

THÈSE
POUR LE DIPLÔME D'ÉTAT
DE DOCTEUR EN PHARMACIE

Soutenue publiquement le 14 septembre 2022

Par Madame Pauline Planté-Bordeneuve

Complémentarité des approches bio-informatiques en séquençage d'exome pour la détection des variations de structure : implications sur les taux de diagnostics dans une cohorte de patients porteurs d'anomalies du neurodéveloppement

Membres du jury :

Président : **Monsieur le Professeur Thierry BROUSSEAU**

Faculté de Pharmacie, Université Lille

Directeur de Thèse : **Monsieur le Docteur Thomas SMOL**

Faculté de Médecine, Université Lille

Assesseur(s) : **Madame la Professeure Claire-Marie DHAENENS**

Faculté de Pharmacie, Université Lille

Madame la Professeure Catherine ROCHE-LESTIENNE

Faculté de Médecine, Université Lille

Monsieur le Docteur Jamal GHOU MID

Faculté de Médecine, Université Lille

REDACTION	VERIFICATION	APPROBATION
Audrey Hennebelle Assistante de direction	Cyrille Responsable Services	Porta des
		Delphine Allorge Doyen

Université de Lille

Président
Premier Vice-président
Vice-présidente Formation
Vice-président Recherche
Vice-présidente Réseaux internationaux et européens
Vice-président Ressources humaines
Directrice Générale des Services

Régis BORDET
Etienne PEYRAT
Christel BEAUCOURT
Olivier COLOT
Kathleen O'CONNOR
Jérôme FONCEL
Marie-Dominique SAVINA

UFR3S

Doyen
Premier Vice-Doyen
Vice-Doyen Recherche
Vice-Doyen Finances et Patrimoine
Vice-Doyen Coordination pluriprofessionnelle et Formations sanitaires
Vice-Doyen RH, SI et Qualité
Vice-Doyenne Formation tout au long de la vie
Vice-Doyen Territoires-Partenariats
Vice-Doyenne Vie de Campus
Vice-Doyen International et Communication
Vice-Doyen étudiant

Dominique LACROIX
Guillaume PENEL
Éric BOULANGER
Damien CUNY
Sébastien D'HARANCY
Hervé HUBERT
Caroline LANIER
Thomas MORGENROTH
Claire PINÇON
Vincent SOBANSKI
Dorian QUINZAIN

Faculté de Pharmacie

Doyen
Premier Assesseur et Assesseur en charge des études
Assesseur aux Ressources et Personnels
Assesseur à la Santé et à l'Accompagnement
Assesseur à la Vie de la Faculté
Responsable des Services
Représentant étudiant

Delphine ALLORGE
Benjamin BERTIN
Stéphanie DELBAERE
Anne GARAT
Emmanuelle LIPKA
Cyrille PORTA
Honoré GUISE

Professeurs des Universités - Praticiens Hospitaliers (PU-PH)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	ALLORGE	Delphine	Toxicologie et Santé publique	81
M.	BROUSSEAU	Thierry	Biochimie	82
M.	DÉCAUDIN	Bertrand	Biopharmacie, Pharmacie galénique et hospitalière	81
M.	DINE	Thierry	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
Mme	DUPONT-PRADO	Annabelle	Hématologie	82
Mme	GOFFARD	Anne	Bactériologie - Virologie	82
M.	GRESSIER	Bernard	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
M.	ODOU	Pascal	Biopharmacie, Pharmacie galénique et hospitalière	80
Mme	POULAIN	Stéphanie	Hématologie	82
M.	SIMON	Nicolas	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
M.	STAELS	Bart	Biologie cellulaire	82

Professeurs des Universités (PU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	ALIOUAT	El Moukhtar	Parasitologie - Biologie animale	87
Mme	AZAROUAL	Nathalie	Biophysique - RMN	85
M.	BLANCHEMAIN	Nicolas	Pharmacotechnie industrielle	85
M.	CARNOY	Christophe	Immunologie	87
M.	CAZIN	Jean-Louis	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
M.	CHAVATTE	Philippe	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	COURTECUISSÉ	Régis	Sciences végétales et fongiques	87

M.	CUNY	Damien	Sciences végétales et fongiques	87
Mme	DELBAERE	Stéphanie	Biophysique - RMN	85
Mme	DEPREZ	Rebecca	Chimie thérapeutique	86
M.	DEPREZ	Benoît	Chimie bioinorganique	85
M.	DUPONT	Frédéric	Sciences végétales et fongiques	87
M.	DURIEZ	Patrick	Physiologie	86
M.	ELATI	Mohamed	Biomathématiques	27
M.	FOLIGNÉ	Benoît	Bactériologie - Virologie	87
Mme	FOULON	Catherine	Chimie analytique	85
M.	GARÇON	Guillaume	Toxicologie et Santé publique	86
M.	GOOSSENS	Jean-François	Chimie analytique	85
M.	HENNEBELLE	Thierry	Pharmacognosie	86
M.	LEBEGUE	Nicolas	Chimie thérapeutique	86
M.	LEMDANI	Mohamed	Biomathématiques	26
Mme	LESTAVEL	Sophie	Biologie cellulaire	87
Mme	LESTRELIN	Réjane	Biologie cellulaire	87
Mme	MELNYK	Patricia	Chimie physique	85
M.	MILLET	Régis	Institut de Chimie Pharmaceutique Albert Lespagnol	86
Mme	MUHR-TAILLEUX	Anne	Biochimie	87
Mme	PERROY	Anne-Catherine	Droit et Economie pharmaceutique	86
Mme	ROMOND	Marie-Bénédicte	Bactériologie - Virologie	87
Mme	SAHPAZ	Sevser	Pharmacognosie	86
M.	SERGHERAERT	Éric	Droit et Economie pharmaceutique	86
M.	SIEPMANN	Juergen	Pharmacotechnie industrielle	85
Mme	SIEPMANN	Florence	Pharmacotechnie industrielle	85

M.	WILLAND	Nicolas	Chimie organique	86
----	---------	---------	------------------	----

Maîtres de Conférences - Praticiens Hospitaliers (MCU-PH)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	BLONDIAUX	Nicolas	Bactériologie - Virologie	82
Mme	DEMARET	Julie	Immunologie	82
Mme	GARAT	Anne	Toxicologie et Santé publique	81
Mme	GENAY	Stéphanie	Biopharmacie, Pharmacie galénique et hospitalière	81
M.	LANNOY	Damien	Biopharmacie, Pharmacie galénique et hospitalière	80
Mme	ODOU	Marie-Françoise	Bactériologie - Virologie	82

Maîtres de Conférences des Universités (MCU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	AGOURIDAS	Laurence	Chimie thérapeutique	85
Mme	ALIOUAT	Cécile-Marie	Parasitologie - Biologie animale	87
M.	ANTHÉRIEU	Sébastien	Toxicologie et Santé publique	86
Mme	AUMERCIER	Pierrette	Biochimie	87
M.	BANTUBUNGI-BLUM	Kadiombo	Biologie cellulaire	87
Mme	BARTHELEMY	Christine	Biopharmacie, Pharmacie galénique et hospitalière	85
Mme	BEHRA	Josette	Bactériologie - Virologie	87
M.	BELARBI	Karim-Ali	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
M.	BERTHET	Jérôme	Biophysique - RMN	85
M.	BERTIN	Benjamin	Immunologie	87
M.	BOCHU	Christophe	Biophysique - RMN	85
M.	BORDAGE	Simon	Pharmacognosie	86
M.	BOSC	Damien	Chimie thérapeutique	86

M.	BRIAND	Olivier	Biochimie	87
Mme	CARON-HOUDE	Sandrine	Biologie cellulaire	87
Mme	CARRIÉ	Hélène	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
Mme	CHABÉ	Magali	Parasitologie - Biologie animale	87
Mme	CHARTON	Julie	Chimie organique	86
M.	CHEVALIER	Dany	Toxicologie et Santé publique	86
Mme	DANEL	Cécile	Chimie analytique	85
Mme	DEMANCHE	Christine	Parasitologie - Biologie animale	87
Mme	DEMARQUILLY	Catherine	Biomathématiques	85
M.	DHIFLI	Wajdi	Biomathématiques	27
Mme	DUMONT	Julie	Biologie cellulaire	87
M.	EL BAKALI	Jamal	Chimie thérapeutique	86
M.	FARCE	Amaury	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	FLIPO	Marion	Chimie organique	86
M.	FURMAN	Christophe	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	GERVOIS	Philippe	Biochimie	87
Mme	GOOSSENS	Laurence	Institut de Chimie Pharmaceutique Albert Lespagnol	86
Mme	GRAVE	Béatrice	Toxicologie et Santé publique	86
Mme	GROSS	Barbara	Biochimie	87
M.	HAMONIER	Julien	Biomathématiques	26
Mme	HAMOUDI-BEN YELLES	Chérifa-Mounira	Pharmacotechnie industrielle	85
Mme	HANNOTHIAUX	Marie-Hélène	Toxicologie et Santé publique	86
Mme	HELLEBOID	Audrey	Physiologie	86

M.	HERMANN	Emmanuel	Immunologie	87
M.	KAMBIA KPAKPAGA	Nicolas	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
M.	KARROUT	Younes	Pharmacotechnie industrielle	85
Mme	LALLOYER	Fanny	Biochimie	87
Mme	LECOEUR	Marie	Chimie analytique	85
Mme	LEHMANN	Hélène	Droit et Economie pharmaceutique	86
Mme	LELEU	Natascha	Institut de Chimie Pharmaceutique Albert Lespagnol	86
Mme	LIPKA	Emmanuelle	Chimie analytique	85
Mme	LOINGEVILLE	Florence	Biomathématiques	26
Mme	MARTIN	Françoise	Physiologie	86
M.	MOREAU	Pierre-Arthur	Sciences végétales et fongiques	87
M.	MORGENROTH	Thomas	Droit et Economie pharmaceutique	86
Mme	MUSCHERT	Susanne	Pharmacotechnie industrielle	85
Mme	NIKASINOVIC	Lydia	Toxicologie et Santé publique	86
Mme	PINÇON	Claire	Biomathématiques	85
M.	PIVA	Frank	Biochimie	85
Mme	PLATEL	Anne	Toxicologie et Santé publique	86
M.	POURCET	Benoît	Biochimie	87
M.	RAVAUX	Pierre	Biomathématiques / Innovations pédagogiques	85
Mme	RAVEZ	Séverine	Chimie thérapeutique	86
Mme	RIVIÈRE	Céline	Pharmacognosie	86
M.	ROUMY	Vincent	Pharmacognosie	86
Mme	SEBTI	Yasmine	Biochimie	87
Mme	SINGER	Elisabeth	Bactériologie - Virologie	87

Mme	STANDAERT	Annie	Parasitologie - Biologie animale	87
M.	TAGZIRT	Madjid	Hématologie	87
M.	VILLEMAGNE	Baptiste	Chimie organique	86
M.	WELTI	Stéphane	Sciences végétales et fongiques	87
M.	YOUS	Saïd	Chimie thérapeutique	86
M.	ZITOUNI	Djamel	Biomathématiques	85

Professeurs certifiés

Civ.	Nom	Prénom	Service d'enseignement
Mme	FAUQUANT	Soline	Anglais
M.	HUGES	Dominique	Anglais
M.	OSTYN	Gaël	Anglais

Professeurs Associés

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	DAO PHAN	Haï Pascal	Chimie thérapeutique	86
M.	DHANANI	Alban	Droit et Economie pharmaceutique	86

Maîtres de Conférences Associés

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	CUCCHI	Malgorzata	Biomathématiques	85
M.	DUFOSSEZ	François	Biomathématiques	85
M.	FRIMAT	Bruno	Pharmacologie, Pharmacocinétique et Pharmacie clinique	85
M.	GILLOT	François	Droit et Economie pharmaceutique	86
M.	MASCAUT	Daniel	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86

M.	MITOUMBA	Fabrice	Biopharmacie, Pharmacie galénique et hospitalière	86
M.	PELLETIER	Franck	Droit et Economie pharmaceutique	86
M.	ZANETTI	Sébastien	Biomathématiques	85

Assistants Hospitalo-Universitaire (AHU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	CUVELIER	Élodie	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
M.	GRZYCH	Guillaume	Biochimie	82
Mme	LENSKI	Marie	Toxicologie et Santé publique	81
Mme	HENRY	Héloïse	Biopharmacie, Pharmacie galénique et hospitalière	80
Mme	MASSE	Morgane	Biopharmacie, Pharmacie galénique et hospitalière	81

Attachés Temporaires d'Enseignement et de Recherche (ATER)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	GEORGE	Fanny	Bactériologie - Virologie / Immunologie	87
Mme	N'GUESSAN	Cécilia	Parasitologie - Biologie animale	87
M.	RUEZ	Richard	Hématologie	87
M.	SAIED	Tarak	Biophysique - RMN	85
M.	SIEROCKI	Pierre	Chimie bioinorganique	85

Enseignant contractuel

Civ.	Nom	Prénom	Service d'enseignement
M.	MARTIN MENA	Anthony	Biopharmacie, Pharmacie galénique et hospitalière



Faculté de Pharmacie de Lille

3, rue du Professeur Laguesse - B.P. 83 - 59006 LILLE CEDEX

03.20.96.40.40 - ☎ : 03.20.96.43.64

<http://pharmacie.univ-lille2.fr>

L'Université n'entend donner aucune approbation aux opinions émises dans les thèses ; celles-ci sont propres à leurs auteurs.

Remerciements

A Monsieur le Professeur Thierry Brousseau,

Professeur des Universités et Praticien Hospitalier, Institut de Biochimie et Biologie Moléculaire, Pôle de Biologie-Pathologie-Génétique, Centre Hospitalier Universitaire de Lille.

Vous me faites l'honneur de présider ce jury et je vous en remercie. Je vous remercie de m'avoir accueillie dans votre institut depuis mon internat.

Soyez assuré de mon profond respect.

A Madame la Professeure Catherine Roche-Lestienne,

Professeure des Universités et Praticien Hospitalier, Institut de Génétique médicale, Hôpital Jeanne de Flandre, Centre Hospitalier Universitaire de Lille.

Je vous remercie grandement d'accepter de faire partie de ce jury de thèse. Vous m'avez accueillie à deux reprises à l'Institut de génétique médicale avec bienveillance et avez su me conseiller dans mon parcours.

Soyez assurée de mon profond respect et de ma gratitude.

A Madame la Professeure Claire-Marie Dhaenens,

Professeure des Universités et Praticien Hospitalier, Institut de Biochimie et Biologie Moléculaire, Pôle de Biologie-Pathologie-Génétique, Centre Hospitalier Universitaire de Lille.

Je vous remercie grandement de l'enthousiasme avec lequel vous avez accepté de faire partie de ce jury de thèse.

Soyez assurée de mon profond respect.

A Monsieur le Docteur Jamal Ghoumid,

Maître de conférences des Universités et Praticien Hospitalier, Clinique de Génétique « Guy Fontaine », Hôpital Jeanne de Flandre, Centre Hospitalier Universitaire de Lille.

Je vous remercie grandement de l'enthousiasme avec lequel vous avez accepté de faire partie de ce jury de thèse. Je tiens à vous exprimer ma gratitude pour m'avoir fait partager votre point de vue clinique au cours de mon internat.

Soyez assuré de mon profond respect.

A Monsieur le Docteur Thomas Smol,

Maître de conférences des Universités et Praticien Hospitalier, Institut de Génétique médicale, Hôpital Jeanne de Flandre, Centre Hospitalier Universitaire de Lille.

Tu me fais l'honneur de diriger ce travail de thèse et je t'en remercie grandement. Ton implication, ta rigueur scientifique, ta disponibilité et ton soutien ont été très précieux tout au long de ce travail. Merci de m'avoir accompagné durant ces années d'internat et ces différents projets où ton avis m'a été d'une aide inestimable, merci de m'avoir transmis la passion de la génétique.

Sois assuré de mon profond respect et de ma gratitude.

Je remercie particulièrement la Docteure Emilie Ait Yahya, bio-informaticienne au laboratoire de biologie moléculaire du CHU de Lille, sans qui ce travail n'aurait pu être effectué. Merci pour le temps que tu as accordé à ce projet, pour ta patience et tes explications.

A Caroline Thuillier pour m'avoir épaulé dans ce travail. Par ton enthousiasme, tes idées, ta disponibilité, tu m'auras tellement appris.

Les docteurs Mélanie Rama et Perrine Brunelle pour m'avoir appris tant de choses en génétique, votre pédagogie et vos conseils m'ont beaucoup apporté.

Aux techniciens du laboratoire des « Puces », Alexis, Delphine, Heidi et Pauline de m'avoir appris sur la biologie moléculaire et ses multiples techniques avec patience et bonne humeur. A Anaïs, Bénédicte et Fadima pour votre aide et votre bonne humeur.

Aux biologistes qui m'ont fait découvrir le monde de la Cytogénétique et de la Biologie Moléculaire : le Professeur Catherine Roche-Lestienne, les Docteurs Elise Boudry, Sonia Bouquillon, Agnès Daudignon, Hélène Guermouche ; et aux techniciens de l'Institut de Génétique Médicales qui m'ont appris les techniques de cytogénétiques.

Aux médecins du service de Génétique Clinique du CHU de Lille, qui m'ont permis d'assister aux consultations de génétique.

Aux biologistes et techniciens de mes différents stages d'internat : Arras, Armentières, l'hématologie du CHU de Lille et le Centre de génétique à la Pitié-Salpêtrière. À toutes les personnes – biologistes, techniciens, ingénieurs, cliniciens, informaticiens et pleins d'autres - qui m'ont accompagnée et formée au cours de mon cursus.

A Pauline, Claire et Simon qui m'ont tant aidé dans mon orientation en génétique.

A mes amis Lillois (et Arrageois), qui sont pour beaucoup à mon affection pour le Nord : Oumaima, Marie-Elise, Mathilde, Simon, Arnaud, Martin. A mes cointernes exceptionnels, avec qui j'ai partagé mon quotidien et mes gardes : Léa, Apolline, Mélanie, Zoé, Amélie, Marie-Amélie, Corentin, Adrien, Benjamin, Khalil, David, Clément, Nicolas, Quentin. A mes cointernes de Paris qui m'ont épaulé dans la rédaction : Elise, Mathieu, Marie, Guillaume et Séverine. A mes amis de longue date qui me supportent depuis bien longtemps, la clique : Cambise, Roxane, Marin, Laura, Coline, Norman, Raphaëlle et Grégoire ; à Salomé et à Claire.

A ma famille qui a su m'épauler lors de la rédaction de cette thèse, Maman, Margot, Papa et Vincent.

A Aurélien, dont la rencontre est la meilleure chose qui me soit arrivée depuis longtemps. Merci pour ces relectures de qualité et surtout pour tout le reste.

À tous celles et ceux qui ont participé de près ou de loin à ce travail et qui ont pris le temps de lire et de corriger cette thèse.

Table des matières

Table des illustrations	22
Table des tableaux	23
Résumé	25
Abréviations	26
I- Introduction	29
A- Les troubles du neurodéveloppement et la déficience intellectuelle	29
B- Classification et méthodes d'évaluation de la DI	30
C- Stratégies d'exploration	32
D- Les plans nationaux	33
E- Bref historique du séquençage	34
F- Enjeux de la détection des SV	36
G- Objectifs	37
II- Matériels et méthodes	39
A- Patients	39
B- Extraction d'ADN	39
C- Exploration de cytogénétique moléculaire par ACPA	40
1) Principe de la technique d'ACPA	40
2) Identification des remaniements cytogénétiques	41
3) Interprétation des CNV	42
D- Exploration par séquençage d'exome	43
1) Préparation des librairies	43
2) Amplification clonale : génération de cluster	46
3) Séquençage	47
4) Traitement bio-informatique	48
5) Etude des CNV/SV par exome	51
6) Outil DRAGEN Copy Number Variant Calling	54
7) Outil DRAGEN Structural Variant Calling	56
8) Outil DRAGEN ROH Caller	56
9) Interprétation des variants	57
E- Validation des CNV/SV et SNV	58
1) Séquençage Sanger	58
2) qPCR	58
III-Résultats	59
A- Détection des CNV : comparaison ACPA versus Exome	59

B-	Identification des CNV sous le seuil de résolution de l'ACPA.....	64
C-	Application de l'outil SV DRAGEN.....	72
D-	ROH DRAGEN	79
1)	Données étudiées.....	79
2)	Préparation du fichier ROH : Mise en forme.....	79
3)	Analyse du fichier ROH : Exploitation.....	81
E-	Diagnostics par exome au CHU de Lille d'octobre 2020 à Mars 2022.....	86
IV-	Discussion	89
A-	Les taux diagnostics.....	89
1)	A l'Institut.....	89
2)	Dans la littérature	90
B-	Détection des CNV en exome versus en ACPA	91
1)	Détection des CNV à la résolution de l'exome	91
2)	Des analyses complémentaires.....	92
3)	Remaniements non caractérisés en exome	93
4)	Amélioration de nos analyses.....	94
C-	SV DRAGEN	94
D-	ROH DRAGEN	95
1)	Validation de méthode de ROH DRAGEN.....	96
2)	Analyse de notre cohorte.....	97
E-	Perspective de la réanalyse d'exome et échanges internationaux	97
F-	Autres pipelines et perspectives.....	98
1)	Repérer les expansions de séquences répétées.....	98
2)	Localiser l'insertion d'éléments mobiles	99
3)	Génome « <i>short read</i> ».....	100
4)	Génome « <i>Long read</i> ».....	100
5)	La cartographie optique.....	101
	Conclusion.....	103
	Bibliographie.....	105
	Annexes.....	113
	Annexe 1 : Tableau de comparaison ACPA versus exome.....	114

Table des illustrations

Figure 1: Schéma de la répartition du QI dans la population générale.	31
Figure 2 : Schéma du PFMG	33
Figure 3 : Schéma de différents SV inspiré de Genome structural variation discovery and genotyping (23).....	36
Figure 4 : Principe de l'ACPA d'après Karampetsou et al (29).	40
Figure 5 : Représentation des ratios d'intensité de fluorescence sur l'outil de visualisation Cytogenomics.	41
Figure 6 : Visualisation du CNV sur UCSC Table Browser.....	42
Figure 7 : Schéma de la préparation des bibliothèques par fragmentation d'après Illumina Nextera flex for Enrichment.	44
Figure 8 : Schéma de capture de la bibliothèque exonique d'après Illumina Nextera flex for Enrichment.	45
Figure 9 : Schéma de la génération de cluster d'après Metzker and al.	46
Figure 10 : Schéma de séquençage par synthèse et image d'un cycle de séquençage par synthèse selon Metzker et al (14).	48
Figure 11 : Schéma du traitement bio-informatique des données d'exome.....	50
Figure 12 : Représentation de la détection de CNV grâce à la comparaison de profondeur de lecture selon Escaramis et al (28).	51
Figure 13 : Illustration selon Escaramis et al de la détection de SV grâce aux paired-end.	52
Figure 14 : Représentation du split read et de l'exome de novo pour identifier des SV selon Escaramis et al (28).	53
Figure 15 : Schéma des différentes méthodes de détection des SV visualisé sur IGV (54).	53
Figure 16 : Schéma du fonctionnement de CNV DRAGEN d'Illumina.	54
Figure 17 : Schéma du raisonnement AnnotSV (25).	55
Figure 18 : Schéma du fonctionnement de l'algorithme ROH mis au point par DRAGEN.	57
Figure 19 : Comparaison de la détection d'anomalie quantitative génomique en ACPA versus Exome.	60
Figure 20 : Duplication arr[GRCh37] 2q24.3(167602904_167992598)x3 identifiée chez le quatuor.	62
Figure 21 : Visualisation sur le Table Browser d'UCSC de la région dupliquée en arr[GRCh37] 3p26.3(1638453_2262692)x3 et la délétion arr[GRCh37] 9p24.3(500525_569970)x1.....	63
Figure 22 : Délétion homozygote de 54 kb emportant la partie 3' du gène KANTB1 chez P1 sur IGV.	65
Figure 23 : A : Délétion hémizygote de survenue de novo, d'environ 35kb, dans le gène HPRT1 responsable du syndrome de Lesch Nyhan chez P2, visualisée sur IGV : seq[GRCh38]del(X)(q26.2) NC_000023.11 :g.(?_134460165)_(134486530_?)del.	66
Figure 24 : Délétion hétérozygote de 4kb de novo dans KMT2C	67
Figure 25 : Délétion en 16q23.1 impliquant le gène WWOX et SNV tronquant dans WWOX en trans, hétérozygote composite.....	69

Figure 26 : Représentation sur Cytogenomics de deux sondes déviantes dans la région de la délétion.	71
Figure 27 : Délétion hétérozygote de 9kb en 9q34.3 emportant la partie 5' du gène EHMT1	72
Figure 28 : délétion intra-exonique, seq[GRCh38]del(19)(q23.1) NC_000019.10 :g.[10356763_10356832]del de SLC25A24.....	73
Figure 29 : Visualisation d'une délétion, seq[GRCh38]del(2)(q23.1) NC_000019.10 :g.[201281733_201281929]del dans CASP8	74
Figure 30 : Visualisation sur IGV d'une insertion de rétrotransposon sans impact clinique sur le patient.	75
Figure 31 : Délétion dans l'exon 4 de MECP2 chez un patient DI	76
Figure 32 : Image d'IGV des délétions de 50, 3 et 2pb dans la partie 5' de l'exon 4 de MECP2	78
Figure 33 : Schémas du fonctionnement d'« Interseq tool » de Galaxy.....	81
Figure 34 : Schéma explicatif de l'application des filtres sur le fichier ROH.	82
Figure 35 : Schéma du raisonnement de confirmation de la délétion/disomie sur le fichier originel de l'exome, application de filtres progressifs pour retrouver l'évènement génétique de disomie/délétion.....	83
Figure 36 : Pourcentages des résultats rendus en exome dans l'UF cytogénétique des déficiences intellectuelle et anomalies du développement du CHU de Lille.....	86
Figure 37 : Participation des différents algorithmes aux rendus de résultat en exome complet à l'institut de génétique médicale au CHU de Lille dans l'UF génétique du neurodéveloppement.	87
Figure 38 : Présentation des capacités d'analyses des STR d'ExpansionHunter sur différentes pathologies d'après Dolzhenko E et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. nov 2017. (83)	99

Table des tableaux

Tableau 1 : tableur ROH initiale issu d'un fichier .bed.	77
Tableau 2 : Coordonnées génomiques des 11 variations identifiées avec la stratégie de filtres sur les données ROH chez 11 patients différents.	83
Tableau 3 : Résultats de l'algorithme ROH DRAGEN sur les 3 témoins.	84

Résumé

L'identification des variations de structures (SV) est essentielle pour l'interprétation des données de séquençages pangénomiques mais est historiquement un sujet complexe. Les SV sont extrêmement variés, >50pb à plusieurs mégabases. Ils sont présents dans le génome humain, affectent les fonctions cellulaires et moléculaires, les fonctions de régulation, les structures 3D et la machinerie transcriptionnelle. Leur détection est un enjeu majeur de la génétique actuelle, autant sur le plan de la connaissance du génome, de ses structures, qu'en pathologie humaine. De multiples stratégies ont émergé pour améliorer la détection des SV, notamment en exome : utilisation d'algorithmes couplés au séquençage haut débit, seul ou en association pour plus de performance, le séquençage troisième génération avec les *Long Reads* et la cartographie optique.

Sur les données d'exome, nous avons évalué 3 approches bio-informatiques pour la détection de SV : CNV DRAGEN, SV DRAGEN et ROH DRAGEN. Nous avons mesuré leur impact sur l'apport diagnostique à l'Institut de Génétique Médicale du CHU de Lille sur une cohorte de patients porteurs de troubles du neurodéveloppement de Septembre 2020 à Mars 2022.

