*) associated with decreased somatic growth, microcephaly, hypotonia, developmental delay, thinning of the corpus callosum, and seizures has already been reported [35].

Nine in silico prediction tools classified this variant as damaging. Furthermore, the algorithms for predicting the impact on the stability of protein, I-Mutant and DUET,

both predicted a strong decrease in the stability of the H147R mutated protein (respectively ΔΔG = -0.75 Kcal/mol; ΔΔG = -1.07 Kcal/mol).

Histidine 147 is located in the CuBD subdomain of the E1 domain (Figure 2). It is one of the histidines involved in the chelation of copper ions. Copper binding to histidine 147 induces histidine-bridging between APP molecules. Previous experimental studies have shown the impact of His147 mutations or CuBD deletion on APP dimerization and processed fragment secretion levels. Collectively, these elements make this variant an interesting candidate for further functional studies.

Individual P14 - Deletion in *STEEP1*

This case involved a female presenting with severe GDD, no language, and axial hypotonia. MRI showed a slight widening of the pericerebral space. The SNP array showed a 6q22.33 deletion inherited from the father. Using GS we identified a *de novo* 1.7kb deletion of *STEEP1* (*CXorf56*) on chromosome X. This deletion in-frame spans exon 6 and parts of its flanking intronic sequences NM_022101.3:c.514-245_607-80del (Sup.figure 2). According to the ENCODE ChIP-seq assay [36], H3K4me1 and H3K2Ac histone mark enrichment in this region suggests the presence of regulatory elements in the vicinity of this deletion. The absence of this SV in the gnomAD database (gnomAD SVs v2.1) and the pLI score of 0.95, support its pathogenicity. Moreover, *STEEP1* loss of function has been associated with ID in several families. In these studies, male phenotypes were described as mild or

severe ID with additional features such as epilepsy, abnormal gait, or abnormal reflexes. In females, penetrance is incomplete, and expressivity is variable. Reported female cases showed non-syndromic mild intellectual disabilities and asymptomatic carriers, depending on the skewed X-inactivation pattern [37,38]. According to the ACMG guidelines, this case was classified as solved.

Individual P32 - Deletion in *CHASERR* lncRNA

This individual is a male presenting with GDD and facial dysmorphism. Clinical examination revealed widely spaced eyes, anteverted nares, a long philtrum, axial hypotonia, peripheral hypertonia, bilateral optic atrophy, recurrent unexplained fevers, and dystonic movements. At 33 months, he could not sit and did not speak, had recurrent unexplained fevers, and dystonic movements. Brain MRI at one month was normal, but MRI at four years showed global dysmyelination, thinning of the corpus callosum, reduced brainstem volume, cortical atrophy, and ventriculomegaly.

By GS we identified a *de novo* heterozygous deletion NC_000015.9:g.93422237_93430600del. This deletion was also detected by RNA-Seq (Figure 3). *CHASERR* is located approximately 1.5kb upstream of *CHD2*, whose haploinsufficiency has already been implicated in developmental and epileptic encephalopathy (MIM:615369). However, the deletion does not cover the *CHD2* promoter and the individual's phenotype is not consistent with the loss of expression of this gene. The regulatory role of *Chaserr* in *Chd2* expression has

been previously reported in a mouse model. The absence of Chaserr results in a significant increase in *Chd2* expression at both the mRNA and protein levels, subsequently causing transcriptional interference [39]. Therefore, a deleterious mechanism by the overexpression of *CHD2* was suspected. RNA-Seq allowed us to confirm skewed allelic expression. We noted an allelic imbalance skewed towards the *CHD2* allele in cis with the *CHASERR* deletion, consistent with what has been found in animal models (Sup.figure 3). No aberrant expression of *CHD2* was detected by Drop. However, the low level of expression of this gene in the blood and fibroblasts could hamper its ability to detect significant variations. We confirmed here that *CHASERR* regulates *CHD2*, an NDD-associated gene. As another case with concording phenotype was described we classified this case as resolved [10]. Nevertheless, further investigation is needed to clarify the underlying pathological mechanisms.

We also observed significant overexpression of *LIMD1* (FC=1.82; padj=4.25E-05). We did not detect any GS variants in this gene, and manual verification with IGV provided no explanation for this event. Moreover, this gene is not associated with neurodevelopment. These elements raise the question of possible regulation by *CHASERR*. However, this hypothesis remains unconfirmed.

Individual P23 - Deletion in *CBX3*

This case involved a female, the first child of non-consanguineous parents, presenting with moderate GDD, infantile muscular hypotonia, talipes valgus,

strabismus, hypertrichosis, and epicanthus. She sat at 11 months, stood at 21 months, and had bi-syllabic language at 18 months.

Trio genome sequencing revealed a balanced event with a breakpoint in the *NUS1* promoter, a gene associated with an autosomal dominant ID. However, karyotyping did not confirm this SV (Sup.figure 4). Subsequent RNA-seq revealed aberrant splicing of *CBX3* (Figure 4. A). A junction was detected between exons 2 and 5, which was absent in the mother and controls (deltaPsi: 0.47). This event was consistent with an in-frame deletion of exons 3-4. This deletion was detected in GS, but it appeared dubious in the IGV overview at first glance. Moreover, this gene was not associated with NDD; therefore, this variant seemed less relevant than the putative translocation involving *NUS1*. Close examination of the IGV overview confirmed that there was indeed a *de novo* 3.8kb deletion of exons 3-4 of *CBX3* in this individual NM_007276.4:c.25-1502_330+141del (Figure 4. B). This deletion results in a highly truncated protein (83 aa vs. 183 aa) with loss of the chromodomain, which is the main functional domain of CBX3 (Figure 4. C). In an animal model, it was shown that CBX3 is required for H4K20me3 epigenetic mark deposition and the regulation of protocadherin genes, which are involved in the mechanisms of neuronal self-avoidance [40]. We classified this event as a strong candidate VUS.

Individual P24 - Intragenic duplication in *PUM1*

This individual is a female born to non-consanguineous parents. Pregnancy was notable for oligohydramnios. Examination revealed severe ID and dysmorphism, including a wide mouth, protruding tongue, and macrocephaly. Unilateral renal hypoplasia and gastrointestinal anomalies, such as esodeviation, gastroesophageal reflux, and chronic constipation, were also observed. Brain MRI showed a periventricular leukomalacia.

Using GS, we identified a *de novo* heterozygous duplication of 13 kb in *PUM1* NM_001020658.1:c.2856+974_3435+489dup involving two L2c retrotransposons with 531bp of homology (Figure 5). OGM allowed us to characterize this event as intragenic tandem duplication. RNA-Seq analysis confirmed that the mRNA contains a 3'exon21 - 5'exon18 junction and a single last exon, as shown by a heterozygous SNP. PUM1 belongs to the PUF protein family. These RNA-binding proteins act as post-transcriptional repressors by binding to specific sequences located in the 3'-UTR of mRNA [41,42]. This in-frame duplication spans five protein domains involved in RNA binding, suggesting a deleterious impact. In addition, haploinsufficiency of *PUM1* has been reported to be responsible for GDD with concordant phenotype [43]. Therefore, we classified this case as resolved.

Discussion:

In this study, 33 individuals with NDD who were negative after ES and array-CGH

were enrolled for trio-GS complemented by RNA-seq for 9 of them and OGM for 2 individuals. GS identified pathogenic or likely pathogenic variants in 16 individuals and achieved an overall diagnostic yield of 48%. Nevertheless, it is interesting to note that a large majority of these pathogenic variants (10/16, 62,5%) were SNVs/indels; therefore, it was theoretically possible to detect them in ES. Seven of the 16 solved cases were supported by variants detected in ES but not reported at the time of the first analysis. Therefore, it is necessary to qualify the net contribution of the trio-GS technique by considering the reanalysis effect. This effect is essentially due to the publication of new gene-disease associations, and its intensity is directly correlated to the time elapsed since the last analysis [44]. Excluding the reanalysis effect, we reported a diagnostic for eight individuals reaching a GS-unique diagnostic rate of 24% which is in line with the 19%-25% reported by previous studies [6–8]. Among them, six cases were solved by SVs (6/8, 75%).

Our study highlights that short-read GS may present limitations in the interpretation of SVs. For example, duplication involving *PUM1* required OGM to be fully characterized. The use of long read GS could also overcome this limitation. Moreover, unlike OGM, it allows the precise identification of breakpoints and provides additional information on DNA methylation status. However, the current cost of these techniques does not facilitate their implementation in diagnostic laboratories.

RNA-Seq contributed to the identification of molecular causes in two probands and a strong candidate in one. This was the case for the *CBX3* deletion in individual P23. Even though it was detected in GS, the IGV preview was dubious, and this variant was eliminated. Doubts regarding the IGV preview can be attributed to three reasons: the overall number of false-positive SVs detected in this individual, the absence of this gene in the NDD databases, and the detection of a more compelling variant. In this individual, RNA-seq confirmed the variant in *CBX3* and eliminated other candidates. Furthermore, its effect on the transcript was immediately delineated, confirming the existence of a truncated transcript rather than a total loss of expression. This case highlights one of the advantages of using RNA-seq in addition to inconclusive. Recent algorithms for the analysis of GS data have shown high sensitivity at the cost of many false positives [45]. Verification of these numerous variants is a time-consuming task prone to error and can therefore lead to discarding the causal variant. RNA-seq appears here as a technique that can counterbalance a low specificity of GS in SVs detection by proposing a short list of variants with proven transcriptional effects.

The deletion in the *CHASERR* lncRNA was detected by GS; however, to our knowledge, no guidelines for the prediction of the impact of variants in non-coding RNAs have been validated for diagnosis. In this case, the identification of an allelic bias in the *CHD2* gene in -cis of the deletion is consistent with a pathogenic mechanism, previously observed in mouse and cell models. As GS implementation becomes progressively more widespread, an increasing number of variants,

especially SVs, will be identified in non-coding RNAs generating a significant number of VUS. Interpretation of these events based on GS data alone is challenging.

However, clinical RNA-seq has several limitations. Notably, RNA-seq provides information that may be unusual for biologists and clinicians. In particular, we identified highly significant aberrant expression or aberrant splicing confirmed in IGV, but it has not been possible to link these to a genomic variant. Understanding these phenomena and their medical relevance is time-consuming and can only be envisaged for a limited number of patients. In addition, the small number of families that accepted to benefit from RNA-seq did not allow us to draw any conclusions on the systematic application of RNA-seq in patients with inconclusive GS. However, it appears from our study that integration on a case-by-case basis is useful for filtering and interpreting SVs.

The choice of the biological tissue must also be considered. The first diagnostic RNA-seq tests were performed on blood samples. However, this tissue offered a low diagnostic yield for most diseases, except for hematological and immune diseases. A comparative study showed better performance of RNA-seq in fibroblasts [10]. Explanations for the low yield of whole-blood assays include the cellular heterogeneity of this tissue and the large transcriptional variations that can affect leukocytes. However, we cannot rule out that the expression and splicing

defects identified in *CBX3*, *PUM1*, and *CHASERR*, which are expressed in blood, could have been identified in blood samples.

We observed that approximately 60% of NDD-associated genes [2] were expressed in fibroblasts with TPM>10, while only 37% were expressed in blood, making fibroblasts more relevant for NDD diagnosis. However, the skin biopsy required to collect fibroblasts may compromise their widespread use. The identification of alternative clinically accessible tissues (CAT), such as buccal cells, hair follicles, or urine cells, could help overcome this limitation. We performed gene expression analysis on one alternative cell type, urine-derived stem cells [46], and observed that they expressed approximately 60% of the NDD-associated genes (TPM >10). This might be a promising alternative CAT for RNA-Seq studies.

It should be noted that the choice of tissue remains a compromise between biological relevance and practical application. Even if a gene associated with neurodevelopment is expressed in a cell type, tissue-specific differences exist in gene expression or splicing. Furthermore, it is illusory to attempt to recapitulate the expression and the splicing pattern complexity of neurodevelopment with a single cell type.

The application of RNA-seq on a case-by-case basis in patients with inconclusive GS presenting with a severe phenotype seems to be an interesting approach. Diagnostic RNA-seq is still in its infancy, and there are several areas for

improvement; however, it is on the way to becoming a standard tool to enhance molecular diagnosis in NDD.

Acknowledgments:

This study was performed thanks to a collaboration between CEA and Nantes hospital. This study was supported by the program “High throughput sequencing and rare diseases” of the Foundation for Rare Diseases. This work has been carried out within the framework of the FHU GenOMedS thanks to the support of the Health cooperation group of University Hospitals of the Great West (GCS HUGO) and the National Alliance for Life Sciences and Health (Aviesan). The CEA-CNRGH sequencing platform was supported by the France Génomique National infrastructure, funded as part of the « Investissements d’Avenir » program managed by the Agence Nationale pour la Recherche (contract ANR-10-INBS-09)". We are most grateful to the Bioinformatics Core Facility of Nantes BiRD, member of Biogenouest, Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013) for the use of its resources and for its technical support.

Funding:

This work was supported by the Foundation for Rare Diseases and Groupama Foundation.

Competing of interests:

None declared.

Contributorship statement:

Conceptualisation, supervision of the overall project and edition of the final manuscript draft : KR, BC, SB.

Validation, writing—original draft: KR, BC, SB.

Data collection : BI, SM, MN, EC, DB, LP, SO, XGH, GG, AT, KR, OP, MDF, BC, SB.

Data analysis and interpretation : KR, OP, MDF, BC.

All authors critically revised the manuscript for important intellectual content. They take full responsibility for the integrity of the data and the accuracy of the data analysis.

Guarantors: KR, SB, BC.

Ethics approval statement :

Informed consent was obtained form participants or the legal guardians according to the French law on Bioethics and following the Helsinki Declaration.This study was approved by the CHU de Nantes ethics committee (Research Programme “Génétique Médicale" DC-2011-1399).

Figures :

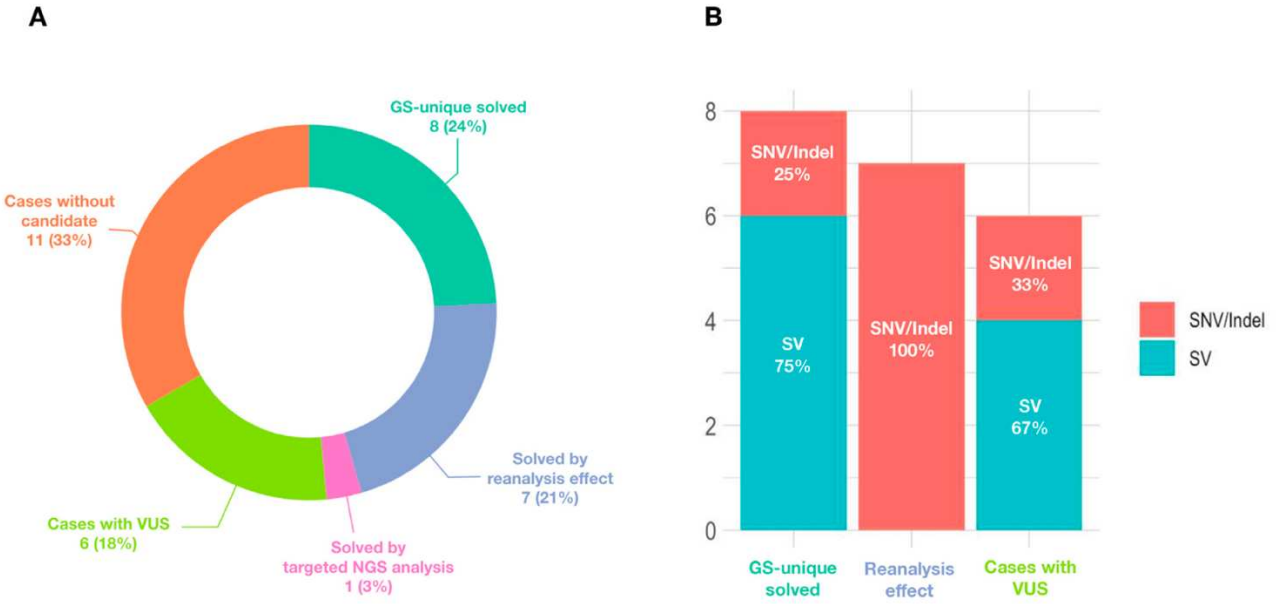


Figure 1: Performance of GS for NDD diagnosis after inconclusive ES. A) Overall diagnostic yield of the combined approach. The GS-unique diagnostic yield was calculated as the percentage of probands in the cohort who received a conclusive molecular diagnosis that would have been unattainable solely through the reanalysis of ES data. Solved by the reanalysis effect are cases solved owing to variants that were detected by ES but were not classified as pathogenic at the time of analysis. Better support for pathogenicity, a recent publication of the genes involved, or additional investigations allowed us to revise their classification. B) Types of variants identified at the case level. No compound heterozygous involving both SNV/indel and SVs were reported.

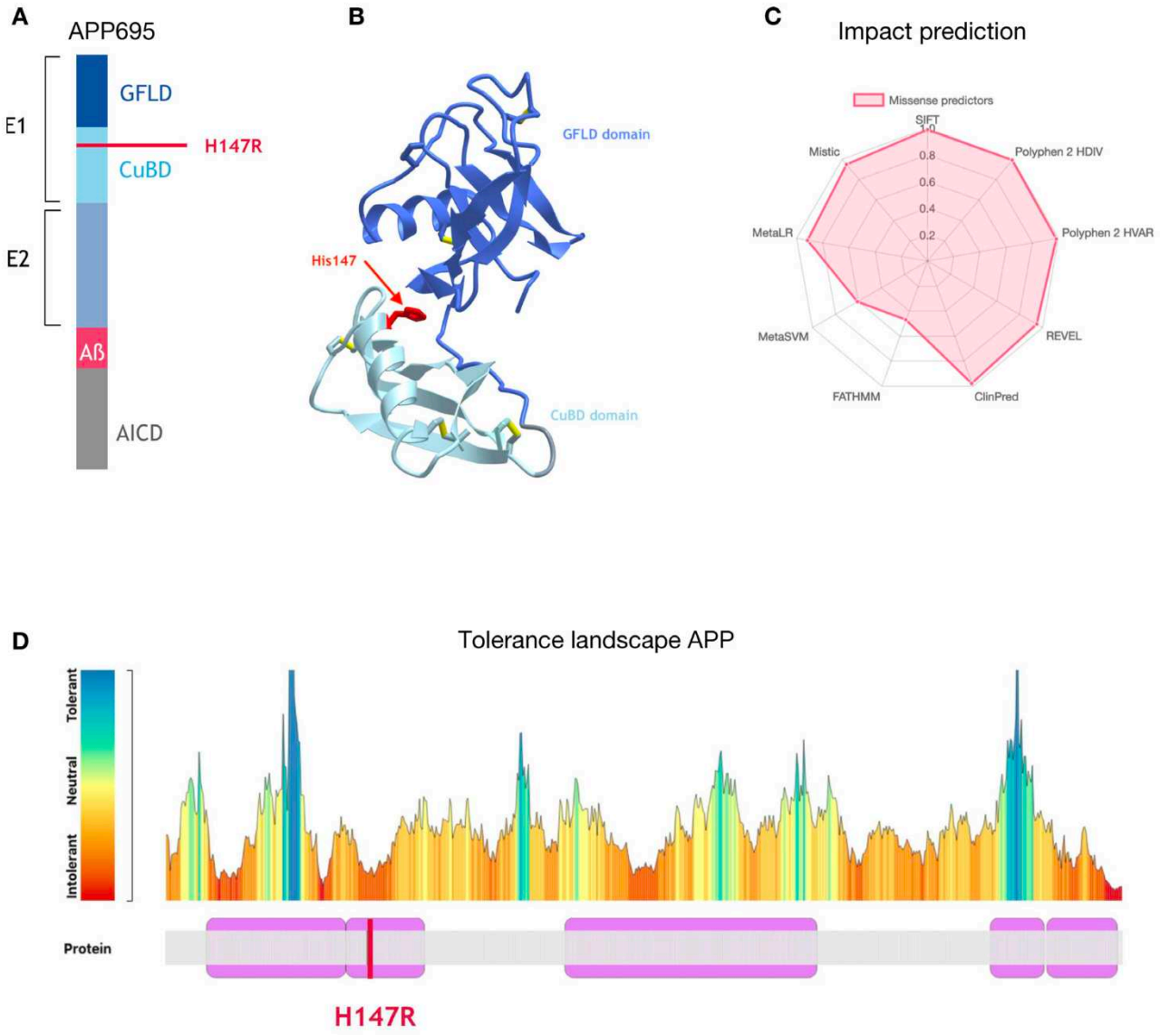


Figure 2: Homozygous VUS in *APP;* NM_000484.3:c.[440A>G] [440A>G p.[H147R] [H147R]. A) The H147 variant is located in the CuBD (copper-binding domain) subdomain of the E1 extracellular domain. Histidine 147 is directly involved in copper binding. B) Protein structure (PDB accession number 4PWQ). C) H147 is predicted pathogenic by nine prediction tools aggregated by Mobidetails. D) Tolerance landscape of APP (695aa) generated by Metadome. H147 is located in a region predicted to be intolerant to mutation.

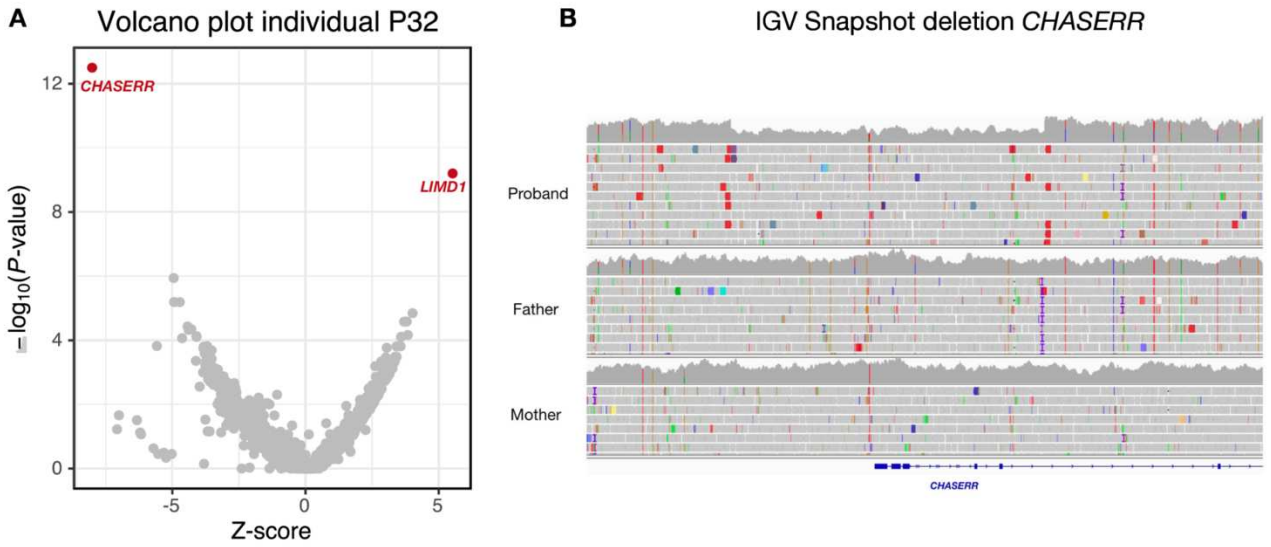


Figure 3: *De novo* heterozygous deletion in lncRNA *CHASERR*; NR_037600.1:g.92,879,007_92,887,374del. A) Volcano plot from OUTRIDER module in individual P32 showing aberrant expression of *CHASERR* in fibroblasts. B) IGV snapshot showing a 8.4 kilobases deletion covering the *CHASERR* promoter and exons 1-3.

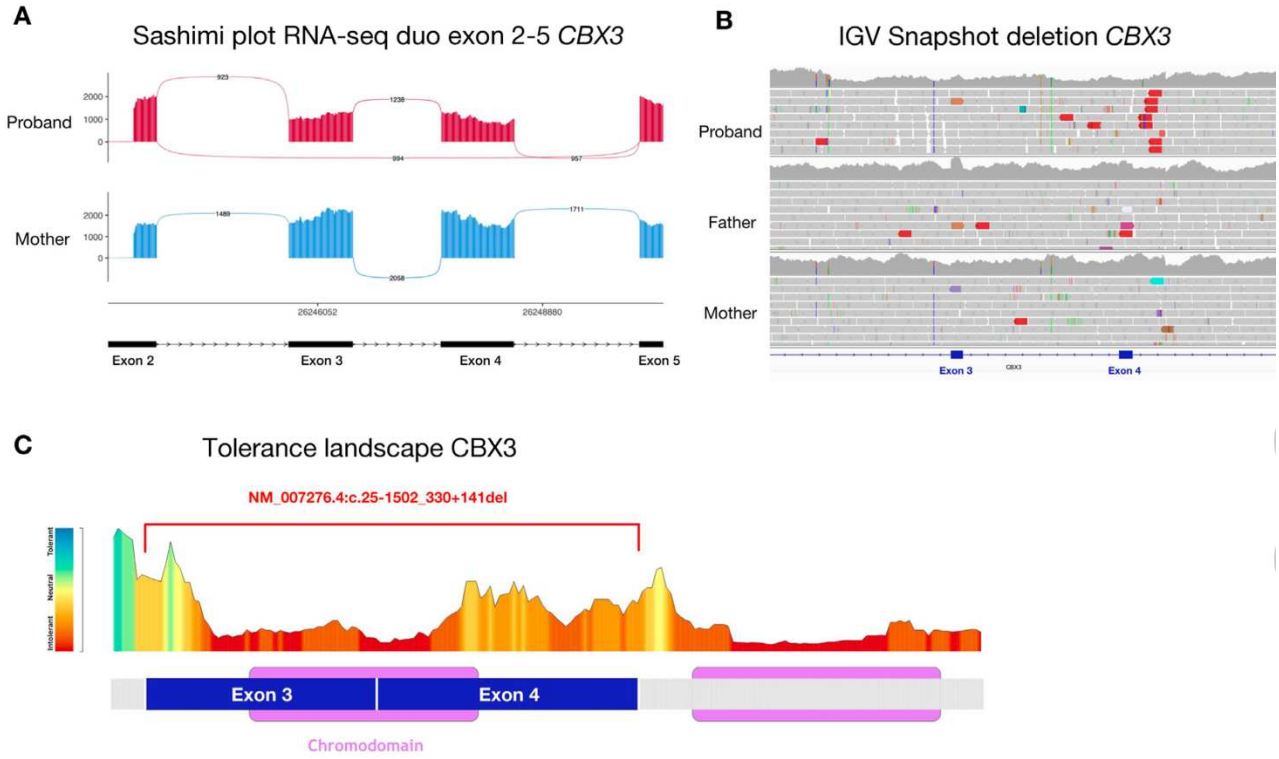


Figure 4: *De novo* heterozygous deletion in *CBX3*; NM_007276.4:c.25-1502_330+141del. A) Sashimi plot RNA-Seq on fibroblasts from P23 individual and healthy mother showing a heterozygous aberrant junction between exons 2 and 5 of CBX3. B) IGV snapshot showing a 3.8 kilobases deletion covering exons 3-4. C) Tolerance landscape of CBX3 (183 aa) generated by Metadome. The exons 3-4 codes for a region predicted to be highly intolerant to mutation. This region contains the functional chromodomain.