La complémentarité de ces approches et leurs limites ont ainsi pu être identifiées : CNV DRAGEN trouve avec une grande efficacité les CNV/SV de plus de 500pb alors que SV DRAGEN à partir de 50pb. Cependant, la couverture fragmentée et partielle de l'exome ne permet pas une utilisation optimale de ces pipelines.

L'utilisation de CNV DRAGEN aura permis de faire 8 diagnostics supplémentaires et SV DRAGEN 2, sur 600 patients, soit un apport diagnostique de 1,7%.

Abréviations

AAIDD : American Association on Intellectual and Developmental Disabilities

ACMG : American College Medical Genetics

ACPA : Analyse Chromosomique par Puce à ADN

ADN : Acide DésoxyriboNucléotide

ARN : Acide RiboNucléotide

BAM : Binary Alignment Map

CAD : Centre de collecte et d'Analyse des Données

CGH-Array : Copy-number Genome Hybridization

CHU : Centre Hospitalo-Universitaire

CIGAR : Concise Idiosyncratic Gapped Alignment Report

CIM : Classification International des Maladies

CNV : Copy Number Variation

ddNTP : Didésoxyribonucléotide

dNTP : Désoxyribonucléotide

DI : Déficience intellectuelle

DGV : Database of Genomic Variant

EDTA : EthylèneDiamineTétraAcétique

GATK : Genome Analysis ToolKit

HLA : Human Leucocyte Antigen

INDEL : Insertion Deletion

IGV : Integrative Genomics Viewer

NAHR : Non Allelic Homologous Recombinaison

NGS : Next Generation Sequencing

OMIM : Online Mendelian Inheritance in Man

OMS : Organisation Mondiale de la Santé

PCR : Polymerase Chain Reaction

PMFG : Plan France Médecine Génomique

QI : Quotient Intellectuel

ROH : Region Of Homozygotie

SBS : Sequencing By Synthesis

SHD : Séquençage Haut Débit

SMB : Streptavidin Magnetic Beads

SNV : Single Nucleotide Variation

SNP : Single Nucleotide Polymorphism

STR : Short Tandem Repeat

SV : Structural Variant

TAD : Topologically Associating Domain
TND : Troubles du Neurodéveloppement
UCSC : University of California Santa Cruz
VCF : Variant Caller Format
VOUS : Variant Of Unknown Significance
VSI : Variant de Signification Indéterminé
WAIS : Wechsler Adult Intelligence Scale
WES : Whole Exome Sequencing
WGS : Whole Genome Sequencing
WISC : Wechsler Intelligence Scale for Children
WPPSI : Wechsler Preschool and Primary Scale of Intelligence

I- Introduction

A- Les troubles du neurodéveloppement et la déficience intellectuelle

Les troubles du neurodéveloppement (TND) sont un groupe de maladies hétérogènes qui regroupe l'ensemble des anomalies de structure ou du fonctionnement cérébral survenant sur un cerveau en développement (1). Cette vaste entité de pathologies, d'une fréquence de 2 à 5% chez les enfants, regroupe les retards développementaux globaux, les déficiences intellectuelles (DI), les épilepsies et les troubles du spectre autistiques. Ces troubles apparaissent parfois très précocement, dans la période néonatale ou plus tardivement, lors de l'entrée à l'école. Ils sont de sévérité très variable et sont d'origine génétique ou environnementale (2,3).

Ces pathologies sont caractérisées par une forte comorbidité, suggérant une étiologie embryonnaire et génétique commune (4). Dans la Classification Internationale des Maladies (CIM-11), elles sont regroupées suivant (5) : les troubles du développement intellectuel ; les troubles développementaux de la parole ou du langage ; les troubles du spectre autistique ; les troubles développementaux des apprentissages ; les troubles de la coordination motrice ; les troubles déficitaires de l'attention avec hyperactivité ; les troubles du mouvement stéréotypé ; les tics ; les syndromes neurodéveloppementaux secondaires.

Particulièrement, les troubles du développement intellectuel ou déficiences intellectuelles (DI) touchent plus d'1% de la population mondiale (6). Les étiologies de DI sont environnementales (traumatismes ou hypoxies à la naissance, toxiques pendant la grossesse, infectieuses en pré et post natal...), génétiques, mais dans plus de 50% des cas, de cause inconnue (7). La déficience peut être de sévérité variable, sa prise en charge est pluridisciplinaire et son dépistage doit être précoce. En effet, plus l'enfant est stimulé, meilleures sont ses chances d'autonomisation futures (7).

D'après l'American Psychiatric Association (APA) qui publie le Manuel diagnostique et statistique des troubles mentaux (DSM-V), révision de 2013 (1) :

« Le handicap intellectuel (trouble du développement intellectuel) est un trouble débutant pendant la période du développement, fait de déficits tant intellectuels qu'adaptatifs dans les domaines conceptuels, sociaux et pratiques. Les trois critères suivants doivent être présents :

- 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, confirmés par l'évaluation clinique et les tests d'intelligence individuels standardisés.
- 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. Sans assistance au long cours, les déficits adaptatifs limitent le fonctionnement dans un ou plusieurs champs d'activité de la vie quotidienne comme la communication, la participation sociale, l'indépendance, dans des environnements variés tels que la maison, l'école, le travail, la collectivité.
- Début du déficit intellectuel et adaptatif pendant la période du développement. »

B- Classification et méthodes d'évaluation de la DI

Le Quotient intellectuel (QI) est un test psychométrique qui fournit une indication quantitative et standardisée de l'intelligence. Le QI standard correspond au rang auquel se situe une personne relativement à une population représentée par une courbe de Gauss (suivant la loi Normale) (8) (Figure 1).

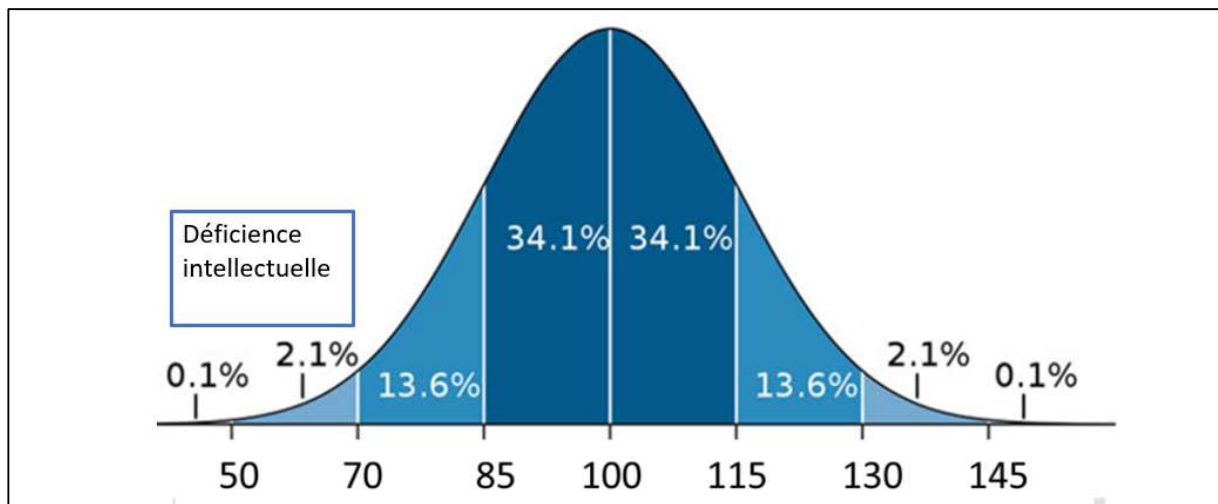


Figure 1: **Schéma de la répartition du QI dans la population générale.**

Selon cette définition, 3% de la population a un QI <70. La définition et la catégorisation de la gravité de la DI est plus complexe que la seule mesure de QI.

Le QI doit être mesurée par une personne qualifiée, dans le cadre d'un examen psychologique et dans des conditions optimales pour une personne. Les tests psychométriques de mesure de l'intelligence sont adaptés à l'âge. On peut évaluer les âges d'acquisitions psychomotrices, le Quotient de Développement ou QD (âge de développement psychomoteur rapporté à l'âge chronologique, chez le petit enfant), ou le QI (âge mental par rapport à l'âge chronologique) (8–10).

De façon plus complète, la « WAIS » pour Wechsler Adult Intelligence Scale est un test de mesure du QI chez l'adulte, à partir de 16 ans. La quatrième version de ce test, révisée de 2008 est en vigueur (10). Elle mesure plusieurs paramètres, qui, ensemble, permettent de calculer le QI : l'indice de compréhension orale ; l'indice de raisonnement perceptif ; l'indice de mémoire du travail ; l'indice de vitesse de traitement. Ce test est composé de 10 sous-tests principaux obligatoires et de 5 sous-tests complémentaires facultatifs, répartis dans ces 4 catégories.

Le « Wechsler Intelligence Scale for Children » ou WISC évalue le QI des enfants de 6 ans à 16 ans et le « Wechsler Preschool and Primary Scale of Intelligence » ou WPPSI, de forme préscolaire, l'évalue pour les enfants de 2 à 6 ans. Ces tests ont les mêmes bases que le WAIS mais sont adaptés à l'âge de leur public. Avant l'âge de 3-5 ans, on parle préférentiellement de retard de développement plutôt que de DI, ce retard pouvant se normaliser avec l'âge (8,10). Ces scores donnent une classification

de la DI : DI légère pour un QI entre 50 et 69 ; modérée entre 35 et 49 ; sévère entre 20 et 34 ; profonde si inférieur à 20.

Ces classifications, basées sur les tests psychométriques, sont progressivement remplacées par des classifications se basant sur les besoins de soutien de la personne dans les domaines conceptuel, social et pratique selon la DSM-V(1).

C- Stratégies d'exploration

La DI ne se définit donc pas uniquement comme un $QI < 70$ mais comme une entité complexe des troubles du neurodéveloppement. Il regroupe un fonctionnement intellectuel déficitaire, un manque d'autonomie dans différents domaines d'activité et doit être présent durant la période de développement (1).

Après évaluation clinique en génétique et en neuropédiatrie, et après exclusion d'un syndrome identifiable en lien avec un phénotype caractéristique, la stratégie exploratoire diagnostique d'une DI, telle que définie dans le Plan France Médecine Génomique 2025 (PFMG), comporte une étude de cytogénétique moléculaire par Analyse Chromosomique par Puce à ADN (ACPA), et une recherche d'un syndrome X-fragile, puis secondairement une étude pangénomique par séquençage haut-débit (SHD). En effet, depuis le développement à large échelle du SHD dans les laboratoires de génétique, particulièrement avec l'étude de l'exome, le nombre de gènes causaux ou candidats dans la DI est estimé à environ 3000 (7). L'hétérogénéité génétique associée aux DI a modifié le paradigme d'interprétation des données génétiques, avec une bascule vers l'ère du « *reverse phenotyping* », c'est-à-dire que les syndromes sont décrits après la mise en cause d'un gène. Ainsi les cohortes syndromiques sont constituées *a posteriori*. Il est classiquement reconnu qu'on identifie une cause génétique dans environ 40% des DI (7). Ces causes se différencient en étiologie cytogénétique et génomique (7). Cependant, la caractérisation des phénomènes physiopathologiques sous-jacent reste un enjeu considérable (7).

D- Les plans nationaux

Le plan de médecine génomique 2025 a mis la génomique au premier plan afin de développer le champ de la médecine personnalisée. Les maladies rares et les cancers sont un enjeu de santé publique. Les patients atteints de pathologies retenues dans ce plan peuvent bénéficier en routine d'un séquençage de leur génome. Cela révolutionne la prise en charge, leur parcours de soin, et évite ainsi l'errance diagnostique (Figure 2) (11).

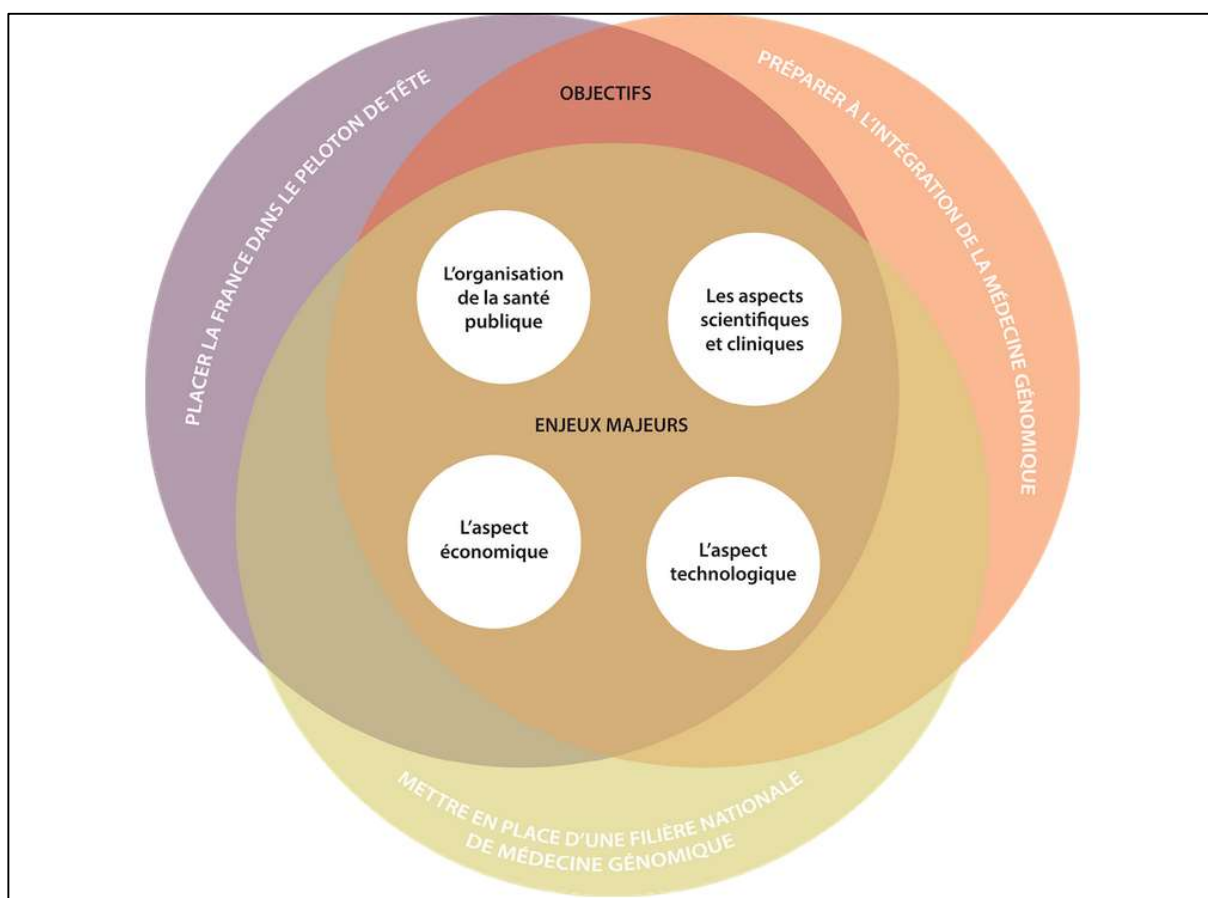


Figure 2 : **Schéma du PFMG**

Il détaille les enjeux et les objectifs au niveau scientifique, économique, technologique et de santé publique pour faire évoluer la génomique en France, sur <https://pfmg2025.aviesan.fr> (9)

Le PFMG a notamment permis l'émergence de deux plateformes de séquençage couvrant l'ensemble du territoire français avec SeqOIA et AURAGEN. De plus, le CRefIX, centre de référence, encadre leurs activités et le Collecteur Analyseur de Données (CAD) garantit la souveraineté des données (stockage, confidentialité, et utilisation éthique de ces données) (11).

Il est constitué de 4 projets pilotes : le programme MULTIPLI dédié au cancer ; le projet DEFIDIAG pour les maladies rares et plus particulièrement la déficience intellectuelle (12) ; le projet GLUCOGEN dédié aux maladies communes ; le projet POPGEN pour identifier les fréquences des variations afin d'établir une base de données de référence de la population française.

Particulièrement, le projet DEFIDIAG, s'inscrit dans le cadre du 3^e Plan National Maladie Rare 2018-2022. La DI est une cause majeure d'errance diagnostique en France, ce qui complique sa prise en charge ; les patients et leur famille doivent avoir accès à un conseil génétique adapté concernant les éventuels risques de transmission. Ce projet a pour but de démontrer la faisabilité et l'efficacité du séquençage du génome complet, en trio, en première intention dans les cas de DI. Il se compare à « la stratégie de référence française », c'est-à-dire la recherche de syndrome de l'X-fragile, l'analyse des chromosomes (ACPA) puis l'étude d'un panel de 44 gènes connus en DI en SHD. Plus de 1200 patients, et leurs deux parents biologiques, participent à l'étude (12,13).

Alors que la stratégie de référence française aboutit à un diagnostic dans 20% des cas, cette étude espère obtenir un rendement diagnostique dans 60% des cas (pour 40% de diagnostic dans la plupart des études de DI en Whole Exome Sequencing (WES) (7). Une étude d'impact permettra de comparer les coûts de recherche diagnostique de ces deux méthodes, et envisagera de modifier les conditions de remboursement de ces analyses (12,13).

E- Bref historique du séquençage

Le séquençage d'ADN (Acide DésoxyriboNucléotide) consiste à déterminer l'enchainement de nucléotides sur des fragments d'ADN ou sur l'ensemble du génome. Il est développé dans les années 70 par les équipes de Frederik Sanger et Walter Gilbert, l'un par synthèse et l'autre par dégradation chimique sélective de nucléotide.

La méthode Sanger consiste en une synthèse d'ADN simple brin par une PCR (Polymerase Chain Reaction) associant de façon aléatoire des didésoxyribonucléotides (ddNTP) terminateurs de chaîne marqués d'un fluorochrome de couleur selon le nucléotide A, T, C et G. Différentes tailles de fragments d'un brin d'ADN sont générées par la PCR, puis séparées selon leur taille en électrophorèse. La lecture par ordre croissant des derniers nucléotides (colorés par un fluorochrome spécifique) des fragments permet de reconstituer la séquence du modèle (14).

A partir des années 2000, cette technologie a considérablement été améliorée par l'avènement du « *Next Generation Sequencing* » (NGS) ou Séquençage Haut Débit (SHD). Ces techniques sont variées mais regroupent des étapes communes : la préparation d'une librairie d'ADN, l'amplification clonale, le séquençage de l'ADN et le traitement informatique des données générées (14). Différentes approches ont été développées, comme, le séquençage Ion Torrent, le pyroséquençage, le séquençage par synthèse, le séquençage par ligation. Ces méthodes permettent une réduction des coûts de séquençage majeure et une grande rapidité. Illumina (San Diego, US-CA), avec le séquençage par synthèse, s'est largement imposée sur le marché mondial (14).

Le premier séquençage entier du génome humain, dans le cadre du « *Human Genome Project* » s'est terminé en 2003, utilisant largement la méthode Sanger, il aura duré presque 15 ans et coûté 2,7 milliards de dollars (15). De nombreux projets de séquençage ont eu lieu depuis, autant dans les domaines des pathologies humaines que pour la connaissance du génome humain (1000 Genomes (16), le projet québécois CARTaGENE (17), le projet britannique 100 000 Genomes (18)). Ces données sont aujourd'hui capitales pour l'interprétation des variants (19).

La compréhension des variations de structure a été améliorée grâce au séquençage de génomes complets. En effet, la majorité des points de cassures sont introniques. Cependant, la longueur des reads ne permet pas encore la meilleure résolution ni la détection de tous les SV ou des grandes séquences répétées (20). Pour s'affranchir de ces limites, de nouvelles technologies, par séquençage de troisième génération, telles que le « Long-read » avec PacBio et Nanopore, ou la cartographie optique ont émergé (20).

Aujourd'hui, les enjeux de séquençage sont différents : réduire les coûts pour donner l'accès au plus grand nombre, accéder à toujours plus de données : les variations de structures, les zones répétées, l'interprétations des variants rares et endémiques à certaines ethnies (20,21). C'est aussi l'avènement du post-génomique avec le méthylome et le transcriptome pour être au plus près de l'impact des variations et de la compréhension des mécanismes génétiques (18).

F- Enjeux de la détection des SV

L'identification des variations de structures (SV) est essentielle pour l'interprétation du génome mais est historiquement un sujet complexe (20,22). Les SV sont extrêmement variés, de plus de 50pb à plusieurs mégabases, et à l'organisation complexe : duplications, délétions, inversions, duplications en tandem, séquences répétées, translocations (Figure 3). Ils comprennent une myriade de sous-classes dont les « *Copy Number Variation* » (CNV). Ils sont présents dans le génome humain, affectent les fonctions cellulaires et moléculaires, les fonctions de régulation, les structures 3D et la machinerie transcriptionnelle (20). Les estimations récentes dénombrent plus de 20 000 SV par individus et sont très majoritairement polymorphiques (13,14).

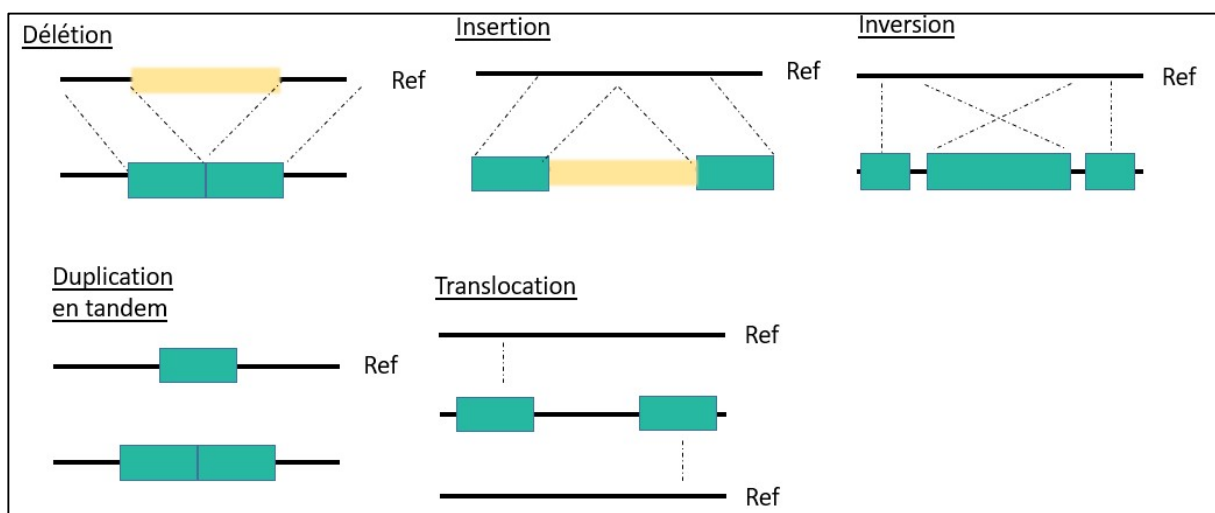


Figure 3 : **Schéma de différents SV inspiré de Genome structural variation discovery and genotyping** (23).

Le premier SV, en haut à gauche de l'image représente une délétion, le fragment jaune sur la séquence de référence est absent de la séquence alignée, les fragments verts, flanquant, se retrouvent accolés. Dans l'insertion, l'inverse se produit, le fragment jaune est inséré entre les reads verts, normalement

jointes sur la séquence de référence. Lors d'une inversion, le fragment vert se retrouve dans l'autre sens de la séquence de référence. Dans la duplication en tandem, le fragment vert se suit à plusieurs reprises au lieu d'une sur la séquence de référence. Dans une translocation, deux fragments censés se suivre sur la séquence de référence sont en fait alignés à deux endroits distincts, par exemple sur deux chromosomes différents.

La détection de ces structures est un enjeu majeur de la génétique actuelle, autant sur le plan de la connaissance du génome, de ses structures, qu'en pathologie humaine où les SV sont impliqués (20,22). De multiples stratégies ont émergé : utilisation d'algorithmes pour la détection des SV, seul ou en les associant pour plus de performance, l'évolution des techniques de séquençage avec notamment le Long-Read et l'augmentation de la profondeur de séquençage des génomes dans les grandes cohortes pour augmenter la sensibilité de détection (20,22).

Actuellement au CHU de Lille dans l'UF Cytogénomique des déficiences intellectuelle et anomalies du développement, la stratégie d'analyse des CNV, selon le type de SV, s'appuie sur le caryotype, la FISH (Fluorescence *In Situ* Hybridization), l'ACPA et en exome complet sur l'algorithme de CNV DRAGEN, basé sur la profondeur de lecture de l'exome et annoté par à AnnotSV (24,25).

G- Objectifs

Les objectifs de ce travail sont d'évaluer les nouveaux outils mis à disposition au CHU de Lille pour l'identification et l'interprétation des SV par séquençage d'exome. Dans un deuxième temps, nous évaluerons le gain diagnostique apporté par ces nouveaux outils. Ce travail se présentera comme suit :

- Comparaison de l'ACPA par rapport à l'exome dans la détection et l'interprétation de CNV, avec l'utilisation de CNV DRAGEN (24) ;
- Détection de variation de structure par l'algorithme « SV » de DRAGEN, Illumina (26) ;

- Détection de disomie uniparentale grâce à l'application d'un algorithme sur les données de l'exome : « ROH » par DRAGEN, Illumina (27) ;
- Statistiques des diagnostics par exome d'octobre 2020 à mars 2022 avec l'apport diagnostic de ces différents algorithmes.

II- Matériels et méthodes

H- Patients

600 patients ont été inclus dans ce travail. Tous les patients ont bénéficié d'un séquençage d'exome complet au CHU de Lille par technologie de séquençage « courts fragments » (14), dans le cadre de l'exploration d'une DI isolée ou syndromique, d'un trouble du neurodéveloppement, d'une épilepsie isolée ou syndromique. Préalablement, la quasi-totalité des patients avaient bénéficié d'une ACPA, au CHU de Lille ou dans un centre extérieur, ne mettant pas en évidence d'anomalie pathogène ou possiblement pathogène en lien avec leurs phénotypes.

Les prescriptions ont été réalisées par un généticien du service de génétique clinique du CHU de Lille, ou un neuropédiatre du service de neuropédiatrie du CHU de Lille. L'ensemble des consentements libres et éclairés ont été recueillis en amont des explorations de génétique.

I- Extraction d'ADN

Les différentes techniques d'analyses cytogénétiques et moléculaires utilisent de l'ADN leucocytaire. Les ADN ont été extraits à partir d'un prélèvement sanguin veineux périphérique sur tube EDTA, de manière automatisée avec le Robot Chemagic STAR (Hamilton, Reno, US-NE) selon le protocole du fournisseur. Les ADN ont été dosés par spectrophotométrie en utilisant le DropSense Chemagic (Hamilton), à 260 nm. Le rendement de l'extraction d'1mL de sang est en moyenne de 150ng/μL dans 150μL.

J- Exploration de cytogénétique moléculaire par ACPA

L'ACPA est une technique de cytogénétique moléculaire pangénomique qui permet d'identifier les remaniements cytogénétiques déséquilibrés. A l'Institut de Génétique Médicale du CHU de Lille, chaque patient a été exploré sur des lames oligonucléotidiques SurePrint G3 Human CGH Microarray (Agilent, Santa Clara, US-CA), possédant environ 60 000 sondes oligonucléotidiques, et permettant une résolution de détection d'environ 50-150kb des variations cytogénétiques déséquilibrées, selon les régions génomiques. C'est la méthode d'analyse pangénomique de référence des CNV (28).

1) Principe de la technique d'ACPA

L'ACPA est basée sur le principe d'hybridation génomique comparative. Les quantités d'ADN associées à aux différentes régions génomiques d'un patient et de témoins sont comparés (Figure 4).

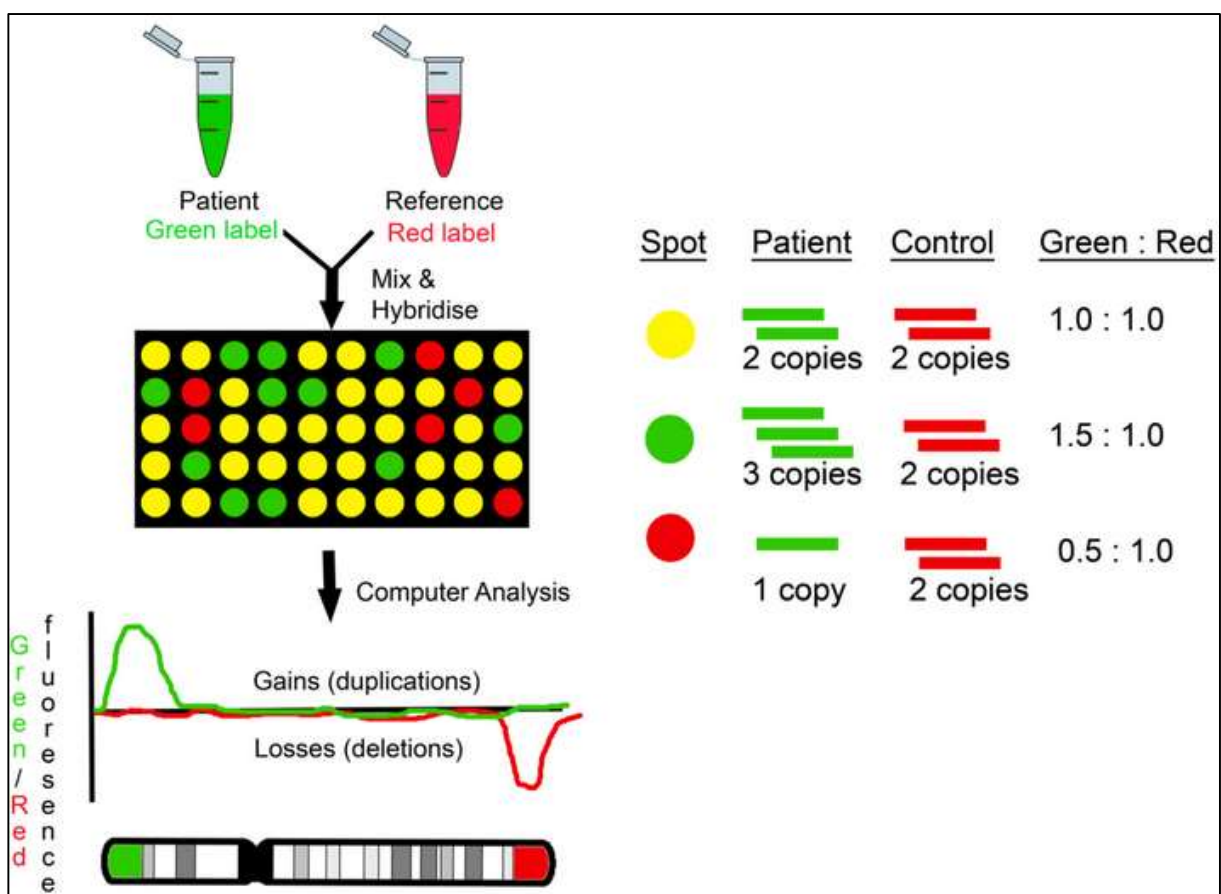


Figure 4 : Principe de l'ACPA d'après Karampetsou et al (29).

L'ADN lymphocytaire du patient et d'un témoin sont marqués par de la Cyanine 3 et 5, l'ADN est hybridé sur des sondes fixées à une lame qui correspondent à des régions chromosomiques précises. La lame avec les hybridations ADN+sondes est lue, un logiciel va localiser sur le génome de référence chaque spot avec la quantité d'ADN correspondant.

Après extraction et fragmentation, les ADN du patient et des témoins vont être marqués par des fluorochromes différents appelés cyanines, puis hybridés sur une lame de verre comportant des séquences oligonucléotidiques simples brins, spécifiques du génome de référence. L'étape d'hybridation sur la lame de verre est réalisée de manière compétitive entre les deux populations d'ADN. Les intensités de fluorescence sont ensuite mesurées et converties en ratio d'intensité de fluorescence entre ADN patient et témoin pour chaque sonde oligonucléotidique. L'intensité de fluorescence est détectée par un lecteur de lames MicroArray Scanner (Agilent) (Figure 4) (30).

2) Identification des remaniements cytogénétiques

Les ratios d'intensités de fluorescence de chaque sonde sont secondairement et informatiquement associés au génome de référence pour chaque chromosome (Figure 5). Ainsi les déviations de ratio seront associées à des déséquilibres cytogénétiques par délétion ou duplication chez le patient. Nous avons utilisé l'outil Cytogenomics (Agilent) comme interface de visualisation des remaniements du nombre de copies pour chaque chromosome.

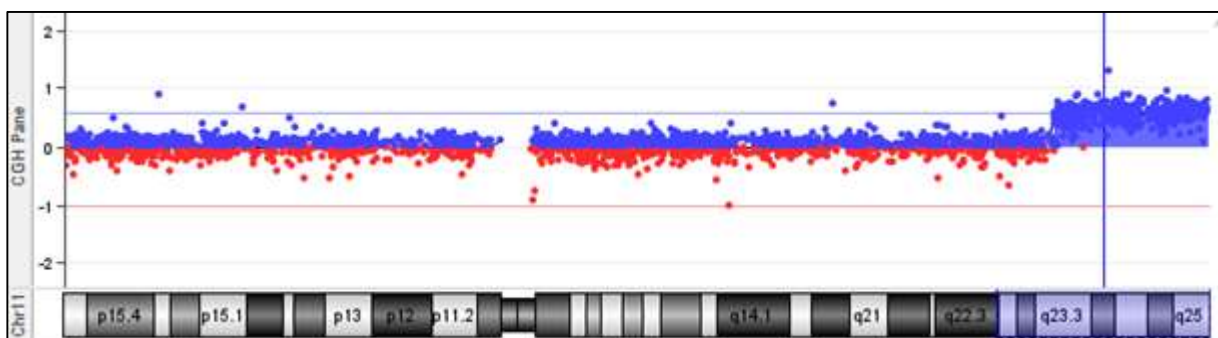


Figure 5 : Représentation des ratios d'intensité de fluorescence sur l'outil de visualisation Cytogenomics.

En ordonnée, les ratios de fluorescence patient versus témoin, en abscisse la position des sondes visualisées. Le patient présente une duplication de la partie terminale du bras long du chromosome 11 en 11q25.

Elle permet également de pré-interpréter en visualisant la récurrence du microremaniement dans notre base de données interne et d'identifier les gènes associés aux régions remaniées (Figure 5)

3) Interprétation des CNV

L'interprétation des CNV fait appel aux recommandations du réseau AChro-Puce mises à jour en 2020, en lien notamment avec les recommandations de l'ACMG et du consortium ClinGen (31). Il s'agit notamment d'évaluer le degré de pathogénicité de chaque microremaniement identifié. Le faisceau d'arguments concerne notamment la récurrence dans les bases de données de populations contrôles : gnomAD (32), Database of Genomic Variant (DGV) (33) ; la récurrence dans les bases de données patients ClinGen (34), DECIPHER (35), Development Delay (DD) (36,37) ; et le contenu génique, notamment en gènes morbides OMIM (38). Ces informations peuvent être regroupées dans l'outil d'annotation génomique « UCSC Genome Browser », permettant d'y associer des informations additionnelles sur la taille des variations, les données de régulation (Figure 6). Chaque CNV est contextualisé selon le phénotype, et interprété selon les critères ACMG en vigueur (39).

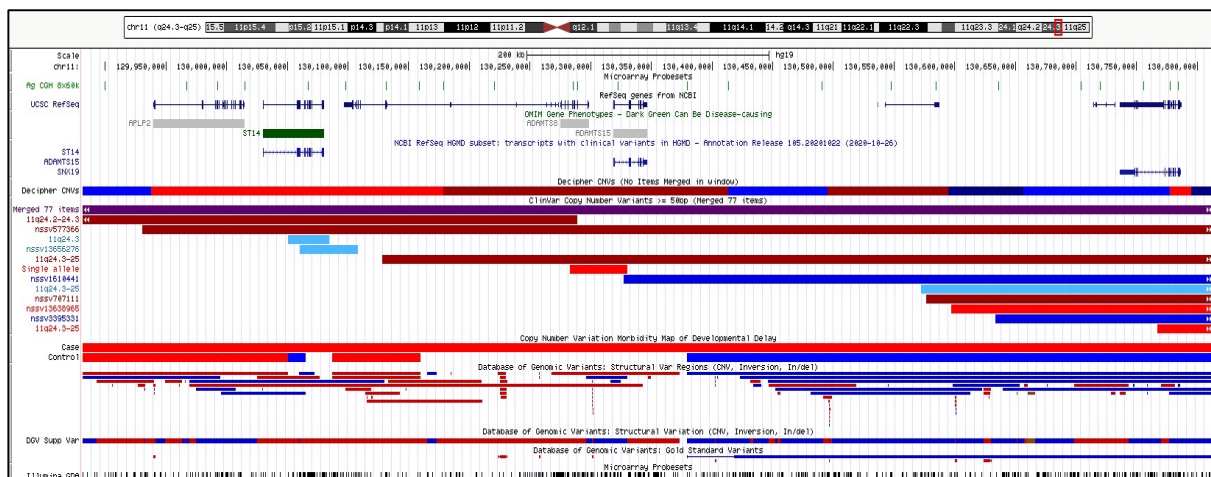


Figure 6 : Visualisation du CNV sur UCSC Table Browser.

Les gènes sont représentés sur UCSC Refseq (40), puis sur OMIM (38). Les gènes en verts sont OMIM morbid. Deux bases de données cliniques (DECIPHER (35) et Developmental Delay (36,37)) rapportent les données de patients recensées dans la littérature, en rouge les délétions et en bleu les duplications. La Database of genomic variant (DGV) rapporte les CNV recensés dans la population générale (33).

Selon les données de la littérature, l'approche par ACPA dans les déficiences intellectuelles – anomalies du développement permet d'aboutir à un diagnostic pour environ 10% des patients (41). La principale limite de l'ACPA concerne sa résolution. Pour les lames SurePrint G3 Human CGH Microarray (Agilent), l'espacement génomique médian entre deux sondes est de 41 kb. On estime ainsi que la taille moyenne minimale d'un remaniement déséquilibré détectable est d'environ 100kb.

K- Exploration par séquençage d'exome

L'approche par séquençage d'exome complet a été développée à partir d'octobre 2020 au CHU de Lille pour les explorations génétiques des thématiques de l'Institut de Génétique Médicale, prenant notamment en charge les troubles du neurodéveloppement. Cette technologie permet d'obtenir les informations génomiques concernant l'ensemble des séquences codantes du patient et de les réanalyser selon les évolutions de la littérature lorsque cela est nécessaire. Les patients ont été analysés par approche de séquençage trio pour 76%, duo pour 20% et solo pour 4% ou selon d'autres combinaisons par inclusion de parents ou autres apparentés dans la stratégie de séquençage. Il est réalisé sur le NovaSeq 6000 (Illumina).

1) Préparation des librairies

La constitution des librairies consiste à préparer une banque d'ADN représentative des échantillons à séquencer. Le terme « librairie » correspond à l'ensemble des régions génomiques d'intérêt comportant à chaque extrémité une paire d'adaptateurs nucléotidiques. Ces derniers sont spécifiques d'une technologie donnée.

L'ADN de chaque patient est coupé en fragments de taille compatible avec la technologie de séquençage. Secondairement, sont ajoutés des index et des amorces universelles. Les index sont un code barre nucléotidique unique associé à un échantillon donné. Ils permettent d'associer chaque séquence produite sur le séquenceur au bon échantillon lors du traitement bioinformatique. Les zones

exoniques et introniques proches sont capturées, représentant 1 à 1,5% du génome, soit 48 à 50Mb (42).

La préparation de la librairie est effectuée par la technique de tagmentation par le kit « *Illumina DNA Prep with Enrichment* » (Illumina) selon les recommandations du fournisseur. L'ADN des patients est découpé en fragments d'environ 400pb via des billes de tagmentation eBLT (enrichment Bead-Linked Transposome) (Figure 7) (43).

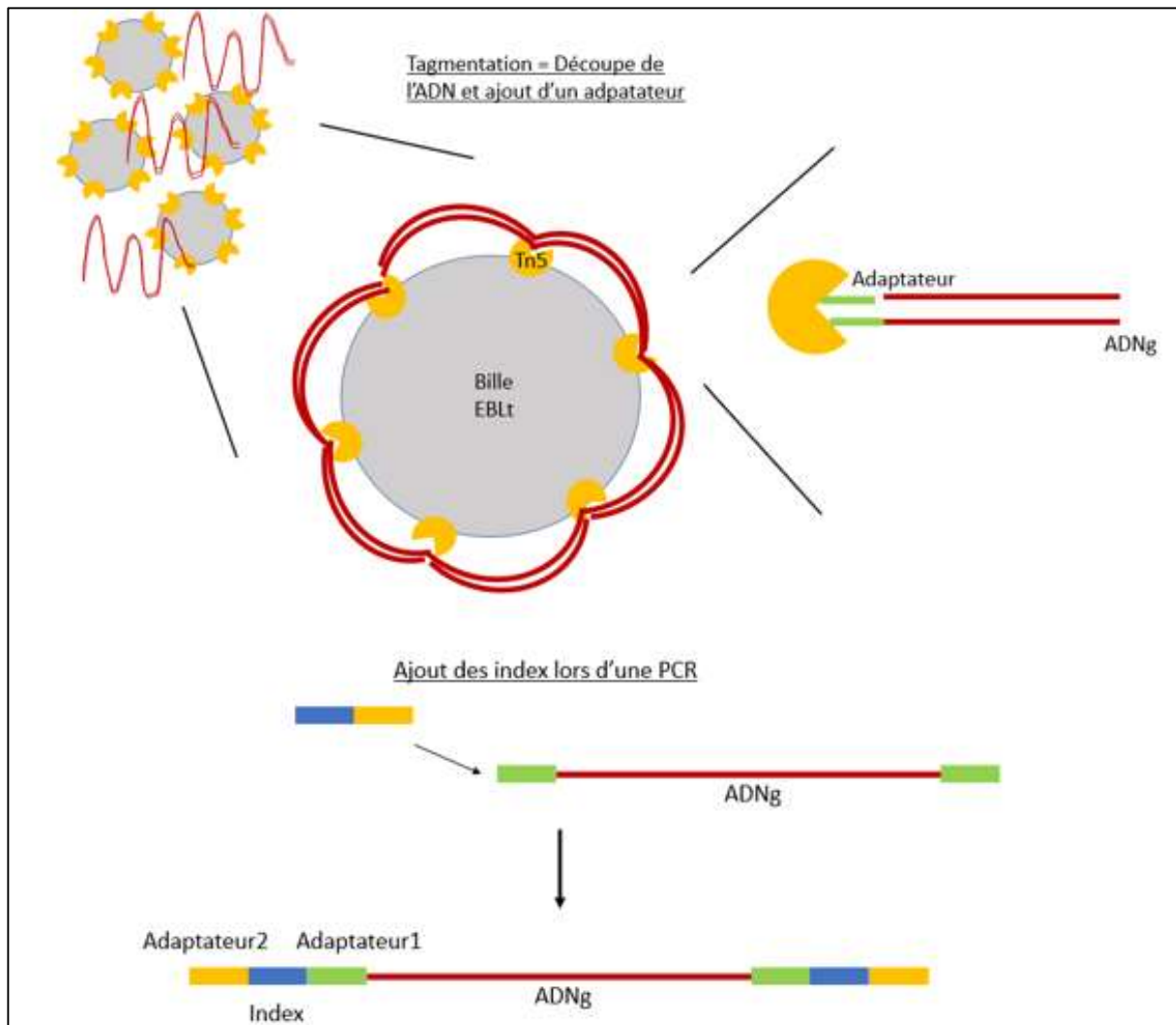


Figure 7 : **Schéma de la préparation des librairies par tagmentation d'après Illumina Nextera flex for Enrichment.**

La tagmentation sur bille permet de générer des fragments d'ADN génomique de 400pb, d'harmoniser les concentrations d'ADN des différents échantillons et d'ajouter un premier adaptateur. Lors d'une seconde réaction, est ajouté l'index, code barre nucléotidique unique pour chaque patient du run, et l'adaptateur du fragment à la flowcell : le bridge PCR.

Cette étape permet d'obtenir des fragments homogènes sur l'ensemble du génome, de rester en concentration constante d'ADN par la saturation des billes (quantité fixe de billes magnétiques ajoutée au mix), et d'ajouter des amorces universelles à chaque fragment d'ADN lors de la même réaction. Une paire de séquences adaptatrices a ensuite été liée à chaque fragment d'ADN, comportant des séquences nucléotidiques consensus qui permettent la reconnaissance de motifs nucléotidiques situés sur la lame de séquençage appelée « FlowCell » (Figure 7). Une mesure aléatoire de la concentration en ADN des librairies pré-enrichie est effectuée sur l'automate Qubit (ThermoFisher Scientific, Waltham, US-MA) en fluorimétrie.

Les régions exoniques sont ensuite hybridées à des sondes biotinylées, et capturées grâce à des billes magnétiques couplées à de streptavidine (Figure 8). On ne retient que les régions codantes et flanquantes des exons. L'ADN est dénaturé et linéarisé progressivement sur thermocycler, en partant de 95°C puis en baissant la température de 2°C toutes les minutes. A 58°C, pendant 16h, les sondes vont s'hybrider à l'équilibre. Puis une PCR d'enrichissement de la librairie exonique purifiée est réalisée ainsi qu'un lavage et une quantification de la librairie sur Qubit.

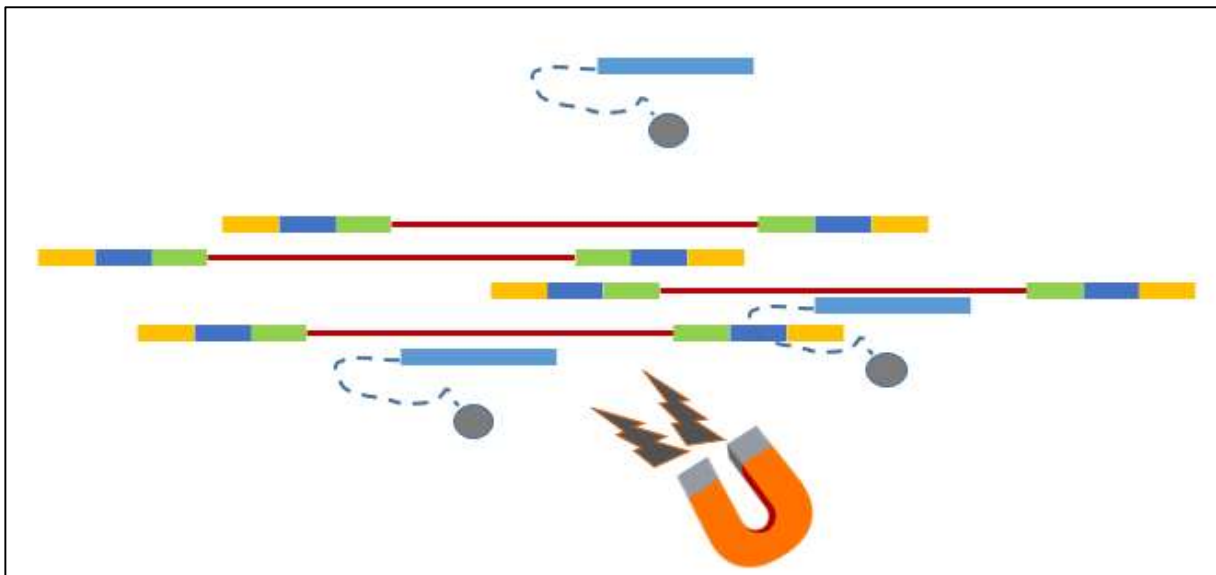


Figure 8 : Schéma de capture de la librairie exonique d'après Illumina Nextera flex for Enrichment. Les régions exoniques sont capturées via des sondes biotinylées, captées par des billes magnétiques avec de la streptavidine. La librairie subit une amplification par PCR puis une purification avant d'être quantifiée.

2) Amplification clonale : génération de cluster

L'étape d'amplification clonale est réalisée directement sur la flowcell par technique de « *Bridge amplification* » (14). Cette génération de clusters est un procédé où chaque fragment de la librairie est amplifié afin d'obtenir un signal de fluorescence optimal (Figure 9).

Les librairies poolées et diluées sont injectées dans la flowcell. Les fragments simples brins comportant des séquences adaptatrices s'hybrident à des séquences oligonucléotidiques complémentaires présentes sur la flowcell (Figure 9) (14). Une fois l'hybridation réalisée, un brin d'ADN complémentaire est synthétisé par une polymérase, le liant de façon covalente au support, puis le brin originel est dénaturé et éliminé. Ce processus se répète plusieurs fois, jusqu'à générer un « cluster » : un grand nombre de fragments d'ADN identiques dans une zone restreinte et identifiable de la flowcell (Figure 9).

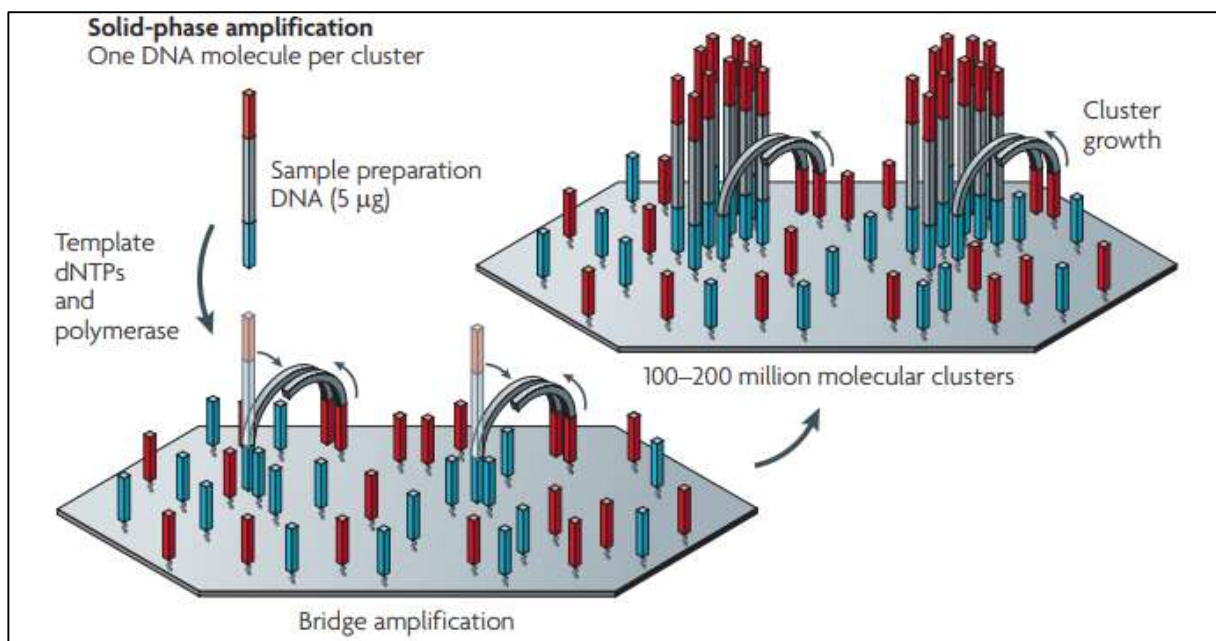


Figure 9 : **Schéma de la génération de cluster d'après Metzker and al.**

La flowcell est tapissée d'oligonucléotides complémentaires aux bridges PCR - le fragment s'hybride sur la flowcell puis le brin complémentaire est synthétisé (liaison covalente au support) - libération du brin originel, liaison en pont via les bridges PCR, synthèse d'un nouveau brin (14)

3) Séquençage

La technique de séquençage appliquée au laboratoire correspond au séquençage par synthèse (SBS). Une amorce universelle est injectée dans la flowcell, et se fixe au néobrin synthétisé complémentaire fixé. A chaque cycle de séquençage, des nucléotides marqués d'un fluorochrome spécifique sont incorporés, excités par une source lumineuse, émettant un signal de fluorescence spécifique du nucléotide. Ce signal est enregistré par le séquenceur (Figure 10) (14). Les nucléotides incorporés possèdent un terminateur de chaine réversible pour éviter l'incorporation de plusieurs nucléotides répétés. Le nucléotide est ensuite clivé pour permettre l'incorporation au cycle suivant d'un autre nucléotide spécifique (Figure 10) (14). Après le séquençage des 10 nucléotides des index puis des 151 premiers nucléotides des fragments, ceux-ci sont retournés, et séquencés en antisens, d'abord les 10 nucléotides des index, puis les 151 premiers nucléotides de la séquence (Figure 10). Cette approche est appelée « séquençage par les deux extrémités (*paired-end*) ».

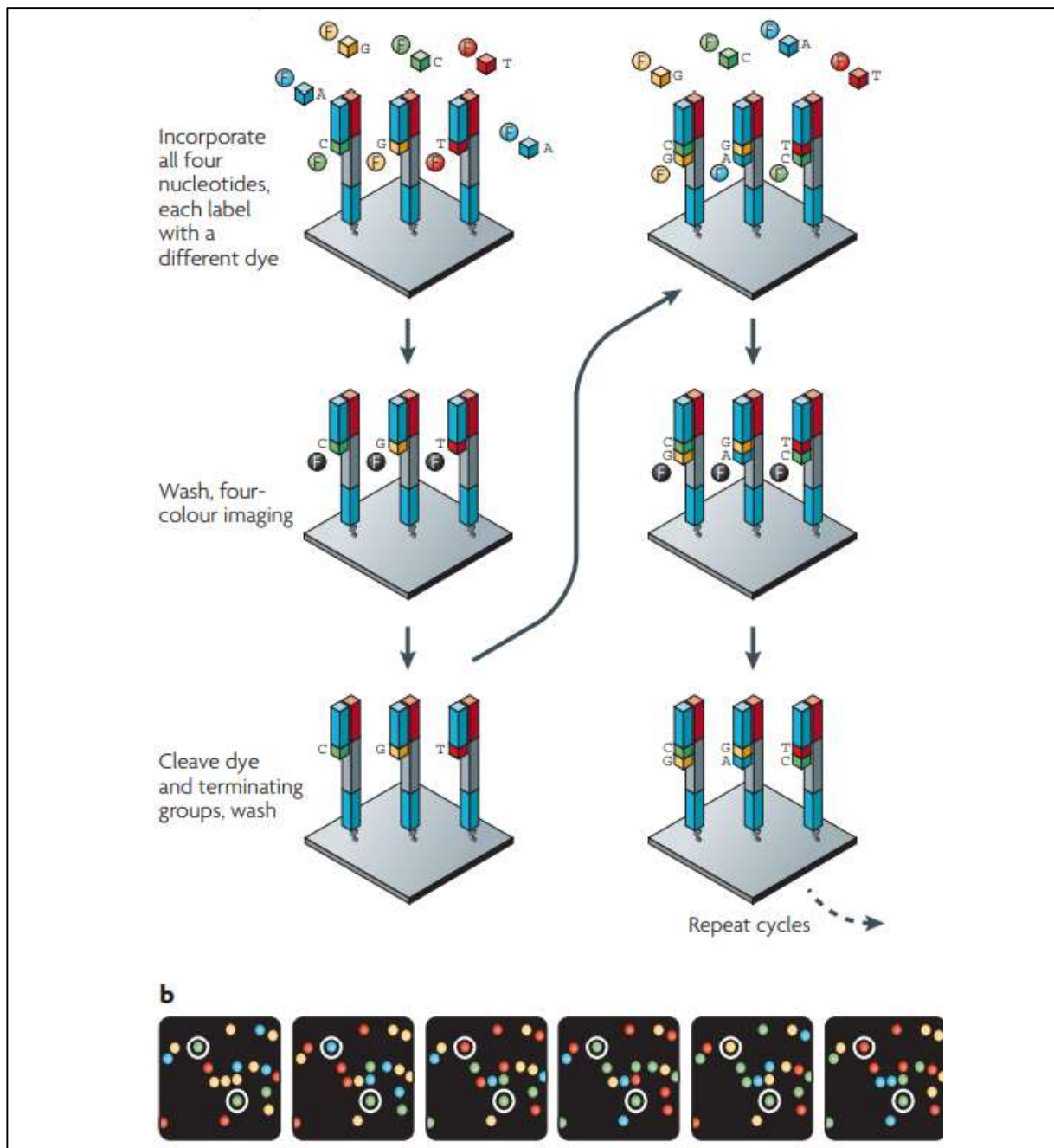


Figure 10 : Schéma de séquençage par synthèse et image d'un cycle de séquençage par synthèse selon Metzker et al (14).