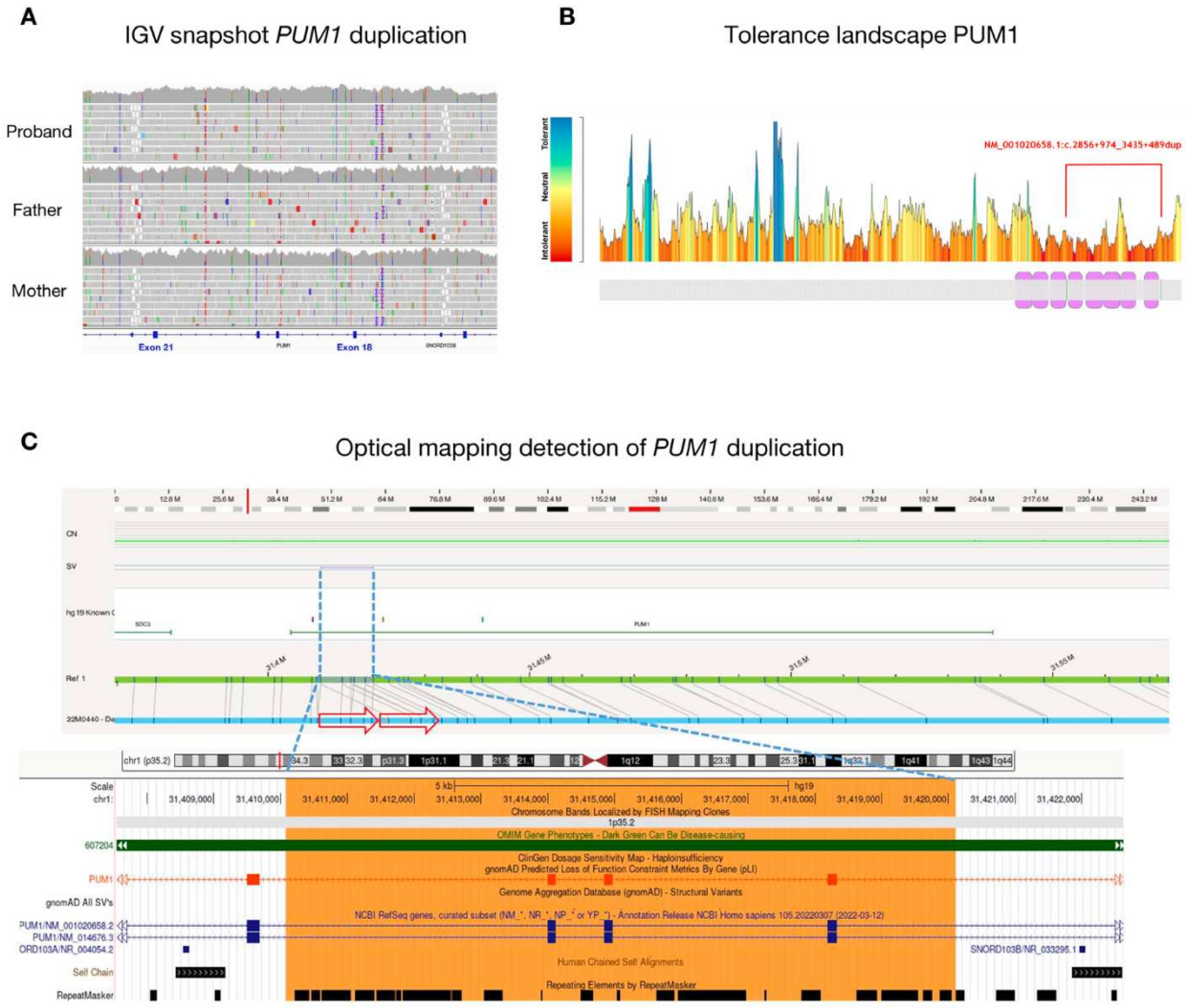
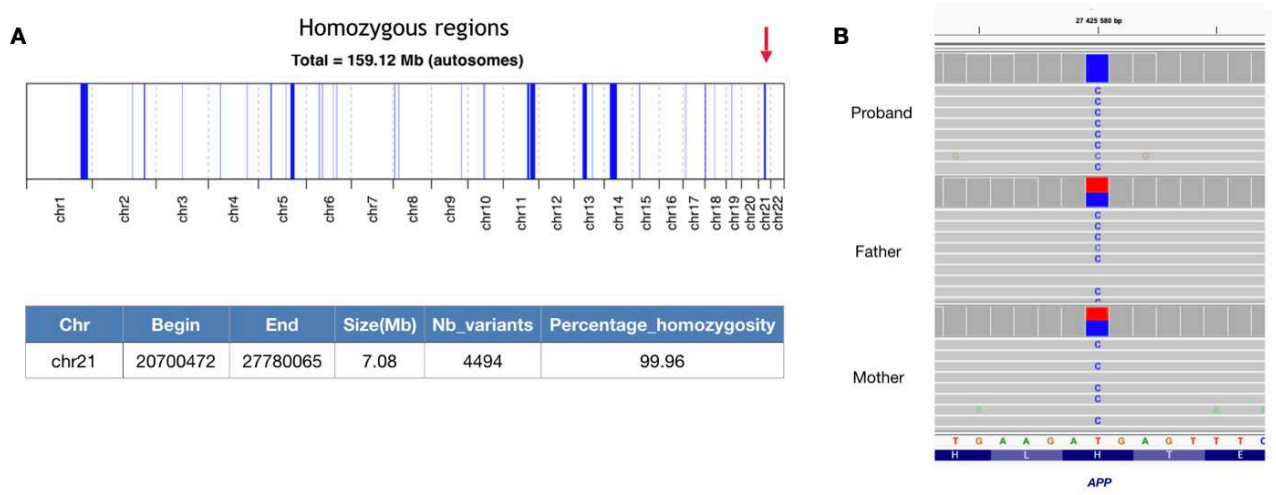


Figure 5: *De novo* heterozygous duplication in *PUM1*; NM_001020658.1:c.2856+974_3435+489dup. A) IGV snapshot showing a 13 kb duplication involving exons 18-21. B) Tolerance landscape of PUM1 (1224 aa) generated by Metadome. The exons 18-21 codes for a region predicted to be highly intolerant to mutation. This region contains functional domains involved in RNA binding. C) Optical map of *PUM1* locus showing an intragenic tandem duplication between two L2c retrotransposons.

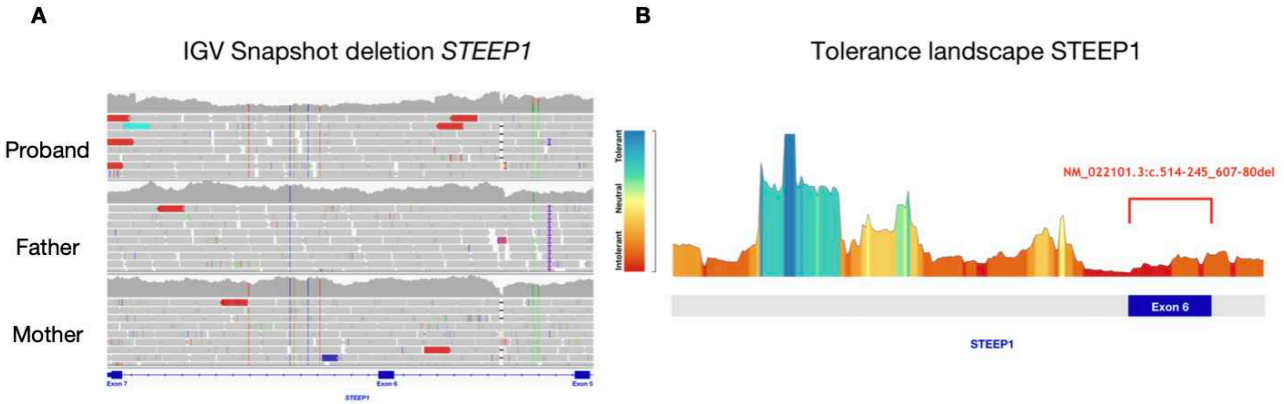
Supplemental material

Supplemental figures :

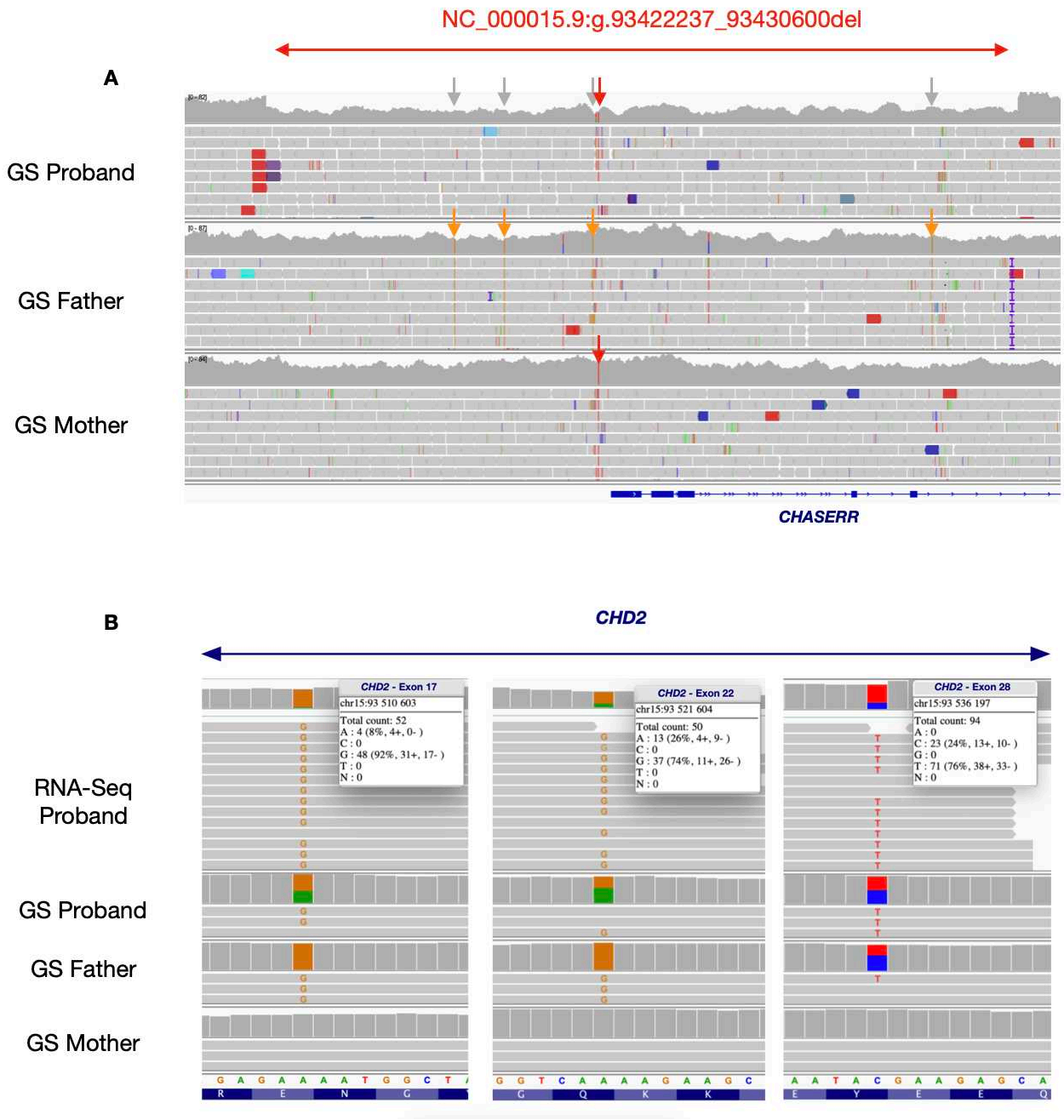


Chr	Begin	End	Size(Mb)	Nb_variants	Percentage_homozygosity
chr21	20700472	27780065	7.08	4494	99.96

Sup Figure 1: P19. A) Automap estimation shows 159.12 Mb of homozygous regions, confirming the consanguinity of the parents. High homozygosity of the *APP* locus (7.08 Mb ; 99.96%). No other homozygous candidate variants were identified in the LOH region. B) IGV snapshot showing homozygous variant NM_000484.3:c.[440A>G] in *APP*.



Sup Figure 2: P14. *De novo* heterozygous deletion in *STEEP1*; NM_022101.3:c.514-245_607-80del A) IGV snapshot showing a 1.7kb deletion leading to the lost of the exon 6 (in-frame). B) Tolerance landscape of STEEP1 generated by Metadome. The exon 6 codes for a region predicted to be highly intolerant to mutation.



Sup Figure 3: P32. A) IGV snapshot of trio-GS at the *CHASERR* locus. Four homozygous SNP (orange arrow) present in the father were absent from the deleted locus of the proband. One homozygous SNP (red arrow) in the mother was found at the deleted locus of the proband. This indicates that the deleted allele was paternal. B) IGV snapshot showing allelic imbalance skewed towards the CHD2 allele in cis- with *CHASERR* deletion.

Sup Figure 4: P23. A) IGV snapshot showing a balanced event in the NUS1 promoter. This event was also observed in the mother. As NUS1 one is already associated with NDD we suspected a mosaicism in the mother. B) Karyotyping of the proband appears normal, invalidating the t(6;7) hypothesis.

Sup Table 1 :

Case/Sex	Type	Gene	Variant	Transmission	Interpretation	ACMG	Reason not reported by ES	Technique applied
P19/M	SNV indel	*APP*	NC_000021.8:g.27425580T>C NM_000484.4:c.440A>G:p.(His147Arg)	Autosomal recessive [HOM]	Strong candidate VUS	NA	More support for pathogenicity	GS
P23/F	SV	*CBX3*	NC_000007.13:g.26244486_26248316del NM_007276.4:c.25-1502_330+141del	*De novo* [HET]	Strong candidate VUS	N/A	Not detected	GS RNA-seq
P31/M	SNV indel	*SYT7*	NC_000011.9:g.61290738C>G NM_001252065.1:c.1141G>C:p.(Val381Leu)	*De novo* [HET]	VUS	N/A	More support for pathogenicity	GS
P3/F	SV	-	NC_000001.10:g.pter_69716317delins[NC_000006.11:g.108848601_qterinv] NC_000006.11:g.108848601_qterdelins[NC_000001.10:g.69716317_pterinv]	*De novo* [HET]	VUS	N/A	Copy neutral translocation not detected by ES	GS
P8/F	SV	-	NC_000005.9:g.105706517_170958911inv	*De novo* [HET]	VUS	N/A	Copy neutral inversion not detected by ES	GS
P25/M	SV	*FGF18*	Complex structural variant duplication and inverted insertion	*De novo* [HET]	VUS	N/A	Complex event not detected by ES	GS RNA-seq OGM

Sup Table 1: Variants of uncertain significance identified in this study. ACMG classe and evidence categories are determined following ACMG/AMP Standards and Guidelines. HOM: Homozygous; HEM: Hemizygous; HET: Heterozygous; COMP HET: Compound heterozygous; MOS: Mosaicism.

Sup Table 2 :

Case/Sex/Age range	Phenotype	ES date	ES kit
P1/F/20-40 y	Intellectual disability severe (IQ 20), global developmental delay, hypotonia, absent speech, autism, epilepsy (since the age of 7 year), cerebral MRI : small periventricular heterotopia and dilated, cortical veins, extensive naevus, height –1.7 SD, weight +0.5 SD	TRIO/2016	1
P2/F/10-20 y	Mild pulmonary valve stenosis, left renal cyst, pyelonephritis, stridor, severe laryngomalacia, small external auditory canals, gastrostomy until the age of 5 years, sitting without support at 1 year, walking acquired at 2.5 years, MRI at 2 months: delayed myelination of brainstem and internal capsules, bruxism, stereotypies, height -1DS , weight -3DS , head circumference -4DS	TRIO/2016	1
P3/F/2-10 y	Axial hypotonia, severe global developmental delay, strabismus, head circumference -2,2 SD	TRIO/2016	1
P4/M/2-10 y	Moderate global developmental delay, severe language delay with first words at 3 years old, behavior troubles, stereotypy	TRIO/2016	1
P5/M/10-20 y	Postnatal growth delay (- 2.5 SD), microcephaly (- 3 SD), global developmental delay, facial dysmorphism, cryptorchidism, inguinal hernia, laryngomalacia, Turricephaly, unilateral pachygyria, patent sagittal suture	TRIO/2016	1
P6/F/20-40 y	Epileptic encephalopathy, facial dysmorphism, ventricular dilatation, myelination delay	TRIO/2016	1
P7/F/10-20 y	Hypotonia, severe global developmental delay, behavior troubles, trunk hyperpigmentation, short stature -3 SD, bilateral hearing impairement, short corpus callosum	TRIO/2016	1
P8/F/10-20 y	Axial hypotonia, peripheral hypertonia, severe global developmental delay, scoliosis, dysmorphic ventricles, thin corpus callosum	TRIO/2016	1
P9/F/10-20 y	Axial hypotonia, facial dysmorphism, severe global developmental delay predominantly on language, nystagmus	TRIO/2016	1
P10/F/10-20 y	Craniosynostosis, mild global developmental delay, bilateral microtia, strabismus, choanal stenosis, short stature (-2 SD)	TRIO/2016	1
P11/M/10-20 y	Epileptic encephalopathy, intellectual disabibility, behavior troubles	TRIO/2016	1
P12/F/10-20 y	Hypotonia, global developmenatal delay, seizures, no speech, facial dysmorphism	TRIO/2016	1
P13/M/10-20 y	Myoclonic epilepsy, encephalopathy, axial hypotonia, languange acquisition then regression, cortical atrophy on cerebral MRI, myelinisation delay, thin corpus callosum	TRIO/2016	1
P14/F/10-20 y	Severe global developmental delay, sat at 15 months, walked at 3.5 years old, no language, axial hypotonia	TRIO/2016	1
P15/F/20-40 y	Severe global developmental delay, walked at 4 years old, short stature, cerebellar atiophy	TRIO/2016	1
P16/M/10-20 y	Severe global developmental delay, macrocephaly +2 SD	TRIO/2016	1
P17/F/10-20 y	Global developmental delay, behavior troubles, clinodactyly, enlarged cerebral ventricles	TRIO/2016	1

Sup Table 2. Exome sequencing kit : 1:SureSelectXT Clinical Research Exome system (Agilent Technologies); 2:SureSelectXT Clinical Research Exome V2 system (Agilent Technologies); 3:Human Comprehensive Exome (Twist Bioscience).

Case/Sex/Age	Phenotype	ES date	ES kit
P18/F/10-20 y	Epilepsy, short stature, microcephaly, facial dysmorphism, cerebral MRI : cortical atrophy	TRIO/2016	1
P19/M/10-20 y	Severe intellectual disability, no language, abnormal behavior, autism, epilepsy	TRIO/2016	1
P20/M/10-20 y	Severe global developmental delay, hypotonia, severe intellectual disability, no speech	TRIO/2016	1
P21/M/10-20 y	Severe global developmental delay, Plagiocephaly, Hypertrichosis, Hypertelorism, Global developmental delay, Epicanthus, Corneal astigmatism, Abnormality of earlobe, Abnormal facial shape	SOLO/2018	2
P22/F/2-10 y	Short stature, Seizures, Moderate global developmental delay, Global developmental delay, Epicanthus, Delayed speech and language development, Abnormal facial shape	SOLO/2020	3
P23/F/2-10 y	Relative microcephaly -1.5SD, microcephaly relative, Talipes valgus, Strabismus, Moderate global developmental delay, Infantile muscular hypotonia, Hypertrichosis, Epicanthus, Delayed speech and language development, Decreased fetal movement	TRIO/2020	3
P24/F/2-10 y	Wide mouth, Unilateral renal hypoplasia, Skin plaque, Protruding tongue, Periventricular leukomalacia, Oligohydramnios, Macrocephaly, Intellectual disability, severe, Gastroesophageal reflux, Failure to thrive, Esodeviation, Chronic constipation	TRIO/2018	1
P25/M/20-40 y	Nystagmus, Nevus, Neonatal hypotonia, Macrocephaly, Intellectual disability, severe, Global developmental delay, Cerebellar hypoplasia, Autism, Absent speech	SOLO/2020	3
P26/F/2-10 y	Severe global developmental delay, Scoliosis, Ptosis, Pachygyria, Neonatal hypotonia, Laryngomalacia, Global developmental delay, Focal clonic seizures	SOLO/2020	3
P27/M/2-10 y	Synophrys, Status epilepticus, Seizures, Recurrent singultus, Moderate global developmental delay, Infantile spasms, Hypoplasia of the corpus callosum, Fused teeth, Delayed speech and language development, Abnormal behavior	SOLO/2018	2
P28/M/10-20 y	Specific learning disability, Spasticity, Postural instability, Loss of speech, Loss of ability to walk, Global developmental delay, Dysphagia, Delayed gross motor development, Delayed ability to walk, Congenital sensorineural hearing impairment, Cerebellar ataxia associated with quadrupedal gait, Autistic behavior, Ataxia, Abnormal pyramidal sign	SOLO/2017	2
P29/F/2-10 y	duplication pituitary, Wide anterior fontanel, Vertebral segmentation defect, Umbilical hernia, Tufted angioma, Tuberous angioma, Thin vermilion border, Sleep apnea, Posterior pharyngeal cleft, Polyhydramnios, moderate intellectual disability, Hypertelorism, Enlarged cisterna magna, Cerebellar dysplasia, Agenesis of cerebellar vermis, Abnormal facial shape	TRIO/2018	1
P30/M/10-20 y	Severe global developmental delay, Seizures, Periventricular white matter hyperdensities, Long fingers, Severe intellectual disability, Global developmental delay, Epileptic encephalopathy, Dystonia, Decreased thalamic volume, Decreased serum ceruloplasmin, Cerebral cortical atrophy	TRIO/2018	1
P31/M/10-20 y	Strabismus, Sleep disturbance, Short stature, Short neck, Retinal detachment, Premature birth, Preaxial polydactyly, Micropenis, Microcephaly, Low-set ears, Involuntary movements, Hypotelorism, Hypermetropia, Duplication of thumb phalanx, Duodenal atresia, Decreased body weight, Clinodactyly of the 5th finger, Arrhythmia	TRIO/2017	1
P32/M/2-10 y	Short 1st metacarpal, Scoliosis, Ptosis, Prominent nose, Optic atrophy, Moderate global developmental delay, Long philtrum, Infantile muscular hypotonia, Global developmental delay, Feeding difficulties, Dystonia, Camptodactyly of toe, Anteverted nares, Adducted thumb, Abnormal facial shape	SOLO/2019	3
P33/F/2-10 y	Short stature, Severe global developmental delay, Retinopathy, Infantile muscular hypotonia, Global developmental delay, Feeding difficulties, Failure to thrive, Abnormal facial shape	TRIO/2020	3

Sup Table 2. Exome sequencing kit : 1:SureSelectXT Clinical Research Exome system (Agilent Technologies); 2:SureSelectXT Clinical Research Exome V2 system (Agilent Technologies); 3:Human Comprehensive Exome (Twist Bioscience).

Sup Table 3: DROP pipeline calls:

OUTRIDER calls

hgncSymbol	geneID	sampleID	pValue	padjust	zScore	l2fc	rawcounts	normcounts	meanCorrected	theta	aberrant	AberrantBySample	AberrantByGene	padj_rank	foldChange
CEP164	ENSG00000110274.16_7	P16	7.91706977897483e-09	0.00112830715723481	-6.12	-0.67	1039	1003.45	1598.74	206.2	TRUE	1	1	1	0.63
ADAMTSL1	ENSG00000178031.17_5	P22	2.693399238081e-08	0.00383851819228388	-6.46	-2.38	651	2074.02	10784.15	19.18	TRUE	1	2	1	0.19
PCGF6	ENSG00000156374.16_7	P23	7.19813936306138e-14	1.02584824800767e-08	-7.83	-0.85	418	271.8	489.56	290.76	TRUE	4	1	1	0.55
AL355802.1	ENSG00000219470.1_5	P23	1.51973280652435e-10	1.08292932268939e-05	-7.93	-3.88	13	2.44	38.56	14.64	TRUE	4	2	2	0.07
SLC39A6	ENSG00000141424.13_7	P23	2.72051890679604e-08	0.0012923893307439	-5.94	-0.77	3576	3669.39	6237.4	135.08	TRUE	4	1	3	0.59
SURF6	ENSG00000148296.7_5	P23	7.59799534826747e-08	0.00270708506158862	-5.68	-0.9	1552	395.06	735.48	96.22	TRUE	4	1	4	0.54
PAIP1	ENSG00000172239.14_9	P24	2.3756323223411e-16	3.38565028108532e-11	-8.71	-0.73	2337	2165.65	3580.37	347.75	TRUE	13	1	1	0.6
MRPL13	ENSG00000172172.8_6	P24	2.04664622206452e-15	1.45839663230934e-10	-8.34	-0.69	1211	881.39	1418.08	406.85	TRUE	13	1	2	0.62
PTMA	ENSG00000187514.16_5	P24	2.34183707520426e-13	1.11249557677174e-08	-7.85	-0.9	24920	8638.92	16134.07	169.44	TRUE	13	1	3	0.54
SAP18	ENSG00000150459.12_6	P24	7.47108499340372e-12	2.66186824977642e-07	-6.91	-0.37	4734	3168.89	4101.33	891.14	TRUE	13	1	4	0.77
BTBD1	ENSG00000064726.10_7	P24	4.29234874458904e-10	1.2234546227518e-05	-6.45	-0.44	3214	2955.35	4015.07	521.56	TRUE	13	1	5	0.74
SNHG7	ENSG00000233016.7_7	P24	6.23995635611959e-10	1.48215345954975e-05	-6.86	-1.67	290	234.66	748.71	44.25	TRUE	13	3	6	0.31
RAD51AP1	ENSG00000111247.14_3	P24	7.93506082024162e-07	0.0161553023431344	3.15	0.45	922	476.65	348.89	323.03	TRUE	13	1	7	1.37
NOL7	ENSG00000225921.7_4	P24	1.23842316079013e-06	0.0220618510859235	-4.68	-0.3	2807	1495.67	1846.92	691.48	TRUE	13	1	8	0.81
HNRNPA3	ENSG00000170144.20_6	P24	2.01649871083776e-06	0.0319314256903383	-4.86	-0.34	14157	9521.67	12033.88	457.98	TRUE	13	1	9	0.79
MPPE1	ENSG00000154889.17_8	P24	2.38569897354456e-06	0.0339999684480061	-5.08	-0.73	660	576.68	955.75	117.52	TRUE	13	1	10	0.6
CLN5	ENSG00000102805.16_7	P24	3.15565370648955e-06	0.0408845784223026	-4.9	-0.62	1014	1026.79	1579.98	151.32	TRUE	13	1	11	0.65
AC026271.1	ENSG00000174977.8_5	P24	3.69323516717925e-06	0.0438620157543023	-3.4	-0.75	363	70.47	117.92	118.15	TRUE	13	1	12	0.59
GNAQ	ENSG00000156052.11_7	P24	4.37610158212878e-06	0.0479741084648332	-4.81	-0.47	2884	2794.9	3864.16	240.77	TRUE	13	1	13	0.72
ZWINT	ENSG00000122952.17_7	P25	3.52101712240169e-12	5.01800404804134e-07	-7.57	-1.02	209	530.46	1079.33	191.25	TRUE	2	1	1	0.49
GLA	ENSG00000102393.12_5	P25	1.16395556966866e-10	8.29410019505052e-06	5.71	0.89	2621	1443.13	780.68	92.67	TRUE	2	1	2	1.85
PAWR	ENSG00000177425.11_7	P26	1.98513234596548e-08	0.00282912630119728	-6.42	-1.46	286	770.33	2118.03	46.7	TRUE	1	1	1	0.36
LINC01578	ENSG00000272888.7_7	P32	2.95917169321535e-13	4.21728580668628e-08	-8.02	-0.97	631	763.15	1500.14	171.14	TRUE	2	1	1	0.51
LIMD1	ENSG00000144791.10_5	P32	5.96167815423907e-10	4.24816524190683e-05	5.53	0.86	7904	2574.43	1422.53	89.17	TRUE	2	1	2	1.82
PSAP	ENSG00000197746.14_6	P33	1.90417407966811e-10	2.71374801876351e-05	-6.88	-0.89	26055	25526.16	47142.42	132.47	TRUE	14	1	1	0.54
SIAE	ENSG00000110013.13_5	P33	3.87722427435622e-09	0.000276282767557415	-6.68	-1.24	384	757.1	1788.13	68.87	TRUE	14	1	2	0.42
SRSF3	ENSG00000112081.17_4	P33	1.22411463731532e-08	0.000581518728990168	-6	-0.53	4770	6494.96	9381.54	284.52	TRUE	14	1	3	0.69
OS9	ENSG00000135506.16_7	P33	9.24079886421774e-08	0.00303969293380115	-5.63	-0.53	9416	9552.66	13753.14	248.28	TRUE	14	1	4.5	0.69
DDOST	ENSG00000244038.10_7	P33	1.06644103554827e-07	0.00303969293380115	-5.54	-0.44	11196	9977.23	13487.14	352.99	TRUE	14	1	4.5	0.74
KPNA6	ENSG00000025800.14_5	P33	4.03359166635939e-07	0.0092147195929882	-5.44	-0.35	1982	3652.53	4653.46	599.04	TRUE	14	1	6.5	0.78
HEXA	ENSG00000213614.10_8	P33	4.52602859784344e-07	0.0092147195929882	-5.39	-0.7	1821	4213.41	6834.76	135.41	TRUE	14	2	6.5	0.62
NPIPA1	ENSG00000183426.17_6	P33	6.95495162257014e-07	0.0123898770521256	-5.44	-0.79	958	2919.55	5053.73	106.47	TRUE	14	1	8	0.58
EVI5	ENSG00000067208.14_5	P33	1.45747241951895e-06	0.0230791976257957	-5.54	-0.76	298	636.39	1078.43	132.05	TRUE	14	1	9	0.59
ZC3H11A	ENSG00000058673.16_7	P33	1.83547322700989e-06	0.0230869143383949	-5.11	-0.45	2329	6438.39	8763.01	294.34	TRUE	14	1	11	0.73
CDCA2	ENSG00000184661.14_4	P33	1.94394631596136e-06	0.0230869143383949	4.64	0.75	245	999.94	592.41	115.35	TRUE	14	1	11	1.68
ZBED6	ENSG00000257315.2_5	P33	1.87895281858377e-06	0.0230869143383949	-5.11	-0.45	2328	6427.63	8753.07	292.65	TRUE	14	1	11	0.73
PRCP	ENSG00000137509.11_8	P33	2.37838874908613e-06	0.0260736817184878	-5.03	-0.63	3184	3352.01	5184.04	140.88	TRUE	14	1	13	0.65
QTRT2	ENSG00000151576.10_5	P33	2.63144340318087e-06	0.0267872954853666	-5.56	-0.55	266	576.13	846.95	341.09	TRUE	14	1	14	0.68

FRASER calls 1/4

sampleID	seqnames	start	end	width	strand	hgncSymbol	addHgncSymbols	type	pValue	padjust	zScore	psiValue	deltaPsi	meanCounts	meanTotalCounts	counts	totalCounts	pValueGene	padjustGene	STRAND_SPECIFIC	PAIRED_END
P16	chr17	16285552	16285553	2	+	UBB		theta	2.1935e-07	0.1036	-2.79	0.77	-0.23	7657.4	7831.8	5750	7494	1.3161e-06	0.0814028970125354	reverse	TRUE
P16	chr2	189864035	189864036	2	+	COL3A1		theta	9.0462e-07	0.1036	-2.73	0.88	-0.12	7119.7	7175.6	3986	4545	4.5231e-06	0.0814028970125354	reverse	TRUE
P16	chr6	29913059	29913227	169	+	HLA-A		psi5	1.7356e-06	0.35532	-2.44	0.8	-0.19	4260.6	4304.4	1608	2002	2.77696e-05	0.121896308925597	reverse	TRUE
P16	chr6	29913059	29977311	64253	+	HLA-J	HCG9;HLA-A;HLA-W;DDX39BP2	psi5	1.7356e-06	0.35532	2.71	0.2	0.2	39.3	4304.4	393	2002	2.6034e-05	0.121896308925597	reverse	TRUE
P16	chr22	33256107	33256108	2	+	TIMP3		theta	2.6081e-06	0.16183	-2.73	0.86	-0.14	2057.3	2094.1	2192	2560	1.04324e-05	0.104825209300007	reverse	TRUE
P16	chr2	69476241	69476242	2	+	ANTXR1		theta	2.7833e-06	0.16183	-2.77	0.19	-0.8	304.1	339	81	430	1.39165e-05	0.104825209300007	reverse	TRUE
P16	chr12	125397202	125398113	912	-	UBC		psi5	3.3066e-06	0.35532	2.63	0.19	0.19	81.5	5046.9	815	4213	0.0002347686	0.456865118361179	reverse	TRUE
P16	chr15	40328574	40328575	2	-	SRP14		theta	4.4688e-06	0.18777	-2.76	0.9	-0.1	2136	2157.1	1939	2149	2.2344e-05	0.121896308925597	reverse	TRUE
P16	chr7	135357554	135357555	2	+	STMP1		theta	5.7539e-06	0.18777	2.76	0.27	0.27	11.1	626.5	110	411	2.87695e-05	0.121896308925597	reverse	TRUE
P16	chr17	55917117	55917118	2	-	MRPS23		theta	7.0723e-06	0.19333	-2.78	0.62	-0.38	352.6	360.3	123	200	2.82892e-05	0.121896308925597	reverse	TRUE
P16	chr1	154574511	154574512	2	-	ADAR		theta	8.1803e-06	0.2117	-2.77	0.47	-0.52	255.1	268.2	118	249	4.09015e-05	0.158618133685881	reverse	TRUE
P16	chr19	1036534	1036535	2	+	CNN2		theta	1.0581e-05	0.23961	-2.77	0.88	-0.12	1033.2	1047.7	1026	1171	9.5229e-05	0.273019937287605	reverse	TRUE
P16	chr17	78183516	78183517	2	-	SGSH		theta	1.1448e-05	0.25395	-2.78	0.4	-0.59	181.4	195.3	91	230	5.724e-05	0.194020604401221	reverse	TRUE
P16	chr6	29857163	29857164	2	+	HLA-H		theta	2.5976e-05	0.43437	-2.16	0.8	-0.2	2342.3	2372.2	687	863	0.000467568	0.7043856352098	reverse	TRUE
P16	chr1	225707024	225707025	2	-	ENAH		theta	2.792e-05	0.45294	-2.68	0.85	-0.15	635.6	649.4	780	917	0.00030712	0.533853034669166	reverse	TRUE
P16	chr14	105235989	105235990	2	-	AKT1		theta	2.9492e-05	0.46872	-2.77	0.89	-0.11	678.8	684.9	482	543	0.000176952	0.383965216894817	reverse	TRUE
P16	chr1	170633237	170633238	2	+	PRRX1		theta	4.3561e-05	0.57046	-2.67	0.82	-0.18	357	361.8	212	260	0.000566293	0.789301850818804	reverse	TRUE
P16	chr2	86737600	86756340	18741	-	CHMP3	RNU6-640P;RNF103-CHMP3	psi5	4.4716e-05	1	2.72	0.15	0.14	7.4	626.2	74	510	0.000178864	0.383965216894817	reverse	TRUE
P16	chr19	51227674	51227675	2	+	CLEC11A		theta	4.5311e-05	0.58632	-2.67	0.88	-0.11	779.4	789.7	750	848	0.000181244	0.383965216894817	reverse	TRUE
P16	chr12	6639125	6639856	732	+	NCAPD2		psi5	4.9871e-05	1	-2.71	0.06	-0.89	127.7	241.2	71	1205	0.000149613	0.34974342377262	reverse	TRUE
P16	chr19	54379026	54379027	2	+	MYADM		theta	6.0653e-05	0.74916	-2.76	0.88	-0.12	414.9	419	289	330	0.000242612	0.456865118361179	reverse	TRUE
P16	chr5	141381205	141381206	2	-	GNPDA1		theta	7.5173e-05	0.88814	-2.77	0.78	-0.21	222.2	228.6	232	296	0.000375865	0.60668159523198	reverse	TRUE
P16	chr10	32307085	32307243	159	-	KIF5B		psi3	9.1017e-05	1	-2.6	0.87	-0.13	664.6	672.2	459	529	0.000364068	0.601972809421552	reverse	TRUE
P16	chr7	75601780	75608768	6989	+	POR		psi5	0.0001428	1	-2.72	0.77	-0.23	230.8	233.7	96	125	0.0005712	0.789301850818804	reverse	TRUE
P16	chr9	5522602	5534744	12143	+	PDCD1LG2		psi5	0.00016276	1	-2.64	0.32	-0.61	59.5	61.9	7	22	0.00048828	0.719596973209888	reverse	TRUE
P16	chr15	42824531	42824532	2	+	SNAP23		theta	0.00019405	1	-2.76	0.85	-0.14	213.5	217	202	237	0.00058215	0.789301850818804	reverse	TRUE
P23	chr15	72500999	72501000	2	-	PKM		theta	7.6951e-07	0.30936	-2.46	0.87	-0.13	18295.1	18724.1	27706	31996	8.46461e-06	0.306615540331036	reverse	TRUE
P23	chr7	100781247	100781248	2	+	SERPINE1		theta	4.3135e-06	0.47991	-2.37	0.89	-0.11	4926.1	4999.5	6008	6742	3.4508e-05	0.313864586210536	reverse	TRUE
P23	chr11	69467908	69467909	2	+	CCND1		theta	8.238e-06	0.47991	-2.43	0.84	-0.16	1909.3	1975.6	3534	4197	9.0618e-05	0.392409792910298	reverse	TRUE
P23	chr2	242194986	242194987	2	-	HDLBP		theta	8.3797e-06	0.47991	-2.45	0.82	-0.17	2163	2207.2	2029	2471	4.18985e-05	0.326643644505013	reverse	TRUE
P23	chr7	26242643	26251281	8639	+	CBX3		psi3	1.0544e-05	1	2.32	0.48	0.47	84.7	1348.8	847	1748	5.272e-05	0.331014901029088	reverse	TRUE
P23	chr7	135613199	135613200	2	-	MTPN		theta	1.4302e-05	0.6327	-2.45	0.62	-0.37	584.4	614.7	500	803	5.7208e-05	0.331014901029088	reverse	TRUE
P23	chr12	63359268	63359269	2	+	RPL14P1		theta	1.5164e-05	0.6327	-2.45	0.89	-0.11	2030.2	2058	2175	2453	6.0656e-05	0.331014901029088	reverse	TRUE
P23	chr1	1735926	1735927	2	-	GNB1		theta	1.5649e-05	0.6327	-2.45	0.89	-0.11	1861.5	1890	2325	2610	7.8245e-05	0.388184050307861	reverse	TRUE
P23	chr9	90342645	90342941	297	+	CTSL		psi5	2.9257e-05	1	-2.25	0.86	-0.13	2007.8	2036.1	1647	1916	0.000175542	0.532209122450543	reverse	TRUE
P23	chr2	110969869	110969870	2	-	MTLN		theta	3.107e-05	0.91274	-2.46	0.49	-0.49	245.9	265.9	193	393	0.00012428	0.477645726928432	reverse	TRUE
P23	chr17	78120803	78120804	2	-	EIF4A3		theta	3.6483e-05	0.94415	-2.43	0.4	-0.56	221.6	244.7	157	388	0.000328347	0.7790745079034 3	reverse	TRUE
P23	chr12	6688051	6688052	2	-	AC006064.6	CHD4	theta	4.6739e-05	0.99612	-2.45	0.78	-0.21	512.9	529.5	589	755	9.3478e-05	0.392409792910298	reverse	TRUE
P23	chr11	66082212	66082213	2	-	CD248		theta	6.2574e-05	1	-2.3	0.87	-0.12	1751.2	1790.4	1197	1371	0.00031287	0.7790745079034 3	reverse	TRUE
P23	chr8	71495514	71495515	2	-	TRAM1		theta	6.3474e-05	1	-2.45	0.89	-0.11	856.9	865.3	653	737	0.00031737	0.7790745079034 3	reverse	TRUE
P24	chr17	48266372	48272592	6221	-	COL1A1		psi5	5.9411e-07	0.23598	2.82	0.13	0.13	258.7	14127.9	2587	19913	3.267605e-05	0.192812324782637	reverse	TRUE
P24	chr11	62292738	62292739	2	-	AHNAK		theta	9.6299e-07	0.18043	-2.81	0.83	-0.17	3037.3	3075.4	1885	2266	0.00022630265	0.662585646488052	reverse	TRUE
P24	chr11	75283172	75283173	2	+	SERPINH1		theta	2.1846e-06	0.18043	-2.8	0.83	-0.17	2266.3	2322.8	2725	3290	1.52922e-05	0.14991259227975	reverse	TRUE
P24	chr22	33256816	33256817	2	+	TIMP3		theta	2.5247e-06	0.18043	-2.79	0.88	-0.12	2641.6	2689	3538	4012	1.00988e-05	0.143016535145336	reverse	TRUE

FRASER calls 2/4

sampleID	seqnames	start	end	width	strand	hgncSymbol	addHgncSymbols	type	pValue	padjust	zScore	psiValue	deltaPsi	meanCounts	meanTotalCounts	counts	totalCounts	pValueGene	padjustGene	STRAND_SPECIFIC	PAIRED_END
P24	chr17	74684246	74684247	2	-	MXRA7		theta	4.117e-06	0.2092	-2.82	0.81	-0.19	1085.2	1105.8	868	1074	1.6468e-05	0.14991259227975	reverse	TRUE
P24	chr15	48784765	48784766	2	-	FBN1		theta	4.2343e-06	0.2092	-2.79	0.81	-0.19	1241.6	1257.1	669	823	1.69372e-05	0.14991259227975	reverse	TRUE
P24	chr14	70352711	70352712	2	-	RPL7AP6		theta	4.7985e-06	0.21043	-2.82	0.86	-0.14	1275	1294.8	1190	1388	1.9194e-05	0.151011306930833	reverse	TRUE
P24	chr19	11275414	11275415	2	-	KANK2		theta	8.3149e-06	0.29352	-2.58	0.12	-0.86	245.1	295.3	34	272	7.48341e-05	0.374507097072762	reverse	TRUE
P24	chr2	47131356	47131357	2	-	MCFD2		theta	9.1837e-06	0.30249	-2.66	0.62	-0.38	465.4	494.5	185	300	5.51022e-05	0.300131840602742	reverse	TRUE
P24	chr12	113828235	113828499	265	+	PLBD2		psi3	2.0113e-05	1	2.81	0.93	0.88	2.8	76.7	28	30	0.00020113	0.61920649448194	reverse	TRUE
P24	chr10	120802132	120802133	2	-	EIF3A		theta	2.3015e-05	0.51052	-2.81	0.9	-0.1	780.6	787.7	637	708	0.000161105	0.543220573462508	reverse	TRUE
P24	chr2	85628759	85628760	2	-	CAPG		theta	2.362e-05	0.51348	-2.73	0.9	-0.1	898.7	906.9	742	824	0.00014172	0.519098417213162	reverse	TRUE
P24	chr6	24977085	24977086	2	+	PPIAP29		theta	2.6445e-05	0.55729	-2.81	0.64	-0.35	193.7	203.6	177	276	7.9335e-05	0.374507097072762	reverse	TRUE
P24	chrX	102841846	102841847	2	+	TCEAL4		theta	2.7393e-05	0.55729	-2.81	0.77	-0.22	318.7	326.2	253	328	0.000136965	0.519098417213162	reverse	TRUE
P24	chr1	170633239	170633240	2	+	PRRX1		theta	3.0415e-05	0.56998	-2.47	0.69	-0.3	340.2	354.5	308	444	0.00030415	0.717881978790448	reverse	TRUE
P24	chr15	90350041	90357195	7155	-	ANPEP		psi3	3.5013e-05	1	2.77	0.14	0.14	11.9	1055.7	119	829	0.000455169	0.856443667672075	reverse	TRUE
P24	chr6	75812371	75812372	2	-	COL12A1		theta	3.6655e-05	0.66404	-2.68	0.9	-0.1	714.5	719.7	459	511	0.00014662	0.519098417213162	reverse	TRUE
P24	chr3	196197186	196197187	2	-	RNF168		theta	4.7389e-05	0.80342	-2.8	0.42	-0.56	79.5	84.9	39	93	0.000236945	0.662585646488052	reverse	TRUE
P24	chr12	120998605	120998606	2	+	RNF10		theta	5.053e-05	0.83218	-2.82	0.84	-0.15	345.3	350.7	293	347	0.00025265	0.662585646488052	reverse	TRUE
P24	chr11	61165448	61165732	285	+	TMEM216		psi5	5.5312e-05	1	-2.82	0	-0.92	33	38.8	0	57	0.000442496	0.856443667672075	reverse	TRUE
P24	chr16	2964291	2979613	15323	+	FLYWCH1		psi5	5.8964e-05	1	-2.72	0.33	-0.55	108.7	128.5	48	144	0.000471712	0.856443667672075	reverse	TRUE
P24	chr16	129505	129506	2	+	MPG		theta	7.2714e-05	1	-2.81	0.87	-0.12	368.6	373.7	351	402	0.000290856	0.717881978790448	reverse	TRUE
P24	chr11	33720046	33720047	2	-	AL049629.2	C11orf91	theta	0.00012507	1	-2.81	0.67	-0.32	107.7	110.2	51	76	0.00025014	0.662585646488052	reverse	TRUE
P24	chr7	6523718	6523719	2	-	DAGLB	KDELR2	theta	0.00014171	1	-2.77	0.89	-0.1	365.7	369.4	315	352	0.00042513	0.856443667672075	reverse	TRUE
P24	chr8	100904280	100904281	2	-	COX6C		theta	0.00016769	1	2.56	0.18	0.17	62	1114.1	160	874	0.00033538	0.766058528145834	reverse	TRUE
P25	chr2	216248908	216249582	675	-	FN1		psi3	8.9082e-07	0.45558	-2.64	0.86	-0.14	12046.4	12404.3	22441	26017	2.048886e-05	0.113650461635756	reverse	TRUE
P25	chr17	40555726	40555727	2	-	CAVIN1		theta	1.0148e-06	0.2001	-2.64	0.72	-0.27	3678.6	3852.6	4573	6312	1.0148e-05	0.108631733641734	reverse	TRUE
P25	chr12	49523506	49525080	1575	-	TUBA1B		psi3	2.0771e-06	0.53114	-2.59	0.66	-0.34	6210.3	6489	5302	8034	1.24626e-05	0.108631733641734	reverse	TRUE
P25	chr12	49580617	49582759	2143	-	TUBA1A		psi5	2.5995e-06	0.33621	-2.61	0.73	-0.27	4824.8	5098.7	7453	10181	7.7985e-06	0.108631733641734	reverse	TRUE
P25	chr9	135896130	135896131	2	+	EEF1A1P5		theta	3.0237e-06	0.20541	-2.65	0.87	-0.13	2849.4	2892.5	2867	3298	1.81422e-05	0.113650461635756	reverse	TRUE
P25	chr20	1350248	1350249	2	-	AL136531.2	FKBP1A	theta	3.2851e-06	0.21004	-2.66	0.83	-0.17	2146.1	2181.6	1775	2130	6.5702e-06	0.108631733641734	reverse	TRUE
P25	chr2	65496065	65496066	2	+	ACTR2		theta	1.2073e-05	0.30165	-2.66	0.9	-0.1	1612.8	1633.1	1792	1995	4.8292e-05	0.196440046042754	reverse	TRUE
P25	chr2	38709444	38709445	2	+	RPLP0P6		theta	1.6535e-05	0.35946	-2.66	0.87	-0.13	939.8	953.4	880	1016	6.614e-05	0.224125189614505	reverse	TRUE
P25	chr8	144993371	144993372	2	-	PLEC		theta	1.7333e-05	0.36714	-2.64	0.87	-0.12	1132.1	1153	1453	1662	0.000242662	0.477624034964986	reverse	TRUE
P25	chrX	51639743	51639744	2	+	MAGED1		theta	2.5098e-05	0.45467	-2.65	0.89	-0.1	934.6	950.4	1323	1481	0.000200784	0.445567403957804	reverse	TRUE
P25	chrX	48436404	48436405	2	+	RBM3		theta	3.1717e-05	0.51197	-2.65	0.89	-0.11	840.3	851.1	867	975	0.000126868	0.368619998600066	reverse	TRUE
P25	chr10	120801787	120801788	2	-	EIF3A		theta	4.0933e-05	0.60947	-2.41	0.82	-0.17	1252.6	1312.6	1419	1729	0.000286531	0.526152544900932	reverse	TRUE
P25	chr11	6340520	6340521	2	-	CAVIN3		theta	4.3281e-05	0.63505	-2.63	0.87	-0.12	672.8	686.2	907	1041	0.000346248	0.555967968941773	reverse	TRUE
P25	chr6	52362713	52362714	2	-	TRAM2		theta	5.1117e-05	0.7018	-2.65	0.86	-0.14	517.6	526.3	525	612	0.000204468	0.445567403957804	reverse	TRUE
P25	chr19	41808779	41808780	2	+	HNRNPUL1		theta	5.6264e-05	0.7458	-2.64	0.89	-0.11	606.4	614.6	636	718	0.000393848	0.600460232094169	reverse	TRUE
P25	chr2	189918675	189949885	31211	-	COL5A2		psi5	6.3497e-05	1	2.68	0.12	0.12	19.5	796.4	195	1618	0.000253988	0.484294251774754	reverse	TRUE
P25	chr1	1431081	1431082	2	+	ATAD3B		theta	7.653e-05	0.85075	-2.56	0.33	-0.62	82.3	87.3	23	69	0.00084183	0.959886742607655	reverse	TRUE
P25	chr8	119123170	119123171	2	-	EXT1		theta	8.3097e-05	0.87918	-2.6	0.89	-0.1	572.5	580.9	712	796	0.000332388	0.54916017805483	reverse	TRUE
P25	chr6	116598279	116598280	2	-	TSPYL1		theta	8.3252e-05	0.87918	-2.64	0.84	-0.15	351	358.6	400	476	0.000333008	0.54916017805483	reverse	TRUE
P25	chr1	32157650	32157651	2	-	COL16A1		theta	0.00010049	0.98651	-2.6	0.88	-0.11	479.9	487.9	598	678	0.00060294	0.782748660372482	reverse	TRUE
P25	chr9	133499029	133499030	2	+	FUBP3		theta	0.00011461	1	-2.65	0.89	-0.1	474.2	481.1	585	654	0.00045844	0.639789519550122	reverse	TRUE
P25	chr16	55513234	55513235	2	+	MMP2		theta	0.00012497	1	-2.63	0.88	-0.12	444.6	459.6	1100	1249	0.00074982	0.933699302819695	reverse	TRUE
P25	chr17	37075287	37075288	2	+	LASP1		theta	0.00018183	1	-2.29	0.08	-0.83	285.4	349.2	41	536	0.00090915	0.973210490676073	reverse	TRUE
P25	chr6	33268335	33268336	2	-	TAPBP		theta	0.00019863	1	-2.63	0.77	-0.22	159	165.4	215	279	0.00079452	0.950562664026005	reverse	TRUE

sampleID	seqnames	start	end	width	strand	hgncSymbol	addHgncSymbols	type	pValue	padjust	zScore	psiValue	deltaPsi	meanCounts	meanTotalCounts	counts	totalCounts	pValueGene	padjustGene	STRAND_SPECIFIC	PAIRED_END
P25	chr2	86732201	86732202	2	-	CHMP3		theta	0.00020833	1	-2.65	0.88	-0.12	296	300.9	353	402	0.00083332	0.959886742607655	reverse	TRUE
P26	chr9	75775203	75775204	2	+	ANXA1		theta	1.1644e-06	0.17294	-2.74	0.9	-0.1	5804.9	5880.1	6884	7636	4.6576e-06	0.162570787228326	reverse	TRUE
P26	chr21	47610354	47610355	2	-	LSS		theta	2.8576e-06	0.17294	-2.71	0.6	-0.4	810.8	863.6	784	1311	6.28672e-05	0.214952997977593	reverse	TRUE
P26	chr19	3055707	3055708	2	-	TLE5		theta	3.4923e-06	0.1898	-2.73	0.86	-0.14	1938.1	1973.9	2118	2476	3.14307e-05	0.193439850166093	reverse	TRUE
P26	chr6	29912394	29912835	442	+	HLA-A		psi3	4.1512e-06	0.70766	-1.82	0.72	-0.26	3097.3	3190.7	1308	1823	0.0001079312	0.272616716363847	reverse	TRUE
P26	chr6	24977114	24977115	2	+	PPIAP29		theta	8.8095e-06	0.2603	-2.74	0.86	-0.14	1111.7	1130	1106	1289	2.64285e-05	0.180726525343925	reverse	TRUE
P26	chr1	89475178	89476586	1409	-	GBP3		psi3	9.5036e-06	1	-2.56	0	-0.86	99.8	101.5	0	7	0.0001520576	0.311945532665151	reverse	TRUE
P26	chr19	4838841	4838842	2	-	PLIN3		theta	1.2786e-05	0.31566	-2.74	0.84	-0.16	837.9	861.1	1207	1439	5.1144e-05	0.204129869122249	reverse	TRUE
P26	chr7	98015303	98015304	2	-	BAIAP2L1	RPS3AP26	theta	1.3247e-05	0.31566	-2.74	0.82	-0.18	749.2	768.5	883	1076	5.2988e-05	0.204129869122249	reverse	TRUE
P26	chr15	73994708	73994709	2	+	CD276		theta	1.4192e-05	0.31566	-2.72	0.74	-0.25	455.7	477.3	622	838	9.9344e-05	0.272616716363847	reverse	TRUE
P26	chr20	18548187	18548188	2	+	SMIM26	AL121900.1	theta	2.1659e-05	0.36672	-2.73	0.71	-0.28	335.5	342.4	170	238	0.000129954	0.285642952914603	reverse	TRUE
P26	chr11	62494370	62494371	2	-	HNRNPUL2-B	HNRNPUL2	theta	4.2083e-05	0.53813	-2.71	0.53	-0.46	132.7	143.8	123	234	0.000126249	0.285642952914603	reverse	TRUE
P26	chr7	44146335	44146336	2	+	AEBP1		theta	5.3051e-05	0.6479	-2.46	0.79	-0.21	857.5	950.6	1382	1754	0.000477459	0.624994485390682	reverse	TRUE
P26	chr2	178082450	178082451	2	+	HNRNPA3		theta	5.6282e-05	0.67973	-2.67	0.8	-0.19	382.7	393.6	445	553	0.000393974	0.551070012970355	reverse	TRUE
P26	chrX	20146155	20146156	2	-	EIF1AX		theta	7.0712e-05	0.79976	-2.72	0.89	-0.11	565	574	721	811	0.000212136	0.354528420566436	reverse	TRUE
P26	chr12	52467687	52470569	2883	+	ATG101		psi3	9.2813e-05	1	-1.11	0.8	-0.14	338	376.6	445	559	0.000371252	0.544015664801658	reverse	TRUE
P26	chr4	57326959	57326960	2	+	PAICS		theta	0.0001017	1	-2.71	0.76	-0.23	204.4	207.9	109	144	0.0003051	0.481470291096863	reverse	TRUE
P26	chr7	128410012	128410013	2	+	CALU		theta	0.0001244	1	-2.08	0.89	-0.11	3740.8	3844.7	4950	5551	0.0004976	0.624994485390682	reverse	TRUE
P26	chr6	33287857	33287858	2	-	DAXX		theta	0.00012763	1	-2.72	0.82	-0.17	237.1	244.3	324	396	0.00051052	0.628397791374635	reverse	TRUE
P26	chr17	4575419	4575420	2	-	PELP1		theta	0.0001436	1	-2.71	0.81	-0.18	210.7	216	226	279	0.0005744	0.68594560646302	reverse	TRUE
P26	chr6	10724867	10749854	24988	+	TMEM14C	TMEM14B;AL024498.2	psi3	0.00020148	1	2.78	0.12	0.12	5.4	376.6	54	453	0.00040296	0.551113840381317	reverse	TRUE
P26	chr6	10749502	10749854	353	+	TMEM14B	AL024498.2	psi3	0.00020148	1	-2.72	0.88	-0.12	371.1	376.6	399	453	0.00020148	0.354528420566436	reverse	TRUE
P26	chr10	89720693	89720694	2	+	PTEN		theta	0.00021309	1	-2.72	0.89	-0.11	299.2	302.7	273	308	0.00085236	0.794824085270282	reverse	TRUE
P26	chr5	864221	864222	2	-	BRD9		theta	0.00023616	1	-2.72	0.64	-0.34	77.8	81.2	61	95	0.0011808	0.995509429621542	reverse	TRUE
P26	chr19	50335415	50335564	150	+	MED25		psi5	0.000613	1	-2.75	0.8	-0.19	125.6	129.5	145	182	0.000613	0.68594560646302	reverse	TRUE
P26	chr19	50335415	50361805	26391	+	PTOV1	MED25;MIR4749	psi5	0.000613	1	2.78	0.2	0.19	3.7	129.5	37	182	0.000613	0.68594560646302	reverse	TRUE
P26	chr2	37431503	37431504	2	+	CEBPZOS		theta	0.0006428	1	-2.71	0.84	-0.15	127.6	129.8	116	138	0.0006428	0.686340100576837	reverse	TRUE
P29	chr21	47424104	47424105	2	+	COL6A1		theta	2.6922e-07	0.16015	-2.73	0.85	-0.15	11667.3	11793	7137	8394	1.373022e-05	0.177392788163101	reverse	TRUE
P29	chr21	47518088	47531371	13284	+	COL6A2		psi3	2.9889e-06	1	-2.41	0	-0.88	736	745.2	0	9	8.66781e-05	0.399953788891534	reverse	TRUE
P29	chr6	29856416	29856417	2	+	HLA-H		theta	3.3632e-06	0.20845	-2.68	0.1	-0.89	1050.7	1143.6	102	1027	8.07168e-05	0.399953788891534	reverse	TRUE
P29	chr16	31201682	31201683	2	+	FUS		theta	6.139e-06	0.27803	-2.71	0.84	-0.16	1336.7	1357.3	1082	1288	5.5251e-05	0.377263065917384	reverse	TRUE
P29	chr17	78183459	78183460	2	-	SGSH		theta	8.5956e-06	0.30917	-2.72	0.27	-0.72	192.5	211.3	68	256	4.2978e-05	0.377263065917384	reverse	TRUE
P29	chr19	13885488	13885489	2	+	C19orf53		theta	9.3504e-06	0.30917	-2.73	0.87	-0.13	1204.1	1221.5	1170	1344	4.6752e-05	0.377263065917384	reverse	TRUE
P29	chr1	1717167	1717168	2	-	GNB1		theta	1.6131e-05	0.41745	-2.73	0.89	-0.11	970	981.4	885	999	0.000145179	0.446594092529209	reverse	TRUE
P29	chr17	6917814	6917815	2	+	RNASEK	RNASEK-C17orf49;C17orf49	theta	2.0803e-05	0.43347	-2.73	0.78	-0.22	399.5	407.1	266	342	0.000104015	0.446594092529209	reverse	TRUE
P29	chr5	151042818	151042819	2	-	SPARC		theta	2.1852e-05	0.43347	-2.26	0.9	-0.1	11317.9	11475.7	10138	11294	0.000262224	0.62738934942049	reverse	TRUE
P29	chr13	31036696	31036697	2	-	HMGB1		theta	2.9075e-05	0.50722	-2.71	0.88	-0.11	782.9	790.1	550	622	0.0001163	0.446594092529209	reverse	TRUE
P29	chr16	418620	418621	2	-	MRPL28		theta	3.4746e-05	0.54734	-2.73	0.81	-0.19	394.4	404.4	429	529	0.000138984	0.446594092529209	reverse	TRUE
P29	chr4	119686019	119686020	2	-	SEC24D		theta	5.4658e-05	0.79433	-2.72	0.83	-0.16	338.1	344.7	327	393	0.00027329	0.630513191718366	reverse	TRUE
P29	chr1	24082876	24083442	567	+	ELOA		psi5	8.074e-05	1	-2.74	0.61	-0.38	158.2	165.4	112	184	0.0004037	0.899268165799337	reverse	TRUE
P29	chr2	183606083	183606084	2	+	DNAJC10		theta	9.9969e-05	1	2.49	0.23	0.22	5.5	324.8	45	198	0.000199938	0.516634974235724	reverse	TRUE
P29	chr15	83041045	83041046	2	-	RPL9P8		theta	0.00012867	1	-2.12	0.56	-0.44	1168.1	1501.2	2344	4187	0.00012867	0.446594092529209	reverse	TRUE
P29	chr11	68777413	68777414	2	-	MRGPRF		theta	0.00012911	1	-2.71	0.85	-0.15	276	279	169	199	0.00051644	0.981226849787763	reverse	TRUE
P29	chr15	82664505	82664506	2	-	RPL9P9		theta	0.00012922	1	-2.12	0.56	-0.44	1167.9	1501	2342	4185	0.00012922	0.446594092529209	reverse	TRUE
P29	chr5	52394292	52394293	2	-	MOCS2		theta	0.00013346	1	-2.72	0.85	-0.15	268.3	272.6	239	282	0.00053384	0.985306926025636	reverse	TRUE