A- A chaque cycle, incorporation d'un nucléotide, lavage, excitation lumineuse et capture de l'image, clivage du terminateur réversible de chaîne, lavage et le cycle est répété. B- Image de cycles de séquençage par synthèse

4) Traitement bio-informatique

Le séquençage haut débit génère des données brutes qui ne sont pas directement interprétables à la sortie du séquenceur. Des outils bio-informatiques sont essentiels au traitement et à l'analyse de ces données. Différents pipelines sont utilisés pour

obtenir une liste de variants annotés propre à chaque patient. L'analyse des données de séquençage est divisée en trois grandes parties :

- l'analyse primaire des données qui correspond à la reconstitution de la séquence d'ADN : du fichier .bcl au fichier .fastq ;
- l'analyse secondaire qui consiste à aligner la séquence retrouvée à une séquence de référence, et à identifier les variants par rapport à cette séquence de référence : du fichier .fastq démultiplexé aux fichiers .bam et .vcf ;
- l'analyse tertiaire qui englobe l'annotation des variants et l'interprétation biologique des résultats : des fichiers .vcf aux fichiers annotés par l'outil maison ANATOLE2.

a) Analyse primaire

Le fichier .bcl est un agrégat d'images du séquençage à chaque cycle de séquençage « *inject-scan-wash* ». Chaque signal lumineux détecté lors de l'incorporation de nucléotide est enregistré à une position, un instant et avec une certaine probabilité d'erreur. Ce fichier est généré par le séquenceur Illumina. L'appel de base (*base calling*) est réalisé par le logiciel Real-Time Analysis d'Illumina. L'outil « DRAGEN BCL Data Conversion » convertit les .bcl en .fastq. Ce logiciel traduit les signaux de fluorescence séquentiels en séquence, reconstituent les reads, associe à chaque base un score Phred qui évalue sa probabilité d'erreur, démultiplexe les données, c'est-à-dire, réattribue à chaque read son patient d'après l'index lu et la SampleSheet (1 index=1 patient) et génère le fichier qualité du Run (Figure 11).

Le .fastq a trois lignes : le header, la séquence, et le score Phred en langage ASCII. DRAGEN BCLconvert génère autant de paire de .fastq qu'il y a d'échantillons sur la flowcell.

b) Analyse secondaire

L'analyse secondaire est réalisée par DRAGEN Join Genotyping (JG). Il réalise le « *Trimming* », c'est-à-dire découpe les index et adaptateurs sur les reads pour ne garder que la séquence et élimine les reads avec de mauvais score Phred. DRAGEN JG aligne les reads sur le génome de référence (GRCh38) selon l'algorithme de Burrows Wheeler avec le logiciel GATK d'Illumina, c'est la génération d'un fichier BAM

(Binary Alignment Map) avant le réalignement autour des INDEL (Figure 11). Ce fichier BAM associe à chaque read ses coordonnées génomiques. Il permet d'apprécier la profondeur du séquençage : nombre de fois qu'un nucléotide est séquencé ; et la couverture : l'étalement des reads sur la zone d'intérêt. Les BAM sont visualisables sur IGV (Integrated Genome Browser) pour évaluer les critères qualités et l'environnement de variant d'intérêt. Enfin, il appelle les variants : chaque différence par rapport au génome de référence va être annotée dans un fichier VCF (Variant Caller Format) et il génère un ensemble de fichiers qualité par patient.

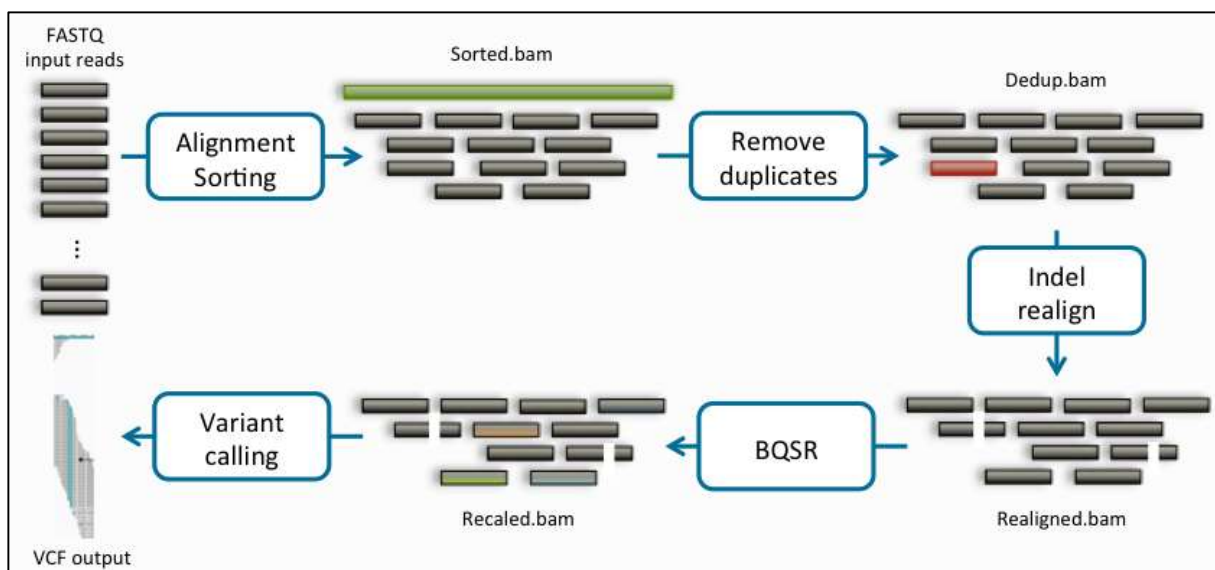


Figure 11 : **Schéma du traitement bio-informatique des données d'exome.**

L'analyse primaire, par BCL Convert traduit chaque image en base, lui associe un score Phred et démultiplexe les données dans des FASTQ. L'analyse secondaire par DRAGEN JG réalise le trimming, l'alignement au génome de référence dans le BAM et l'appel de variant dans le VCF. Le fichier Pedigree reconstitue les liens de parentés. L'analyse tertiaire agrège les différentes bases de données sur le VCF pour permettre l'interprétation des variants. Schéma issu de « Van der Auwera GA & O'Connor BD. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition)*. O'Reilly Media. » (44)

c) Analyse tertiaire

Cette analyse permet l'agrégation d'annotation des variants. Elle est réalisée par Variant Effect Predictor (VEP), avec prédiction de l'effet du variant détecté (faux-sens, épissage, non-sens...) (45) et l'interrogation de différentes bases de données (OMIM (38), CADD (46), GnomAD (32), ClinVar (47)...) (Figure 11). Au CHU de Lille, ces données sont ensuite assemblées par un outil maison appelé ANATOLE2.

5) Etude des CNV/SV par exome

Les « Structural Variant » ou SV sont des événements de structure >50pb, des CNV mais aussi des réarrangements génomiques (translocations, inversions) (Figure 3). Ils sont courants dans la population générale et sont une cause majeure de variations polymorphiques (48). Ils sont aussi mis en cause en pathologie humaine (7,12,37).

Pour détecter les SV, quatre approches existent. Le plus souvent les algorithmes les cumulent pour optimiser leur performance de détection.

La première possibilité est une approche quantitative, appelée « Read Depth » ou comparaison de profondeur de lecture. Cette analyse permet d'identifier les CNV par comparaison de la densité de lectures/reads alignés sur le génome de référence pour l'ensemble des intervalles préalablement définis. Ces intervalles correspondent le plus souvent aux bornes exoniques pour le séquençage d'exome (Figure 12) (28).

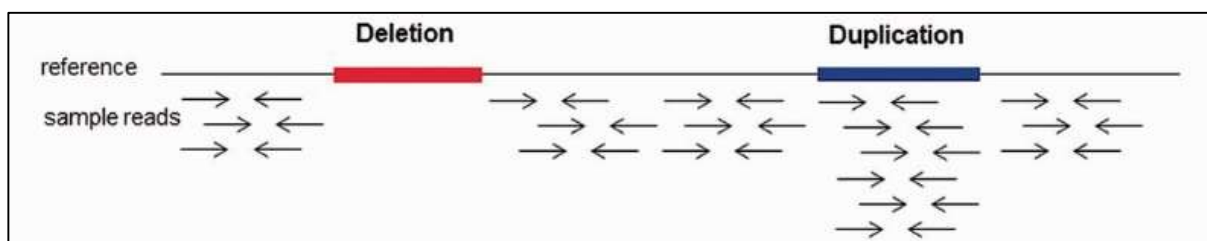


Figure 12 : Représentation de la détection de CNV grâce à la comparaison de profondeur de lecture selon Escaramis et al (28).

Il y a moins de reads alignés dans les délétions, et à l'inverse, plus dans les duplications. Les bornes en analyse d'exome correspondent souvent aux limites de l'exon.

Cette approche présente quelques limites, comme la faible résolution et localisation des points de cassures, et la sensibilité aux biais d'enrichissement en bases AT ou GC du génome. Elle est utilisée par un grand nombre d'outils appliqués à l'exome ou aux panels tels que CANOES (49), CNVnator (détecte les variations nettes de profondeur) (50) ou CNVkit (51). Les outils « Read Depth » présentent une stratégie de détection des CNV similaires à l'approche utilisée en ACPA.

La deuxième possibilité implique le séquençage par paire, appelé « Paired-Read ». Il s'agit d'identifier les alignements aberrants entre les deux lectures/reads d'un même couple, suggérant la présence d'un remaniement entre ces deux lectures/reads (Figure 13). Cette approche utilise les propriétés du séquençage. Au CHU de Lille, nous réalisons un séquençage en paire de 2x150 pb de chaque fragment d'ADN. La fragmentation découpant le génome en fragment d'environ 400pb, nous attendons une distance entre lecture/read sens et lecture/read antisens d'environ 100 pb en moyenne.

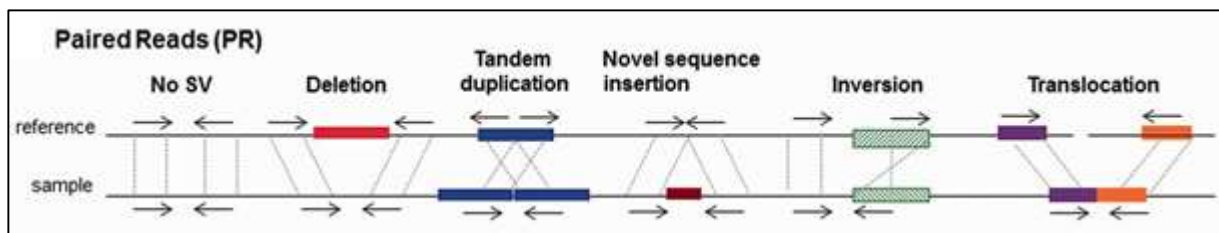


Figure 13 : *Illustration selon Escaramis et al de la detection de SV grâce aux paired-end.*

Selon l'écartement et l'orientation des reads, nous pouvons déduire de quel type de SV il s'agit (28).

En présence d'un éloignement entre les deux lectures/reads, nous évoquons une délétion. En présence d'un rapprochement, nous évoquons une duplication. En présence d'un changement d'orientation entre échantillon et référence, nous évoquons une inversion ou une duplication en tandem (Figure 13). L'outil historique Breakdancer utilise cette approche (52).

La troisième approche concerne l'étude des lectures/reads chimériques, appelée « *Split Reads* » ou « *Clip Reads* ». Ces reads imparfaitement alignés mettent en évidence des points de cassure (Figure 14). Split-read alignments : certains reads ne sont pas parfaitement alignés sur la séquence de référence, les cibler peut permettre de détecter des points de cassures, et éventuellement de localiser l'évènement en mappant le split read (Figure 14).

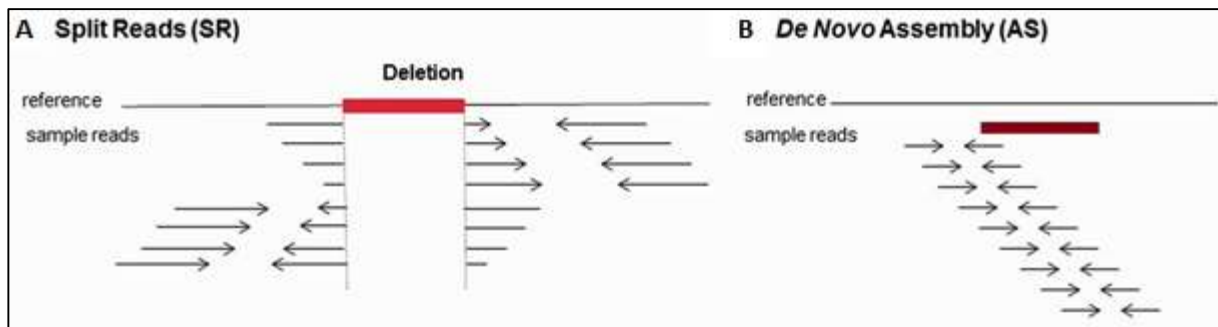


Figure 14 : **Représentation du split read et de l'exome de novo pour identifier des SV selon Escaramis et al (28).**

A. Représentation de la détection de SV grâce au split reads. En identifiant la séquence splitée, on peut retrouver de quoi est composé le SV. B. Schéma du De novo assembly.

La dernière approche, moins consensuelle dans le cadre du séquençage court fragment, concerne l'assemblage de novo du génome. Cette approche très consommatrice en ressources informatiques, reconstitue le génome *de novo*, sans alignement sur un génome de référence, grâce à la continuité des reads. L'outil TIGRA utilise cette approche (Figure 14) (53). Cette approche est plus adaptée au génome qu'à l'exome.

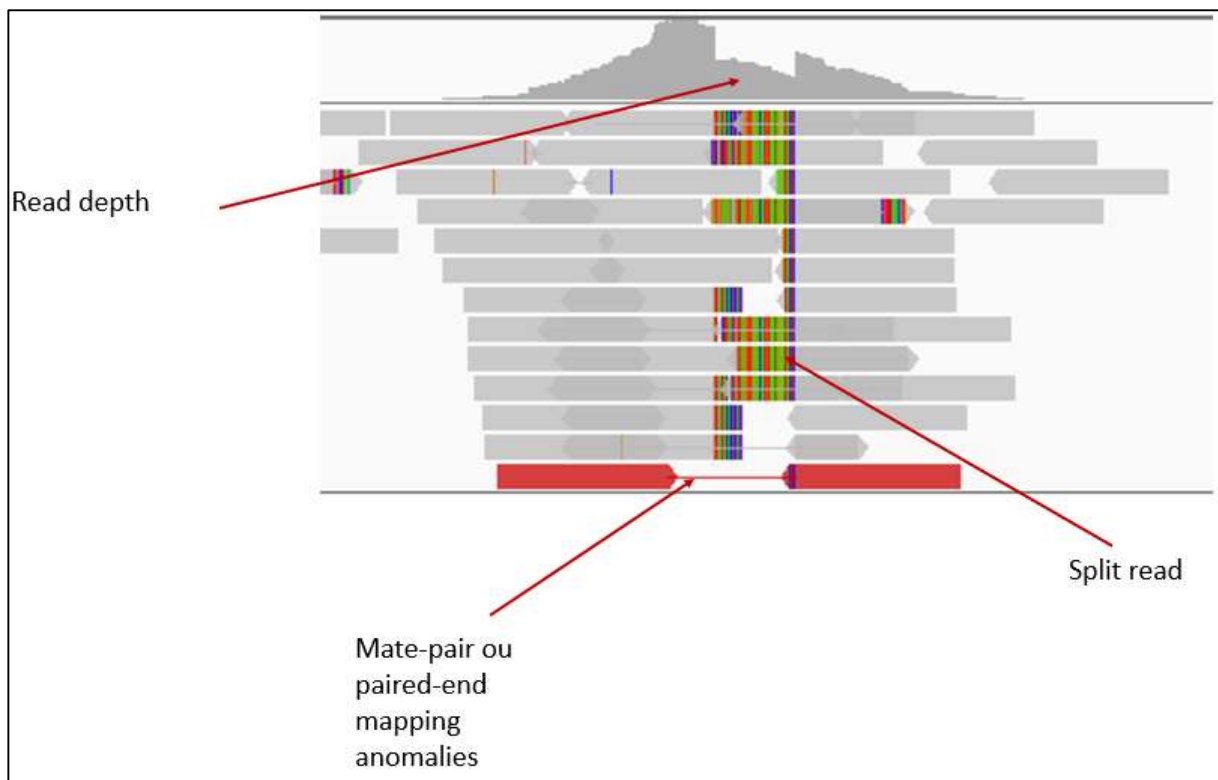


Figure 15 : **Schéma des différentes méthodes de détection des SV visualisé sur IGV (54).**

La chute de profondeur de couverture repérée par la diminution du nombre de reads dans la zone montre le fonctionnement de CNV DRAGEN. Le soft-clipping, soit les reads imparfaitement alignés à la

séquence de référence sont représentés par les reads chimériques : partie grise bien alignée au génome de référence, chaque segment de couleur représente un nucléotide divergent de la séquence de référence ; le paired-end mapping correspond aux reads reverse et forward d'un même fragment qui ne sont pas séparés d'une distance adéquate : un rapprochement signe une duplication alors qu'un éloignement signe une délétion. Le soft-clipping et le paired-end sont utilisés par SV DRAGEN dans la détection des CNV.

Integrative Genomic Viewer (54) est un outil essentiel pour la visualisation des SV (Figure 15). Il permet de discriminer grâce à un code couleur avec le pair-end des inversions, délétions, duplications, d'observer visuellement des variations de profondeur dans des régions, de mettre en évidence des rétrotransposons... Des exemples illustreront l'utilisation d'IGV au cours de ce travail.

6) Outil DRAGEN Copy Number Variant Calling

DRAGEN CNV est un pipeline bio-informatique, proposé par Illumina, utilisant les comparaisons de profondeurs de lecture pour l'ensemble des régions d'intérêt (Figure 16). Il s'adapte à l'exome en ne s'intéressant qu'aux régions capturées (24).

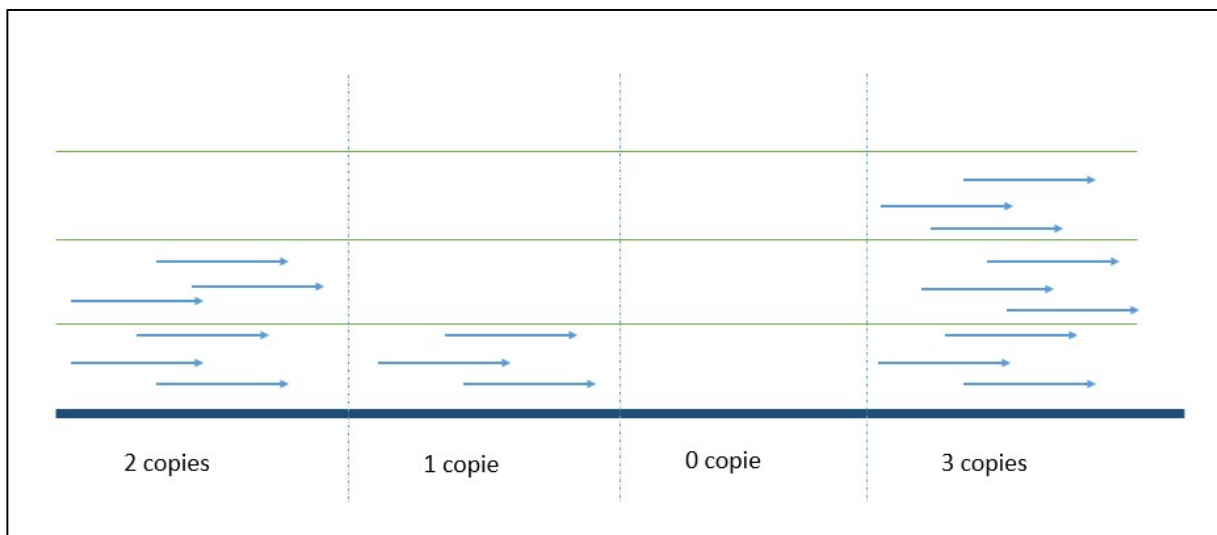


Figure 16 : **Schéma du fonctionnement de CNV DRAGEN d'Illumina.**

La profondeur d'une région est comparée à la profondeur moyenne de l'individu et à ces pairs sur la même flowcell pour déterminer s'il y existe un CNV à cet endroit.

DRAGEN CNV est paramétré pour identifier les CNV de plus de 500pb, les autres CNV de plus petite taille doivent être appelés avec les SNV/INDEL sur le VCF Caller. Ce pipeline travaille sur les .bam d'une même flowcell. Chaque échantillon va subir un pré-traitement de ses données avant l'appel de CNV. D'abord, l'algorithme compte les profondeurs de lecture des cibles, puis, il corrige les biais intrinsèques (biais d'enrichissement en GC, défaut de couverture...), il normalise ensuite la profondeur par la ploidie attendue de l'échantillon et lisse la profondeur de lecture sur l'ensemble des zones capturées. En effet, le programme prend en compte le nombre de reads de chaque individu, les différences de profondeur de séquençage en fonction des régions (régions plus difficiles à séquencer que d'autres) (figure 16 et 17). Les zones d'intérêt sont ensuite délimitées par segmentation anormale /normale et les CNV sont appelés, scorés et qualifiés.

Les CNV identifiés sont ensuite annotés par AnnotSV (25). Cet outil compile les informations de différentes bases de données : les bases de population générale, de patients, les gènes inclus dans l'anomalie, les éléments régulateurs, les TADs... Et attribue une note de -1 à 1 traduite en classe ACMG de 1 à 5 selon les critères ACMG et Clingen (Figure 17) (34,39).

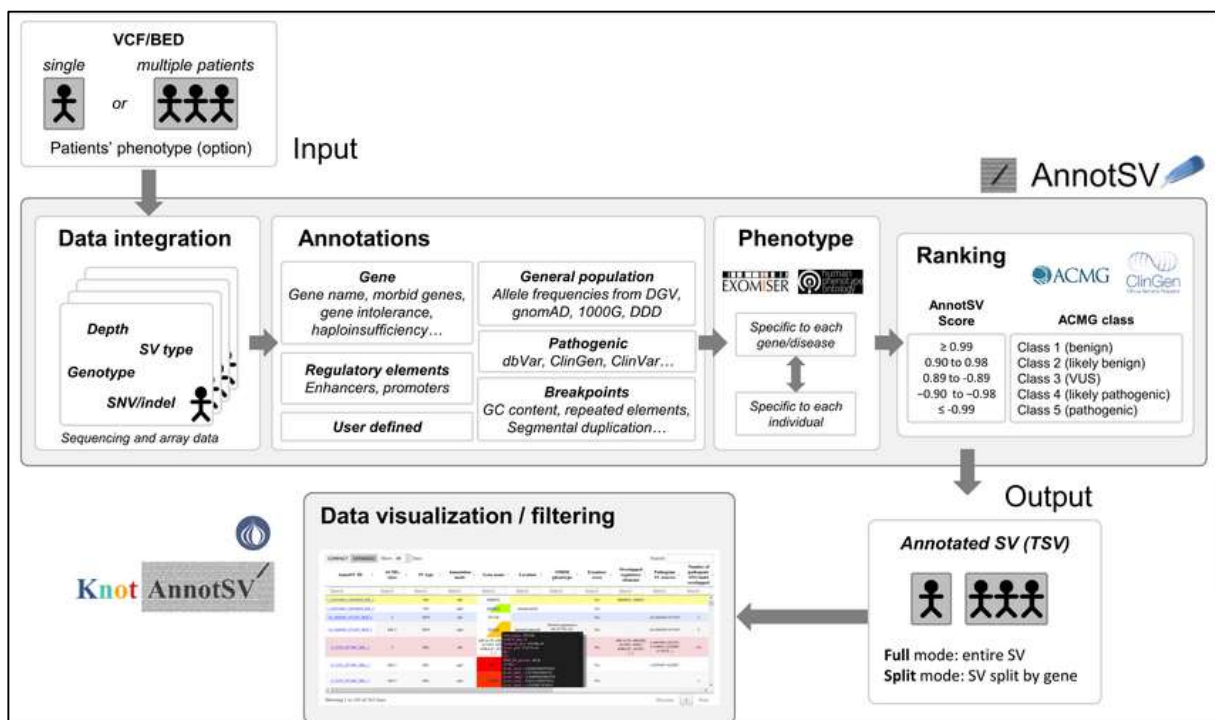


Figure 17 : **Schéma du raisonnement AnnotSV** (25).

AnnotSV annote les CNV en cherchant les gènes inclus dans la variation et leurs scores *in silico* d'haploinsuffisances / de triplosensibilité, les gènes OMIM Morbid, les éléments régulateurs, la

réurrence dans la population générale ou dans des populations de patients, attribue un phénotype aux gènes inclus dans la région grâce aux termes HPO recensés et propose une classification type ACMG.

7) Outil DRAGEN Structural Variant Calling

SV DRAGEN se base sur l'analyse du paired-end et sur le soft clipping pour repérer les SV de plus de 50pb et les points de cassures (Figure 14 & Figure 15). Il recherche les SV en deux étapes : d'abord, il scanne le génome pour construire toutes les associations break-end possibles puis il les analyse et les annote. Il reconstruit diverses structures sur l'exome et attribue un score qualité qui correspond à la probabilité d'existence du SV (26) (Figure 14 & Figure 15). Ces variations sont rapportées sur un fichier format VCF.

L'analyse des fichiers issus de SV DRAGEN est réalisée selon le mode d'hérédité, le score qualité, les scores d'haploinsuffisance des gènes d'intérêt et la récurrence des événements. Les critères ACMG sont appliqués (39).

8) Outil DRAGEN ROH Caller

L'analyse « Regions of homozygosity » (ROH) ou « Régions d'homozygotie » développée par DRAGEN Illumina tool est un outil qui permet de faire ressortir les zones d'homozygoties contigües entre les deux allèles d'un chromosome (27,55).

Ce fichier ROH nous permet de mettre en évidence les isodisomies uniparentales ainsi que les délétions. Les hétérodisomies ne sont pas mises en évidence car cette analyse n'est pas effectuée en trio, elle ne repère pas les zones d'homologies avec les parents, seulement les zones de perte d'hétérozygotie. Les patients issus de parents apparentés possèdent plus de zones d'homologies en raison de leur identité par descendance.

Cette recherche, qui s'appuie sur les VCF (Variant Call Format), passe par le « Small Variant Caller ». Elle exclue les INDEL, les variations complexes ou multi alléliques et les chromosomes sexuels. L'algorithme cible d'abord une région qui contient au moins 50 SNV homozygotes consécutifs sans variant hétérozygote ou trou de plus de 500 kb

entre les variants. Il réalise ensuite un score en ajoutant 0,025 à chaque SNV homozygote, et d'importantes pénalités pour chaque variant hétérozygote rencontré. Cela permet une tolérance pour les erreurs de séquençages. La région s'élargie jusqu'à rencontrer une lacune de plus de 500 kb. Les régions qui se chevauchent sont fusionnées, il n'y a pas de taille maximale mais les régions se terminent aux limites des chromosomes. Il n'y a pas de chevauchement sur des zones de plus de 3Mb d'écart (27) (Figure 18). Comme nous travaillons sur exome, les bornes de la perte d'hétérozygotie ne sont qu'estimées : les points de cassures sont le plus souvent introniques.