FRASER calls 4/4

sampleID	seqnames	start	end	width	strand	hgncSymbol	addHgncSymbols	type	pValue	padjust	zScore	psiValue	deltaPsi	meanCounts	meanTotalCounts	counts	totalCounts	pValueGene	padjustGene	STRAND_SPECIFIC	PAIRED_END
P32	chr15	60639860	60639861	2	-	ANXA2		theta	3.3373e-07	0.33185	-2.65	0.89	-0.11	22180.2	22438.3	20944	23525	3.3373e-06	0.170177005784621	reverse	TRUE
P32	chr12	49578888	49578889	2	-	TUBA1A		theta	2.4866e-06	0.33185	-2.61	0.89	-0.11	5525	5577.7	4313	4840	9.9464e-06	0.170177005784621	reverse	TRUE
P32	chr6	29925272	29925273	2	+	HLA-W		theta	5.1993e-06	0.33185	2.06	0.79	0.73	4.6	974	11	14	1.03986e-05	0.170177005784621	reverse	TRUE
P32	chr2	47130231	47130232	2	-	MCFD2		theta	1.2109e-05	0.43734	-2.64	0.76	-0.23	634.6	653.8	617	809	4.8436e-05	0.274686128757168	reverse	TRUE
P32	chr11	33729536	33729537	2	-	AL049629.2	CD59	theta	1.3574e-05	0.43734	-2.6	0.6	-0.39	387.7	416.7	437	727	5.4296e-05	0.284232796591204	reverse	TRUE
P32	chr3	154900965	154900966	2	+	MME		theta	1.368e-05	0.43734	-2.58	0.78	-0.21	1173	1208.4	1288	1642	0.0001368	0.489984340668497	reverse	TRUE
P32	chr15	72492984	72494794	1811	-	PKM		psi3	1.8613e-05	1	-1.28	0	-0.57	68.5	582.7	0	74	0.000260582	0.611499518486965	reverse	TRUE
P32	chr3	127410913	127410914	2	-	MGLL		theta	1.9989e-05	0.53046	-2.63	0.74	-0.26	417.4	435.5	504	685	7.9956e-05	0.362751740208244	reverse	TRUE
P32	chr7	35840881	35871104	30224	+	SEPTIN7		psi5	2.0677e-05	1	2.52	0.35	0.34	56.3	903.2	339	974	0.000227447	0.552804902440613	reverse	TRUE
P32	chr6	122765253	122765254	2	-	SERINC1		theta	2.6291e-05	0.5832	-2.65	0.89	-0.11	904.1	914.3	789	891	0.000105164	0.420986218971093	reverse	TRUE
P32	chr1	55352610	55352611	2	-	DHCR24		theta	3.4951e-05	0.71678	-2.54	0.77	-0.23	527.5	531.5	131	171	0.000104853	0.420986218971093	reverse	TRUE
P32	chr20	53691404	53691405	2	+	RPL12P4		theta	3.7685e-05	0.74476	-2.64	0.78	-0.21	388.7	402.9	508	650	0.00037685	0.827287628946788	reverse	TRUE
P32	chr2	61719334	61719459	126	-	XPO1		psi5	3.9413e-05	1	-2.64	0.75	-0.24	394.7	410.8	482	639	0.000157652	0.496365025638091	reverse	TRUE
P32	chr6	18258631	18258632	2	-	DEK		theta	3.9924e-05	0.76132	-2.63	0.84	-0.15	477.3	487.2	531	630	0.00019962	0.522492917116695	reverse	TRUE
P32	chr8	146016696	146017257	562	-	RPL8		psi3	4.3404e-05	1	2.7	0.65	0.63	14.7	145.3	147	225	0.000173616	0.496365025638091	reverse	TRUE
P32	chr16	56660967	56660968	2	+	MT1E		theta	5.3894e-05	0.89195	-2.53	0.88	-0.12	545	551.3	443	506	0.000431152	0.862980916862352	reverse	TRUE
P32	chr17	7485153	7485154	2	+	AC016876.3	CD68	theta	7.376e-05	1	-2.61	0.89	-0.11	602	610.5	695	780	0.00029504	0.669282314216821	reverse	TRUE
P32	chr13	114287601	114288204	604	+	TFDP1		psi3	8.2223e-05	1	-2.7	0.85	-0.15	447.8	457.7	544	643	0.000411115	0.862980916862352	reverse	TRUE
P32	chr5	52388758	52388759	2	+	ITGA2		theta	0.00012473	1	-2.31	0	-0.9	131.2	164.5	0	247	0.00049892	0.94314424091947	reverse	TRUE
P32	chr20	35826803	35826804	2	+	RPN2		theta	0.00023955	1	-2.6	0.45	-0.51	56.5	63.3	56	124	0.0004791	0.931553561961416	reverse	TRUE
P33	chr2	38709324	38709325	2	+	RPLP0P6		theta	8.2868e-06	0.70064	-2.34	0.72	-0.27	1501.6	1544.2	1121	1547	3.31472e-05	0.369706619764706	reverse	TRUE
P33	chr13	113976750	113976751	2	+	LAMP1		theta	1.0286e-05	0.70064	-2.32	0.87	-0.13	2309.7	2338.6	1921	2210	5.143e-05	0.422178548822565	reverse	TRUE
P33	chr11	69468818	69468819	2	+	CCND1		theta	1.0432e-05	0.70064	-2.31	0.81	-0.19	2071.8	2117.9	1914	2374	0.000135616	0.555738712105515	reverse	TRUE
P33	chr19	10504000	10504001	2	-	CDC37		theta	3.114e-05	1	-2.33	0.84	-0.15	935.6	953.1	937	1112	0.00012456	0.555738712105515	reverse	TRUE
P33	chr14	75598980	75598981	2	-	TMED10		theta	3.1896e-05	1	-2.33	0.56	-0.41	299.7	307.7	101	181	0.000127584	0.555738712105515	reverse	TRUE
P33	chr8	144995405	144995406	2	-	PLEC		theta	4.7404e-05	1	-2.31	0.89	-0.11	971.8	987.8	1281	1441	0.000379232	0.713544177357312	reverse	TRUE
P33	chr3	47958377	47958378	2	-	MAP4		theta	5.3937e-05	1	-2.33	0.9	-0.1	956.7	967.1	930	1034	0.000323622	0.683449403528176	reverse	TRUE
P33	chr18	8376043	8376044	2	+	PTPRM		theta	5.596e-05	1	2.32	0.56	0.52	5.9	154.7	58	103	0.0002798	0.627699736224168	reverse	TRUE
P33	chr9	139687451	139687452	2	+	TMEM141	AL355987.3	theta	5.9099e-05	1	-2.33	0.57	-0.4	203.8	213	122	214	0.000236396	0.593403019787128	reverse	TRUE
P33	chr18	346892	346893	2	-	COLEC12		theta	6.4203e-05	1	-2.3	0.88	-0.11	873.5	877.7	301	343	0.000256812	0.599173901520976	reverse	TRUE
P33	chr19	13885389	13885443	55	+	C19orf53		psi5	6.4232e-05	1	2.3	0.19	0.18	18	1100.8	180	970	0.000513856	0.833041452650438	reverse	TRUE
P33	chr6	170627851	170627852	2	+	FAM120B		theta	0.00011279	1	-2.32	0.41	-0.53	108.3	116.4	56	137	0.00056395	0.833041452650438	reverse	TRUE
P33	chr10	70742383	70742384	2	+	DDX21		theta	0.00012355	1	-2.31	0.72	-0.26	247.3	253.7	162	226	0.0004942	0.833041452650438	reverse	TRUE
P33	chr19	11777223	11777224	2	-	HNRNPA1P10		theta	0.00013595	1	-2.32	0.77	-0.21	286.9	291.8	162	211	0.0005438	0.833041452650438	reverse	TRUE
P33	chr9	130648354	130648355	2	-	AL157935.3	ST6GALNAC6	theta	0.00023198	1	-2.32	0.82	-0.17	265.7	272.4	297	364	0.00046396	0.820057271436598	reverse	TRUE

MAE calls

gene_name	ID	contig	position	variantID	refAllele	altAllele	refCount	altCount	totalCount	pvalue	padj	log2FC	altRatio	AF	AF_afr	AF_amr	AF_eas	AF_nfe	MAX_AF	rare	gene_type	other_names	N_var	cohort_freq	MAE	MAE_ALT
SNHG14	P16	chr15	25304758	.	C	T	0	29	29	0.000446527622348894	0.0114957111285566	7.40017797023297	1	1E-04	3E-04	0	0	0	3E-04	FALSE	lncRNA		1	0.1	TRUE	TRUE
FLNB	P16	chr3	58104626	.	G	T	9	104	113	8.52570006153225e-08	1.57364192661163e-05	3.63004788189686	0.92	7E-04	0	0	0	0.001	0.001	FALSE	protein_coding		1	0.1	TRUE	TRUE
HGS	P22	chr17	79660663	.	A	G	4	132	136	2.11598843118188e-09	4.42818669870973e-07	5.16637994719267	0.971	7E-05	0	0	0	1E-04	1E-04	FALSE	protein_coding		1	0.1	TRUE	TRUE
AIFM1	P24	chrX	129299753	.	C	A	11	58	69	0.000125679530701733	0.00335863029808936	2.52407833497455	0.841	5E-05	0	0	0	9E-05	9E-05	FALSE	protein_coding		1	0.1	TRUE	TRUE
COLEC12	P24	chr18	318029	.	T	A	2	21	23	0.00244550996681754	0.0346201110523429	3.51784265785704	0.913	0.002	7E-04	0.001	0	0.004	0.004	FALSE	protein_coding	AP000915.1	1	0.1	TRUE	TRUE
MRI1	P24	chr19	13875821	.	G	T	7	30	37	0.00333868088230662	0.0425943532312332	2.22506440597205	0.811	0.003	9E-04	0.001	0	0.005	0.005	FALSE	protein_coding		1	0.1	TRUE	TRUE
PCSK9	P24	chr1	55505651	.	C	T	0	12	12	0.00389602023969395	0.0478888895578904	6.1531594501359	1	0.008	0.002	0.01	0	0.01	0.01	FALSE	protein_coding		1	0.1	TRUE	TRUE
MEST	P25	chr7	130145727	.	A	C	0	13	13	0.00327964644383008	0.0455053706289042	6.258582626624	1	0.001	7E-04	0.002	0	0.002	0.002	FALSE	protein_coding		1	0.1	TRUE	TRUE
MMP1	P25	chr11	102660962	.	T	G	1557	7366	8923	2.63744488010503e-07	3.05636173866895e-05	2.35758782913522	0.826	0.003	0.002	0.002	0	0.003	0.003	FALSE	protein_coding	WTAPP1	1	0.1	TRUE	TRUE
NGFR	P25	chr17	47590459	.	G	T	2	128	130	5.41376804562646e-08	8.11592576300303e-06	6.11546863161224	0.985	0.002	7E-04	0.006	0	0.003	0.006	FALSE	protein_coding	AC006487.1	1	0.1	TRUE	TRUE
AL669831.3	P26	chr1	565283	.	T	C	1	1799	1800	5.17069099079231e-13	1.03252583215058e-10	10.9284305115231	0.999	3E-04	3E-04	0	0	5E-04	5E-04	FALSE	transcribed_processed_pseudogene	MTND2P28	1	0.1	TRUE	TRUE
SSNA1	P26	chr9	140083122	.	G	A	14	73	87	5.71830943664438e-05	0.00174089712373644	2.497945062888	0.839	0.006	0.002	0	0	0.008	0.008	FALSE	protein_coding		1	0.1	TRUE	TRUE
DNAJC10	P29	chr2	183606088	.	G	C	0	45	45	0.000130148652156937	0.00351256290670052	8.04142698755107	1	7E-04	0	0	0	0.001	0.001	FALSE	protein_coding		1	0.1	TRUE	TRUE
PRMT1	P29	chr19	50190833	.	G	C	1	23	24	0.00268652132820438	0.0374456003221514	4.63046524888601	0.958	0.004	8E-04	0.001	0	0.005	0.005	FALSE	protein_coding		1	0.1	TRUE	TRUE
BLZF1	P29	chr1	169337581	.	G	A	7	70	77	2.773715937123 68e-06	0.000161796696463842	3.4288406279806	0.909	0.006	0.002	0.007	0	0.01	0.01	FALSE	protein_coding		1	0.1	TRUE	TRUE
MEG3	P32	chr14	101298291	.	T	G	1	160	161	9.10468391365233e-07	8.27885061218866e-05	7.45049428130272	0.994	0.007	0.002	0.007	0	0.01	0.01	FALSE	lncRNA		1	0.1	TRUE	TRUE
MUC20-OT1	P32	chr3	195393086	.	G	A	1	161	162	8.82971447182377e-07	8.06669386274919e-05	7.45948304398127	0.994	0.007	0.004	0.01	0	0.01	0.01	FALSE	lncRNA	SDHAP2	1	0.1	TRUE	TRUE
MSMO1	P33	chr4	166249087	.	G	A	23	99	122	8.23941554571518e-05	0.00239190233292112	2.2260873492767	0.811	9E-04	0.002	0.002	0	6E-04	0.002	FALSE	protein_coding		1	0.1	TRUE	TRUE

Supplemental Method:

Urine derived stem cell isolation and culture.

Approximately 200 ml of fresh urine samples were collected from donors without NDD. Each urine sample was centrifuged at 400g for 10 min at room temperature and the supernatant was discarded. Next, 5 ml of washing buffer (Dulbecco's phosphate-buffered saline buffer supplemented with 100 U/ml penicillin, 100 µg/ml streptomycin, and 500 ng/ml amphotericin B) was added and centrifuged at 400 × g for 10 min at room temperature. After carefully discarding the supernatant, 3 ml of proliferation medium (1:1 mixture of DMEM high glucose (Gibco, Thermo Fisher Scientific, Waltham, USA) / REGM with SingleQuot Kit (LONZA, Basel, Switzerland), supplemented with 5% (v/v) FBS, 0.5% (v/v) NEAA, 0.5% (v/v) GlutaMax, 50 U/ml penicillin, 50µg/ml streptomycin, and 1.5 µg/ml amphotericin B) was added to suspend the cell pellet, and the volume was transferred into a single well of a 6-well plate (coated beforehand with 0.1% (w/v) gelatin). The cells were then incubated at 37 °C. Approximately 96 h after plating, most of the medium was aspirated, leaving approximately 1 ml, and then 2 ml of proliferation medium was added. Half of the proliferation medium was replaced daily.

References:

1. International Classification of Diseases, Eleventh Revision (ICD-11), World Health Organization (WHO) 2019/2021.https://icd.who.int/browse11.

2. Kochinke, K., Zweier, C., Nijhof, B., Fenckova, M., Cizek, P., Honti, F., Keerthikumar, S., Oortveld, M.A.W., Kleefstra, T., Kramer, J.M., et al. (2016). Systematic Phenomics Analysis Deconvolutes Genes Mutated in Intellectual Disability into Biologically Coherent Modules. Am. J. Hum. Genet. *98*, 149–164. 10.1016/j.ajhg.2015.11.024.

3. Leblond, C.S., Le, T.-L., Malesys, S., Cliquet, F., Tabet, A.-C., Delorme, R., Rolland, T., and Bourgeron, T. (2021). Operative list of genes associated with autism and neurodevelopmental disorders based on database review. Mol. Cell. Neurosci. *113*, 103623. 10.1016/j.mcn.2021.103623.

4. Srivastava, S., Love-Nichols, J.A., Dies, K.A., Ledbetter, D.H., Martin, C.L., Chung, W.K., Firth, H.V., Frazier, T., Hansen, R.L., Prock, L., et al. (2019). Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. Genet. Med. *21*, 2413–2421. 10.1038/s41436-019-0554-6.

5. Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W.M., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature *511*, 344–347. 10.1038/nature13394.

6. Palmer, E.E., Sachdev, R., Macintosh, R., Melo, U.S., Mundlos, S., Righetti, S., Kandula, T., Minoche, A.E., Puttick, C., Gayevskiy, V., et al. (2021). Diagnostic Yield of Whole Genome Sequencing After Nondiagnostic Exome Sequencing or Gene Panel in Developmental and Epileptic Encephalopathies. Neurology *96*, e1770–e1782. 10.1212/WNL.0000000000011655.

7. Sun, Y., Peng, J., Liang, D., Ye, X., Xu, N., Chen, L., Yan, D., Zhang, H., Xiao, B., Qiu, W., et al. (2022). Genome sequencing demonstrates high diagnostic yield in children with undiagnosed global developmental delay/intellectual disability: A prospective study. Hum. Mutat. *43*, 568–581. 10.1002/humu.24347.

8. Ewans, L.J., Minoche, A.E., Schofield, D., Shrestha, R., Puttick, C., Zhu, Y., Drew, A., Gayevskiy, V., Elakis, G., Walsh, C., et al. (2022). Whole exome and genome sequencing in mendelian disorders: a diagnostic and health economic analysis. Eur. J. Hum. Genet. *30*, 1121–1131. 10.1038/s41431-022-01162-2.

9. Gonorazky, H.D., Naumenko, S., Ramani, A.K., Nelakuditi, V., Mashouri, P., Wang, P., Kao, D., Ohri, K., Viththiyapaskaran, S., Tarnopolsky, M.A., et al. (2019). Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. Am. J. Hum. Genet. *104*, 466–483. 10.1016/j.ajhg.2019.01.012.

10. Murdock, D.R., Dai, H., Burrage, L.C., Rosenfeld, J.A., Ketkar, S., Müller, M.F., Yépez, V.A., Gagneur, J., Liu, P., Chen, S., et al. (2021). Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. J. Clin. Invest. *131*, e141500. 10.1172/JCI141500.

11. Colin, E., Duffourd, Y., Tisserant, E., Relator, R., Bruel, A.-L., Tran Mau-Them, F., Denommé-Pichon, A.-S., Safraou, H., Delanne, J., Jean-Marçais, N., et al. (2022). OMIXCARE: OMICS technologies solved about 33% of the patients with heterogeneous rare neuro-developmental disorders and negative exome sequencing results and identified 13% additional candidate variants. Front. Cell Dev. Biol. *10*, 1021785. 10.3389/fcell.2022.1021785.

12. Dekker, J., Schot, R., Bongaerts, M., de Valk, W.G., van Veghel-Plandsoen, M.M., Monfils, K., Douben, H., Elfferich, P., Kasteleijn, E., van Unen, L.M.A., et al. (2023). Web-accessible application for identifying pathogenic transcripts with RNA-seq: Increased sensitivity in diagnosis of neurodevelopmental disorders. Am. J. Hum. Genet. *110*, 251–272. 10.1016/j.ajhg.2022.12.015.

13. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222. 10.1093/bioinformatics/btv710.

14. Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339. 10.1093/bioinformatics/bts378.

15. Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., Emig-Agius, D., Gross, A., Narzisi, G., Bowman, B., et al. (2019). ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics *35*, 4754–4756. 10.1093/bioinformatics/btz431.

16. Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., Mills, R.E., The 1000 Genomes Project Consortium, and Devine, S.E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. *27*, 1916–1929. 10.1101/gr.218032.116.

17. Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., and Muller, J. (2018). AnnotSV: an integrated tool for structural variations annotation. Bioinformatics *34*, 3572–3574. 10.1093/bioinformatics/bty304.

18.; on behalf of the ACMG Laboratory Quality Assurance Committee, Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., et al. (2015). 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. *17*, 405–423. 10.1038/gim.2015.30.

19. Quinodoz, M., Peter, V.G., Bedoni, N., Royer Bertrand, B., Cisarova, K., Salmaninejad, A., Sepahi, N., Rodrigues, R., Piran, M., Mojarrad, M., et al. (2021). AutoMap is a high performance homozygosity mapping tool using next-generation sequencing data. Nat. Commun. *12*, 518. 10.1038/s41467-020-20584-4.

20. Baux, D., Van Goethem, C., Ardouin, O., Guignard, T., Bergougnoux, A., Koenig, M., and Roux, A.-F. (2021). MobiDetails: online DNA variants interpretation. Eur. J. Hum. Genet. *29*, 356–360. 10.1038/s41431-020-00755-z.

21. Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. *33*, W306–W310. 10.1093/nar/gki375.

22. Pires, D.E.V., Ascher, D.B., and Blundell, T.L. (2014). DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res. *42*, W314–W319. 10.1093/nar/gku411.

23. Hoefgen, S., Dahms, S.O., Oertwig, K., and Than, M.E. (2015). The Amyloid Precursor Protein Shows a pH-Dependent Conformational Switch in Its E1 Domain. J. Mol. Biol. *427*, 433–442. 10.1016/j.jmb.2014.12.005.

24. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. 10.1093/bioinformatics/bts635.

25. Yépez, V.A., Mertes, C., Müller, M.F., Klaproth-Andrade, D., Wachutka, L., Frésard, L., Gusic, M., Scheller, I.F., Goldberg, P.F., Prokisch, H., et al. (2021). Detection of aberrant gene expression events in RNA sequencing data. Nat. Protoc. *16*, 1276–1296. 10.1038/s41596-020-00462-5.

26. Brechtmann, F., Mertes, C., Matusevičiūtė, A., Yépez, V.A., Avsec, Ž., Herzog, M., Bader, D.M., Prokisch, H., and Gagneur, J. (2018). OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. Am. J. Hum. Genet. *103*, 907–917. 10.1016/j.ajhg.2018.10.025.

27. Mertes, C., Scheller, I.F., Yépez, V.A., Çelik, M.H., Liang, Y., Kremer, L.S., Gusic, M., Prokisch, H., and Gagneur, J. (2021). Detection of aberrant splicing events in RNA-seq data using FRASER. Nat. Commun. *12*, 529. 10.1038/s41467-020-20573-7.

28. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. 10.1186/s13059-014-0550-8.

29. Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., et al. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat. Commun. *8*, 15824. 10.1038/ncomms15824.

30. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527. 10.1038/nbt.3519.

31. Aicher, J.K., Jewell, P., Vaquero-Garcia, J., Barash, Y., and Bhoj, E.J. (2020). Mapping RNA splicing variations in clinically accessible and nonaccessible tissues to facilitate Mendelian disease diagnosis using RNA-seq. Genet. Med. *22*, 1181–1190. 10.1038/s41436-020-0780-y.

32. Mak, C.C.Y., Doherty, D., Lin, A.E., Vegas, N., Cho, M.T., Viot, G., Dimartino, C., Weisfeld-Adams, J.D., Lessel, D., Joss, S., et al. (2020). MN1 C-terminal truncation syndrome is a novel neurodevelopmental and craniofacial disorder with partial rhombencephalosynapsis. Brain *143*, 55–68. 10.1093/brain/awz379.

33. Vuillaume, M.-L., Cogné, B., Jeanne, M., Boland, A., Ung, D.-C., Quinquis, D., Besnard, T., Deleuze, J.-F., Redon, R., Bézieau, S., et al. (2018). Whole genome sequencing identifies a de novo 2.1 Mb balanced paracentric inversion disrupting FOXP1 and leading to severe intellectual disability. Clin. Chim. Acta *485*, 218–223. 10.1016/j.cca.2018.06.048.

34. Tepe, B., Macke, E.L., Niceta, M., Weisz Hubshman, M., Kanca, O., Schultz-Rogers, L., Zarate, Y.A., Schaefer, G.B., Granadillo De Luque, J.L., Wegner, D.J., et al. (2023). Bi-allelic variants in INTS11 are associated with a complex neurological disorder. Am. J. Hum. Genet. *110*, 774–789. 10.1016/j.ajhg.2023.03.012.