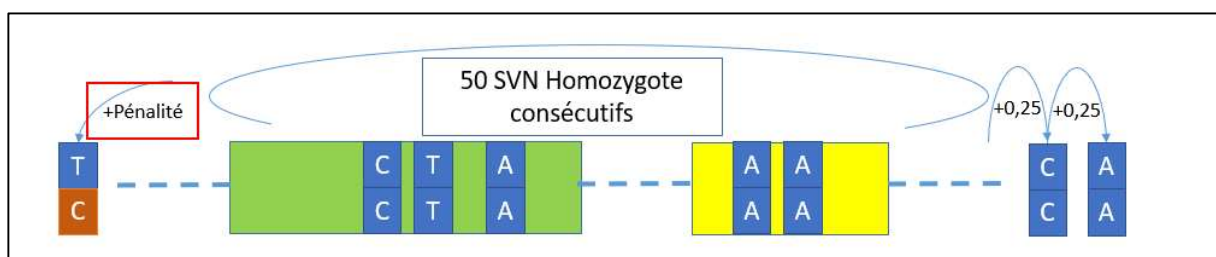


Figure 18 : **Schéma du fonctionnement de l'algorithme ROH mis au point par DRAGEN.**

L'algorithme cible 50 SNV homozygotes consécutifs puis réalise un score en ajoutant 0,025 à chaque SNV homozygote, et d'importantes pénalités pour chaque variant hétérozygote rencontré. Il s'arrête lorsqu'il rencontre une lacune de plus de 500 kb et fusionne les zones chevauchantes.

9) Interprétation des variants

Les variants sont ensuite interprétés selon les critères ACMG (39). Les variants ponctuels et CNV sont interprétés selon les modes de transmission, la fréquence dans la population générale (gnomAD (32)), la présence dans les bases de données pathogènes (ClinVar (47), HGMD), les prédictions de pathogénicité *in silico* (CADD (46), REVEL (56), Mistic (57)), l'impact sur l'épissage (SpliceAI (58), SPIP (59)), les modélisations protéiques (AlphaFold (60), Missense3D (61)), la corrélation entre les terminologies HPO associées à chaque patient et les différents gènes candidats (OMIM (38), SysID, Pubmed) (62).

E- Validation des CNV/SV et SNV

Selon le type et la taille du remaniement de structure, une validation par séquençage Sanger ou par qPCR a été réalisée. Les amorces Sanger et qPCR ont été designées grâce à différents sites : UCSC pour repérer la zone et la séquence d'intérêt, Primer3 pour designer automatiquement des amorces compatibles dans la zone et dans les conditions demandées, SNPCheck de Genetools pour vérifier l'absence de SNP dans la zone et la parfaite identité de la séquence.

1) Séquençage Sanger

Les SNP d'intérêt qui ont des critères qualités trop faible (couverture <100X, zone avec des erreurs de séquençage, biais de brins...), les ségrégations familiales, sont confirmés par Sanger. Le séquençage Sanger est une technique par synthèse. Il est réalisé au laboratoire selon les recommandations du fournisseur, BigDyeTM v3.1 Cycle sequencing kit Thermofisher.

2) qPCR

Les variations du nombre de copie sont systématiquement vérifiées par qPCR (PCR quantitative). On réalise un mix de qPCR : amorces, solution tampon avec des nucléotides, enzyme polymérase et un agent qui s'incorpore à l'ADN double brin (SYBR Green). S'ensuit des cycles de dénaturation-hybridation-élongation où la quantité d'ADN est mesurée en temps réel via un spectrophotomètre (63). La normalisation de la quantité d'ADN se fait grâce au gène *RPPH1*, gène de ménage (housekeeping gene).

III-Résultats

L- Détection des CNV : comparaison ACPA versus Exome

Dans un premier temps, nous avons comparé l'approche de détection des CNV par exome par rapport à l'ACPA qui est actuellement la méthode de référence. Nous avons comparé les données de 68 patients porteurs d'un CNV caractérisé en ACPA, et ayant secondairement bénéficiés d'un exome. Les coordonnées de CNV sont reprises dans « l'Annexe 1 : Tableau de comparaison ACPA versus exome ».

Les puces d'ACPA utilisées ont une résolution de détection minimale de 60K. L'analyse des CNV en exome se fait avec l'outil CNV DRAGEN à partir des fichiers d'alignements de séquences .bam. Les CNV appelés ont été annotés avec l'outil AnnotSV (25). Les CNV-ACPA ont secondairement été recherchés manuellement parmi les CNV-exome appelés par l'outil DRAGEN CNV pour chaque patient.

76 CNV ou déséquilibres chromosomiques avaient été préalablement identifiés en ACPA (Annexe 1 : Tableau de comparaison ACPA versus exome). On recense 6 aneuploïdies et 70 CNV. Parmi les CNV, la taille médiane est de 484 kb, le minimum à 43 kb et le maximum à 15150 kb. Il y a 60% de duplications et 40% de délétions, portées à 81% sur les autosomes (n= 57) et à 19% sur les gonosomes (n= 13). Les aneuploïdies détectées sont toutes sur les gonosomes.

Au niveau chromosomique, 5/6 des déséquilibres (83%) impliquant des gonosomes identifiés en ACPA sont appelés par séquençage d'exome (Figure 19) :

- Deux patients porteurs d'un syndrome de Jacob, 47, XYY, homogène : $\text{arr}(Y)x2$;
- Deux patients porteurs d'un syndrome de Klinefelter 47, XXY homogène : $\text{arr}(X)x2,(Y)x1$;
- Une patiente porteuse d'un triple 47, XXX homogène : $\text{arr}(X)x3$.

Seule la perte d'un chromosomique X d'une patiente avec un Syndrome de Turner en mosaïque entre 5 et 10% dans le sang n'est pas appelée en exome : arr(X)x1[0.1].

Un caryotype et de la FISH ont été réalisé en parallèle :

mos 45,X[1]/46,XX[18].ish X(wcpX,SHOX,DXZ1)x1[5]/X(wcpX,SHOX,DXZ1)x2[79]
 .nuc ish(SHOX,DXZ1)x1[12/206]

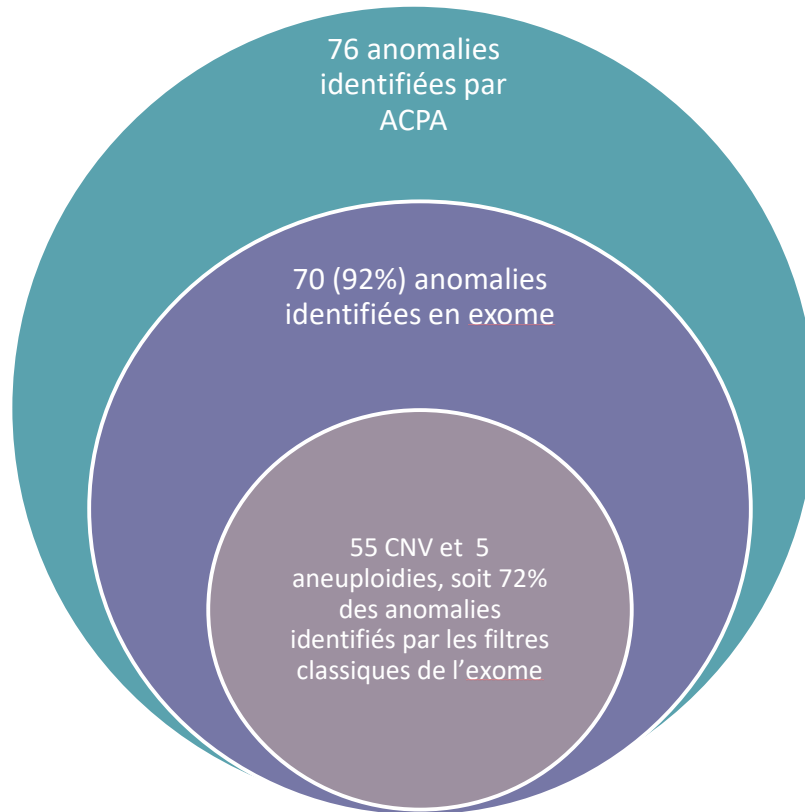


Figure 19 : Comparaison de la détection d'anomalie quantitative génomique en ACPA versus Exome.

Sur les 76 anomalies identifiées en ACPA, 92 % sont identifiées en exome et 72% sont identifiées avec les « filtres classiques ».

Sur les 70 CNV identifiés en ACPA, 65 (93%) sont également observés parmi les variations appelées en exome (Figure 19). Les cinq CNV non retrouvés en exome concernaient :

- Deux CNV, une duplication et une délétion, en mosaïques à 20% dans le sang sur le chromosome Y, chez un même patient :
 arr[GRCh37]Yp11.31q11.221(2656461_16057257)x2[0.2],
 arr[GRCh37]Yq11.221q12(16188682_59335913)x0[0.2] ;

- Une duplication hétérozygote de 390kb en 2q24, chez un quatuor, 2 frères et le père porteurs, la mère indemne : arr[GRCh37] 2q24.3(167602904_167992598)x3.
- Une duplication hétérozygote de 624kb en 3p26, non héritée maternelle, chez un patient DI dans un cadre syndromique : arr[GRCh37] 3p26.3(1638453_2262692)x3.
- Une délétion de 69kb en 9p24, non héritée maternelle, chez un patient présentant un décalage des acquisitions : arr[GRCh37] 9p24.3(500525_569970)x1.

Les évènements : arr[GRCh37]Yp11.31q11.221(2656461_16057257)x2[0.2], arr[GRCh37]Yq11.221q12(16188682_59335913)x0[0.2] en mosaïque à 20% dans le sang chez un patient ne sont pas retrouvés en exome. La normalisation de couverture sur l'Y est plus délicate car environ la moitié des individus sur la flowcell sont des garçons, et le nombre de cibles exoniques sur le chromosome Y est plus restreint que sur d'autres chromosomes. Il y a 14 286 cibles pour le chromosome X contre 1048 cibles sur Y.

La duplication arr[GRCh37] 2q24.3(167602904_167992598)x3 n'est pas retrouvée chez le frère du cas index. Après analyse des fichiers .bam sur IGV, la région dupliquée ne comprend que deux exons : l'exon 1 mal couvert et l'exon 2 de *XIRP2* de 410pb avec une grande zone intronique (Figure 20A). En faisant manuellement les ratios de profondeur de lecture l'exon 2 de *XIRP2* du quatuor et des d'exons, bien couverts, flanquant la duplication : l'exon 2 de *SCN7A* et exon 3 de *XIRP2* (Figure 20B), nous retrouvons un évènement à type de duplication. Cependant, nous sommes en dessous de la résolution de CNV DRAGEN de 500pb.

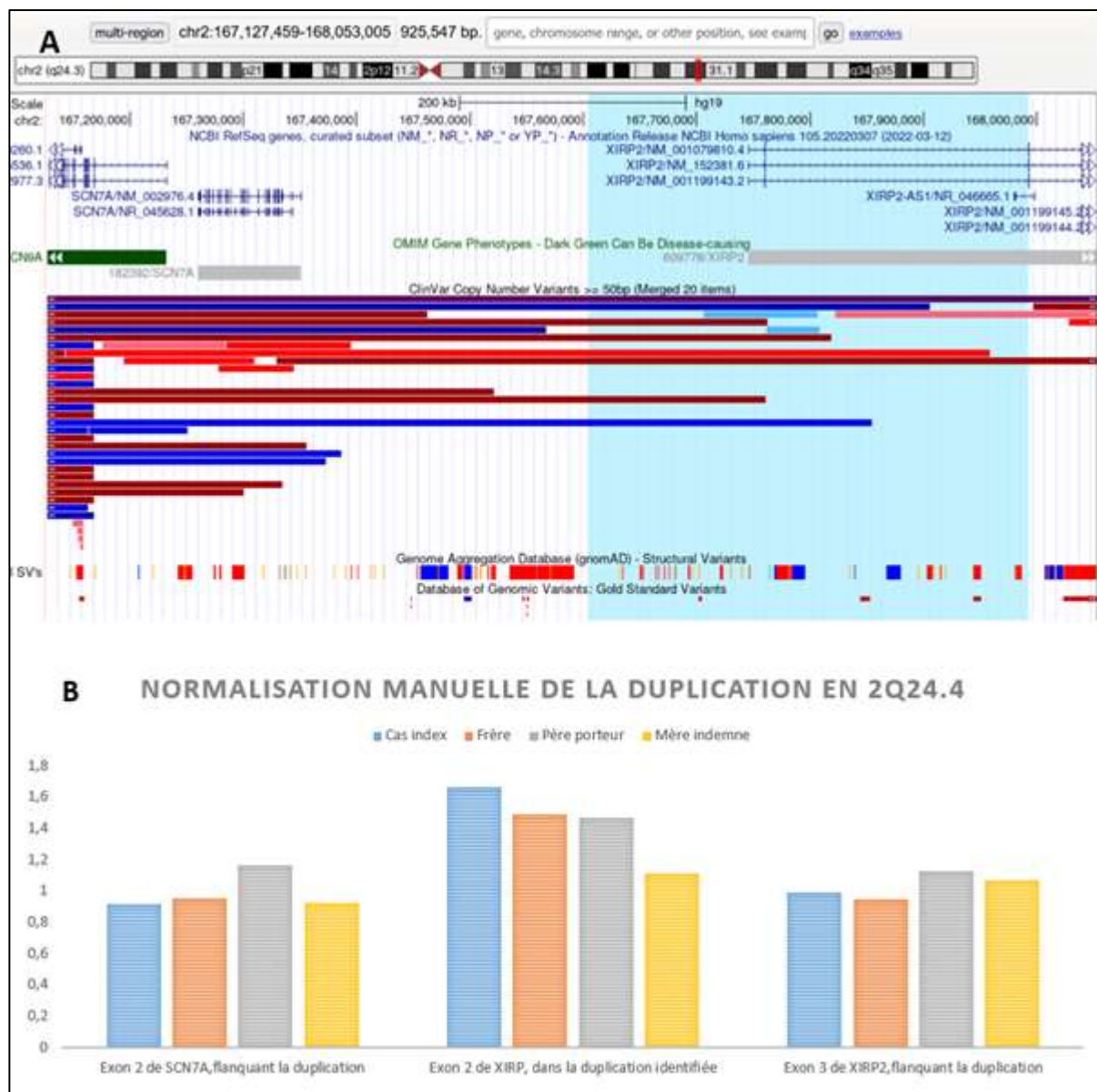


Figure 20 : **Duplication arr[GRCh37] 2q24.3(167602904_167992598)x3 identifiée chez le quatuor.**
A : Visualisée sur le Table Browser d'UCSC, surlignée en bleu, la duplication arr[GRCh37] 2q24.3(167602904_167992598)x3 identifiée chez le quatuor, comprenant les exons 1 et 2 de XIRP2. L'environnement de cette duplication n'est pas surligné. Cette duplication n'est pas retrouvée dans la population générale ou dans une population de patients (gnomAD et ClinVar), XIRP2 n'est pas OMIM Morbid. L'exon 2 de SNC7A et l'exon de 3 de XIRP2 flanquent la duplication. B : Normalisation manuelle de la région de XIRP2 dupliquée et capturée en exome, ainsi que des exons flanquants la duplication : l'exon 2 de SNC7A et le 3 de XIRP2. Le cas index, son frère et son père, porteurs de la duplication en 2q24.4 présentent un ratio à 1,5 de l'exon 2 de XIRP2 et des ratios normaux des exons flanquants, et leur mère, indemne de la duplication, présente des ratios normaux pour l'exon 2 de XIRP2 et les régions flanquantes.

La duplication en arr[GRCh37] 3p26.3(1638453_2262692)x3 n'est pas retrouvée chez le cas index en exome. La région, surlignée en bleu sur le Table browser d'UCSC, ne

M-Identification des CNV sous le seuil de résolution de l'ACPA

L'exome a permis l'identification de CNV, avec l'outil unique CNV DRAGEN, sous le seuil de résolution de l'ACPA. Il a aussi permis de trouver des combinaisons composites CNV/SNV expliquant le phénotype du patient (SNV en regard, CNV sous le seuil de résolution de l'ACPA).

Parmi 7 CNV, identifiés par CNV DRAGEN, toutes des délétions, nous avons une taille médiane de 35kb, avec maximum à 54kb et un minimum à 4kb.

Patient 1 : Délétion homozygote de 54kb en 16q21 impliquant le gène KATNB1

Ce patient, dont les parents sont apparentés, a présenté, en anténatal, une microcéphalie avec un aspect court du corps calleux et un retard de gyration. La naissance a été déclenchée à 33 SA dans un contexte de prééclampsie. Agé de 3 ans, il présente un retard neurodéveloppemental global avec une hypotonie axiale importante, un nystagmus pendulaire, une microcéphalie, des malformations des extrémités et des malformations génitales.

Ce patient a bénéficié d'une analyse d'exome complet en trio. Une délétion homozygote impliquant *DRC7* et la partie 3' de *KATNB1*, de 54kb, a été retrouvée : seq[GRCh38]del(16)(q21) NC_000016.10:g.57688349_57741861del (Figure 22). *KATNB1* est impliqué dans un syndrome provoquant une lissencéphalie et une microcéphalie #MIM616212. Cette délétion n'est jamais retrouvée à l'état homozygote dans la population générale, des anomalies dans ce gène sont décrites dans la littérature scientifique (64,65) ; KTNB1 est une protéine associée à la libération des microtubules, phénomène capital pour la croissance axonale. Cette délétion a été classée en probablement pathogène. La région génomique impliquée par la délétion n'est couverte que par une sonde en ACPA, et n'était pas détectable par l'algorithme utilisée au laboratoire.

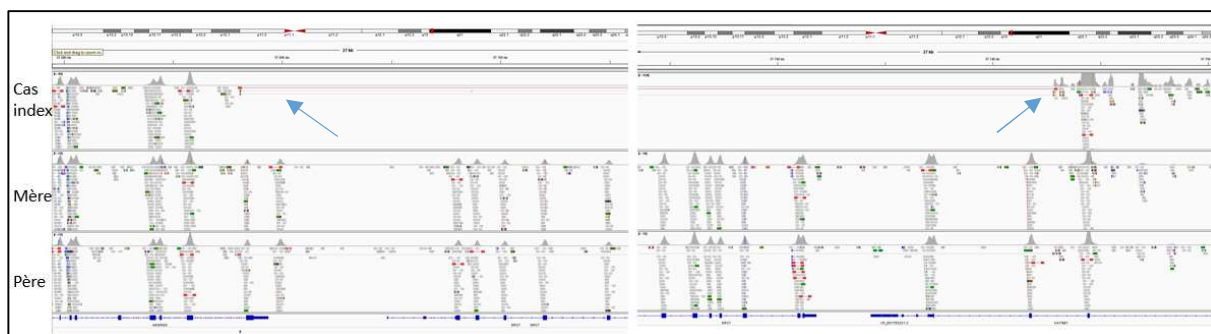


Figure 22 : Délétion homozygote de 54 kb emportant la partie 3' du gène *KATNB1* chez P1 sur IGV.

Un read sur le point de cassure a été capturé, permettant de borner précisément cette délétion. La délétion est suspectée devant la combinaison d'une profondeur à 0 comprenant le premier exon du gène *ADGRG3*, les 17 exons de *DRC7* et les trois premiers exons de *KATNB1*, associée à une augmentation de la taille de l'insert pour 3 paires de reads capturés et chevauchant le point de cassure. Les paires de reads du paired-end bornant la délétion sont signalées par des flèches oranges.

Patient 2 : Délétion hémizygote de 40kb en Xq26.2 impliquant le gène *HPRT1*

Ce garçon de 3 ans est issu d'une fratrie de 7, sans antécédent médicaux. Il est né eutrophe à 40 SA après une grossesse sans particularité. Dans un contexte de retard des acquisitions et de tétraparésie spastique, un bilan métabolique avait révélé un syndrome de Lesch Nyhan chez ce garçon. Un exome trio a retrouvé une délétion hémizygote d'environ 40kb en seq[GRCh38]del(X)(q26.2)dn NC_000023.11:g.134460165_134486530del, de survenue *de novo* dans le gène *HPRT1* (#MIM300322), responsable de ce syndrome (Figure 23A). La délétion des quatre premiers exons n'est pas retrouvée dans la population générale, une délétion chevauchante, en NC_000023.11:g.134460292_134500097del, de 40kb, comprenant 9 exons du gène *HPRT1* est décrite dans la base ClinVar chez un patient atteint du syndrome de Lesch Nyhan (RCV000631416). Cette délétion est classée pathogène.

L'ACPA n'était pas informative pour ce patient, la délétion était sous le seuil de résolution. Après lecture *a posteriori* des profils d'ACPA, nous observons une seule sonde déviante sur ce gène (Figure 23B)

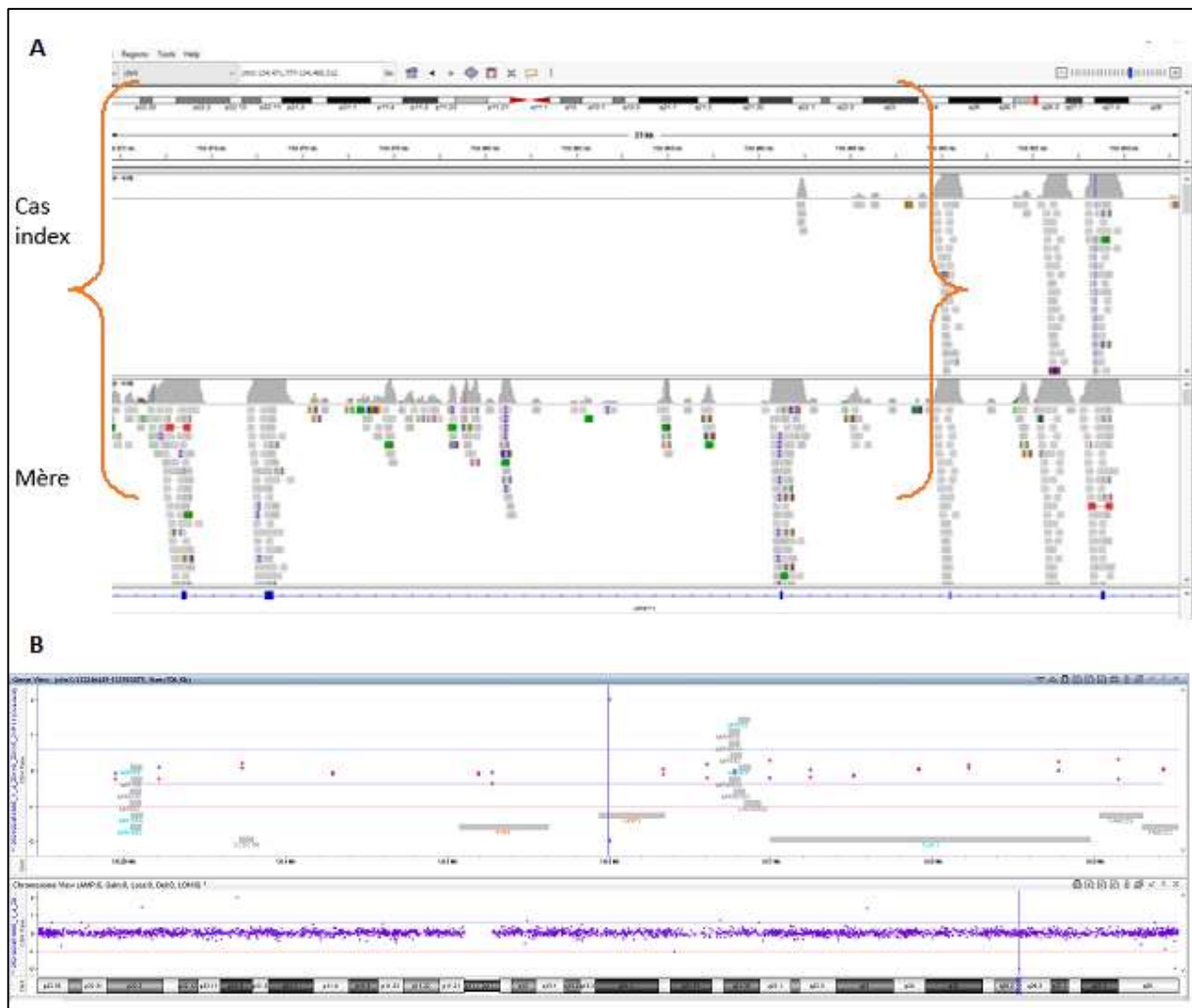


Figure 23 : A : Délétion hémizygote de survenue de novo, d'environ 35kb, dans le gène *HPRT1* responsable du syndrome de Lesch Nyhan chez P2, visualisée sur IGV : seq[GRCh38]del(X)(q26.2) NC_000023.11 :g.(?_134460165)_ (134486530_?)del.

La délétion est suspectée devant l'absence de reads de l'exon 1 à 4 dans le gène *HPRT1*. B : Mise en évidence d'une sonde déviante située dans l'intron 2 de *HPTR1* en chr(X):134473833-134473893 sur Cytogenomics, le logiciel de visualisation de l'ACPA. Les sondes n+1 et n-1 sont associées à un log2 du nombre de copies de la région à 0, et considérées comme non déviantes.

Patient 3 : Délétion hétérozygote de novo de 4kb en 7q36.1 de 4 exons du gène *KMT2C*

Ce garçon de 6 ans présente un trouble du spectre autistique. Il est né à 36+6 SA d'une grossesse sans particularité. Il n'y a pas d'antécédents médicaux particuliers dans sa famille. Les inquiétudes commencent vers 2 ans devant une absence de langage et une régression du développement. On retrouve en exome complet une délétion partielle de *KMT2C* (#MIM617768) responsable du syndrome de Kleefstra.

Cette délétion de 4kb emporte 4 exons (du 52 au 55/59) de *KMT2C* (Figure 24), en seq[GRCh38]del(7)(q36.1) NC_000007.14:g.152144710_152149030del. Cette délétion n'a jamais été rapportée dans les bases de données de population générale ni celle de trouble du neurodéveloppement (ClinVar (47), DECIPHER (35), Developmental Delay (36,37)). Ce patient avait bénéficié d'une ACPA rendue négative au préalable, aucune sonde informative ne couvre cette région d'après UCSC table Browser. Une qPCR a confirmé cette délétion, classée pathogène.

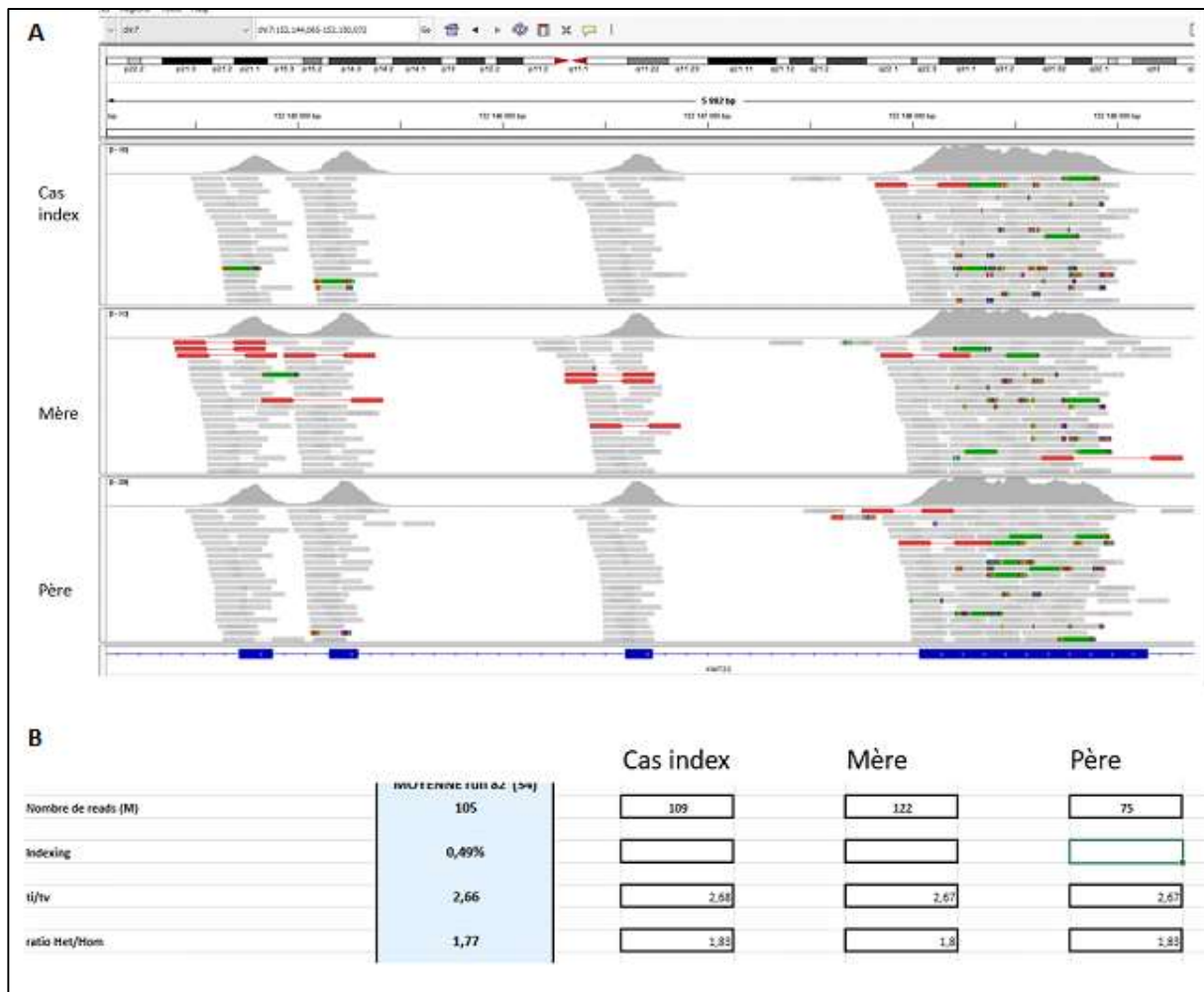


Figure 24 : **Délétion hétérozygote de 4kb de novo dans *KMT2C***

A) Délétion hétérozygote de 4kb de novo dans *KMT2C*, responsable d'un syndrome de Kleefstra chez P3. Cette délétion emporte les exons 52 à 55 sur 59 du gène *KMT2C*. Les points de cassures n'ont pas été capturés, ce qui ne permet pas de délimiter exactement cette anomalie. La profondeur de couverture chez le cas index est de 153X contre 253X chez sa mère et 200X chez son père. B) Nombre de reads chez le cas Index son père et sa mère. Le père est nettement moins couvert que le cas index et sa mère. C'est un exemple de la double normalisation inter et intra-échantillon qu'exécute CNV DRAGEN pour la détection des anomalies

Sur la Figure 24, on remarque une différence de profondeur sur le gène entre le cas index à 153X, son père, à 200X et sa mère à 253X. Cependant, le père a un nombre de read total inférieur à 75 millions de reads contre 109 millions pour le cas index et 122 chez sa mère. Après normalisation de la profondeur intra et inter échantillons, le cas index apparaît bien délété des 4 exons de *KMT2C* (Figure 24 A & B),

Patient 4 : Délétion en 16q23.1 impliquant le gène WWOX et SNV tronquant dans WWOX en trans, hétérozygote composite

Après une grossesse marquée par une clarté nucale augmentée, cette petite fille évolue en période néonatale vers une encéphalopathie épileptique. Devant un tableau précoce catastrophique, cette patiente a bénéficié d'un exome complet en urgence avant tout autre analyse génétique. La découverte d'anomalies hétérozygotes composites dans le gène *WWOX* : une délétion et un variant ponctuel en regard, a permis la mise en place d'un régime cétogène et d'une nutrition parentérale. En effet, les anomalies à transmission récessive de *WWOX* (#MIM616211) sont responsables d'encéphalopathie épileptique et développementale. Cette enfant de 12 mois présente par ailleurs des troubles neurologiques, digestifs et une scoliose.

Cette anomalie de 64kb, seq[GRCh38]del(16)(q23.1) NC_000016.10 :g.(?_78099778)_(78115155_78164183)del, emporte les exons 2 et 3 sur 8 dans le gène *WWOX*, n'a jamais été rapporté ni dans les bases de données de population générale ni de patients. Le variant NM_016373.4(*WWOX*):c.[790C>T], est tronquant, jamais rapporté homozygote dans la population générale (gnomAD) et décrite pathogène dans la base ClinVar (VCV000241105.3) (Figure 25).

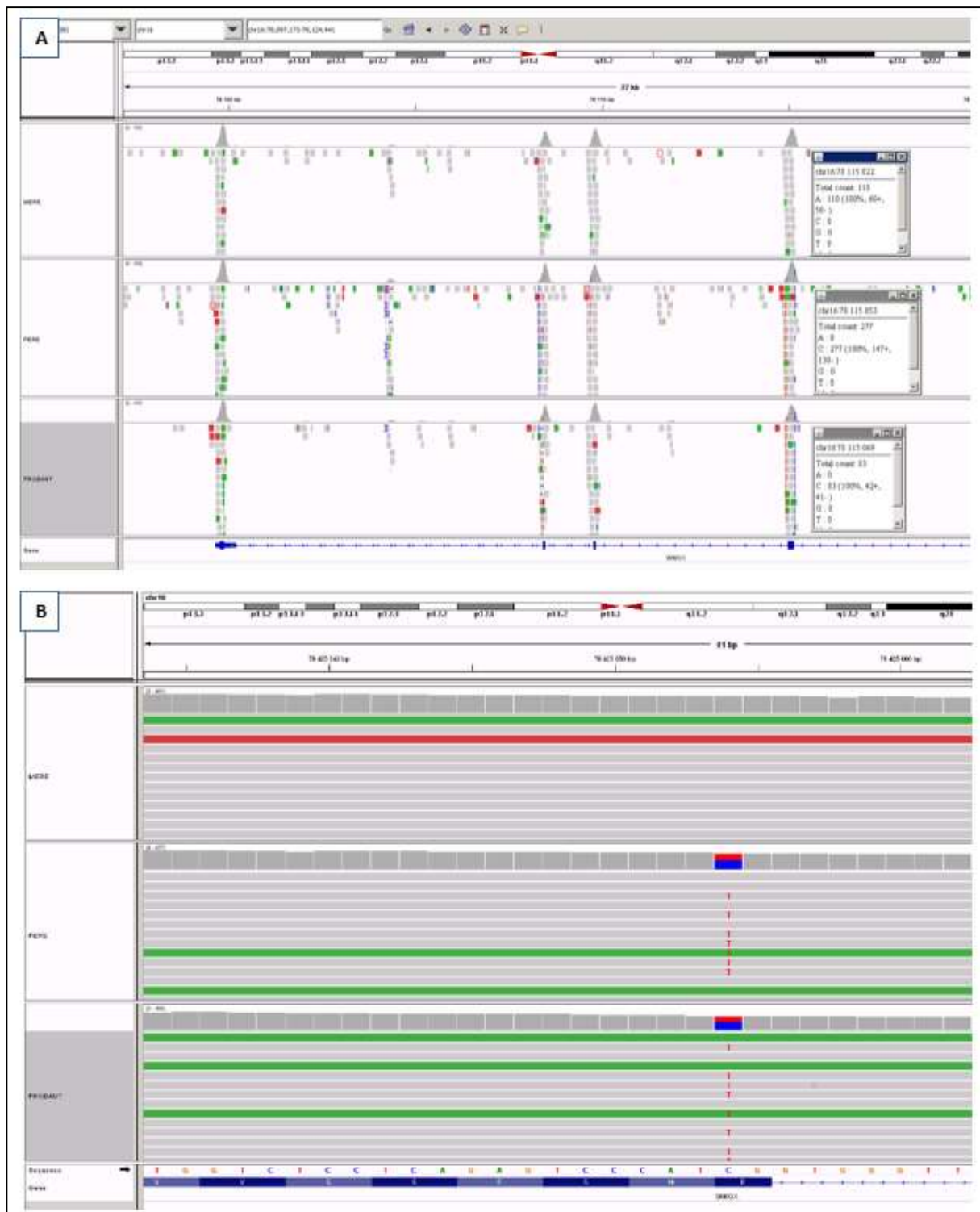


Figure 25 : Délétion en 16q23.1 impliquant le gène WWOX et SNV tronquant dans WWOX en trans, hétérozygote composite

A- 1 anomalie de 64kb, seq[GRCh38]del(16)(q23.1) NC_000016.10 :g.(?_78099778)_ (78115155_78164183)del, emporte les exons 2 et 3 sur 8 dans le gène WWOX, héritée maternelle B- En regard une mutation ponctuelle héritée paternelle NM_016373.4(WWOX):c.[790C>T].

Une sonde couvre 60pb de l'intron 2 de *WWOX* en position chr16:78114687-78114746, l'anomalie n'aurait pas été identifiée en ACPA car sous le seuil de détection de 3 sondes déviantes.

P5 et P6 : Délétions hétérozygotes composites en trans dans la région 15q15.3, héritées biparentales

Les patient P6 et P7 sont issus d'une grossesse gémellaire monochoriale, monoamiotique et de parents non apparentés. Les antécédents médicaux familiaux notables sont une macrocéphalie chez le père et des migraines épisodiques chez la mère. L'accouchement a été déclenché à 34 SA en raison d'une stagnation pondérale chez l'un des jumeaux

Les jumeaux, âgés de 6 ans, présentent un phénotype similaire avec notamment une ataxie épisodique, une macrocéphalie, un retard du langage, une surdité, un retard des acquisitions et un syndrome cérébelleux. L'IRM cérébrale montre un élargissement du quatrième ventricule et des citernes de la base, une atrophie du mésencéphale, une atrophie vermiennne avec fissure cérébelleuse élargie, un élargissement des ventricules et des espaces sous-arachnoïdiens plus marqué en frontal et au niveau des vallées sylviennes.

Ni l'ACPA ni les panels surdité et cervelet n'ont identifié d'anomalie. En exome quatuor, deux délétions hétérozygotes composites, en *trans*, très proche, dans les régions 15q15.3 : seq[GRCh38]del(15)(q15.3) NC_000015.10:g.43600482_43611025del ; seq[GRCh38]del(15)(q15.3) NC_000015.10:g.43600482_43648877del ont été détectées. La première délétion de 10kb, héritée de la mère, n'est jamais rapportée, ni en population générale ni de patients, et emporte la partie 5' du gène, de l'exon 14 à 26 sur 28 de *STRC* (MIM#603720) responsable de surdité autosomique récessive. La délétion de 48kb, héritée du père, rapportée une fois dans la population générale à l'état hétérozygote et une fois à l'état homozygote dans un contexte de surdité récessive emporte la partie 5' de *STRC*, de l'exon 1 à 27, et les 13 exons de *CASTPER2* (Figure 26). Cette association de remaniements explique partiellement le phénotype des patients.

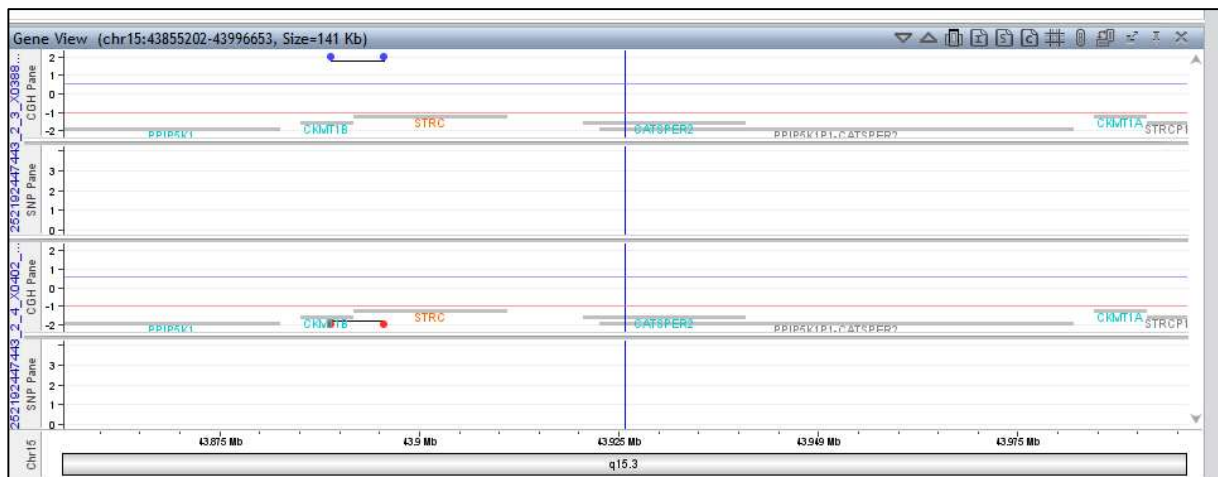


Figure 26 : Représentation sur Cytogenomics de deux sondes déviantes dans la région de la délétion.

L'ACPA met en évidence deux sondes déviantes dans la région après réanalyse. Elles se situent dans l'intron 22 et 23 de STRC aux positions chr15:43602485-43602539 et chr15:43603440-43603507, respectivement de 55 et 68pb. Le gène CATSPER2 n'est pas couvert en ACPA de résolution 60K.

P7 : Délétion hétérozygote de 9kb en 9q34.3 emportant la partie 5' du gène EHMT1

Né à 37 SA, ce garçon a présenté en période néonatale une hypotonie axiale persistante, des troubles de l'oralité, une hypotrophie modérée et une hypoglycémie. Par la suite, une surdité bilatérale a été diagnostiquée. L'IRM cérébrale met en évidence des anomalies morphologiques du corps calleux et du tronc cérébral associées à des anomalies de signal de la substance blanche et des malformations internes bilatérales des oreilles, notamment intéressant la cochlée.

Le syndrome de Kleefstra a pu être diagnostiqué grâce à la caractérisation d'une délétion hétérozygote de 9kb dans le gène *EHMT1*, non héritée maternelle, lors d'un séquençage en exome duo. Cette délétion, seq[GRCh38]del(9)(q34.3) NC_000009.12:g.[?_137775057]_[137784197_ ?]del, emporte les exons 11 à 16 du gène *EHMT1* (Figure 27A). Elle n'a pas été vu en ACPA, il n'y a pas de sondes couvrant cette région (UCSC Table Browser) (Figure 27 B). La délétion a été confirmée *de novo* en qPCR et classée probablement pathogène selon les critères ACMG (39).

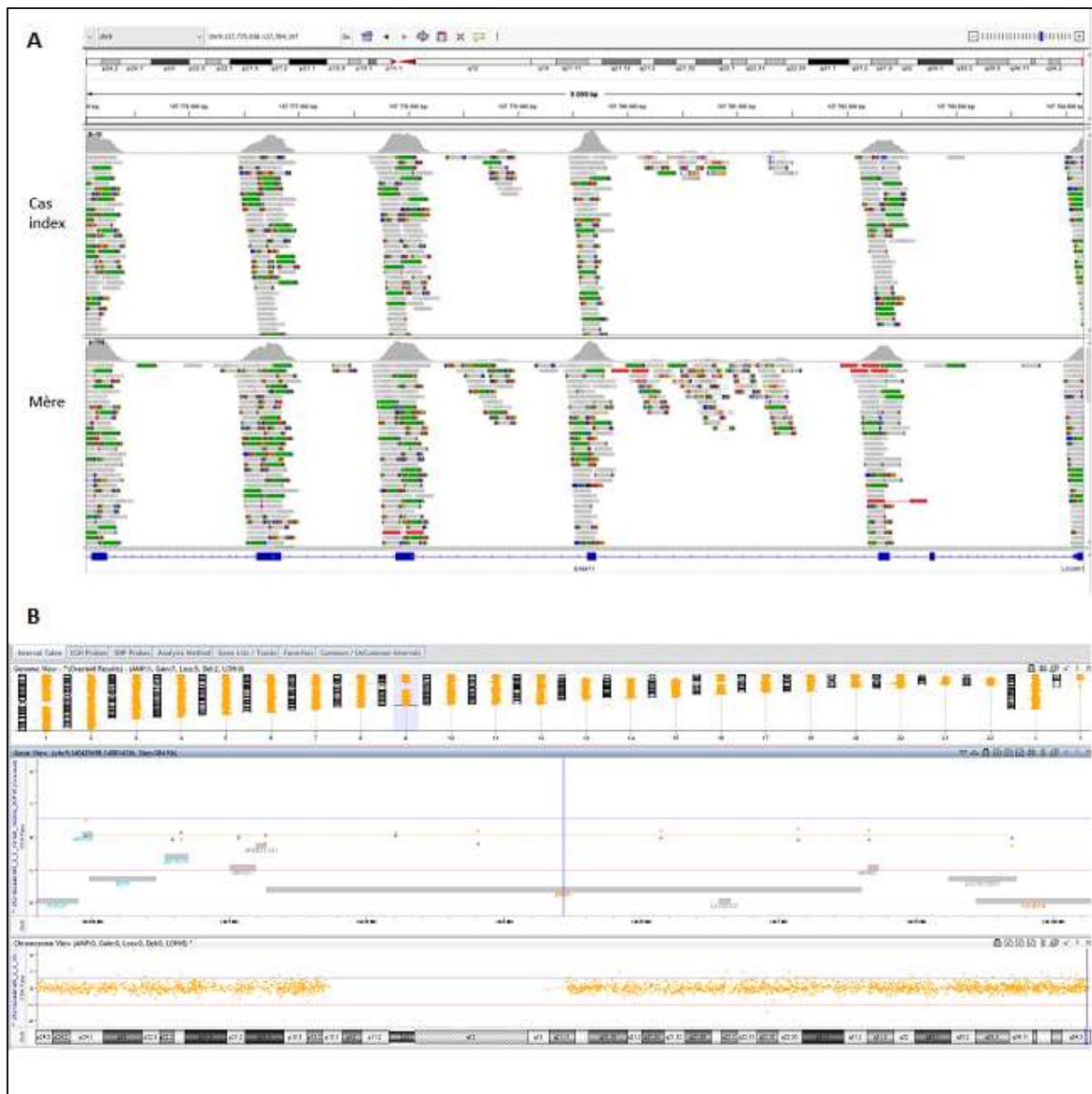


Figure 27 : **Délétion hétérozygote de 9kb en 9q34.3 emportant la partie 5' du gène EHMT1**

A- Région de la délétion en 9q34.3 emportant les 6 derniers exons d'EHMT1, chute de la profondeur chez le cas index, le point de cassure n'a pas été capturé. B- Région du 9q34.3 visualisé sur Cytogénomique, il n'y a pas de sonde déviante dans cette zone, la délétion de 9kb n'est en effet pas couverte par une sonde en lame d'ACPA de résolution 60K.

N- Application de l'outil SV DRAGEN

Les données d'appel de variation de structure par l'outil SV DRAGEN ont été exploitées de la même façon que celles générées par CNV DRAGEN. Elles ont été annotées avec AnnotSV (25), la récurrence et la taille de chaque événement ont été calculées. Les SV appelés ont été filtrés selon le mode d'héritabilité, le score qualité,

la récurrence et la classification proposée par AnnotSV (25), en analysant uniquement les cas index. Les événements détectés après application de ces filtres étaient analysés par leur pertinence clinique (terme HPO (62), OMIM (38), SysID) et moléculaire (pLi et mode d'hérédité, cassure dans un exon, fréquence dans DGV (33) et gnomAD (32)). Cet algorithme repère parfaitement les points de cassure grâce aux split-reads. Il a un seuil de détection des SV à 50 pb.

A l'analyse de ces données, nous nous rendons compte que SV DRAGEN ne peut réaliser les chevauchements entre les événements en exome car il ne bénéficie pas de la continuité de couverture du génome. Il détecte uniquement les anomalies intra-exoniques.

Les SV sont fréquents en exome comme le montre les quelques exemples suivants (Figure 28, Figure 29, Figure 30). Ces SV sont sans impact clinique, ils illustrent les polymorphismes de structure appelés par la stratégie de séquençage en exome. Des délétions hétérozygotes (Figure 28) ou homozygotes (Figure 29) intra-exoniques sont identifiées dans des gènes comme *SLC25A24* ou *CASP8*.

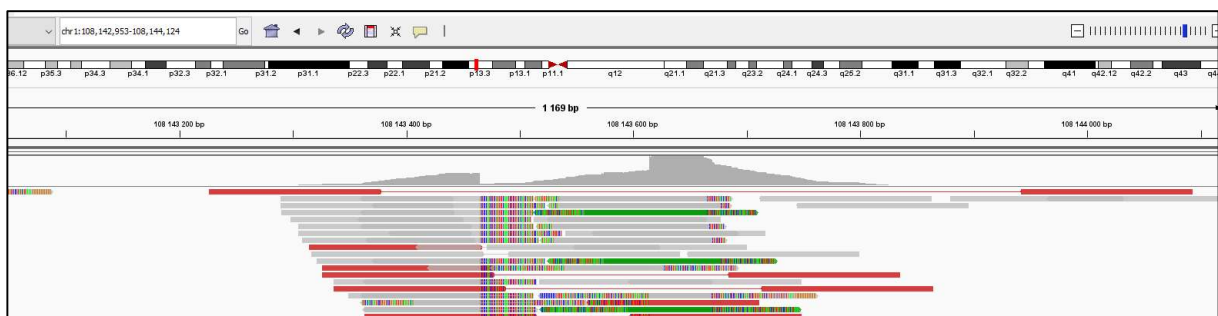


Figure 28 : délétion intra-exonique, seq[GRCh38]del(19)(q23.1) NC_000019.10 :g.[10356763_10356832]del de SLC25A24

Gène impliqué dans le syndrome progéroïde Fontaine #MIM612289, sans lien avec le phénotype du patient. Visualisation sur IGV des reads de paired-end rouge, montrant une délétion, et des reads soft-clippés. Après alignement de la séquence, la région soft-clippée correspond à l'extrémité de l'exon bornant la délétion.

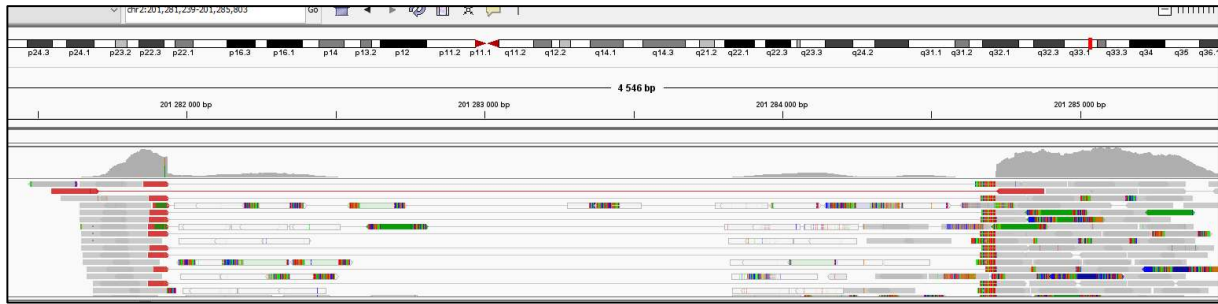


Figure 29 : Visualisation d'une délétion, seq[GRCh38]del(2)(q23.1) NC_000019.10 :g.[201281733_201281929]del dans CASP8

Gène OMIM Morbid décrit en oncologie, mis en évidence par des cassures brutales dans les exons, des reads paired-end rouge et du soft-clipping. Les reads blancs sont des reads dont l'alignement est douteux, et associés à des scores faibles de qualité.

Nous avons également identifié à plusieurs reprises la présence de rétrotransposons. Il s'agit d'un élément nucléotidique appartenant à la famille des éléments transposables, c'est-à-dire capable de se déplacer et de se multiplier dans le génome. Ils donnent naissance à des séquences répétées dispersées. Le rétrotransposon, à l'inverse du transposon, a un intermédiaire ARN retro-transcrit en ADNc puis intégré dans le génome (66). A plusieurs reprises, nous avons identifié une insertion du rétrotransposon *PRKRA* (Figure 30). Un rétrotransposon est composé uniquement d'exons, puis il est inséré dans le génome. De ce fait, le paired-end sur cette image lie les exons entre eux, comme s'il existait une délétion des introns. Le soft-clipping en fin d'exons correspond aux exons adjacents. Le soft-clipping aux extrémités du rétrotransposon, trop court, n'a pas permis de localiser son insertion dans le génome.

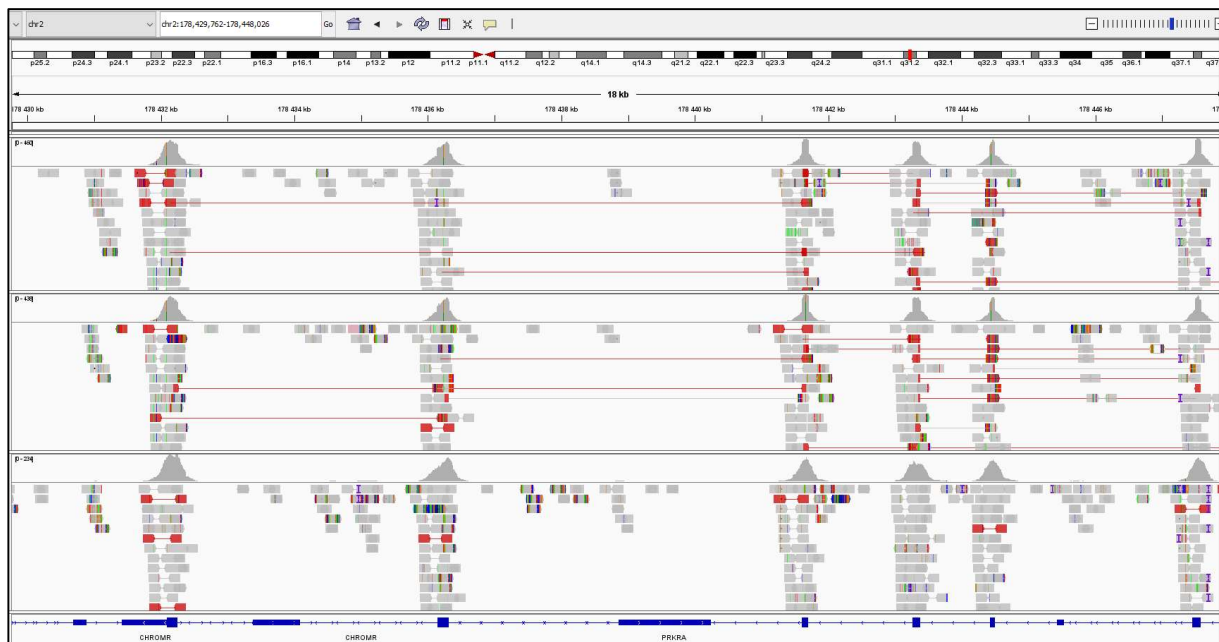


Figure 30 : **Visualisation sur IGV d'une insertion de rétrotransposon sans impact clinique sur le patient.**

Un rétrotransposon est une intégration/insertion génomique d'une séquence d'ARN. Ce SV est récurrent. Les reads font des ponts entre les exons du gène, les zones introniques apparaissent déléetées : le rétrotransposon n'a pas d'intron. Les soft clips correspondent aux exons proches.

Un très grand nombre d'évènements SV est appelé par exome. Après tri des SV en lien avec des gènes impliqués dans les troubles du neurodéveloppement, nous retrouvons 2 variations dans l'exon 4 de *MECP2* (Figure 31 & Figure 32).