35. Klein, S., Goldman, A., Lee, H., Ghahremani, S., Bhakta, V., UCLA Clinical Genomics Center, Nelson, S.F., and Martinez-Agosto, J.A. (2016). Truncating mutations in APP cause a distinct neurological phenotype. Ann. Neurol. *80*, 456–460. 10.1002/ana.24727.

36. The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. 10.1038/nature11247.

37. Verkerk, A.J.M.H., Zeidler, S., Breedveld, G., Overbeek, L., Huigh, D., Koster, L., van der Linde, H., de Esch, C., Severijnen, L.-A., de Vries, B.B.A., et al. (2018). CXorf56, a dendritic neuronal protein, identified as a new candidate gene for X-linked intellectual disability. Eur. J. Hum. Genet. *26*, 552–560. 10.1038/s41431-017-0051-9.

38. Rocha, M.E., Silveira, T.R.D., Sasaki, E., Sás, D.M., Lourenço, C.M., Kandaswamy, K.K., Beetz, C., Rolfs, A., Bauer, P., Reardon, W., et al. (2020). Novel clinical and genetic insight into CXorf56-associated intellectual disability. Eur. J. Hum. Genet. *28*, 367–372. 10.1038/s41431-019-0558-3.

39. Rom, A., Melamed, L., Gil, N., Goldrich, M.J., Kadir, R., Golan, M., Biton, I., Perry, R.B.-T., and Ulitsky, I. (2019). Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability. Nat. Commun. *10*, 5092. 10.1038/s41467-019-13075-8.

40. Newman, A.G., Sharif, J., Bessa, P., Zaqout, S., Brown, J., Mueller, S., Böhm-Sturm, P., Ohara, O., Koseki, H., Singh, P.B., et al. HP1 deficiency results in De-Repression of Endogenous Retroviruses and Induction of Neurodegeneration via Complement. 55.

41. Wickens, M., Bernstein, D.S., Kimble, J., and Parker, R. (2002). A PUF family portrait: 3′UTR regulation as a way of life. Trends Genet. *18*, 150–157. 10.1016/S0168-9525(01)02616-6.

42. Lu, G., and Hall, T.M.T. (2011). Alternate Modes of Cognate RNA Recognition by Human PUMILIO Proteins. Structure *19*, 361–367. 10.1016/j.str.2010.12.019.

43. Gennarino, V.A., Palmer, E.E., McDonell, L.M., Wang, L., Adamski, C.J., Koire, A., See, L., Chen, C.-A., Schaaf, C.P., Rosenfeld, J.A., et al. (2018). A Mild PUM1 Mutation Is Associated with Adult-Onset Ataxia, whereas Haploinsufficiency Causes Developmental Delay and Seizures. Cell *172*, 924-936.e11. 10.1016/j.cell.2018.02.006.

44. Li, J., Gao, K., Yan, H., Xiangwei, W., Liu, N., Wang, T., Xu, H., Lin, Z., Xie, H., Wang, J., et al. (2019). Reanalysis of whole exome sequencing data in patients with epilepsy and intellectual disability/mental retardation. Gene *700*, 168–175. 10.1016/j.gene.2019.03.037.

45. Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I., Mounier, N., Dessimoz, C., and Sedlazeck, F.J. (2019). Structural variant calling: the long and the short of it. Genome Biol. *20*, 246. 10.1186/s13059-019-1828-7.

46. Lang, R., Liu, G., Shi, Y., Bharadwaj, S., Leng, X., Zhou, X., Liu, H., Atala, A., and Zhang, Y. (2013). Self-Renewal and Differentiation Capacity of Urine-Derived Stem Cells after Urine Preservation for 24 Hours. PLoS ONE *8*, e53980. 10.1371/journal.pone.0053980.

DISCUSSION

Nos travaux décrivent l'utilité d'une approche diagnostique intégrant le RNA-Seq et une technique long fragments, tel que l'OGM, au séquençage de génome. Notre cohorte était composée de 33 individus présentant une déficience intellectuelle ou un retard global de développement et ayant tous reçus un SE et une ACPA non conclusifs. A l'issue de ce SG, tous les patients sans diagnostic se sont vus proposer un prélèvement de fibroblastes pour un RNA-seq complémentaire. Neuf familles ont accepté (9 cas index et une mère non symptomatique). Enfin, un OGM a été effectué pour deux individus en confirmation de variants. Cette approche a permis d'atteindre un rendement diagnostique global de 48% sur la cohorte. Ce rendement brut est toutefois difficile à comparer, en l'état, à ceux rapportés dans la littérature. Il est nécessaire pour cela de catégoriser les types de variants détectés pour tenir compte des variations de méthodologies des diverses études.

1. Apport et limites du SG short read

Comme nous l'avons présenté en introduction, le SG offre globalement de meilleures performances pour le diagnostic des TND que les techniques standards, notamment le SE combiné avec la ACPA. Toutefois, le premier constat de notre étude est que 62,5% des cas résolus (10/16) l'ont été par identification de SNV/Indels dans des exons ou sites canoniques d'épissages. Ce résultat peut paraitre surprenant car, ces régions étant théoriquement couvertes en SE, ces variants auraient pu être détectés lors des analyses précédentes. Rétrospectivement, nous relevons que les variants causaux rapportés chez 7 patients (21%) avaient effectivement été détectés en SE mais que leur caractère

pathogène n'avait pas été retenu lors de cette première analyse. Pour ces 7 cas, 4 années se sont écoulées entre le SE et le SG. Au cours de cet intervalle de temps, de nouvelles associations gène-phénotype ont été publiées et ont ainsi conduit à une révision de ces variants. Ce phénomène a été documenté dans la littérature sous le nom « d'effet de ré-analyse ». Nous pouvons définir cet effet comme le rendement diagnostique obtenu par ré-analyse des données de SE avec de nouveaux outils bio-informatiques et à la lumière des nouvelles associations gène-phénotype publiées [245–247]. Il existe une corrélation directe entre le temps écoulé et le rendement imputable à l'effet de ré-analyse. Une étude sur la ré-analyse mensuelle de 240 cas non résolus a montré un rendement de 0,57% par mois [248]. Nous notons qu'une extrapolation de ce taux sur 4 ans (27%) est cohérent avec les 21% que nous observons sur notre cohorte. Est-il donc préférable de ré-analyser systématiquement les données de SE avant un SG ? Nous avons un temps émis l'hypothèse que le SG pouvait avoir une contribution indirecte dans la révision des SNV/Indels. En effet, le fait de ne pas détecter d'autres événements tels que des SV aurait pu renforcer l'intérêt pour des variants à première vue peu convaincants. Nous ne pouvons pas exclure que cela puisse participer au processus décisionnel lors du choix de candidats recherche parmi les VUS. En revanche, rien ne nous permet de soutenir cette hypothèse en ce qui concerne la classification des cas résolus. La ré-analyse d'exome est une tâche au coût marginal pouvant s'inscrire dans une routine et qui, selon nous, devrait être un préalable à toute nouvelle analyse par SG.

2. Importance des SV dans les cas résolus et candidats recherche

Pour quantifier la contribution exclusive de notre approche multi-omique nous avons calculé le rendement propre au SG en excluant les cas résolus par effet de ré-analyse et un cas résolu par panel sur prélèvement salivaire (SNV en mosaïque absent du sang).

Nous avons rapporté un diagnostic pour 8 patients soit un rendement de 24% ce qui est comparable aux 19%-25% rapportés dans la littérature [249–251]. Comme nous pouvions nous y attendre, une part importante de ces cas résolus grâce au SG impliquaient des SV (6/8, 75%). Parmi ces SV, 4 étaient des CNV qui impliquaient des exons codants et qui n'avaient été détectés ni par SE ni par ACPA (délétion dans *SLC6A8*, *RSPRY1*, *STEEP1* et duplication intragénique de *PUM1*). L'explication la plus probable à l'échec du SE pour ces cas est la faible sensibilité de cette technique sur les événements ayant cette petite taille. Un autre cas est dû à un événement équilibré avec points de cassure introniques, une inversion dans *FOXP1*, non détectable en SE. Enfin, le dernier cas résolu par un SV implique la délétion dans le long ARN non-codant *CHASERR* dont les exons ne sont pas couverts en SE. Notre étude a également apporté plusieurs candidats recherche dont une délétion dans le gène *CBX3*. Bien que la combinaison d'algorithmes de *calling* des SV que nous avons utilisé ait détecté tous ces événements sur la base des données de SG seules, l'apport du RNA-seq et de l'OGM a été déterminant pour 3 cas. Deux cas de figures se sont présentés : 1°/ Difficultés à confirmer le variant lors de la validation manuelle sur IGV. 2°/ Interprétation difficile d'un variant dans un ARN non-codant.

L'OGM nous a permis de confirmer un variant pour lequel un doute subsistait. En effet, la duplication dans *PUM1*, bien que détectée par les algorithmes basés sur la *read depth*, n'était pas totalement convaincante sur IGV. Les larges régions homologues présentes à proximité de cet événement rendaient les algorithmes basés sur les *split read* et discordance de paires inopérants. Par ailleurs, l'OGM a apporté une aide précieuse dans la compréhension d'un événement complexe impliquant le gène *FGF18*. Les données de SG *short read* nous laissaient penser à une inversion suivie d'une duplication au niveau

d'un des points de cassure. L'OGM a infirmé cette hypothèse. Sur la base des nouvelles informations, l'hypothèse privilégiée serait une insertion inversée. Il est tentant de se questionner sur le choix de l'OGM alors que le SG *long read* est de plus en plus accessible. Notre choix méthodologique pour l'approche long fragments reposait sur une confirmation de variant. La résolution de l'OGM étant suffisante pour nos candidats, et son coût étant moins élevé, nous avons privilégié cette technique. L'approche SG *long read*, plus coûteuse, serait, selon nous préférable pour des cas sans candidats après SG *short read* et RNA-Seq.

Parmi tous les SV identifiés, 8 étaient *de novo*. Nous avons cherché à déterminer s'il était possible d'inférer un mécanisme d'apparition de ces événements (Tableau 6). Avant tout, il est intéressant de noter que tous les SV en apparence équilibrés s'accompagnaient d'indels aux points de cassure et, qu'un seul d'entre eux est apparu dans une région pouvant expliquer la survenue d'un remaniement (régions riches en AT). Pour 3 CNV, la nature des séquences aux points de cassure était informative. En effet, la délétion dans *CHASERR* implique une microhomologie de 8 pb (TCACTTCA), la délétion dans *CBX3* implique deux éléments Alu (AluSp et AluSq2) et la duplication dans *PUM1* implique deux éléments LINE-2 (L2c).

Variant	Type de SV	Gène	Nature des points de cassure
NC_000003.11:g.68954396_71064931inv	Inversion	*FOXP1*	-
NC_000023.10:g.118673832_118675628del	Délétion	*STEEP1*	-
NC_000001.10:g.31409001_31422000dup	Duplication	*PUM1*	LINE-2 (L2c ; L2c)
NC_000015.9:g.93422237_93430600del	Délétion	*CHASERR*	Microhomologie (8 pb : TCACTTCA)
NC_000007.13:g.26244486_26248316del	Délétion	*CBX3*	Alu (AluSp ; AluSq2)
NC_000001.10:g.pter_69716317delins[NC_000006.11:g.108848601_qterinv] NC_000006.11:g.108848601_qterdelins[NC_000001.10:g.69716317_pterinv]	Translocation	-	-
NC_000005.9:g.105706517_170958911inv	Inversion	-	AT-rich
Complex structural variant duplication and inverted insertion	Complexe	*FGF18*	-

Tableau 6 : Nature des points de cassure des SV identifiés

En ce qui concerne la délétion dans *CHASERR*, nous pouvons donc émettre l'hypothèse d'un mécanisme de MMEJ. Le MMBIR est également une hypothèse possible, mais sans pouvoir totalement l'exclure, elle serait plus probable dans le cas d'un événement complexe, ce qui n'est pas le cas ici. Les deux autres CNV impliquent des rétrotransposons partageant une forte homologie ce qui suggère plutôt des mécanismes de NAHR. Le remaniement entre Alu dans *CBX3* n'est pas quelque chose d'inhabituel, les variants structuraux entre éléments de ce type sont connus et fréquents. En revanche, la duplication dans *PUM1* impliquant deux LINE-2 se révèle plus surprenante. En effet, elle implique deux régions avec une homologie parfaite de 531pb alors même que les L2 sont

considérés comme des éléments ayant fortement dérivés et ne conservant qu'une homologie très faible entre eux. On constate néanmoins que la région d'homologie concernée est plus complexe qu'un simple L2c. Sur les 531 pb seulement 82 correspondent à la signature L2c, les 449 autres bases de la séquence sont annotées comme appartenant à deux snoRNA (*SNORD103A* et *SNORD103B*). L'homologie inhabituelle de ces éléments L2c pourrait donc s'expliquer par deux hypothèses non exclusives l'une de l'autre : Une mobilisation relativement récente de ce L2 ; ou, plus probablement, une pression de sélection sur les snoRNA fonctionnels.

Bien que le cas du L2c dans *PUM1* que nous venons d'évoquer soit vraisemblablement un cas particulier, la question de la récurrence des SV impliquant des rétrotransposons de grandes tailles et de leur difficulté à être détectés en génome *short read* se pose. Nous avons cherché à évaluer quels types de retrotranspons sont régulièrement rapportés en lien avec des SV. Pour cela, nous avons extrait de la base ClinVar [252] les CNV pour lesquels les points de cassure exacts étaient connus et nous avons isolés ceux pour lesquels deux retrotranspons de même famille étaient présents aux points de cassure (Figure 28). Il apparait bien, comme nous l'attendions, que les L2 ne sont que marginalement impliqués comparativement aux L1 et Alu. Les Alu sont les ET plus nombreux dans le génome humain et aussi ceux qui semblent le plus souvent associés aux CNV. Une méthode de prédiction des gènes à risque de recombinaison médiées par des Alu a d'ailleurs été publiée [159].

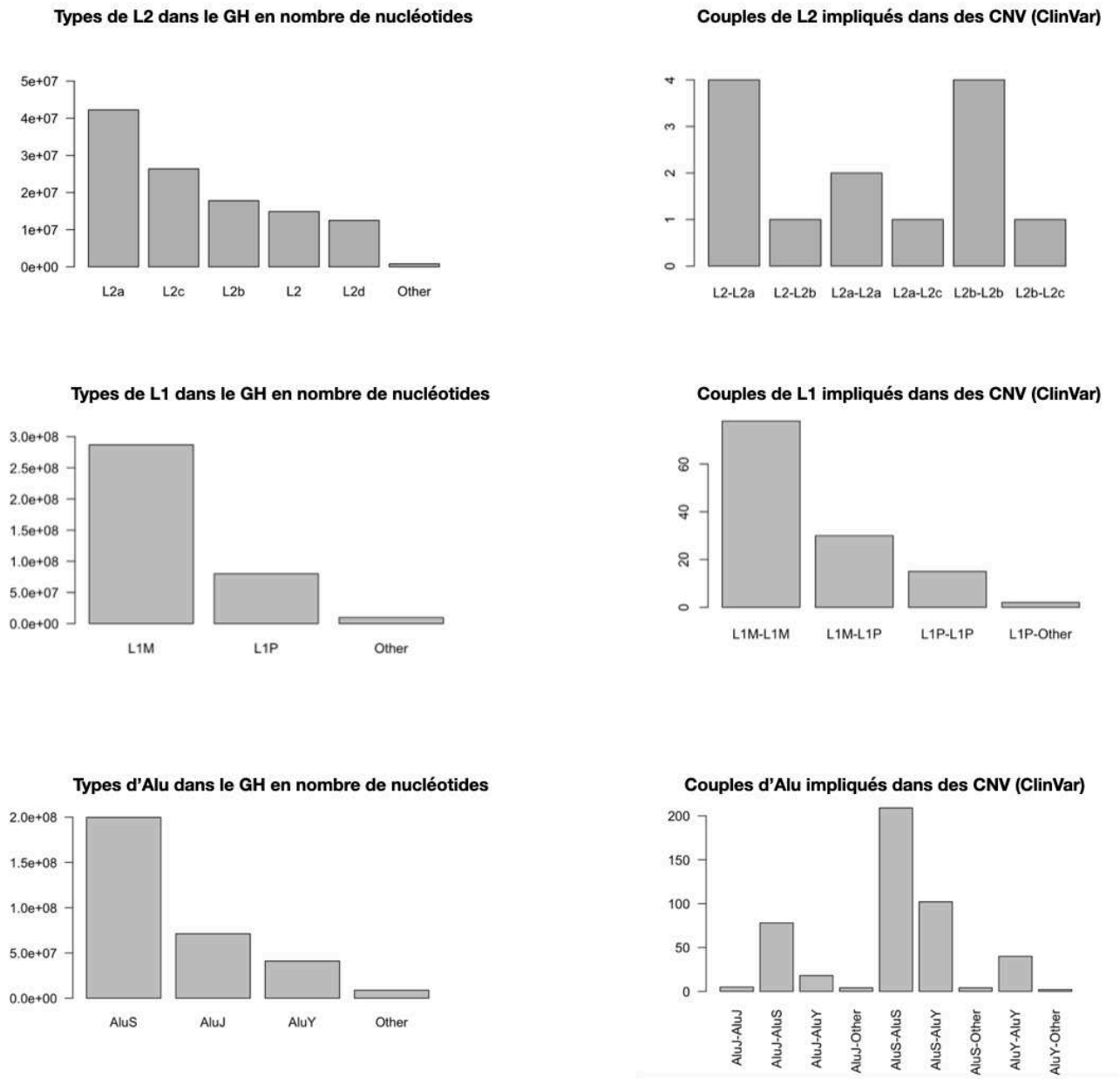


Figure 32 : Rétrotransposons impliqués dans des CNV pathogènes ou probablement pathogènes rapportés dans ClinVar. GH : Génome humain. Base de données de retrotransposons : Repeat Masker (https://www.repeatmasker.org)

La difficulté que nous avons eu à identifier par SG *short read* la duplication impliquant des régions homologues de 531pb nous a incité à entreprendre une démarche similaire, mais pour les L1. En effet, même s'ils sont moins fréquemment en cause que les Alu, ils sont, comme eux, des éléments relativement récents avec une forte homologie. Mais, avec leur taille moyenne de 6kb, ils sont beaucoup plus grand que les Alu. De fait, un SV dont les points de cassure tomberaient dans des L1 de très forte homologie pourraient être

difficile à détecter sur la base de *split read* ou de discordance entre paires. Cette difficulté sera donc particulièrement marquée si le SV est équilibré. Nous avons donc cherché à déterminer le nombre de gènes associés aux TND (base SysNDD) susceptibles d'être concernés par ce problème. Pour cela, nous avons extrait les coordonnées de tous les L1 présents dans des gènes associés aux TND et nous n'avons conservé que ceux situés dans des régions de faible qualité d'alignement (MAPQ moyen < 30). Puis, nous avons calculé les coordonnées d'hypothétiques SV médiés par des L1 de mêmes types ayant une homologie supérieure ou égale à 98%. Plus de 30 000 paires possibles ont été calculées. Pour la majorité d'entre elles, il s'agissait de translocations. Nous avons choisi de limiter cette étude aux inversions, moins nombreuses. Nous prédisons donc 808 inversions possibles, potentiellement impossibles à détecter en SG *short read*, dans 56 gènes. Ces prédictions pourraient servir de base pour justifier une approche complémentaire par une technique long fragments (OGM, SG *long read*) quand un phénotype est évocateur. Par exemple, nous avons eu le cas d'un fœtus présentant une holoprosencéphalie pour lequel un variant non-sens dans le gène *PLCH1* a été détecté par SG *short read*. Ce gène est impliqué dans des holoprosencéphalies récessives mais aucun événement n'a été détecté sur le deuxième allèle et aucun autre variant candidat n'a été identifié. Toutefois, nous prédisons qu'un L1 (L1PA2) dans l'intron 12 de *PLCH1* et situé dans une région de MAPQ moyen inférieur à 20 est susceptible de former 16 inversions avec des éléments partageant 98% d'homologie. Face à ce type de phénotype évocateur, une recherche de SV par une méthode complémentaire paraitrait justifiée. Nous n'avons malheureusement pas pu aller plus loin sur ce cas précis. Nous n'avions à notre disposition que de l'ADN fœtal, partiellement dégradé, et donc inutilisable pour une approche long fragments.

3. Apport et limites du RNA-Seq pour la confirmation de certains SV

Le RNA-Seq a été informatif pour trois patients. La délétion de deux exons en phase dans le gène *CBX3*, identifiée en SG, mais pour laquelle un doute subsistait après vérification IGV a ainsi été confirmée. L'analyse du cas impliquant une duplication du gène *PUM1* a également confirmé les résultats du SG et de l'OGM. En effet, l'analyse a montré que son ARNm présentait une jonction anormale entre l'extrémité 3′ de l'exon 21 et 5′ de l'exon 18 (Figure 29). Par ailleurs, on constate que l'ARNm n'est pas tronqué. Il est néanmoins important de noter que cette jonction aberrante n'a été constatée que par vérification manuelle. En effet, l'algorithme FRASER n'a pas détecté cette aberration dans la version utilisée lors de notre première analyse. Plus surprenant, lors du renouvellement de l'analyse avec une version plus récente de FRASER, cette aberration a bien été détectée mais d'autres événements parfaitement détectés jusque-là, comme la jonction aberrante dans *CBX3*, ne l'étaient plus.

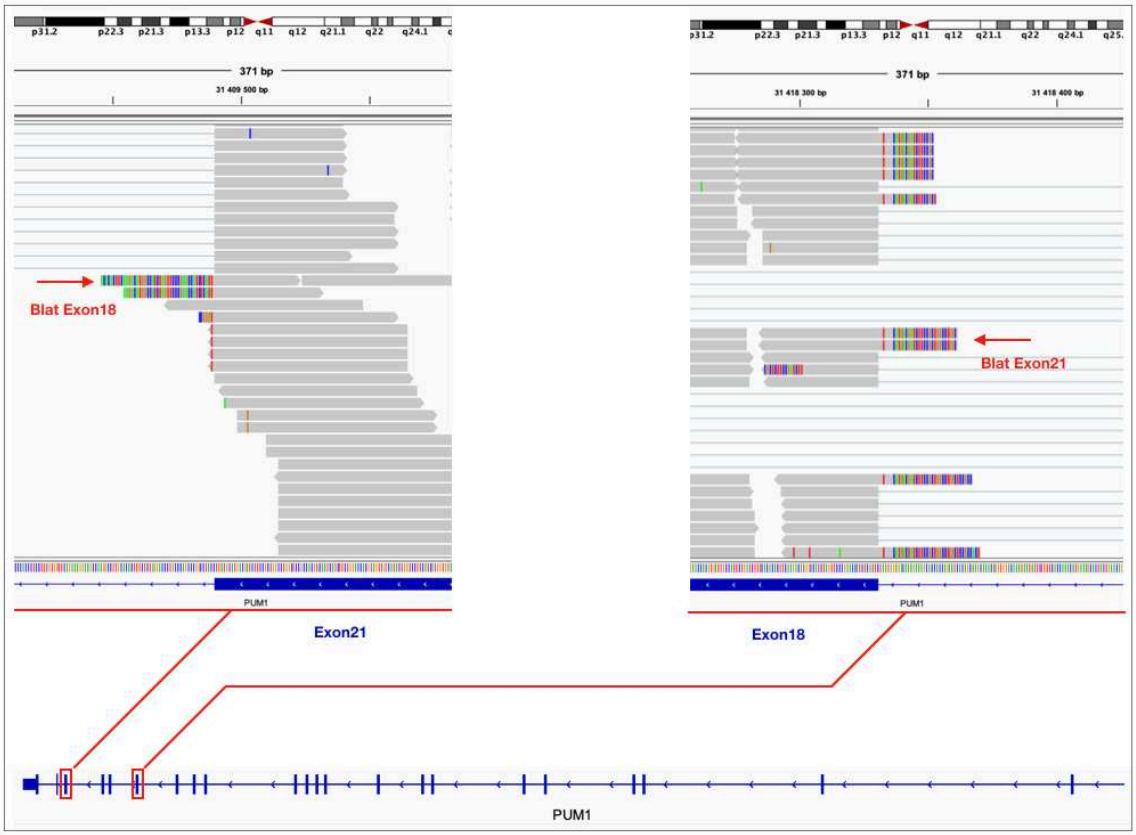


Figure 33 : Aperçu IGV de la jonction aberrante entre les exons 21 et 18 de *PUM1*

Le troisième cas pour lequel le RNA-Seq a été déterminant est la délétion dans le lncRNA *CHASERR*. Bien que cette délétion ait été identifiée sans équivoque en SG, il était difficile, en l'état, de conclure quant à sa pathogénicité. Ce variant a retenu notre attention pour plusieurs raisons. Tout d'abord, CHASERR fait partie des rares lncRNA dont le locus est très conservé. Ensuite, une délétion comparable avait déjà été publiée pour une patiente avec un phénotype similaire [89]. Enfin, une étude fonctionnelle sur des modèles murins et cellulaires a montré un rôle régulateur sur le gène voisin *CHD2*. Dans cette étude, il est précisément suggéré que CHASERR est un inhibiteur de *CHD2*. Cette inhibition en *-cis* est indépendante d'un effet positionnel [253]. Les haploinsuffisances de *CHD2* sont déjà associées à des TND mais leurs phénotypes cliniques d'encéphalopathies épileptiques développementales (MIM : 615369) sont différents de ceux des patients avec délétion de *CHASERR*. Toutefois, l'effet attendu des

haploinsuffisances de *CHASERR* étant une surexpression de *CHD2*, il n'était pas exclu que les manifestations cliniques soient propres à ce mécanisme. Par ailleurs, aucune duplication complète de *CHD2* pouvant mimer les effets d'une surexpression, et ainsi nous servir de point de comparaison, n'a été rapportée à ce jour. L'analyse par RNA-Seq des fibroblastes du patient a parfaitement détecté une perte d'expression d'environ 50% de *CHASERR*. Nous savions, grâce à la présence exclusive de SNP maternels au sein de la délétion, que la copie délétée (*de novo*) était l'allèle paternel. Grâce à cette information, et la présence d'autres SNP dans *CHD2*, nous avons pu mettre en évidence par RNA-seq un biais allélique dans ce gène en faveur de l'allèle paternel, donc en *-cis* de la délétion *CHASERR* comme rapporté dans les études fonctionnelles. Cet impact confirmé au niveau transcriptionnel a été, pour nous, déterminant dans le classement pathogène de ce variant. En effet, bien que les phénotypes soient très proches, le fait que seulement un patient soit publié à ce jour invitait à la prudence. Aujourd'hui, un troisième patient a été identifié (non publié) et des travaux de confirmation de l'impact de *CHASERR* au niveau protéique sont en cours. Grâce à notre étude, *CHASERR* rejoint la très courte liste des lncRNA impliqués dans des pathologies humaines à transmission mendélienne. Il nous semble toutefois important de souligner que, là encore, il s'agissait d'une vérification ciblée, reposant sur une hypothèse préalable du rôle de *CHASERR* sur un gène voisin. Le biais allélique étant d'un rapport d'environ 80/20% il aurait été théoriquement possible de le détecter, à l'aide du module recherchant les expressions monoalléliques du pipeline DROP, en abaissant le seuil de détection sous les 80%. Nous avons testé cette configuration et même abaissé le seuil à 70%. Même avec ces paramètres, le biais allélique n'a jamais été mis en évidence par DROP. Il s'agit là d'une des limites du RNA-Seq clinique que nous avons relevée. L'apport du RNA-Seq pour la compréhension des variants dans les régions non-codantes connait toutefois quelques limites. Nous avons

choisi d'isoler les ARN par capture poly-A ce qui limite le nombre d'ARN non-codant que nous pouvons étudier. Les miRNA ne nous étaient pas accessibles et près de la moitié des lncRNA ne le sont pas non plus. Il est possible de réaliser l'analyse sur ARN total mais au prix d'une baisse de sensibilité globale sur les ARNm, ce qui n'est pas souhaitable. Un enrichissement en ARN non-codants comme les miRNA pourrait être envisagé au cas par cas si un variant génomique était identifié dans un de ces gènes en SG. Toutefois, à ce jour les miRNA ne sont pas rapportés comme fréquemment impliqués dans des TND. Les variants dans ces ARN ne sont peut-être pas rapportés par manque de preuves pour soutenir leur pathogénicité ou bien ils ne jouent pas de rôle majeur dans ces pathologies. L'intensification du recours au diagnostic par SG nous permettra peut-être de clarifier ce point.