Le premier patient a eu un exome pour une déficience intellectuelle légère et un retard de langage (Figure 32).

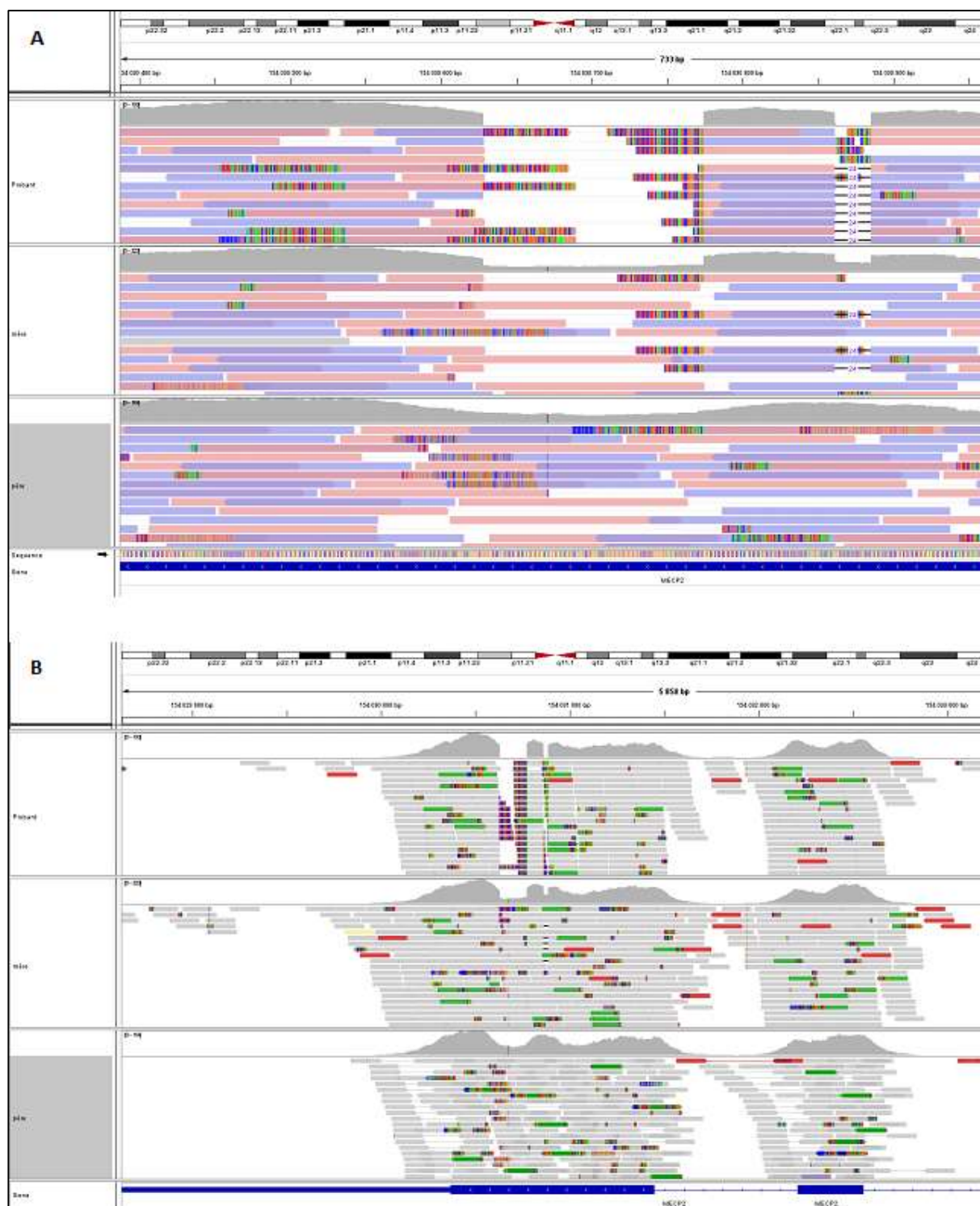


Figure 31 : Délétion dans l'exon 4 de MECP2 chez un patient DI

A- Délétions dans l'exon 4 du gène MECP2, seq[GRCh38]del(X)(q28) NC_0000023.11 :g.[154030886del;154030629_154030781del], ne provoque pas de décalage de cadre de lecture. Le SV est classé VSI. B- Visualisation de la zone délétée de MECP2 élargie. Ces SV sont hérités maternels. Les séquences soft clippées s'alignent aux bornes de la délétion de 153pb.

Nous avons identifié via SV DRAGEN deux délétions intra-exoniques successives, hémizygotes, héritées de la mère, de 153 et 24 pb, seq[GRCh38]del(X)(q28)

NC_0000023.11:g.[154030886del;154030629_154030781del],
NM_001110792.2:c.[979_1002del;1047_1199del]. Cela provoque une délétion en phase dans le quatrième exon de *MECP2*.

Ce SV est classé VSI car il n'est jamais identifié dans la population générale, ni dans les bases de patient. L'étude de corrélation génotype-phénotype a lieu actuellement à l'hôpital Cochin.

Le second patient, un garçon de 11 ans, né eutrophe à 41 + 6 SA avec une grossesse sans particularité, a eu un exome pour retard psychomoteur et hypotonie initiale. Il a des troubles du comportement, une déficience intellectuelle modérée. Il n'y a pas d'antécédents particuliers dans sa famille, hormis sa mère qui a une surdité congénitale bilatérale. Il est pris en charge dans un IME et en orthophonie, il ne présente pas de signes évocateurs d'épilepsie, ni troubles alimentaires ni de sommeil. Son IRM est normale, il présente une avance staturopondérale harmonieuse à +3DS.

Nous retrouvons une anomalie classée probablement pathogène (ACMG classe 4) dans l'exon 4 de *MECP2* (Figure 32), seq[GRCh38]del(X)(q28) NC_000023.11:g.[1540300768_154030769del;154030626_154030675del], hémizygote, héritée maternelle, jamais identifiée dans la population générale, ni dans les bases de données de patients *MECP2*. Cette anomalie provoque un décalage du cadre de lecture, à priori sans intervention du NMD. Nous avons identifié une délétion classée probablement pathogène grâce à SV DRAGEN de 50pb suivie de deux délétions de 3pb et 2pb. Ces délétions provoqueraient un décalage du cadre de lecture avec apparition prématuré d'un codon stop. Sa mère est conductrice pour l'anomalie, le conseil génétique peut être adapté grâce à ces informations.

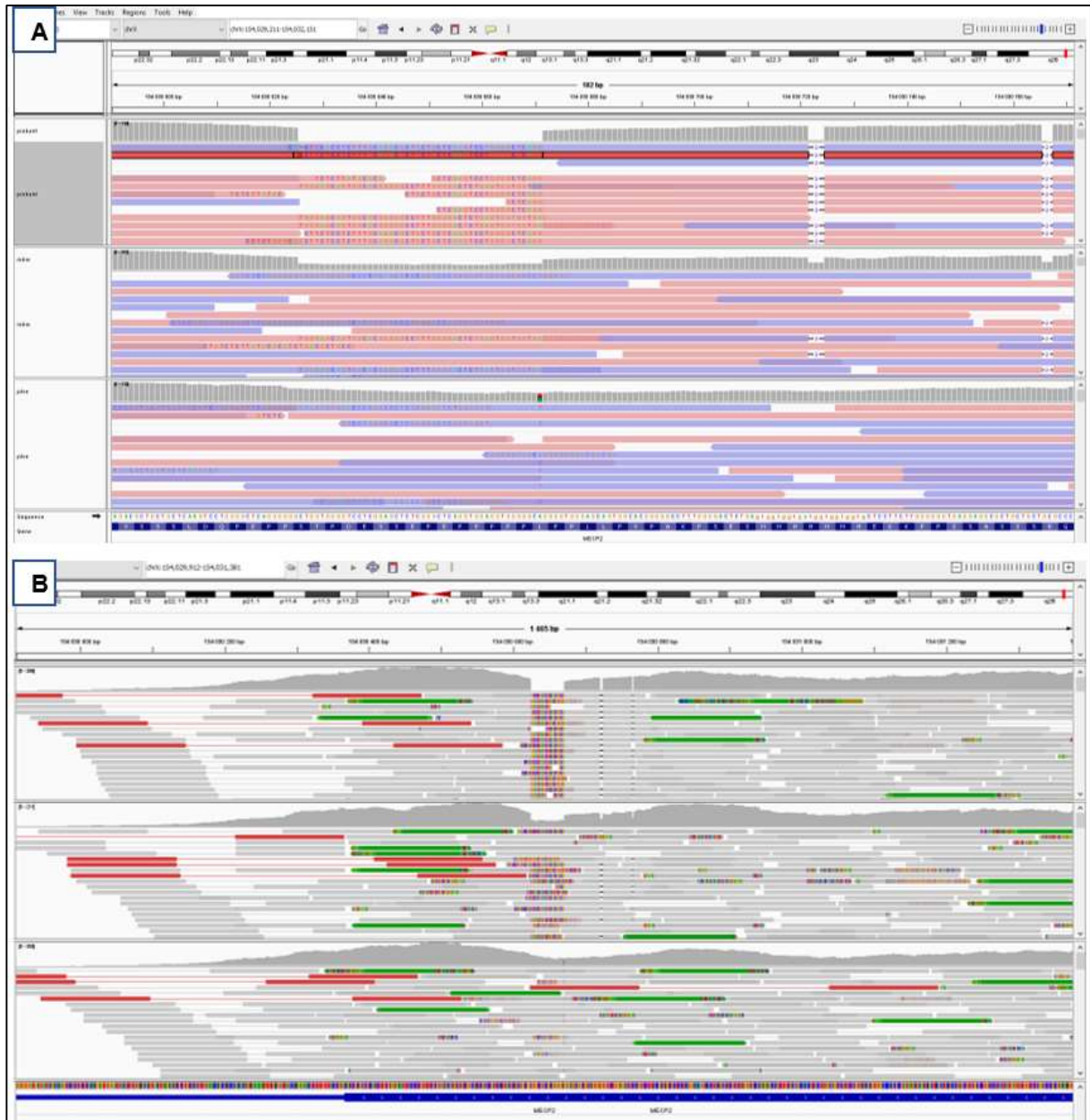


Figure 32 : Image d'IGV des délétions de 50, 3 et 2pb dans la partie 5' de l'exon 4 de MECP2
 A-Délétions de 50, 3 et 2pb dans la partie 5' de l'exon 4 de MECP2, seq[GRCh38]del(X)(q28) NC_000023.11 :g.[1540300768_154030769del;154030626_154030675del], hémizygotes, héritées maternelles. B-Délétion dans l'exon 4 du gène MECP2 provoquant un décalage du cadre de lecture. Les séquences soft-clippées correspondent aux bornes de la délétion.

O- ROH DRAGEN

1) Données étudiées

L'analyse ROH recherche les régions génomiques de perte d'hétérozygotie chez un patient. Ces zones sont extraites par l'algorithme ROH DRAGEN d'un fichier.bed.

Les données ROH ont été générées pour 717 patients et leurs parents. 168 patients ont au moins une région d'homozygotie repérée par l'algorithme. Les fichiers .bed ou « Browser Extensible Data » sont un format utilisé pour stocker des régions génomiques sous forme de coordonnées avec des annotations associées.

2) Préparation du fichier ROH : Mise en forme

Tableau 1 : **tableur ROH initial issu d'un fichier .bed.**

La région ROH est indiquée par son chromosome, le début et la fin de la région par « start » et « end », le score du ROH, le nombre de SNP homozygotes et hétérozygotes dans la région.

Chromosome	Start	End	Score	SNP homozygotes	SNP hétérozygote
chr1	201891497	247921805	6.90	354	2
chr3	44242032	47577463	1.25	50	0
chr3	127604987	131228591	1.27	51	0
chr5	176093307	181155604	2.40	96	0
chr7	149027909	151408980	1.30	52	0

Tout d'abord, les données des 168 individus ont été agrégées dans un même tableur Excel, chaque ligne correspond à une région d'homozygotie chez un individu. Chaque région a été secondairement annotée : la taille de la région homozygote, la somme de la taille des régions ROH par patient, la récurrence de chaque région pour éliminer les artefacts, le statut de l'individu (proband, père, mère) et son identité par un numéro d'identification. Le ratio SNP hétérozygotes/ SNP homozygotes (het/hom), le nombre de SNP identifiés et le nombre d'INDEL ont été renseignés pour chaque région.

Ce fichier a d'abord pour vocation de mettre en évidence des isodisomies et les délétions. La disomie uniparentale est la présence chez une personne de deux chromosomes d'une même paire chromosomique provenant d'un seul de ses parents.

Les isodisomies sont des anomalies chromosomiques où les deux chromosomes d'une paire proviennent du même chromatide de l'un des parents ; les hétérodisomies sont des anomalies où les deux chromosomes d'une paire proviennent des deux chromatides différents d'une paire de chromosome de l'un des parents. Ces anomalies peuvent devenir pathogènes si elles touchent une région à empreinte ou si un SNV hétérozygote pathogène pour une maladie récessive est transmis de façon homozygote (67).

Dans le cas d'une délétion, les SNP d'une région seront appelés comme homozygotes car sans copie allélique alternative en face, exactement comme les isodisomies. C'est pour cela que l'algorithme ROH DRAGEN appelle les délétions.

Après une extraction de tous les gènes à empreintes connus via le site <https://www.geneimprint.com/>, les coordonnées génomiques en GRCh38 ont été obtenues pour chaque gène à empreinte afin de les croiser avec les coordonnées génomiques des régions ROH. Cette étape a été réalisée grâce à l'outil « bedtools intersect » de Galaxy (<https://usegalaxy.org/>) et le Table browser d'UCSC (<https://genome.ucsc.edu/cgi-bin/hgTables>) (Figure 33). Ces gènes ont ensuite été identifiés ainsi que leur mécanisme d'empreinte (maternelle ou paternelle).

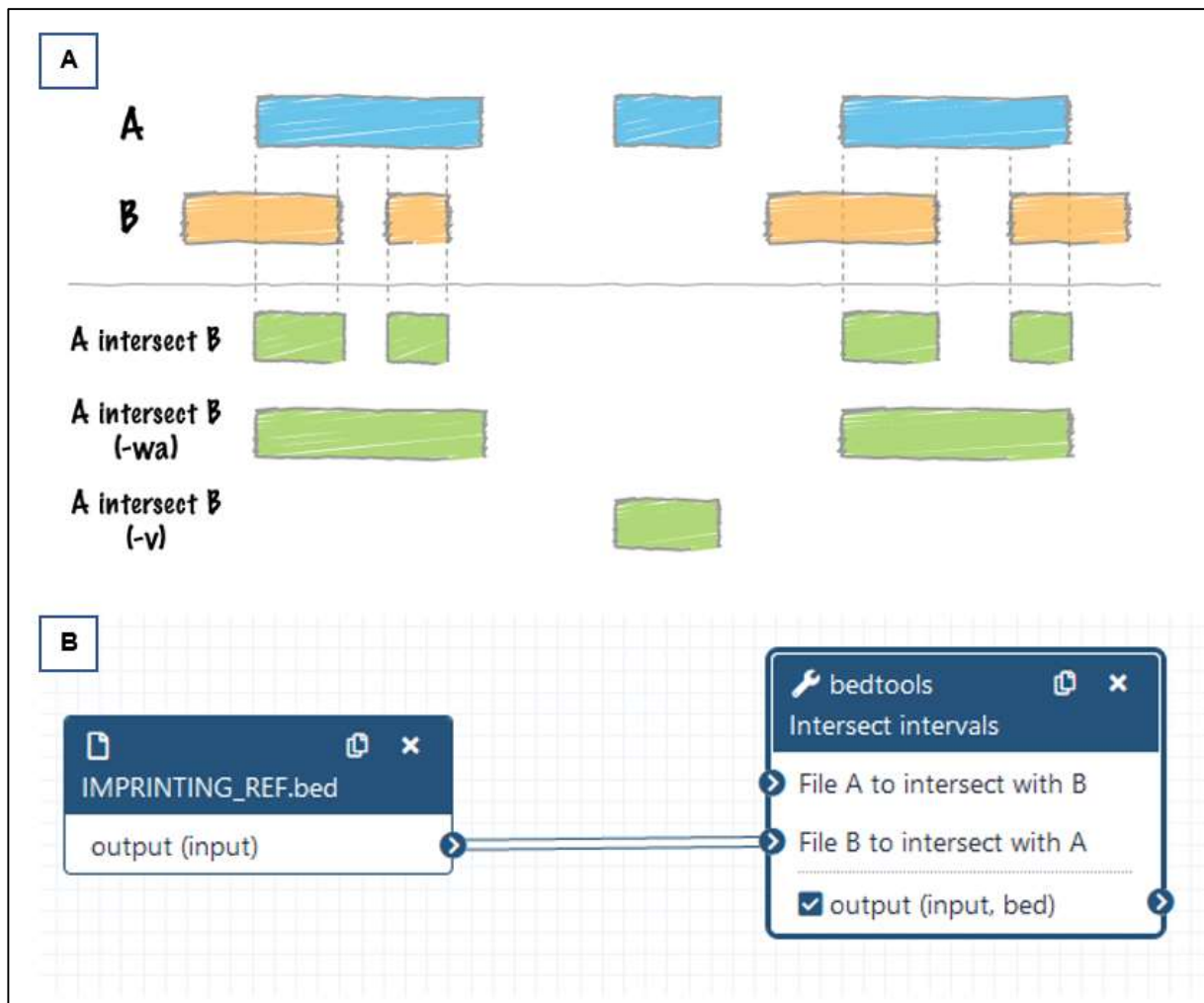


Figure 33 : Schémas du fonctionnement d'« Interseq tool » de Galaxy

A- Ce logiciel permet de croiser les bornes génomiques de deux fichiers pour vérifier s'il y a des intersections. Ici, les gènes à empreintes avec les coordonnées génomiques des régions ROH ont été croisés. B- Workflow de l'analyse Galaxy intersection entre les bornes des gènes à empreinte (file A) et les zones ROH (file B).

3) Analyse du fichier ROH : Exploitation

Notre analyse de ce fichier se concentre uniquement sur les probants (717 patients). Avec ce fichier, nous espérons mettre en évidence les isodisomies et les délétions. D'après la littérature, le taux de disomie détectée (Isodisomie + hétérodisomie) par exome est de 0,31% et aboutit à un diagnostic clinique dans 0,14% des cas (67).

Recherche d'isodisomie/délétion

Les patients issus d'union entre apparentés ou avec de multiples régions ROH, signature d'une identité par descendance, sont exclus de l'analyse, selon les pratiques rapportées dans la littérature scientifique (29). En analysant individuellement chaque

région ROH restantes dans le Table Browser d'UCSC, nous éliminons les régions récurrentes dans la cohorte, les régions contenant les gènes du HLA et les zones intergéniques. En effet, au vu de la faible fréquence des disomies uniparentales, il est peu probable qu'une région ROH d'intérêt soit identifiée chez plusieurs patients ; les régions HLA sont des régions extrêmement polymorphiques avec des sets d'alignement adaptés qui ne correspondent pas au génome de référence, il n'est pas étonnant de trouver de larges zones homozygotes dans les régions du HLA ; les régions intergéniques ne sont, par essence, pas capturées en exome (ou très faiblement), les SNP homozygotes captés dans ces zones ne sont pas d'une qualité fiable.

Les patients identifiés avec un diagnostic en SNV/INDEL selon les critères ACMG de classe 4 ou 5 sont exclus. Les patients sans diagnostic identifié ou avec un variant de classe 3 sont conservés.

Après application de ces filtres, il ne reste que 11 patients. Seuls 2 patients ont une zone ROH contenant des gènes à empreintes (Figure 34).

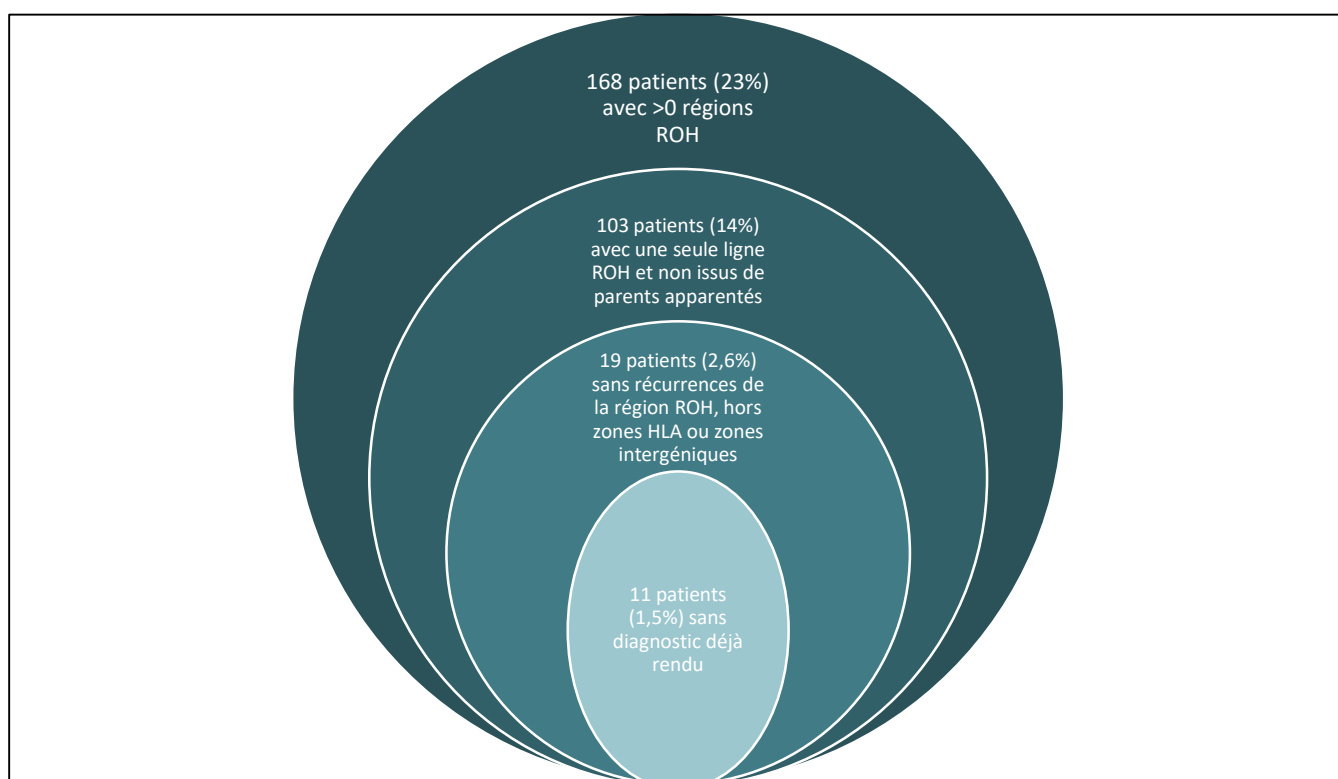


Figure 34 : **Schéma explicatif de l'application des filtres sur le fichier ROH.**

23% des patients ont au moins un ROH, 14% sont éligibles à l'analyse car sans identité par descendance, 2,6% ont un ROH d'intérêt (hors HLA, zone intergénique et ROH non récurrent), et 1,5% des patients n'ont pas de diagnostic déjà identifié.

Pour chacun de ces patients, nous analysons le fichier original comportant l'ensemble des variants identifiés pour chaque patient. Il comprend l'ensemble des SNV/INDEL d'un individu sur son exome, soit environ 40 000 SNV/INDEL.

Cette analyse va permettre de confirmer la région ROH, d'identifier s'il s'agit d'une suite de SNP/SNV homozygotes avec des variants très fréquents dans la population générale, ou d'un réel événement type disomie/délétion. La « région ROH » ne doit pas être héritée des deux parents : un des parents devrait être indemne de SNP sur la région où son enfant est homozygote (figure 31).

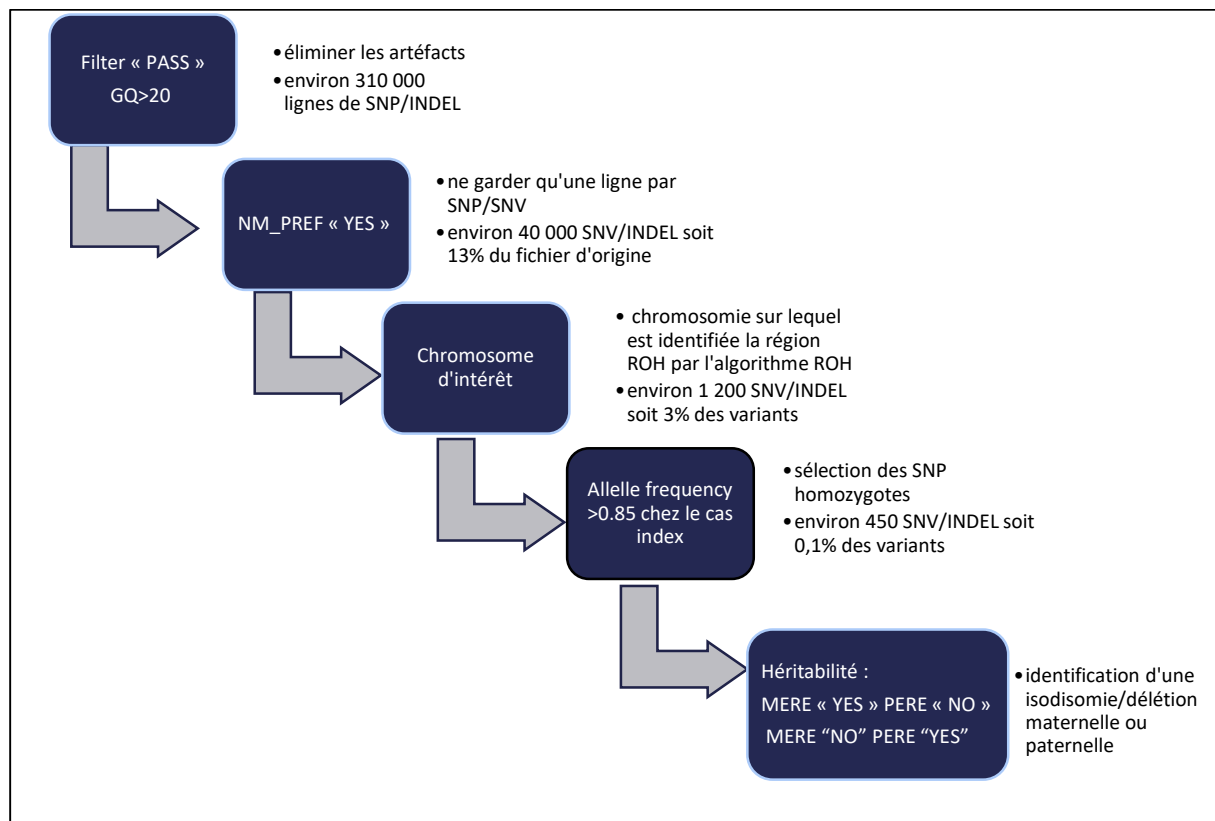


Figure 35 : Schéma du raisonnement de confirmation de la délétion/disomie sur le fichier original de l'exome, application de filtres progressifs pour retrouver l'évènement génétique de disomie/délétion.

Dans le fichier regroupant l'ensemble des variations d'un individu (environ 320 000 lignes de variations), nous sélectionnons les SNP avec un score qualité « PASS » (310 000 lignes de SNP), nous ne gardons qu'une seule ligne par SNP en sélectionnant le transcrit MANE (environ 40 000 SNP), nous identifions le chromosome porteur de la région ROH identifiée par l'algorithme (1 300 SNP), puis nous

sélectionnons les SNP avec une fréquence allélique >0,85 chez le cas index (450 SNP). Enfin, nous cherchons alternativement les SNP présent uniquement chez l'un des deux parents pour identifier une disomie uniparentale

Sur les 11 patients, aucune zone de ROH non héritée de l'un des parents n'a été identifiée (Tableau 1). Ces zones existaient bel et bien, cependant leur présence est expliquée par la grande fréquence des SNP et la ressemblance des parents sur ces régions chromosomiques. Nous avons réitéré cette analyse sur l'ensemble des patients en cherchant notamment une délétion. Nous n'avons pas identifié de résultat supplémentaire.