Nous avons été confrontés à plusieurs des limites du RNA-Seq. Tout d'abord, comme nous l'avons dit, les outils bio-informatiques pour l'analyse des données sont peu nombreux et encore à leurs débuts. Malgré l'augmentation du nombre de publications, il est encore difficile d'évaluer précisément le nombre d'événements ratés. Le fait que des variants confirmés soient détectés par une version antérieure d'un algorithme mais pas par une version plus récente montre qu'une analyse plus approfondie des performances s'impose avant une implémentation en routine. De plus, le principe même de l'identification d'événements par recherche *d'outliers* nécessite de constituer une cohorte conséquente dont les conditions de séquençage doivent être les plus proches possibles. Des matrices externes de plusieurs centaines d'individus sont mises à disposition de la communauté pour les analyses sur sang total ou fibroblastes. L'utilisation de ces matrices est indispensable pour de petites cohortes comme la nôtre, tout particulièrement pour la détection d'expression aberrante. Nous avons en effet pu

constater que nous n'aurions pas pu identifier nos événements d'intérêt si nous avions lancé l'analyse avec moins de 50 échantillons au total (cohorte + matrice externe). Néanmoins, l'utilisation de la matrice externe génère un fort effet batch pouvant potentiellement diminuer les performances du pipeline. La filtration des événements peu également présenter un challenge pour le biologiste. Le nombre d'événements aberrant retournés par le pipeline n'est pas aussi important que ce que l'on peut avoir avec le SG mais il nécessiterait tout de même d'être réduit pour être facilement exploitable. Toutefois, même en gardant les événements statistiquement hautement significatifs, l'interprétation n'est pas toujours aisée. Beaucoup d'anomalies d'expression n'ont pas pu être mises en lien avec un variant détecté en SG ni même après vérification manuelle du gène dans IGV. Nous ne pouvons qu'émettre l'hypothèse d'une modification d'un régulateur inconnu, sans pouvoir vérifier cela pour le moment.

Une autre limite du RNA-Seq clinique est d'ordre biologique. L'expression des gènes étant tissu spécifique, est-il pertinent de travailler avec des échantillons sanguins ou des fibroblastes de peau ? Pour les cas que nous avons résolus, les fibroblastes exprimaient suffisamment les gènes concernés. D'ailleurs ces gènes sont également exprimés dans le sang total. Il n'est toutefois pas possible d'affirmer que les événements auraient été détectés par une analyse sur échantillon sanguin. En effet, outre le niveau d'expression des gènes d'intérêt, il faut prendre en compte l'homogénéité d'expression dans le tissu. Les niveaux d'expression de gènes dans le sang sont très hétérogènes d'un individu à l'autre, ce qui crée une variabilité préjudiciable à la performance des algorithmes. Ce qui explique, en partie, les rendements inférieurs du RNA-Seq sur sang total comparé aux fibroblastes, plus homogènes [89]. Hélas, les impacts transcriptionnels de certains variants, comme la duplication de *FGF18* par exemple, ne sont pas vérifiables sur

prélèvement de fibroblastes, car les gènes ne sont pas exprimés dans ces derniers. La pertinence des fibroblastes pour le diagnostic des TND repose sur trois critères : 1°/Le pourcentage de gènes associés au TND exprimés ; 2°/Leur homogénéité ; 3°/Leur relative facilité d'obtention, par simple biopsie cutanée. En effet, environ 60% des gènes impliqués dans des TND sont exprimés dans les fibroblastes que nous avons analysés (TPM > 10) alors qu'il y en a moins de 40% dans le sang total. Aujourd'hui, à notre connaissance, nous n'avons pas trouvé de taux significativement plus élevé dans un tissu cliniquement accessible. Si les fibroblastes semblent, de toute évidence, un bon compromis, il n'est toutefois pas rare que le prélèvement soit refusé par les familles. En effet, bien que peu invasive, la biopsie cutanée nécessaire à leur obtention peut contribuer à une moindre adhésion des patients.

Plusieurs alternatives peuvent, *a priori*, être envisagées comme les cellules buccales, folliculaires ou des cultures de lymphocytes. Leur potentiel, tant en taux de gènes associés aux TND exprimé qu'en homogénéité n'a pas été spécifiquement étudié dans un contexte de RNA-Seq clinique. En ce qui nous concerne, nous avons choisi de nous intéresser au potentiel des cellules souches urinaires (*urine derived stem cells* : USC). Il s'agit d'un type cellulaire retrouvé dans les urines (1 à 5 cellules pour 250 ml d'urine) et qui présente des caractéristiques des cellules souches notamment la multipotence et les capacités de prolifération [254,255]. Il est possible de générer des populations clonales en quantité suffisante pour une analyse RNA-Seq en moins de 15 jours (Figure 30).

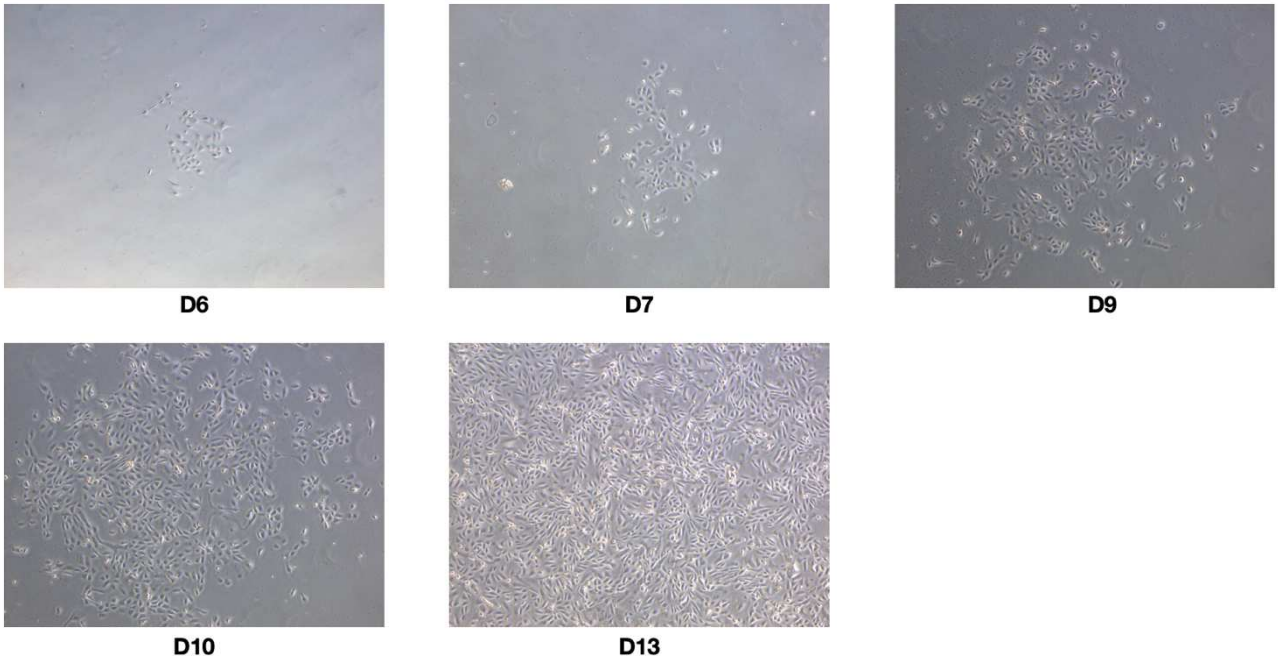


Figure 34 : Mise en culture d'USC jusqu'à J13

Deux types sont généralement présents dans les prélèvements. Il est facile de les isoler sur la base de leur morphologie pour générer des cultures homogènes (Figure 31). Outre leur différence de morphologie, les types diffèrent dans leur vitesse de prolifération et leurs nombres de passages possibles.

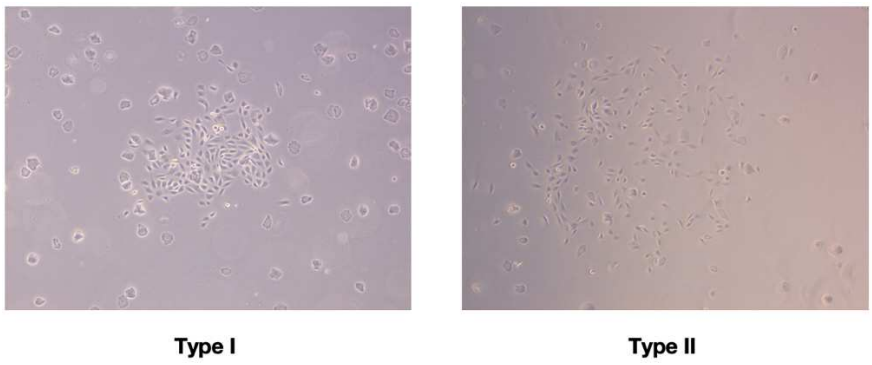


Figure 35 : Deux types d'USC rencontrés dans les prélèvements

Ce choix de tissu nous semble pertinent à plus d'un titre. Tout d'abord, le prélèvement ne requiert pas de geste invasif et l'adhésion du patient est donc excellente. Deuxièmement,

le taux de gènes associés aux TND exprimé (TPM > 10) est comparable à celui des fibroblastes (60%). Il est intéressant de noter que certains gènes, comme plusieurs gènes synaptiques, sont exprimés dans les USC mais pas dans les fibroblastes (Figure 32). Ce type cellulaire présente des avantages pour un déploiement en routine du RNA-Seq, mais il y a tout de même des contraintes à prendre en compte en pratique. Notamment, il convient de ne pas sous-estimer les difficultés de prélèvement sur des patients très jeunes et présentant des formes sévères de TND. De plus, les risques de contamination ne sont pas négligeables lors du recueil et près de 40% des cultures échouent. Enfin, pour pouvoir évaluer le potentiel de ces USC, il faut dans un premier temps constituer une cohorte conséquente que nous estimons à pas moins de 70 individus. Nous n'avons malheureusement pas eu le temps d'aller jusqu'à ce stade au cours de cette thèse mais, il me semble intéressant de poursuivre ces investigations.

La littérature suggère que USC pourraient avoir un autre atout. Comme cela a été montré avec d'autres types cellulaires, comme les lymphocytes ou les fibroblastes [256,257], il est possible de transdifférencier les USC en cellules *neuron-like* appelées iNeurons (iN) [258,259]. L'intérêt de ces iN pour notre question biologique est double. Ces iN expriment jusqu'à 78% de gènes associés au TND (TPM > 10), et beaucoup de gènes synaptiques (Figure 32). Ils pourraient donc être utilisés pour confirmer des variants au cas par cas, tout particulièrement les variants d'épissage. En revanche, étant donné le temps nécessaire et la complexité du protocole, il parait peu réaliste d'envisager une application en routine sur un grand nombre de patients dans une optique exploratoire. De plus, il faut également rester prudent lorsque nous concluons à une anomalie d'épissage sur un type cellulaire non-naturel comme celui-ci. En effet, il est courant d'observer une différence d'expression des isoformes d'un type cellulaire physiologique à un autre (par exemple

l'isoforme APP695 que nous avons rencontrée au cours de nos travaux et dont l'expression est principalement neuronale alors que d'autres isoformes sont ubiquitaires). Il parait donc probable que des variations d'expression d'isoformes entre ces iN et les différents types de neurones humains puissent là aussi exister.

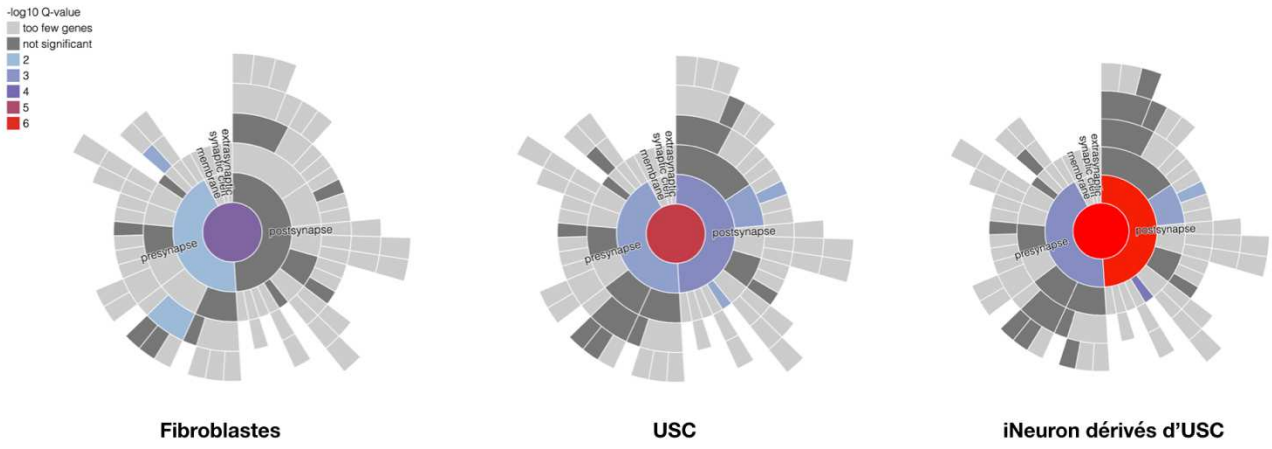


Figure 36 : Comparaisons des ontologies associées aux synapses enrichies dans les gènes exprimés par les fibroblastes, USC et iN. Généré à l'aide de SynGO [260]

Enfin, au-delà du diagnostic, la question de l'impact fonctionnel d'un variant pathogène se pose. Des iN ont montré qu'ils pouvaient former des dendrites et des synapses potentiellement fonctionnelles [256]. Ils pourraient donc constituer un nouvel outil pour ces validations fonctionnelles directement à partir des USC déjà cultivées sans recourir à une biopsie ni mettre en œuvre de lourds protocoles impliquant des cellules souches induites (iPSC). Peu de protocoles ont été publiés pour la transdifférenciation des USC en iN [258,259]. Leur avantage par rapport à ceux impliquant les lymphocytes et les fibroblastes est leur simplicité de mise en œuvre et l'absence de recours à des cellules gliales murines nourricières. Nous avons tenté de reproduire un de ces protocoles [258],

mais sans succès à ce jour. Un changement de morphologie a bien été constaté, mais, ces cellules n'ayant pas survécu, il est impossible de confirmer si cela était dû à la transdifférenciation ou à d'autres facteurs. De plus, ces essais ayant été réalisés sur des cellules de patients, il n'est pas exclu que les variants pathogènes aient un impact négatif sur la *fitness* des cellules. Malgré ces difficultés, nous pensons qu'il est opportun de continuer à développer et tester ce modèle pour répondre aux nouveaux challenges dans l'interprétation des variants, tout particulièrement ceux dans les régions non-codantes.

CONCLUSION ET PERSPECTIVES

Partout dans le monde, le SG devient progressivement la technique de première intention pour le diagnostic des TND. La France, en accord avec le plan France Médecine Génomique 2025, suit cette tendance. Notre étude a confirmé les avantages de cette technique comparée au standard actuel SE + ACPA. L'apport est, comme nous l'avions anticipé, particulièrement évident pour les variants structuraux et pour les variants non-codants. Toutefois, comme nous l'avons constaté, certains SV restent difficiles à confirmer et leur impact pas toujours simple à appréhender. Cette difficulté d'interprétation est particulièrement marquée pour les variants dans les régions non-codantes. Les techniques complémentaires RNA-Seq et OGM ont montré leur utilité en complément du SG *short read*. Toutefois, certaines de ces techniques en sont encore à leurs débuts et nécessitent d'être évaluées dans des études plus larges. Nous avons vu également que le RNA-Seq a des performances qui sont tissus dépendantes et, là encore, il est indispensable de comparer ces tissus dans des études solides.

Cette approche multi-omique au cas par cas est un pas de plus vers une médecine personnalisée. D'ailleurs, la piste de validation fonctionnelle sur iN que nous avons évoqué en fin de cette thèse s'inscrit, elle aussi, pleinement dans ce nouveau paradigme. Toutes ces nouvelles connaissances acquises et ces évolutions technologiques font que la génétique médicale est, et restera encore longtemps, un monde complexe et passionnant.

Références

1. Morris-Rosendahl, D.J., and Crocq, M.-A. (2020). Neurodevelopmental disorders—the history and future of a diagnostic concept. Dialogues in Clinical Neuroscience *22*, 65–72. 10.31887/DCNS.2020.22.1/macrocq.

2. La Malfa, G., Lassi, S., Bertelli, M., Salvini, R., and Placidi, G.F. (2004). Autism and intellectual disability: a study of prevalence on a sample of the Italian population. J Intellect Disabil Res *48*, 262–267. 10.1111/j.1365-2788.2003.00567.x.

3. Socanski, D., Aurlien, D., Herigstad, A., Thomsen, P.H., and Larsen, T.K. (2013). Epilepsy in a large cohort of children diagnosed with attention deficit/hyperactivity disorders (ADHD). Seizure *22*, 651–655. 10.1016/j.seizure.2013.04.021.

4. Brikell, I., Ghirardi, L., D'Onofrio, B.M., Dunn, D.W., Almqvist, C., Dalsgaard, S., Kuja-Halkola, R., and Larsson, H. (2018). Familial Liability to Epilepsy and Attention-Deficit/Hyperactivity Disorder: A Nationwide Cohort Study. Biological Psychiatry *83*, 173–180. 10.1016/j.biopsych.2017.08.006.

5. Robertson, J., Hatton, C., Emerson, E., and Baines, S. (2015). Prevalence of epilepsy among people with intellectual disabilities: A systematic review. Seizure *29*, 46–62. 10.1016/j.seizure.2015.03.016.

6. Zablotsky, B. (2017). Estimated Prevalence of Children With Diagnosed Developmental Disabilities in the United States, 2014–2016.

7. Yang, Y., Zhao, S., Zhang, M., Xiang, M., Zhao, J., Chen, S., Wang, H., Han, L., and Ran, J. (2022). Prevalence of neurodevelopmental disorders among US children and adolescents in 2019 and 2020. Front. Psychol. *13*, 997648. 10.3389/fpsyg.2022.997648.

8. Rh, B., Jr, H., Lr, R., Jw, K., Oj, L., Ae, C., and Ke, Z. (2020). National Health Statistics Reports, Number 139, February 19, 2020.

9. Maulik, P.K., Mascarenhas, M.N., Mathers, C.D., Dua, T., and Saxena, S. (2011). Prevalence of intellectual disability: A meta-analysis of population-based studies. Research in Developmental Disabilities *32*, 419–436. 10.1016/j.ridd.2010.12.018.

10. Geschwind, D.H. (2011). Genetics of autism spectrum disorders. Trends in Cognitive Sciences *15*, 409–416. 10.1016/j.tics.2011.07.003.

11. D.D.M.N.S.Y.P. Investigators, Prevalence of autism spectrum disorder among children aged 8 years – Autism and developmental disabilities monitoring network, 11 sites, United States, 2010, MMWR Surveill Summ 63 (2) (2014) 1–21.

12. Gonzalez-Mantilla, A.J., Moreno-De-Luca, A., Ledbetter, D.H., and Martin, C.L. (2016). A Cross-Disorder Method to Identify Novel Candidate Genes for Developmental Brain Disorders. JAMA Psychiatry *73*, 275. 10.1001/jamapsychiatry.2015.2692.

13. Simon, V., Czobor, P., Bálint, S., Mészáros, Á., and Bitter, I. (2009). Prevalence and correlates of adult attention-deficit hyperactivity disorder: meta-analysis. Br J Psychiatry *194*, 204–211. 10.1192/bjp.bp.107.048827.

14. Polanczyk, G., de Lima, M.S., Horta, B.L., Biederman, J., and Rohde, L.A. (2007). The

Worldwide Prevalence of ADHD: A Systematic Review and Metaregression Analysis. Am J Psychiatry.

15. Kessler, R.C., Adler, L., Barkley, R., Biederman, J., Conners, C.K., Demler, O., Faraone, S.V., Greenhill, L.L., Howes, M.J., Secnik, K., *et al.* (2006). The Prevalence and Correlates of Adult ADHD in the United States: Results From the National Comorbidity Survey Replication. Am J Psychiatry.

16. Frazier, T.W., Youngstrom, E.A., Glutting, J.J., and Watkins, M.W. (2007). ADHD and Achievement: Meta-Analysis of the Child, Adolescent, and Adult Literatures and a Concomitant Study With College Students. J Learn Disabil *40*, 49–65. 10.1177/00222194070400010401.

17. Faraone, S.V., and Larsson, H. (2019). Genetics of attention deficit hyperactivity disorder. Mol Psychiatry *24*, 562–575. 10.1038/s41380-018-0070-0.

18. Stiles, J., and Jernigan, T.L. (2010). The Basics of Brain Development. Neuropsychol Rev *20*, 327–348. 10.1007/s11065-010-9148-4.

19. Lange, S., Probst, C., Gmel, G., Rehm, J., Burd, L., and Popova, S. (2017). Global Prevalence of Fetal Alcohol Spectrum Disorder Among Children and Youth: A Systematic Review and Meta-analysis. JAMA Pediatr *171*, 948. 10.1001/jamapediatrics.2017.1919.

20. Déficiences et handicaps d'origine périnatale: dépistage et prise en charge (2004). (Paris: INSERM).

21. LEJEUNE, J., GAUTIER, M., and TURPIN, R. (1959). [Study of somatic chromosomes from 9 mongoloid children]. C R Hebd Seances Acad Sci *248*, 1721–1722.

22. Shin, M., Besser, L.M., Kucik, J.E., Lu, C., Siffel, C., and Correa, A. (2009). Prevalence of Down syndrome among children and adolescents in 10 regions of the United States. Pediatrics *124*, 1565–1571. 10.1542/peds.2009-0745.

23. Hassold, T., and Hunt, P. (2001). To err (meiotically) is human: the genesis of human aneuploidy. Nat Rev Genet *2*, 280–291. 10.1038/35066065.

24. Goel, N., Morris, J.K., Tucker, D., Walle, H.E.K., Bakker, M.K., Kancherla, V., Marengo, L., Canfield, M.A., Kallen, K., Lelong, N., *et al.* (2019). Trisomy 13 and 18—Prevalence and mortality—A multi-registry population based analysis. Am J Med Genet *179*, 2382–2392. 10.1002/ajmg.a.61365.

25. Martin, J.P., and Bell, J. (1943). A PEDIGREE OF MENTAL DEFECT SHOWING SEX-LINKAGE. Journal of Neurology, Neurosurgery & Psychiatry *6*, 154–157. 10.1136/jnnp.6.3-4.154.

26. Verkerk, J.M.H., and Sutcliffe, J.S. Identification of a Gene (HIM?-1) Containing a CGG Repeat Coincident with a Breakpoint Cluster Region Exhibiting Length Variation in Fragile X Syndrome.

27. Ciaccio, C., Fontana, L., Milani, D., Tabano, S., Miozzo, M., and Esposito, S. (2017). Fragile X syndrome: a review of clinical and molecular diagnoses. Ital J Pediatr *43*, 39. 10.1186/s13052-017-0355-y.

28. Nolin, S.L., Brown, W.T., Glicksman, A., Houck, Jr., G.E., Gargano, A.D., Sullivan, A., Biancalana, V., Bröndum-Nielsen, K., Hjalgrim, H., Holinski-Feder, E., *et al.* (2003). Expansion of the Fragile X CGG Repeat in Females with Premutation or Intermediate Alleles. The

American Journal of Human Genetics *72*, 454–464. 10.1086/367713.

29. Kochinke, K., Zweier, C., Nijhof, B., Fenckova, M., Cizek, P., Honti, F., Keerthikumar, S., Oortveld, M.A.W., Kleefstra, T., Kramer, J.M., *et al.* (2016). Systematic Phenomics Analysis Deconvolutes Genes Mutated in Intellectual Disability into Biologically Coherent Modules. The American Journal of Human Genetics *98*, 149–164. 10.1016/j.ajhg.2015.11.024.

30. Deciphering Developmental Disorders Study (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature *542*, 433–438. 10.1038/nature21062.

31. Kahler, S.G., and Fahey, M.C. (2003). Metabolic disorders and mental retardation. Am. J. Med. Genet. *117C*, 31–41. 10.1002/ajmg.c.10018.

32. Kline, A.D., Moss, J.F., Selicorni, A., Bisgaard, A.-M., Deardorff, M.A., Gillett, P.M., Ishman, S.L., Kerr, L.M., Levin, A.V., Mulder, P.A., *et al.* (2018). Diagnosis and management of Cornelia de Lange syndrome: first international consensus statement. Nat Rev Genet *19*, 649–666. 10.1038/s41576-018-0031-0.

33. Cheon, C.-K., and Ko, J.M. (2015). Kabuki syndrome: clinical and molecular characteristics. Korean J Pediatr *58*, 317. 10.3345/kjp.2015.58.9.317.

34. Aggarwal, A., Rodriguez-Buritica, D.F., and Northrup, H. (2017). Wiedemann-Steiner syndrome: Novel pathogenic variant and review of literature. European Journal of Medical Genetics *60*, 285–288. 10.1016/j.ejmg.2017.03.006.

35. Silbereis, J.C., Pochareddy, S., Zhu, Y., Li, M., and Sestan, N. (2016). The Cellular and Molecular Landscapes of the Developing Human Central Nervous System. Neuron *89*, 248–268. 10.1016/j.neuron.2015.12.008.

36. Jayaraman, D., Bae, B.-I., and Walsh, C.A. (2018). The Genetics of Primary Microcephaly. Annu. Rev. Genom. Hum. Genet. *19*, 177–200. 10.1146/annurev-genom-083117-021441.

37. Gilmore, E.C., and Walsh, C.A. (2013). Genetic causes of microcephaly and lessons for neuronal development. WIREs Dev Biol *2*, 461–478. 10.1002/wdev.89.

38. Bahi-Buisson, N., Poirier, K., Boddaert, N., Saillour, Y., Castelnau, L., Philip, N., Buyse, G., Villard, L., Joriot, S., Marret, S., *et al.* (2008). Refinement of cortical dysgeneses spectrum associated with TUBA1A mutations. Journal of Medical Genetics *45*, 647–653. 10.1136/jmg.2008.058073.

39. Zhou, Y., Song, H., and Ming, G. (2023). Genetics of human brain development. Nat Rev Genet. 10.1038/s41576-023-00626-5.

40. Binquet, C., Lejeune, C., Faivre, L., Bouctot, M., Asensio, M.-L., Simon, A., Deleuze, J.-F., Boland, A., Guillemin, F., Seror, V., *et al.* (2022). Genome Sequencing for Genetics Diagnosis of Patients With Intellectual Disability: The DEFIDIAG Study. Front. Genet. *12*, 766964. 10.3389/fgene.2021.766964.

41. HOLLEY, R.W., APGAR, J., EVERETT, G.A., MADISON, J.T., MARQUISEE, M., MERRILL, S.H., PENSWICK, J.R., and ZAMIR, A. (1965). STRUCTURE OF A RIBONUCLEIC ACID. Science *147*, 1462–1465. 10.1126/science.147.3664.1462.

42. Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van Den Berghe, A., *et al.* (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature *260*, 500–507. 10.1038/260500a0.

43. Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. *74*, 5463–5467. 10.1073/pnas.74.12.5463.

44. Sboner, A., Mu, X., Greenbaum, D., Auerbach, R.K., and Gerstein, M.B. (2011). The real cost of sequencing: higher than you think! Genome Biol *12*, 125. 10.1186/gb-2011-12-8-125.

45. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., *et al.* (2022). The complete sequence of a human genome.

46. Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., *et al.* (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol *27*, 182–189. 10.1038/nbt.1523.

47. Srivastava, S., Love-Nichols, J.A., Dies, K.A., Ledbetter, D.H., Martin, C.L., Chung, W.K., Firth, H.V., Frazier, T., Hansen, R.L., Prock, L., *et al.* (2019). Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. Genetics in Medicine *21*, 2413–2421. 10.1038/s41436-019-0554-6.

48. Ko, M.H.-J., and Chen, H.-J. (2023). Genome-Wide Sequencing Modalities for Children with Unexplained Global Developmental Delay and Intellectual Disabilities—A Narrative Review. Children *10*, 501. 10.3390/children10030501.

49. Lelieveld, S.H., Spielmann, M., Mundlos, S., Veltman, J.A., and Gilissen, C. (2015). Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. Human Mutation *36*, 815–822. 10.1002/humu.22813.

50. Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q.B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.-L., and Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc. Natl. Acad. Sci. U.S.A. *112*, 5473–5478. 10.1073/pnas.1418631112.

51. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760. 10.1093/bioinformatics/btp324.

52. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359. 10.1038/nmeth.1923.

53. Wilton, R., and Szalay, A.S. (2020). Arioc: High-concurrency short-read alignment on multiple GPUs. PLoS Comput Biol *16*, e1008383. 10.1371/journal.pcbi.1008383.

54. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research *38*, e164–e164. 10.1093/nar/gkq603.

55. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol *17*, 122. 10.1186/s13059-016-0974-4.

56. Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., and Muller, J. (2018). AnnotSV: an integrated tool for structural variations annotation. Bioinformatics *34*, 3572–3574. 10.1093/bioinformatics/bty304.

57. The 1000 Genomes Project Consortium, Corresponding authors, Auton, A., Abecasis, G.R.,

Steering committee, Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., *et al.* (2015). A global reference for human genetic variation. Nature *526*, 68–74. 10.1038/nature15393.

58. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., *et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216–220. 10.1038/nature11690.

59. Exome Aggregation Consortium, Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291. 10.1038/nature19057.

60. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., *et al.* (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. 10.1038/s41586-020-2308-7.

61. Gudmundsson, S., Singer-Berk, M., Watts, N.A., Phu, W., Goodrich, J.K., Solomonson, M., Genome Aggregation Database Consortium, Rehm, H.L., MacArthur, D.G., and O'Donnell-Luria, A. (2022). Variant interpretation using population databases: Lessons from gnomAD. Human Mutation *43*, 1012–1030. 10.1002/humu.24309.

62. Ng, P.C. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research *31*, 3812–3814. 10.1093/nar/gkg509.

63. Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L.A., Edwards, K.J., Day, I.N.M., and Gaunt, T.R. (2013). Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Human Mutation *34*, 57–65. 10.1002/humu.22225.

64. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Methods *7*, 248–249. 10.1038/nmeth0410-248.

65. De Baets, G., Van Durme, J., Reumers, J., Maurer-Stroh, S., Vanhee, P., Dopazo, J., Schymkowitz, J., and Rousseau, F. (2012). SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. Nucleic Acids Research *40*, D935–D939. 10.1093/nar/gkr996.

66. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., *et al.* (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. The American Journal of Human Genetics *99*, 877–885. 10.1016/j.ajhg.2016.08.016.

67. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet *46*, 310–315. 10.1038/ng.2892.

68. Feng, B.-J. (2017). PERCH: A Unified Framework for Disease Gene Prioritization: HUMAN MUTATION. Human Mutation *38*, 243–251. 10.1002/humu.23158.

69. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N., and Karchin, R. (2013). Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics *14*, S3. 10.1186/1471-2164-14-S3-S3.

70. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., *et al.* (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell *176*, 535-548.e24. 10.1016/j.cell.2018.12.015.

71. Garcia, F.A.D.O., Andrade, E.S.D., and Palmero, E.I. (2022). Insights on variant analysis in silico tools for pathogenicity prediction. Front. Genet. *13*, 1010327. 10.3389/fgene.2022.1010327.

72. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., *et al.* (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun *10*, 1784. 10.1038/s41467-018-08148-z.

73. Logsdon, G.A., Vollger, M.R., and Eichler, E.E. (2020). Long read human genome sequencing and its applications. Nat Rev Genet *21*, 597–614. 10.1038/s41576-020-0236-x.

74. Wenger, A.M., Guturu, H., Bernstein, J.A., and Bejerano, G. (2017). Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genet Med *19*, 209–214. 10.1038/gim.2016.88.

75. Lin, B., Hui, J., and Mao, H. (2021). Nanopore Technology and Its Applications in Gene Sequencing. Biosensors *11*, 214. 10.3390/bios11070214.

76. Cuber, P., Chooneea, D., Geeves, C., Salatino, S., Creedy, T.J., Griffin, C., Sivess, L., Barnes, I., Price, B., and Misra, R. (2023). Comparing the accuracy and efficiency of third generation sequencing technologies, Oxford Nanopore Technologies, and Pacific Biosciences, for DNA barcode sequencing applications. Ecological Genetics and Genomics *28*, 100181. 10.1016/j.egg.2023.100181.

77. Levy-Sakin, M., and Ebenstein, Y. (2013). Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy. Current Opinion in Biotechnology *24*, 690–698. 10.1016/j.copbio.2013.01.009.

78. Yuan, Y., Chung, C.Y.-L., and Chan, T.-F. (2020). Advances in optical mapping for genomic research. Computational and Structural Biotechnology Journal *18*, 2051–2062. 10.1016/j.csbj.2020.07.018.

79. Jeffet, J., Margalit, S., Michaeli, Y., and Ebenstein, Y. (2021). Single-molecule optical genome mapping in nanochannels: multidisciplinarity at the nanoscale. Essays in Biochemistry *65*, 51–66. 10.1042/EBC20200021.

80. Lam, E.T., Hastie, A., Lin, C., Ehrlich, D., Das, S.K., Austin, M.D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M., *et al.* (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol *30*, 771–776. 10.1038/nbt.2303.

81. Savara, J., Novosád, T., Gajdoš, P., and Kriegová, E. (2021). Comparison of structural variants detected by optical mapping with long read next-generation sequencing. Bioinformatics *37*, 3398–3404. 10.1093/bioinformatics/btab359.

82. Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., *et al.* (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat Commun *8*, 15824. 10.1038/ncomms15824.

83. Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O'Grady, G.L., *et al.* (2017). Improving genetic

diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. *9*, eaal5209. 10.1126/scitranslmed.aal5209.

84. Kernohan, K.D., Frésard, L., Zappala, Z., Hartley, T., Smith, K.S., Wagner, J., Xu, H., McBride, A., Bourque, P.R., Consortium, C.C., *et al.* (2017). Whole-transcriptome sequencing in blood provides a diagnosis of spinal muscular atrophy with progressive myoclonic epilepsy: KERNOHAN et al. Human Mutation *38*, 611–614. 10.1002/humu.23211.

85. Hamanaka, K., Miyatake, S., Koshimizu, E., Tsurusaki, Y., Mitsuhashi, S., Iwama, K., Alkanaq, A.N., Fujita, A., Imagawa, E., Uchiyama, Y., *et al.* (2019). RNA sequencing solved the most common but unrecognized NEB pathogenic variant in Japanese nemaline myopathy. Genetics in Medicine *21*, 1629–1638. 10.1038/s41436-018-0360-6.

86. Frésard, L., Smail, C., Ferraro, N.M., Teran, N.A., Li, X., Smith, K.S., Bonner, D., Kernohan, K.D., Marwaha, S., Zappala, Z., *et al.* (2019). Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat Med *25*, 911–919. 10.1038/s41591-019-0457-8.

87. Wai, H., Douglas, A.G.L., and Baralle, D. (2019). RNA splicing analysis in genomic medicine. The International Journal of Biochemistry & Cell Biology *108*, 61–71. 10.1016/j.biocel.2018.12.009.

88. Karam, R., Conner, B., LaDuca, H., McGoldrick, K., Krempely, K., Richardson, M.E., Zimmermann, H., Gutierrez, S., Reineke, P., Hoang, L., *et al.* (2019). Assessment of Diagnostic Outcomes of RNA Genetic Testing for Hereditary Cancer. JAMA Netw Open *2*, e1913900. 10.1001/jamanetworkopen.2019.13900.

89. Murdock, D.R., Dai, H., Burrage, L.C., Rosenfeld, J.A., Ketkar, S., Müller, M.F., Yépez, V.A., Gagneur, J., Liu, P., Chen, S., *et al.* (2021). Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. Journal of Clinical Investigation *131*, e141500. 10.1172/JCI141500.

90. Lord, J., Gallone, G., Short, P.J., McRae, J.F., Ironfield, H., Wynn, E.H., Gerety, S.S., He, L., Kerr, B., Johnson, D.S., *et al.* (2019). Pathogenicity and selective constraint on variation near splice sites. Genome Res. *29*, 159–170. 10.1101/gr.238444.118.

91. Rivas, M.A., Pirinen, M., Conrad, D.F., Lek, M., Tsang, E.K., Karczewski, K.J., Maller, J.B., Kukurba, K.R., DeLuca, D.S., Fromer, M., *et al.* (2015). Effect of predicted protein-truncating genetic variants on the human transcriptome. Science *348*, 666–669. 10.1126/science.1261877.

92. Zhang, Q. (2020). Human genetics of life-threatening influenza pneumonitis. Hum Genet *139*, 941–948. 10.1007/s00439-019-02108-3.

93. Blakes, A.J.M., Wai, H.A., Davies, I., Moledina, H.E., Ruiz, A., Thomas, T., Bunyan, D., Thomas, N.S., Burren, C.P., Greenhalgh, L., *et al.* (2022). A systematic analysis of splicing variants identifies new diagnoses in the 100,000 Genomes Project. Genome Med *14*, 79. 10.1186/s13073-022-01087-x.

94. Grodecká, L., Buratti, E., and Freiberger, T. (2017). Mutations of Pre-mRNA Splicing Regulatory Elements: Are Predictions Moving Forward to Clinical Diagnostics? IJMS *18*, 1668. 10.3390/ijms18081668.

95. Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet *50*,

151–158. 10.1038/s41588-017-0004-9.

96. Mertes, C., Scheller, I.F., Yépez, V.A., Çelik, M.H., Liang, Y., Kremer, L.S., Gusic, M., Prokisch, H., and Gagneur, J. (2021). Detection of aberrant splicing events in RNA-seq data using FRASER. Nat Commun *12*, 529. 10.1038/s41467-020-20573-7.

97. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550. 10.1186/s13059-014-0550-8.

98. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140. 10.1093/bioinformatics/btp616.

99. Brechtmann, F., Mertes, C., Matusevičiūtė, A., Yépez, V.A., Avsec, Ž., Herzog, M., Bader, D.M., Prokisch, H., and Gagneur, J. (2018). OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. The American Journal of Human Genetics *103*, 907–917. 10.1016/j.ajhg.2018.10.025.

100. Mohammadi, P., Castel, S.E., Cummings, B.B., Einson, J., Sousa, C., Hoffman, P., Donkervoort, S., Jiang, Z., Mohassel, P., Foley, A.R., *et al.* (2019). Genetic regulatory variation in populations informs transcriptome analysis in rare disease. Science *366*, 351–356. 10.1126/science.aay0256.

101. Albers, C.A., Paul, D.S., Schulze, H., Freson, K., Stephens, J.C., Smethurst, P.A., Jolley, J.D., Cvejic, A., Kostadima, M., Bertone, P., *et al.* (2012). Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. Nat Genet *44*, 435–439. 10.1038/ng.1083.

102. Gonorazky, H.D., Naumenko, S., Ramani, A.K., Nelakuditi, V., Mashouri, P., Wang, P., Kao, D., Ohri, K., Viththiyapaskaran, S., Tarnopolsky, M.A., *et al.* (2019). Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. The American Journal of Human Genetics *104*, 466–483. 10.1016/j.ajhg.2019.01.012.

103. Yépez, V.A., Mertes, C., Müller, M.F., Klaproth-Andrade, D., Wachutka, L., Frésard, L., Gusic, M., Scheller, I.F., Goldberg, P.F., Prokisch, H., *et al.* (2021). Detection of aberrant gene expression events in RNA sequencing data. Nat Protoc *16*, 1276–1296. 10.1038/s41596-020-00462-5.

104. the 1000 Genomes Project (2011). Variation in genome-wide mutation rates within and between human families. Nat Genet *43*, 712–714. 10.1038/ng.862.

105. Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat Rev Genet *12*, 363–376. 10.1038/nrg2958.

106. Belyeu, J.R., Brand, H., Wang, H., Zhao, X., Pedersen, B.S., Feusier, J., Gupta, M., Nicholas, T.J., Brown, J., Baird, L., *et al.* (2021). De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. The American Journal of Human Genetics *108*, 597–607. 10.1016/j.ajhg.2021.02.012.

107. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., *et al.* (2019). An open resource of structural variation for medical and population genetics (Genomics) 10.1101/578674.

108. Lupski, J.R., De Oca-Luna, R.M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B.J., Saucedo-Cardenas, O., Barker, D.F., Killian, J.M., Garcia, C.A., *et al.* (1991). DNA duplication

associated with Charcot-Marie-Tooth disease type 1A. Cell *66*, 219–232. 10.1016/0092-8674(91)90613-4.

109. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., *et al.* (2006). Global variation in copy number in the human genome. Nature *444*, 444–454. 10.1038/nature05329.

110. Warburton, D. De Novo Balanced Chromosome Rearrangements and Extra Marker Chromosomes Identified at Prenatal Diagnosis: Clinical Significance and Distribution of Breakpoints.

111. Schluth-Bolard, C., Labalme, A., Cordier, M.-P., Till, M., Nadeau, G., Tevissen, H., Lesca, G., Boutry-Kryza, N., Rossignol, S., Rocas, D., *et al.* (2013). Breakpoint mapping by next generation sequencing reveals causative gene disruption in patients carrying apparently balanced chromosome rearrangements with intellectual deficiency and/or congenital malformations. J Med Genet *50*, 144–150. 10.1136/jmedgenet-2012-101351.

112. Schluth-Bolard, C., Delobel, B., Sanlaville, D., Boute, O., Cuisset, J.-M., Sukno, S., Labalme, A., Duban-Bedu, B., Plessis, G., Jaillard, S., *et al.* (2009). Cryptic genomic imbalances in de novo and inherited apparently balanced chromosomal rearrangements: Array CGH study of 47 unrelated cases. European Journal of Medical Genetics *52*, 291–296. 10.1016/j.ejmg.2009.05.011.

113. Gribble, S.M. (2005). The complex nature of constitutional de novo apparently balanced translocations in patients presenting with abnormal phenotypes. Journal of Medical Genetics *42*, 8–16. 10.1136/jmg.2004.024141.

114. Gijsbers, A.C.J., Bosch, C.A.J., Dauwerse, J.G., Giromus, O., Hansson, K., Hilhorst-Hofstee, Y., Kriek, M., Van Haeringen, A., Bijlsma, E.K., Bakker, E., *et al.* (2010). Additional cryptic CNVs in mentally retarded patients with apparently balanced karyotypes. European Journal of Medical Genetics *53*, 227–233. 10.1016/j.ejmg.2010.06.003.

115. Hanlon, V.C.T., Lansdorp, P.M., and Guryev, V. (2022). A survey of current methods to detect and genotype inversions. Human Mutation *43*, 1576–1589. 10.1002/humu.24458.

116. Osborne, L.R., Li, M., Pober, B., Chitayat, D., Bodurtha, J., Mandel, A., Costa, T., Grebe, T., Cox, S., Tsui, L.-C., *et al.* (2001). A 1.5 million–base pair inversion polymorphism in families with Williams-Beuren syndrome. Nat Genet *29*, 321–325. 10.1038/ng753.

117. Bayés, M., Magano, L.F., Rivera, N., Flores, R., and A. Pérez Jurado, L. (2003). Mutational Mechanisms of Williams-Beuren Syndrome Deletions. The American Journal of Human Genetics *73*, 131–151. 10.1086/376565.

118. Hobart, H.H., Morris, C.A., Mervis, C.B., Pani, A.M., Kistler, D.J., Rios, C.M., Kimberley, K.W., Gregg, R.G., and Bray-Ward, P. (2010). Inversion of the Williams syndrome region is a common polymorphism found more frequently in parents of children with Williams syndrome. Am. J. Med. Genet. *154C*, 220–228. 10.1002/ajmg.c.30258.

119. Koolen, D.A., Vissers, L.E.L.M., Pfundt, R., De Leeuw, N., Knight, S.J., Regan, R., Kooy, R.F., Reyniers, E., Romano, C., Fichera, M., *et al.* (2006). A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. Nat Genet *38*, 999–1001. 10.1038/ng1853.

120. Visser, R., Shimokawa, O., Harada, N., Kinoshita, A., Ohta, T., Niikawa, N., and Matsumoto, N. (2005). Identification of a 3.0-kb Major Recombination Hotspot in Patients with Sotos

Syndrome Who Carry a Common 1.9-Mb Microdeletion. The American Journal of Human Genetics *76*, 52–67. 10.1086/426950.

121. Gimelli, G. (2003). Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. Human Molecular Genetics *12*, 849–858. 10.1093/hmg/ddg101.

122. Sharp, A.J., Mefford, H.C., Li, K., Baker, C., Skinner, C., Stevenson, R.E., Schroer, R.J., Novara, F., De Gregori, M., Ciccone, R., *et al.* (2008). A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat Genet *40*, 322–328. 10.1038/ng.93.

123. Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V.G., *et al.* (2005). A common inversion under selection in Europeans. Nat Genet *37*, 129–137. 10.1038/ng1508.

124. Quinlan, A.R., and Hall, I.M. (2012). Characterizing complex structural variation in germline and somatic genomes. Trends in Genetics *28*, 43–53. 10.1016/j.tig.2011.10.002.

125. Brandler, W.M., Antaki, D., Gujral, M., Noor, A., Rosanio, G., Chapman, T.R., Barrera, D.J., Lin, G.N., Malhotra, D., Watts, A.C., *et al.* (2016). Frequency and Complexity of De Novo Structural Mutation in Autism. The American Journal of Human Genetics *98*, 667–679. 10.1016/j.ajhg.2016.02.018.

126. Kidd, J.M., Graves, T., Newman, T.L., Fulton, R., Hayden, H.S., Malig, M., Kallicki, J., Kaul, R., Wilson, R.K., and Eichler, E.E. (2010). A Human Genome Structural Variation Sequencing Resource Reveals Insights into Mutational Mechanisms. Cell *143*, 837–847. 10.1016/j.cell.2010.10.027.

127. Shaw, C.J. (2004). Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Human Molecular Genetics *13*, 57R – 64. 10.1093/hmg/ddh073.

128. Waye, J.S., and Willard, H.F. (1989). Human j3 satellite DNA: Genomic organization and sequence definition of a class of highly repetitive tandem DNA. Proc. Natl. Acad. Sci. USA.

129. Richard, G.-F., Kerrest, A., and Dujon, B. (2008). Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. Microbiol Mol Biol Rev *72*, 686–727. 10.1128/MMBR.00011-08.

130. International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research:, Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. 10.1038/35057062.

131. Sun, J.X., Helgason, A., Masson, G., Ebenesersdóttir, S.S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., *et al.* (2012). A direct characterization of human mutation based on microsatellites. Nat Genet *44*, 1161–1165. 10.1038/ng.2398.

132. Chintalaphani, S.R., Pineda, S.S., Deveson, I.W., and Kumar, K.R. (2021). An update on the neurological short tandem repeat expansion disorders and the emergence of long read sequencing diagnostics. acta neuropathol commun *9*, 98. 10.1186/s40478-021-01201-x.

133. Dolzhenko, E., van Vugt, J.J.F.A., Shaw, R.J., Bekritsky, M.A., van Blitterswijk, M., Narzisi, G., Ajay, S.S., Rajan, V., Lajoie, B.R., Johnson, N.H., *et al.* (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. *27*, 1895–1903.

10.1101/gr.225672.117.

134. Tankard, R.M., Bennett, M.F., Degorski, P., Delatycki, M.B., Lockhart, P.J., and Bahlo, M. (2017). Detecting tandem repeat expansions in cohorts sequenced with short-read sequencing data (Bioinformatics) 10.1101/157792.

135. Dashnow, H., Lek, M., Phipson, B., Halman, A., Lonsdale, A., Davis, M., Lamont, P., Clayton, J.S., Laing, N.G., MacArthur, D.G., *et al.* STRetch: detecting and discovering pathogenic short tandem repeat expansions.

136. Tang, H., Kirkness, E.F., Lippert, C., Biggs, W.H., Fabani, M., Guzman, E., Ramakrishnan, S., Lavrenko, V., Kakaradov, B., Hou, C., *et al.* (2017). Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. The American Journal of Human Genetics *101*, 700–715. 10.1016/j.ajhg.2017.09.013.

137. Dolzhenko, E., Bennett, M.F., Richmond, P.A., Trost, B., Chen, S., Van Vugt, J.J.F.A., Nguyen, C., Narzisi, G., Gainullin, V.G., Gross, A.M., *et al.* (2020). ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. Genome Biol *21*, 102. 10.1186/s13059-020-02017-z.

138. McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. U.S.A. *36*, 344–355. 10.1073/pnas.36.6.344.

139. Finnegan, D.J. (1989). Eukaryotic transposable elements and genome evolution. Trends Genet *5*, 103–107. 10.1016/0168-9525(89)90039-5.

140. Pace, J.K., and Feschotte, C. (2007). The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. Genome Res. *17*, 422–432. 10.1101/gr.5826307.

141. Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat Rev Genet *10*, 691–703. 10.1038/nrg2640.

142. Moyes, D., Griffiths, D.J., and Venables, P.J. (2007). Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. Trends in Genetics *23*, 326–333. 10.1016/j.tig.2007.05.004.

143. Kazazian, H.H., and Moran, J.V. (2017). Mobile DNA in Health and Disease. N Engl J Med *377*, 361–370. 10.1056/NEJMra1510092.

144. Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V., and Kazazian, H.H. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. U.S.A. *100*, 5280–5285. 10.1073/pnas.0831042100.

145. Morales, M.E., White, T.B., Streva, V.A., DeFreece, C.B., Hedges, D.J., and Deininger, P.L. (2015). The Contribution of Alu Elements to Mutagenic DNA Double-Strand Break Repair. PLoS Genet *11*, e1005016. 10.1371/journal.pgen.1005016.

146. Carroll, M.L., Roy-Engel, A.M., Nguyen, S.V., Salem, A.-H., Vogel, E., Vincent, B., Myers, J., Ahmad, Z., Nguyen, L., Sammarco, M., *et al.* (2001). Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. Journal of Molecular Biology *311*, 17–40. 10.1006/jmbi.2001.4847.

147. Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. Nat Genet *35*, 41–48. 10.1038/ng1223.

148. Bennett, E.A., Keller, H., Mills, R.E., Schmidt, S., Moran, J.V., Weichenrieder, O., and Devine, S.E. (2008). Active *Alu* retrotransposons in the human genome. Genome Res. *18*, 1875–1883. 10.1101/gr.081737.108.

149. Beck, C.R., Garcia-Perez, J.L., Badge, R.M., and Moran, J.V. (2011). LINE-1 Elements in Structural Variation and Disease. Annu. Rev. Genom. Hum. Genet. *12*, 187–215. 10.1146/annurev-genom-082509-141802.

150. Wang, H., Xing, J., Grover, D., Hedges, D.J., Han, K., Walker, J.A., and Batzer, M.A. (2005). SVA Elements: A Hominid-specific Retroposon Family. Journal of Molecular Biology *354*, 994–1007. 10.1016/j.jmb.2005.09.085.

151. Kazazian, H.H., Wong, C., Youssoufian, H., Scottt, A.F., and Phillips, D.G. (1988). Haemophilia A resulting from de novo insertion of Ll sequences represents a novel mechanism for mutation in man.

152. Feusier, J., Watkins, W.S., Thomas, J., Farrell, A., Witherspoon, D.J., Baird, L., Ha, H., Xing, J., and Jorde, L.B. (2019). Pedigree-based estimation of human mobile element retrotransposition rates. Genome Res. *29*, 1567–1577. 10.1101/gr.247965.118.

153. Yang, F., and Wang, P.J. (2016). Multiple LINEs of retrotransposon silencing mechanisms in the mammalian germline. Seminars in Cell & Developmental Biology *59*, 118–125. 10.1016/j.semcdb.2016.03.001.

154. Hancks, D.C., and Kazazian, H.H. (2016). Roles for retrotransposon insertions in human disease. Mobile DNA *7*, 9. 10.1186/s13100-016-0065-9.

155. Wallace, M.R., Andersen, L.B., Saulino, A.M., Gregoryt, P.E., Glovert, T.W., and Collinst, F.S. (1991). A de nova Alu insertion results in neurofibromatosis type. *353*.

156. Alesi, V., Genovese, S., Lepri, F.R., Catino, G., Loddo, S., Orlando, V., Di Tommaso, S., Morgia, A., Martucci, L., Di Donato, M., *et al.* (2023). Deep Intronic LINE-1 Insertions in NF1: Expanding the Spectrum of Neurofibromatosis Type 1-Associated Rearrangements. Biomolecules *13*, 725. 10.3390/biom13050725.

157. Wimmer, K., Callens, T., Wernstedt, A., and Messiaen, L. (2011). The NF1 Gene Contains Hotspots for L1 Endonuclease-Dependent De Novo Insertion. PLoS Genet *7*, e1002371. 10.1371/journal.pgen.1002371.

158. Gu, S., Yuan, B., Campbell, I.M., Beck, C.R., Carvalho, C.M.B., Nagamani, S.C.S., Erez, A., Patel, A., Bacino, C.A., Shaw, C.A., *et al.* (2015). Alu-mediated diverse and complex pathogenic copy-number variants within human chromosome 17 at p13.3. Human Molecular Genetics *24*, 4061–4077. 10.1093/hmg/ddv146.

159. Song, X., Beck, C.R., Du, R., Campbell, I.M., Coban-Akdemir, Z., Gu, S., Breman, A.M., Stankiewicz, P., Ira, G., Shaw, C.A., *et al.* (2018). Predicting human genes susceptible to genomic instability associated with *Alu* / *Alu* -mediated rearrangements. Genome Res. *28*, 1228–1242. 10.1101/gr.229401.117.

160. Startek, M., Szafranski, P., Gambin, T., Campbell, I.M., Hixson, P., Shaw, C.A., Stankiewicz, P., and Gambin, A. (2015). Genome-wide analyses of LINE–LINE-mediated nonallelic homologous recombination. Nucleic Acids Research *43*, 2188–2198. 10.1093/nar/gku1394.

161. Legoix, P., Sarkissian, H.D., Cazes, L., Giraud, S., Sor, F., Rouleau, G.A., Lenoir, G., Thomas, G., and Zucman-Rossi, J. (2000). Molecular Characterization of Germline NF2 Gene Rearrangements. Genomics *65*, 62–66. 10.1006/geno.2000.6139.

162. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J., and Eichler, E.E. (2001). Segmental Duplications: Organization and Impact Within the Current Human Genome Project Assembly. Genome Res. *11*, 1005–1017. 10.1101/gr.187101.

163. Cheung, J., Estivill, X., Khaja, R., MacDonald, J.R., Lau, K., Tsui, L.-C., and Scherer, S.W. (2003). Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol *4*, R25. 10.1186/gb-2003-4-4-r25.

164. Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. (2002). Recent Segmental Duplications in the Human Genome. Science *297*, 1003–1007. 10.1126/science.1072047.

165. Zhang, L. (2004). Patterns of Segmental Duplication in the Human Genome. Molecular Biology and Evolution *22*, 135–141. 10.1093/molbev/msh262.

166. Vollger, M.R., Guitart, X., Dishuck, P.C., Mercuri, L., Harvey, W.T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K.M., Lewis, A.P., *et al.* (2022). Segmental duplications and their variation in a complete human genome. Science *376*, eabj6965. 10.1126/science.abj6965.

167. Chen, K.-S., Manian, P., Koeuth, T., Potocki, L., Zhao, Q., Chinault, A.C., Lee, C.C., and Lupski, J.R. (1997). Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. Nat Genet *17*, 154–163. 10.1038/ng1097-154.

168. Amos-Landgraf, J.M., Ji, Y., Gottlieb, W., Depinet, T., Wandstrat, A.E., Cassidy, S.B., Driscoll, D.J., Rogan, P.K., Schwartz, S., and Nicholls, R.D. (1999). Chromosome Breakage in the Prader-Willi and Angelman Syndromes Involves Recombination between Large, Transcribed Repeats at Proximal and Distal Breakpoints. The American Journal of Human Genetics *65*, 370–386. 10.1086/302510.

169. Christian, S. (1999). Large genomic duplicons map to sites of instability in the Prader-Willi/Angelman syndrome chromosome region (15q11-q13). Human Molecular Genetics *8*, 1025–1037. 10.1093/hmg/8.6.1025.

170. Perez Jurado, L. (1998). A duplicated gene in the breakpoint regions of the 7q11.23 Williams-Beuren syndrome deletion encodes the initiator binding protein TFII-I and BAP-135, a phosphorylation target of BTK. Human Molecular Genetics *7*, 325–334. 10.1093/hmg/7.3.325.

171. Peoples, R., Franke, Y., Wang, Y.-K., Pérez-Jurado, L., Paperna, T., Cisco, M., and Francke, U. (2000). A Physical Map, Including a BAC/PAC Clone Contig, of the Williams-Beuren Syndrome–Deletion Region at 7q11.23. The American Journal of Human Genetics *66*, 47–68. 10.1086/302722.

172. Edelmann, L. (1999). A common molecular basis for rearrangement disorders on chromosome 22q11. Human Molecular Genetics *8*, 1157–1167. 10.1093/hmg/8.7.1157.

173. Edelmann, L., Pandita, R.K., and Morrow, B.E. (1999). Low-Copy Repeats Mediate the Common 3-Mb Deletion in Patients with Velo-cardio-facial Syndrome. The American Journal of Human Genetics *64*, 1076–1086. 10.1086/302343.

174. Shaikh, T.H. (2000). Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. Human Molecular Genetics *9*, 489–501. 10.1093/hmg/9.4.489.

175. Lieber, M.R. (2010). The Mechanism of Double-Strand DNA Break Repair by the

Nonhomologous DNA End-Joining Pathway. Annu. Rev. Biochem. *79*, 181–211. 10.1146/annurev.biochem.052308.093131.

176. Porubsky, D., Höps, W., Ashraf, H., Hsieh, P., Rodriguez-Martin, B., Yilmaz, F., Ebler, J., Hallast, P., Maria Maggiolini, F.A., Harvey, W.T., *et al.* (2022). Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. Cell *185*, 1986-2005.e26. 10.1016/j.cell.2022.04.017.

177. Chiang, C., Jacobsen, J.C., Ernst, C., Hanscom, C., Heilbut, A., Blumenthal, I., Mills, R.E., Kirby, A., Lindgren, A.M., Rudiger, S.R., *et al.* (2012). Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. Nat Genet *44*, 390–397. 10.1038/ng.2202.

178. Carvalho, C.M.B., and Lupski, J.R. (2016). Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet *17*, 224–238. 10.1038/nrg.2015.25.

179. Burssed, B., Zamariolli, M., Bellucco, F.T., and Melaragno, M.I. (2022). Mechanisms of structural chromosomal rearrangement formation. Mol Cytogenet *15*, 23. 10.1186/s13039-022-00600-6.

180. Bétermier, M., Bertrand, P., and Lopez, B.S. (2014). Is Non-Homologous End-Joining Really an Inherently Error-Prone Process? PLoS Genet *10*, e1004086. 10.1371/journal.pgen.1004086.

181. Seol, J.-H., Shim, E.Y., and Lee, S.E. (2018). Microhomology-mediated end joining: Good, bad and ugly. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis *809*, 81–87. 10.1016/j.mrfmmm.2017.07.002.

182. Gu, W., Zhang, F., and Lupski, J.R. (2008). Mechanisms for human genomic rearrangements. Pathogenetics *1*, 4. 10.1186/1755-8417-1-4.

183. Stankiewicz, P., and Lupski, J.R. (2002). Genome architecture, rearrangements and genomic disorders. Trends in Genetics *18*, 74–82. 10.1016/S0168-9525(02)02592-1.

184. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222. 10.1093/bioinformatics/btv710.

185. Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339. 10.1093/bioinformatics/bts378.

186. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol *15*, R84. 10.1186/gb-2014-15-6-r84.

187. Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. *21*, 974–984. 10.1101/gr.114876.110.

188. Roller, E., Ivakhno, S., Lee, S., Royce, T., and Tanner, S. (2016). Canvas: versatile and scalable detection of copy number variants. Bioinformatics *32*, 2375–2377. 10.1093/bioinformatics/btw163.

189. Thung, D.T., de Ligt, J., Vissers, L.E., Steehouwer, M., Kroon, M., de Vries, P., Slagboom,

E.P., Ye, K., Veltman, J.A., and Hehir-Kwa, J.Y. (2014). Mobster: accurate detection of mobile element insertions in next generation sequencing data.

190. Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., Mills, R.E., The 1000 Genomes Project Consortium, and Devine, S.E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. *27*, 1916–1929. 10.1101/gr.218032.116.

191. Ohno, S. (1972). So much "junk" DNA in our genome. Brookhaven Symp Biol *23*, 366–370.

192. The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. 10.1038/nature11247.

193. Proudfoot, N. (1980). Pseudogenes. Nature *286*, 840–841. 10.1038/286840a0.

194. Khachane, A.N., and Harrison, P.M. (2009). Assessing the genomic evidence for conserved transcribed pseudogenes under selection. BMC Genomics *10*, 435. 10.1186/1471-2164-10-435.

195. Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., *et al.* (2021). GENCODE 2021. Nucleic Acids Research *49*, D916–D923. 10.1093/nar/gkaa1087.

196. Matera, A.G., Terns, R.M., and Terns, M.P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat Rev Mol Cell Biol *8*, 209–220. 10.1038/nrm2124.

197. Turunen, J.J., Niemelä, E.H., Verma, B., and Frilander, M.J. (2013). The significant other: splicing by the minor spliceosome: Splicing by the minor spliceosome. WIREs RNA *4*, 61–76. 10.1002/wrna.1141.

198. Elsaid, M.F., Chalhoub, N., Ben-Omran, T., Kumar, P., Kamel, H., Ibrahim, K., Mohamoud, Y., Al-Dous, E., Al-Azwani, I., Malek, J.A., *et al.* (2017). Mutation in noncoding RNA RNU12 causes early onset cerebellar ataxia: *RNU12* in Cerebellar Ataxia. Ann Neurol. *81*, 68–78. 10.1002/ana.24826.

199. Farach, L.S., Little, M.E., Duker, A.L., Logan, C.V., Jackson, A., Hecht, J.T., and Bober, M. (2018). The expanding phenotype of *RNU4ATAC* pathogenic variants to Lowry Wood syndrome. Am J Med Genet *176*, 465–469. 10.1002/ajmg.a.38581.

200. Merico, D., Roifman, M., Braunschweig, U., Yuen, R.K.C., Alexandrova, R., Bates, A., Reid, B., Nalpathamkalam, T., Wang, Z., Thiruvahindrapuram, B., *et al.* (2015). Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing. Nat Commun *6*, 8718. 10.1038/ncomms9718.

201. Edery, P., Marcaillou, C., Sahbatou, M., Labalme, A., Chastang, J., Touraine, R., Tubacher, E., Senni, F., Bober, M.B., Nampoothiri, S., *et al.* (2011). Association of TALS Developmental Disorder with Defect in Minor Splicing Component *U4atac* snRNA. Science *332*, 240–243. 10.1126/science.1202205.

202. Huang, Z., Du, Y., Wen, J., Lu, B., and Zhao, Y. (2022). snoRNAs: functions and mechanisms in biological processes, and roles in tumor pathophysiology. Cell Death Discov. *8*, 259. 10.1038/s41420-022-01056-8.

203. Jenkinson, E.M., Rodero, M.P., Kasher, P.R., Uggenti, C., Oojageer, A., Goosey, L.C., Rose, Y., Kershaw, C.J., Urquhart, J.E., Williams, S.G., *et al.* (2016). Mutations in SNORD118 cause the cerebral microangiopathy leukoencephalopathy with calcifications and cysts. Nat Genet

48, 1185–1192. 10.1038/ng.3661.

204. Duker, A.L., Ballif, B.C., Bawle, E.V., Person, R.E., Mahadevan, S., Alliman, S., Thompson, R., Traylor, R., Bejjani, B.A., Shaffer, L.G., *et al.* (2010). Paternally inherited microdeletion at 15q11.2 confirms a significant role for the SNORD116 C/D box snoRNA cluster in Prader–Willi syndrome. Eur J Hum Genet *18*, 1196–1201. 10.1038/ejhg.2010.102.

205. Honda, S., Kawamura, T., Loher, P., Morichika, K., Rigoutsos, I., and Kirino, Y. (2017). The biogenesis pathway of tRNA-derived piRNAs in Bombyx germ cells. Nucleic Acids Research *45*, 9108–9120. 10.1093/nar/gkx537.

206. Ozata, D.M., Gainetdinov, I., Zoch, A., O'Carroll, D., and Zamore, P.D. (2019). PIWI-interacting RNAs: small RNAs with big functions. Nat Rev Genet *20*, 89–108. 10.1038/s41576-018-0073-3.

207. Kim, Y.-K., and Kim, V.N. (2007). Processing of intronic microRNAs. EMBO J *26*, 775–783. 10.1038/sj.emboj.7601512.

208. Ha, M., and Kim, V.N. (2014). Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol *15*, 509–524. 10.1038/nrm3838.

209. Denli, A.M., Tops, B.B.J., Plasterk, R.H.A., Ketting, R.F., and Hannon, G.J. (2004). Processing of primary microRNAs by the Microprocessor complex. Nature *432*, 231–235. 10.1038/nature03049.

210. Agarwal, V., Bell, G.W., Nam, J.-W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. eLife *4*, e05005. 10.7554/eLife.05005.

211. Jo, M.H., Shin, S., Jung, S.-R., Kim, E., Song, J.-J., and Hohng, S. (2015). Human Argonaute 2 Has Diverse Reaction Pathways on Target RNAs. Molecular Cell *59*, 117–124. 10.1016/j.molcel.2015.04.027.

212. Vasudevan, S., Tong, Y., and Steitz, J.A. (2007). Switching from Repression to Activation: MicroRNAs Can Up-Regulate Translation. Science *318*, 1931–1934. 10.1126/science.1149460.

213. Thomson, D.W., and Dinger, M.E. (2016). Endogenous microRNA sponges: evidence and controversy. Nat Rev Genet *17*, 272–283. 10.1038/nrg.2016.20.

214. Mencía, Á., Modamio-Høybjør, S., Redshaw, N., Morín, M., Mayo-Merino, F., Olavarrieta, L., Aguirre, L.A., Del Castillo, I., Steel, K.P., Dalmay, T., *et al.* (2009). Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. Nat Genet *41*, 609–613. 10.1038/ng.355.

215. Soldà, G., Robusto, M., Primignani, P., Castorina, P., Benzoni, E., Cesarani, A., Ambrosetti, U., Asselta, R., and Duga, S. (2012). A novel mutation within the MIR96 gene causes non-syndromic inherited hearing loss in an Italian family by altering pre-miRNA processing. Human Molecular Genetics *21*, 577–585. 10.1093/hmg/ddr493.

216. Hughes, A.E., Bradley, D.T., Campbell, M., Lechner, J., Dash, D.P., Simpson, D.A., and Willoughby, C.E. (2011). Mutation Altering the miR-184 Seed Region Causes Familial Keratoconus with Cataract. The American Journal of Human Genetics *89*, 628–633. 10.1016/j.ajhg.2011.09.014.

217. Iliff, B.W., Riazuddin, S.A., and Gottsch, J.D. (2012). A Single-Base Substitution in the Seed Region of miR-184 Causes EDICT Syndrome. Invest. Ophthalmol. Vis. Sci. *53*, 348.

10.1167/iovs.11-8783.

218. Bykhovskaya, Y., Seldin, M.F., Liu, Y., Ransom, M., Li, X., and Rabinowitz, Y.S. (2015). Independent Origin of c.57 C > T Mutation in MIR184 Associated with Inherited Corneal and Lens Abnormalities. Ophthalmic Genetics *36*, 95–97. 10.3109/13816810.2014.977491.

219. Conte, I., Hadfield, K.D., Barbato, S., Carrella, S., Pizzo, M., Bhat, R.S., Carissimo, A., Karali, M., Porter, L.F., Urquhart, J., *et al.* (2015). MiR-204 is responsible for inherited retinal dystrophy associated with ocular coloboma. Proc. Natl. Acad. Sci. U.S.A. *112*. 10.1073/pnas.1401464112.

220. Grigelioniene, G., Suzuki, H.I., Taylan, F., Mirzamohammadi, F., Borochowitz, Z.U., Ayturk, U.M., Tzur, S., Horemuzova, E., Lindstrand, A., Weis, M.A., *et al.* (2019). Gain-of-function mutation of microRNA-140 in human skeletal dysplasia. Nat Med *25*, 583–590. 10.1038/s41591-019-0353-2.

221. Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., *et al.* (2015). The landscape of long noncoding RNAs in the human transcriptome. Nat Genet *47*, 199–208. 10.1038/ng.3192.

222. Chen, J., Shishkin, A.A., Zhu, X., Kadri, S., Maza, I., Guttman, M., Hanna, J.H., Regev, A., and Garber, M. (2016). Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome Biol *17*, 19. 10.1186/s13059-016-0880-9.

223. Quinn, J.J., Zhang, Q.C., Georgiev, P., Ilik, I.A., Akhtar, A., and Chang, H.Y. (2016). Rapid evolutionary turnover underlies conserved lncRNA–genome interactions. Genes Dev. *30*, 191–207. 10.1101/gad.272187.115.

224. Hennessy, E.J., Van Solingen, C., Scacalossi, K.R., Ouimet, M., Afonso, M.S., Prins, J., Koelwyn, G.J., Sharma, M., Ramkhelawon, B., Carpenter, S., *et al.* (2018). The long noncoding RNA CHROME regulates cholesterol homeostasis in primates. Nat Metab *1*, 98–110. 10.1038/s42255-018-0004-9.

225. Zhang, X., Xue, C., Lin, J., Ferguson, J.F., Weiner, A., Liu, W., Han, Y., Hinkle, C., Li, W., Jiang, H., *et al.* (2018). Interrogation of nonconserved human adipose lincRNAs identifies a regulatory role of *linc-ADAL* in adipocyte metabolism. Sci. Transl. Med. *10*, eaar5987. 10.1126/scitranslmed.aar5987.

226. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., *et al.* (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789. 10.1101/gr.132159.111.

227. Kopp, F., and Mendell, J.T. (2018). Functional Classification and Experimental Dissection of Long Noncoding RNAs. Cell *172*, 393–407. 10.1016/j.cell.2018.01.011.

228. Zhao, Z., Sentürk, N., Song, C., and Grummt, I. (2018). lncRNA PAPAS tethered to the rDNA enhancer recruits hypophosphorylated CHD4/NuRD to repress rRNA synthesis at elevated temperatures. Genes Dev. *32*, 836–848. 10.1101/gad.311688.118.

229. Uroda, T., Anastasakou, E., Rossi, A., Teulon, J.-M., Pellequer, J.-L., Annibale, P., Pessey, O., Inga, A., Chillón, I., and Marcia, M. (2019). Conserved Pseudoknots in lncRNA MEG3 Are Essential for Stimulation of the p53 Pathway. Molecular Cell *75*, 982-995.e9. 10.1016/j.molcel.2019.07.025.

230. Gong, C., Li, Z., Ramanujan, K., Clay, I., Zhang, Y., Lemire-Brachat, S., and Glass, D.J.

(2015). A Long Non-coding RNA, LncMyoD, Regulates Skeletal Muscle Differentiation by Blocking IMP2-Mediated mRNA Translation. Developmental Cell *34*, 181–191. 10.1016/j.devcel.2015.05.009.

231. Wang, L., Yu, X., Zhang, Z., Pang, L., Xu, J., Jiang, J., Liang, W., Chai, Y., Hou, J., and Li, F. (2017). Linc-ROR promotes esophageal squamous cell carcinoma progression through the derepression of SOX9. J Exp Clin Cancer Res *36*, 182. 10.1186/s13046-017-0658-2.

232. Tsagakis, I., Douka, K., Birds, I., and Aspden, J.L. (2020). Long non-coding RNAs in development and disease: conservation to mechanisms. J. Pathol. *250*, 480–495. 10.1002/path.5405.

233. Brown, J., Hendrich, B.D., and Rupert, J.L. The Human X/ST Gene: Analysis of a 17 kb Inactive X-Specific RNA That Contains Conserved Repeats and Is Highly Localized within the Nucleus.

234. Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., and Brockdorfft, B.N. Requirement for Xist in X chromosome inactivation.

235. Marahrens, Y., Panning, B., Dausman, J., Strauss, W., and Jaenisch, R. (1997). Xist-deficient mice are defective in dosage compensation but not spermatogenesis. Genes Dev. *11*, 156–166. 10.1101/gad.11.2.156.

236. Almeida, M., Pintacuda, G., Masui, O., Koseki, Y., Gdula, M., Cerase, A., Brown, D., Mould, A., Innocent, C., Nakayama, M., *et al.* (2017). PCGF3/5–PRC1 initiates Polycomb recruitment in X chromosome inactivation. Science *356*, 1081–1084. 10.1126/science.aal2512.

237. Van deVondervoort, I.I.G.M., Gordebeke, P.M., Khoshab, N., Tiesinga, P.H.E., Buitelaar, J.K., Kozicz, T., Aschrafi, A., and Glennon, J.C. (2013). Long non-coding RNAs in neurodevelopmental disorders. Front. Mol. Neurosci. *6*. 10.3389/fnmol.2013.00053.

238. Aliperti, V., Skonieczna, J., and Cerase, A. (2021). Long Non-Coding RNA (lncRNA) Roles in Cell Biology, Neurodevelopment and Neurological Disorders. ncRNA *7*, 36. 10.3390/ncrna7020036.

239. Liaci, C., Prandi, L., Pavinato, L., Brusco, A., Maldotti, M., Molineris, I., Oliviero, S., and Merlo, G.R. (2022). The Emerging Roles of Long Non-Coding RNAs in Intellectual Disability and Related Neurodevelopmental Disorders. IJMS *23*, 6118. 10.3390/ijms23116118.

240. Ang, C.E., Ma, Q., Wapinski, O.L., Fan, S., Flynn, R.A., Lee, Q.Y., Coe, B., Onoguchi, M., Olmos, V.H., Do, B.T., *et al.* (2019). The novel lncRNA lnc-NR2F1 is pro-neurogenic and mutated in human neurodevelopmental disorders. eLife *8*, e41770. 10.7554/eLife.41770.

241. Talkowski, M.E., Maussion, G., Crapper, L., Rosenfeld, J.A., Blumenthal, I., Hanscom, C., Chiang, C., Lindgren, A., Pereira, S., Ruderfer, D., *et al.* (2012). Disruption of a Large Intergenic Noncoding RNA in Subjects with Neurodevelopmental Disabilities. The American Journal of Human Genetics *91*, 1128–1134. 10.1016/j.ajhg.2012.10.016.

242. Dornelles-Wawruk, H., Soledad Heredia, R., De Paula-Junior, M.R., Cardoso, M.T.O., Bonadio, R.S., Dos Reis, B.F., Pic-Taylor, A., De Oliveira, S.F., and Mazzeu, J.F. (2019). A Balanced Reciprocal Translocation t(2;9)(p25;q13) Disrupting the LINC00299 Gene in a Patient with Intellectual Disability. Mol Syndromol *10*, 234–238. 10.1159/000500397.

243. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380. 10.1038/nature11082.

244. Kentepozidou, E., Aitken, S.J., Feig, C., Stefflova, K., Ibarra-Soria, X., Odom, D.T., Roller, M., and Flicek, P. (2020). Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. Genome Biol *21*, 5. 10.1186/s13059-019-1894-x.

245. Ewans, L.J., Schofield, D., Shrestha, R., Zhu, Y., Gayevskiy, V., Ying, K., Walsh, C., Lee, E., Kirk, E.P., Colley, A., *et al.* (2018). Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. Genetics in Medicine *20*, 1564–1574. 10.1038/gim.2018.39.

246. Al-Nabhani, M., Al-Rashdi, S., Al-Murshedi, F., Al-Kindi, A., Al-Thihli, K., Al-Saegh, A., Al-Futaisi, A., Al-Mamari, W., Zadjali, F., and Al-Maawali, A. (2018). Reanalysis of exome sequencing data of intellectual disability samples: Yields and benefits. Clin Genet *94*, 495–501. 10.1111/cge.13438.

247. Li, J., Gao, K., Yan, H., Xiangwei, W., Liu, N., Wang, T., Xu, H., Lin, Z., Xie, H., Wang, J., *et al.* (2019). Reanalysis of whole exome sequencing data in patients with epilepsy and intellectual disability/mental retardation. Gene *700*, 168–175. 10.1016/j.gene.2019.03.037.

248. Baker, S.W., Murrell, J.R., Nesbitt, A.I., Pechter, K.B., Balciuniene, J., Zhao, X., Yu, Z., Denenberg, E.H., DeChene, E.T., Wilkens, A.B., *et al.* (2019). Automated Clinical Exome Reanalysis Reveals Novel Diagnoses. The Journal of Molecular Diagnostics *21*, 38–48. 10.1016/j.jmoldx.2018.07.008.

249. Sun, Y., Peng, J., Liang, D., Ye, X., Xu, N., Chen, L., Yan, D., Zhang, H., Xiao, B., Qiu, W., *et al.* (2022). Genome sequencing demonstrates high diagnostic yield in children with undiagnosed global developmental delay/intellectual disability: A prospective study. Human Mutation *43*, 568–581. 10.1002/humu.24347.

250. Palmer, E.E., Sachdev, R., Macintosh, R., Melo, U.S., Mundlos, S., Righetti, S., Kandula, T., Minoche, A.E., Puttick, C., Gayevskiy, V., *et al.* (2021). Diagnostic Yield of Whole Genome Sequencing After Nondiagnostic Exome Sequencing or Gene Panel in Developmental and Epileptic Encephalopathies. Neurology *96*, e1770–e1782. 10.1212/WNL.0000000000011655.

251. Ewans, L.J., Minoche, A.E., Schofield, D., Shrestha, R., Puttick, C., Zhu, Y., Drew, A., Gayevskiy, V., Elakis, G., Walsh, C., *et al.* (2022). Whole exome and genome sequencing in mendelian disorders: a diagnostic and health economic analysis. Eur J Hum Genet *30*, 1121–1131. 10.1038/s41431-022-01162-2.

252. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., *et al.* (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research *46*, D1062–D1067. 10.1093/nar/gkx1153.

253. Rom, A., Melamed, L., Gil, N., Goldrich, M.J., Kadir, R., Golan, M., Biton, I., Perry, R.B.-T., and Ulitsky, I. (2019). Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability. Nat Commun *10*, 5092. 10.1038/s41467-019-13075-8.

254. Lang, R., Liu, G., Shi, Y., Bharadwaj, S., Leng, X., Zhou, X., Liu, H., Atala, A., and Zhang, Y. (2013). Self-Renewal and Differentiation Capacity of Urine-Derived Stem Cells after Urine Preservation for 24 Hours. PLoS ONE *8*, e53980. 10.1371/journal.pone.0053980.

255. Zhou, T., Benda, C., Dunzinger, S., Huang, Y., Ho, J.C., Yang, J., Wang, Y., Zhang, Y., Zhuang, Q., Li, Y., *et al.* (2012). Generation of human induced pluripotent stem cells from urine samples. Nat Protoc *7*, 2080–2089. 10.1038/nprot.2012.115.

256. Tanabe, K., Ang, C.E., Chanda, S., Olmos, V.H., Haag, D., Levinson, D.F., Südhof, T.C., and

Wernig, M. (2018). Transdifferentiation of human adult peripheral blood T cells into neurons. Proc Natl Acad Sci USA *115*, 6470–6475. 10.1073/pnas.1720273115.

257. Mollinari, C., Zhao, J., Lupacchini, L., Garaci, E., Merlo, D., and Pei, G. (2018). Transdifferentiation: a new promise for neurodegenerative diseases. Cell Death Dis *9*, 830. 10.1038/s41419-018-0891-4.

258. Cheng, L., Lei, Q., Yin, C., Wang, H.-Y., Jin, K., and Xiang, M. (2017). Generation of Urine Cell-Derived Non-integrative Human iPSCs and iNSCs: A Step-by-Step Optimized Protocol. Front. Mol. Neurosci. *10*, 348. 10.3389/fnmol.2017.00348.

259. Kang, P.J., Son, D., Ko, T.H., Hong, W., Yun, W., Jang, J., Choi, J.-I., Song, G., Lee, J., Kim, I.Y., *et al.* (2019). mRNA-Driven Generation of Transgene-Free Neural Stem Cells from Human Urine-Derived Cells. 18.

260. Koopmans, F., Van Nierop, P., Andres-Alonso, M., Byrnes, A., Cijsouw, T., Coba, M.P., Cornelisse, L.N., Farrell, R.J., Goldschmidt, H.L., Howrigan, D.P., *et al.* (2019). SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. Neuron *103*, 217-234.e4. 10.1016/j.neuron.2019.05.002.

Titre : Identification de variants structuraux et non-codants par approche multi-omique chez des patients atteints de troubles du neurodéveloppement.

Mots clés : Séquençage, génome, RNA-seq, ARN non codants, troubles du neurodéveloppement

Résumé : Le diagnostic moléculaire des troubles neurodéveloppementaux (TND) repose encore principalement sur le séquençage d'exome, avec un rendement diagnostique plafonnant à 31 % pour les TND isolés et 53 % pour les TND syndromiques. Le séquençage de génome (SG) a montré de meilleures performances mais sa mise en œuvre systématique en diagnostic pose de nouveaux défis, notamment en ce qui concerne la détection et l'interprétation des variants structuraux (SV) ou des variants localisés dans des régions non codantes.Nous avons montré que l'application de techniques omiques complémentaires, telles que le RNA-Seq clinique et la cartographie optique du génome, contribuent à améliorer le rendement diagnostique du séquençage du génome pour des cas complexes ou difficiles à interpréter. Notre étude confirme que, sur la base du SG seul, certains SV n'auraient pas été retenus comme pathogènes/candidats et qu'il aurait été difficile de poser les hypothèses mécanistiques pour les événements complexes. Par ailleurs, nous avons identifié et rapporté une délétion du long ARN non codant régulateur *CHASERR* que nous associons à un nouveau TND syndromique.Le RNA-Seq clinique et les techniques de séquençage long fragments sont des sujets de recherche en pleine expansion. Leur application en diagnostic de routine n'est pas encore standardisée et nécessitera encore une longue mise au point. Avec nos résultats nous apportons un éclairage nouveau sur cette introduction progressive, en association avec le SG.

Title : Identification of structural and non-coding variants by multi-omics approach in patients with neurodevelopmental disorders

Keywords : Sequencing, genome, RNA-seq, non-coding RNA, neurodevelopmental disorders

Abstract : Molecular diagnosis of neurodevelopmental disorders (NDD) relies mainly on exome sequencing, with a diagnostic yield of 31% for isolated NDD and 53% for syndromic NDD. Genome sequencing (GS) has shown better performance; however, its systematic implementation in diagnostics poses new challenges, particularly in the detection and interpretation of structural variants (SV) or variants located in non-coding regions. We have shown that the application of complementary omics techniques, such as clinical RNA-Seq and optical genome mapping, helps improve the diagnostic yield of genome sequencing for complex or difficult-to-interpret cases. Our study confirms that, based on GS alone, some SVs would not have been retained as pathogens/candidates and that it would have been difficult to pose mechanistic hypotheses for complex events. Furthermore, we have identified and reported a deletion of the long non-coding RNA regulator *CHASERR*, which we associate with a new syndromic NDD. Currently, clinical RNA-Seq and long-fragment sequencing techniques are the subject of intense research. Their application in routine diagnostics has not yet been standardized and will take a long time to improve. Our results shed new light on this gradual introduction in association with GS.