Tableau 2 : Coordonnées génomiques des 11 variations identifiées avec la stratégie de filtres sur les données ROH chez 11 patients différents.

Ce tableau a été implémenté de la taille du ROH en kb, du nombre de gènes à empreintes et du type d'empreinte, du diagnostic rendu en exome, du ratio hétérozygote/homozygote, du nombre de SNP et INDEL par patient.

Chr	start	end	taille en Kb	score	homoz ygote	heteroz ygote	files	Statut	Rec urren ce	Nb de gènes à empreinte	empreint e maternelle	empreint e paternell e	rendu	het/hom	snp	indel
19	35372239	45315576	9943	4.03	239	2		PROBAN	1	2	ZNF229	ZNF225	non rendu	1,66	35184	44
12	5494466	11309341	5815	2.6	104	0		PROBAN	1	1	RBP5		VSI	1,76	35887	46
3	44242032	47670712	3429	1.4	56	0		PROBAN	1	0			VSI	1,82	36139	37
14	19713771	22902058	3188	2.67	107	0		PROBAN	1	0			NEG	1,78	35729	48
17	41610252	44398188	2788	2.1	84	0		PROBAN	1	0			non rendu	1,75	34941	45
10	69880499	72210693	2330	1.32	53	0		PROBAN	1	0			NEG	1,97	37547	45
19	40742104	42527059	1785	1.52	61	0		PROBAN	1	0			non rendu	1,83	36635	63
16	89233482	90095778	862	1.5	60	0		PROBAN	1	0			VSI	1,73	34710	44
19	8903893	9758009	854	1.25	50	0		PROBAN	1	0			VSI	1,6	33889	45
22	49927295	50744826	818	1.35	54	0		PROBAN	1	0			VSI	1,68	34437	46
2	2,19E+08	2,2E+08	608	1.27	51	0		PROBAN	1	0			Neg	1,74	34568	48

Passage de patients témoins

En parallèle, trois patients issus d'une analyse de microsatellites effectuée au laboratoire pour identifier les disomies du chromosome 15 ont été sélectionnés en témoins pour notre analyse avec ROH DRAGEN. Le premier patient est porteur d'une isodisomie d'origine paternelle du chromosome 15 (témoin +), le deuxième patient est porteur d'un phénomène mixte : une isodisomie de 15q11.2 à 15q13.3 (10Mb) et une hétérodisomie paternelle sur le reste du chromosome 15 (témoin +/-), et enfin, le troisième patient est porteur d'une hétérodisomie du chromosome 15 (témoin -) (Tableau 2). Tous ont des résultats confirmés par analyse de microsatellites. Théoriquement, ROH DRAGEN doit identifier les isodisomies mais pas les hétérodisomies, d'où l'appellation de ces témoins.

Ces trois patients témoins ont été séquencés en exome pour valider l'approche ROH DRAGEN. Le témoin + a un ROH sur le chr15 avec un score à « 8,32 », les SNP homozygotes recensés couvrent une grande partie du chr15. Le témoin +/-, avec une isodisomie et une hétérodisomie associée, a un ROH sur le chr19 avec un score qualité faible, à « 1,42 », et aucune région appelée sur le chromosome 15. Le témoin – a un ROH sur le chr15 avec un score qualité faible à « 1,82 ». Les deux derniers ROH, en raison de leur localisation et de leur score qualité faible ne semblent pas pertinent. Tous les ROH ont été analysés manuellement pour chercher leur pertinence

Tableau 3 : Résultats de l'algorithme ROH DRAGEN sur les 3 témoins.

Le témoin +, porteur d'une isodisomie du chromosome 15 a un ROH du chr15 avec un score à 8,32, les SNP homozygotes recensés couvrent une grande partie du chr15. Le témoin +/- avec une isodisomie et une hétérodisomie partielles a un ROH sur le chr19 avec un score qualité faible, à 1,42. Le témoin -, porteur d'une hétérodisomie du chr15 a un ROH sur le chr15 avec un score qualité faible à 1,82.

Patient	Numéro DI	Chromosome	Start	End	Score	Hom	Het	Het/Hom	SNP	CNV
Témoin +		chr15	39934293	90981611	8.32	450	3	1,64	34964	34
Témoin +/-		chr19	8889463	9341631	1.42	57	0	1,97	37356	70
Témoin -		chr15	87880231	90981611	1.82	73	0	1,7	35121	48

Comme précédemment, le fichier originel comportant l'ensemble des variants identifiés pour chaque patient avec l'ensemble des SNP/SNV est analysé (Figure 34).

Le témoin + a environ 900 SNP/ INDEL sur le chr15, dont 820 homozygotes. Il y a 131 SNP homozygotes chez le probant de qualité acceptable (GQ>20) hétérozygotes chez son père et non retrouvé chez sa mère, repartis sur l'ensemble du chromosome 15. Nous confirmons l'isodisomie paternelle du chr15, homogène.

Le témoin +/- a 1200 SNP/INDEL sur le chr15 dont 500 homozygotes. Pour le témoin +/-, il y a 30 SNP homozygotes chez le probant de qualité acceptable hétérozygote chez son père et non retrouvé chez sa mère, localisé dans la région 15q11.2q13.3. Comme ce témoin a un événement de disomie mixte, nous nous attendions à retrouver moins de SNP homozygote que chez le témoin positif. Cependant, c'est une surprise que l'algorithme ROH ne détecte pas l'isodisomie de 10Mb.

Le témoin – a environ 1200 SNP/INDEL sur le chr15 dont 660 homozygotes. Pour ce témoin, nous retrouvons 26 SNP homozygotes chez le probant hétérozygote chez l'un de ses parents en 15q26. Ce patient peut avoir une zone isodisomie en plus de l'hétérodisomie diagnostiquée par microsatellite. En effet, un crossing-over entre les chromatides a pu avoir lieu.

P- Diagnostics par exome au CHU de Lille d'octobre 2020 à Mars 2022

A la date du 25 mars 2022, 600 patients ont reçu un résultat d'exome (Figure 36). Cette population de patient était composée de 76% de trios, 20% de duos et 4% de solos, avec une médiane d'âge de de 10 et 11 mois. Les pipelines CNV et SV DRAGEN ont permis de détecter 10 SV (Figure 37), soit 1,7% de diagnostic additionnels à l'appel unique des variations ponctuelles SNV/INDEL. Inclus dans ces SV, a été mis en évidence un PIEV chez une petite fille de 10 mois n'ayant pas eu d'ACPA au préalable. Néanmoins, ce facteur de prédisposition n'explique pas entièrement son phénotype.

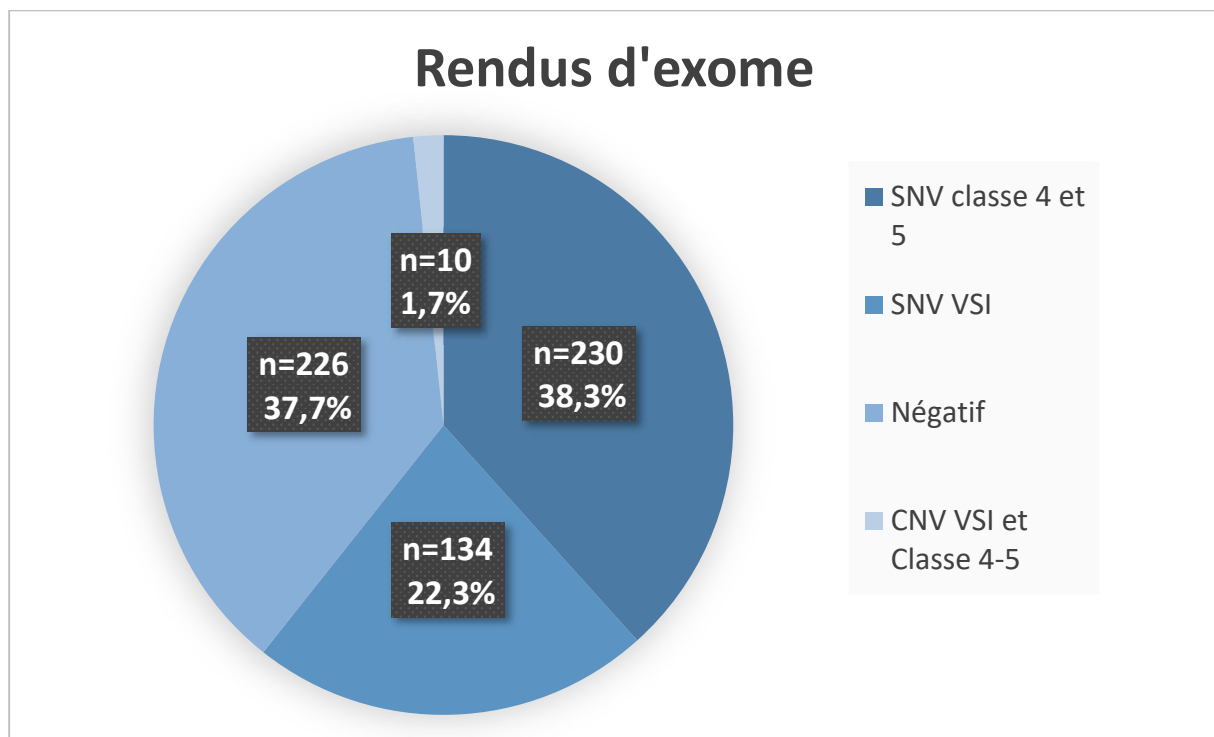


Figure 36 : Pourcentages des résultats rendus en exome dans l'UF cytogénétique des déficiences intellectuelle et anomalies du développement du CHU de Lille.

38,3% des patients ont un résultat de classe pathogène ou probablement pathogène selon les critères ACMG. 60,0% des patients ont eu un résultat d'exome négatif mais 37,7% d'entre ont un variant de signification inconnu détecté, qui sera reclassé en fonction de la ségrégation familiale et des évolutions de la littérature. 1,7% des résultats d'exome sont des SV détectés par CNV DRAGEN ou SV DRAGEN.

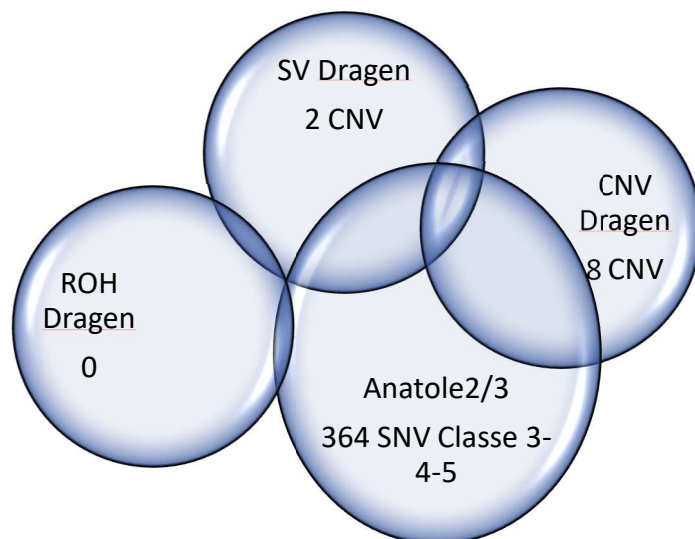


Figure 37 : **Participation des différents algorithmes aux rendus de résultat en exome complet à l'institut de génétique médicale au CHU de Lille dans l'UF génétique du neurodéveloppement.**

IV- Discussion

Cette étude monocentrique visait à déterminer l'intérêt de l'addition d'outils d'appel de variations de structure par 3 approches algorithmiques différentes sur un jeu de données d'exome établi en routine diagnostique. Nous rapportons dans ce travail les données de 600 patients atteints de troubles de neurodéveloppement qui ont reçu un résultat d'analyse génétique en exome complet, à l'institut génétique médical au CHU de Lille sur la période d'Octobre 2020 à Mars 2022. L'ensemble des exomes a été analysé avec un appel de variations issus des pipelines DRAGEN-GATK, CNV DRAGEN, ROH DRAGEN afin de cibler à la fois les SNV/INDEL, et les SV/CNV. Les patients ont été analysés par approche de séquençage trio pour 76%, duo pour 20% et solo pour 4% ou selon d'autres combinaisons par inclusion de parents ou autres apparentés dans la stratégie de séquençage.

479 patients (80%) ont eu une analyse des SV avec le pipeline SV DRAGEN, en complément de CNV DRAGEN.

L'ensemble des patients analysés et des patients en attente d'analyse, soit 717 patients, ont bénéficié d'une analyse avec le pipeline ROH DRAGEN à la recherche d'une disomie ou d'une délétion.

A- Les taux diagnostics

1) A l'Institut

L'analyse des variations ponctuelles par exome a initialement permis d'identifier des variants de classe 4 ou 5 par SNV/INDEL selon la classification ACMG (39) chez 230 patients soit un taux diagnostique de 38,3%. Ce taux diagnostique atteint 40% en ajoutant les variations de structure identifiées chez 10 patients supplémentaires.

L'analyse SV DRAGEN n'a pas été réalisée sur l'intégralité des patients de la cohorte (80%), elle a permis de d'identifier 2 diagnostics (un VSI et un variant de classe 4, des études fonctionnelles sur des anomalies proches sont en cours dans le laboratoire de génétique médicale de l'hôpital Cochin pour aider à classer ces anomalies).

370 patients, soit 60,0%, ont reçu un résultat de génétique négatif. Cependant, un variant de signification indéterminé (VSI) a été trouvé chez 22,3% d'entre eux (134 patients). Ces variants seront reclassés grâce à la ségrégation familiale, les discussions clinico-biologiques et l'évolution de la littérature.

2) Dans la littérature

La comparaison du taux de diagnostic avec les cohortes de la littérature est délicate. En effet, les effectifs, le recrutement mono ou multicentrique des patients et les stratégies d'analyses moléculaires ne sont pas toujours comparables (68). Les taux diagnostics dans les troubles du neurodéveloppement sont très différents : selon la méta-analyse, les taux diagnostics d'ASD sont de 17%, d'épilepsie de 24% et de DI de 28% (68). De plus, en fonction des centres, de la stratégie d'analyse et du recrutement des patients, les taux diagnostics des troubles du neurodéveloppement varient de 28% à 36% (68,69).

Le taux diagnostic dans l'UF de génétique du neurodéveloppement du CHU de Lille est similaire à ceux rapportés dans la littérature : 40% (68,69).

Notre recrutement est monocentrique, entre octobre 2020 et 2022. De nombreux gènes ont déjà été découverts avec l'utilisation du SHD depuis les premières études DI. Nous ne connaissons pas le recrutement ou la stratégie d'analyse des patients : utilisation de séquençage ciblé dans certains centres avant de passer à l'exome complet, variants non classés selon les critères ACMG (39), ou encore des approches stratégiques qui varient selon les pays (68).

En France, la stratégie d'exploration génétique des patients dépend de la pathologie, et est très évolutive sur les dernières années. L'exploration par séquençage pangénomique prend une place de plus en plus importante dans les stratégies appliquées par les filières. Ainsi pour la déficience intellectuelle, les premiers résultats de l'étude DEFIDIAG pourraient changer la stratégie X-fragile, ACPA puis SHD, en orientant les patients en première intention vers le génome complet (12). Les recommandations de l'American College of Medical Genetics 2021, par Manickam et al, orientent déjà en première intention sur du séquençage d'exome ou de génome complet chez les patients pédiatriques porteurs d'anomalies congénitales ou de DI (70).

Ces études illustrent bien l'hétérogénéité des causes d'anomalie du neurodéveloppement et la complexité de leur diagnostic moléculaire (12,70).

B- Détection des CNV en exome versus en ACPA

Dans un premier temps, nous avons comparé la capacité de détection de CNV connus en ACPA (n= 68) chez des patients secondairement explorés en exome. Sur les 76 anomalies détectées en ACPA, 83% des aneuploïdies (n=6/7) et 92% des CNV (n=65/70) ont été retrouvé en exome.

1) Détection des CNV à la résolution de l'exome

92% des CNV détectés en ACPA sont aussi retrouvés par exome dans notre centre. Nous retrouvons 10 diagnostics supplémentaires en SV, sous le seuil de résolution de l'ACPA.

La résolution de l'exome est supérieure à l'ACPA. Nous obtenons logiquement un nombre supérieur de SV à analyser, c'est une contrainte additionnelle à l'interprétation. Les bases de données actuelles concernent des CNV plus grands ou des variations ponctuelles (SNV/INDEL) dont le mécanisme pathogène peut être différent. Les bases de données partagées de cytogénétiques, telles que DGV, reprenant des données issues d'ACPA ou de SNP-array, ne sont pas toujours assez résolutive pour permettre la comparaison des données identifiées à la résolution de l'exome.

Les performances de diagnostic de l'exome sont autour de 35-40% quand elle est de 10% en ACPA (7,41,68).

Les différents exemples rapportés dans les **résultats** illustrent ces remaniements génétiques qui aboutissent à des diagnostics, soit par identification de variations dont la taille est située sous le seuil de résolution de l'ACPA, soit par identification de combinaisons de variations obtenues par des pipelines différents.

2) Des analyses complémentaires

L'interprétation des données générées en exome est nécessairement plus longue devant le grand nombre de cibles analysées par rapport à l'ACPA. A ce jour les coûts sont également plus élevés et l'analyse ne se concentre que sur 1 à 3% du génome, là où l'ACPA est pangénomique (7). La sensibilité du SHD dans la détection d'anomalie génétique est très dépendante de la profondeur de couverture. En effet, l'augmentation de la profondeur de lecture permet d'augmenter le seuil de détection des variations. Ainsi le suivi de variations en mosaïques, en maladie résiduel, pourrait s'envisager par exome (32). Néanmoins, l'augmentation de profondeur se fait au détriment du nombre d'échantillons séquencés dans la même technique et est associée à des coûts plus élevés (18).

L'ACPA permet d'éliminer des syndromes connus. Ces syndromes impliquant des microremaniements sont parfois récurrents, reconnaissable cliniquement ; par exemple le Syndrome Di George ou la microdélétion 22q11, le syndrome du « Cri du chat avec la délétion partielle du bras court du chromosome 5 (71,72).

Certains syndromes ne sont mis en évidence que par une recherche très ciblée, grâce à l'expertise des cliniciens. Par exemple, le syndrome Cornelia de Lange mis en évidence sur un prélèvement salivaire, la trisomie 22 sur biopsie de peau, le syndrome de Pallister-Killian recherché en FISH car ces anomalies génétiques sont présentes en faible mosaïque (73–75).

L'avènement et les progrès du SHD ont mis en évidence la fréquence des variations de structures, jusqu'à 20 000 par génome (22). L'expertise des cytogénéticiens dans l'interprétation des SV est essentielle : la connaissance des mécanismes de remaniement et de leur complexité, l'interprétation des SV, la détection parfois délicates d'anomalies de structure (28). La limite entre biologie moléculaire et cytogénétique devient plus ténue, la collaboration entre ces disciplines avec l'avènement du SHD est importante.

Les techniques de détection de SV ont beaucoup évoluées au cours des années : le caryotype, la FISH, l'ACPA puis le SHD avec ses multiples algorithmes (28). Chaque technique a des avantages mais aussi des limites. Le SHD peut induire un biais d'interprétation en se concentrant uniquement sur les régions codantes alors que l'ACPA est pangénomique. Le génome complet allie des avantages des deux techniques avec son aspect pangénomique et à sa continuité de lecture.

3) Remaniements non caractérisés en exome

Premièrement, une perte de X en mosaïque à 10% dans le sang chez une patiente avec un Syndrome de Turner ainsi que deux CNV en mosaïques sur l'Y, une duplication de 13Mb et une délétion de 43MB n'ont pas été appelés en exome. Deux hypothèses s'offrent à nous : la technique d'ACPA ne nécessite pas d'amplification préalable, les mosaïques ne sont pas « perdues », contrairement à l'exome où un biais d'amplification peut avoir lieu (76), mais surtout, les pipelines en génétique constitutionnelle sont calibrés pour identifier des variations constitutionnelles. La détection de variations somatiques fait appel à des stratégies bio-informatiques dédiées non appliquées dans ce travail. De plus, la sensibilité de détection des SV en SHD dépend de la profondeur sur la région (18,32).

Deuxièmement, chez un quatuor, nous n'avons pas détecté une duplication héritée paternelle de 390kb, arr[GRCh37] 2q24.3(167602904_167992598)x3, chez le frère du cas index. Comme montré dans les « Détection des CNV : comparaison ACPA versus Exome », cette duplication ne comprend qu'un exon de 409pb dans une large zone intronique. Nous avons démontré que la délétion était détectable manuellement (Figure 20), cependant sous le seuil de significativité de CNV DRAGEN à 500pb. Cette duplication comprend une partie du gène *XIRP2*, impliqué dans des cas de cardiopathie hypertrophique (77).

Cette duplication est visible chez le cas index et son père, ces deux individus avaient un plus grand nombre de reads totaux que le frère. Cette exemple illustre les limites de détection de CNV DRAGEN par la taille de l'anomalie et l'hétérogénéité des qualités de séquençage entre les différents patients (24) (Figure 20).

Deux autres CNV sont non détectés par CNV DRAGEN : la duplication 3p26.3(1638453_2262692)x3 et la délétion arr[GRCh37] 9p24.3(500525_569970)x1. Ces deux CNV sont dans des zones indemnes de sondes d'exome avec le design Illumina mais avec des sondes sur le design Agilent GenetiSure Cyto CGH 8x60K (Agilent) (Figure 21).

L'analyse en exome complet ne permet pas de trouver des SV dans des zones introniques pouvant jouer un rôle dans la régulation génique, à l'inverse de l'ACPA. Par exemple la rupture d'un « Topological Associated Domain » (TAD), structure

pouvant être dans une région intergénique mais influant sur l'expression génique, est, par exemple, une cause connue d'anomalie des membres (78).

4) Amélioration de nos analyses

La comparaison ACPA/Exome dans la détection des CNV nous a permis d'affiner nos connaissances des outils à disposition du CHU de Lille, d'en connaître les limites, mais aussi d'améliorer nos capacités d'analyses des CNV en exome. Par exemple, les PIEV, remaniements fréquents, étaient trop souvent filtrés sur leurs récurrences. Les limites de CNV DRAGEN ont été exacerbées par l'absence de détection de la duplication en 2q24.4. Les remaniements de grande taille et les aneuploïdies sont souvent découpés par l'algorithme CNV DRAGEN : les exons qui permettent leur détection sont loin les uns des autres (24).

Les recommandations actuelles en matière de détection de SV suggèrent d'utiliser plusieurs algorithmes différents pour allier leurs avantages et leurs limites. En effet, certains sont performants pour détecter certains types de SV mais pas d'autre, ils peuvent être influencés par les biais de GC, les régions d'ADN répétées, les variabilités dans la préparation de la librairie (79). Il faut trouver un équilibre sensibilité/ spécificité avec un nombre raisonnable de pipelines à utiliser (80). Au laboratoire, nous avons utilisé les algorithmes proposés par DRAGEN pour leur disponibilité et leur facilité à s'implanter en routine. Une alliance de plusieurs pipelines, de différents développeurs pourrait être envisagée.

C- SV DRAGEN

Le pipeline SV DRAGEN a permis de réaliser un diagnostic chez deux patients non apparentés porteurs d'une délétion hémizygote dans *MECP2*.

Le seuil de détection de SV DRAGEN est de 50pb soit 10 fois inférieur à celui de CNV DRAGEN (24,26). Ces anomalies étaient appelées par GATK mais annotées de façon incorrecte. Elles peuvent être masquées après réalignement des INDELs, donc non appelées par GATK.

SV DRAGEN appelle les anomalies aux deux points de cassure (via le soft clipping) et par read discordant (via le paired-end). L'identification des points de cassure permet une annotation par AnnotSV des gènes impliqués aux extrémité du SV (25,26).

La limite principale de ce pipeline sur exome, avec la discontinuité de la couverture, est qu'il ne peut détecter que les SV intra-exoniques : il repère des points de cassure ou des paired-end incohérents. Il semble complémentaire de CNV DRAGEN par sa résolution.

Ce pipeline génère des données à partir des fichiers .bam, avant réaligement des INDEL. Ces fichiers nécessitent du temps de calcul sur les serveurs DRAGEN, une analyse approfondie et un stockage.

Des microduplications non récurrentes qui induisent des cassures et des réparations de l'ADN double brin au niveau de *MECP2* par le NAHR (Non Allelic Homologous Recombinaison) ont été décrite par Bauters et al. Ces anomalies ont été décrites chez des garçons porteurs de microduplication de 0,3 à 2,3 Mb. Un STR localisé à proximité de la zone minimale critique de ce réarrangement pourrait expliquer la récurrence de ces remaniements (81).

L'exon 4 de *MECP2* a des segments de séquences répétées. On pourrait imaginer un mécanisme proche dans les réarrangements de *MECP2* étudiés dans **Résultats**.

L'étude SV DRAGEN pourrait être envisagée lors de rendu négatif, en réanalyse d'exome. Théoriquement GATK retrouve le même type d'anomalie en première intention mais les annote de façon incorrecte, mais il pourrait ne pas les détecter à cause du réaligement des INDELs (26,82).

D-ROH DRAGEN

Les « runs of homozygotie » ou ROH sont une suite de SNP homozygotes dans une région du génome. Ils sont utilisés en génétique des populations pour évaluer la proximité génétique d'un groupe d'individus, l'existence d'un ancêtre commun. En

génétique mendélienne, ces ROH sur SHD sont utilisés pour trouver des délétions et des disomies (29,83).

1) Validation de méthode de ROH DRAGEN

ROH DRAGEN est un outil de détection de ces ROH qui s'applique sur exome et génome. Sa condition principale est de détecter au moins 50 SNP homozygotes consécutifs chez un individu. La plupart des SNP sont localisés dans les régions introniques du génome (27). Cette condition, facilement accessible en génome, est très difficile à remplir en exome. En effet, les régions exoniques, plus courtes, ne sont pas contiguës, et sont moins susceptibles d'être modifiées (84).

L'analyse ROH DRAGEN a été réalisée deux fois sur des individus témoins : un témoin porteur d'une isodisomie du chromosome 15 (témoin +), un témoin porteur d'une iso et d'une hétérodisomie du chromosome 15 (témoin +/-) et un témoin porteur d'une hétérodisomie du 15 (témoin -). La première analyse a été réalisée sur l'ensemble des reads générés par le séquençage d'exome complet, alors qu'à la seconde analyse, un « blacklist files » a éliminé les régions de mauvaise qualité (liste des artéfacts récurrents sur exome réalisées par Illumina).

La première analyse ne s'est pas avérée concluante : seul le témoin + a bien été mis en évidence, le pipeline détecte un ROH avec un score de 8,32, on retrouve manuellement sur l'ensemble du chromosome 15 des SNP homozygotes absents chez l'un des parents. La région d'isodisomie de 10Mb du témoin +/- n'est pas détectée par ROH DRAGEN : la détection manuelle permet de mettre en évidence 30 SNP homozygotes en 15q11q13.3 absents de chez l'un des parents ; le témoin - a un ROH a type d'isodisomie, retrouvé sur l'analyse manuelle en 15q26.

La seconde analyse « blacklist file » n'avait pas apporté de résultats plus concluants. L'analyse est plus stricte, élimine de nombreux artéfacts mais n'a pas assez de variants homozygotes pour détecter des ROH de plus de 10Mb sur exome, sa résolution est à l'échelle d'un chromosome : on ne détecte que le témoin + mais toujours pas le témoin +/-.

2) Analyse de notre cohorte

D'une part, les isodisomies provoquent des syndromes cliniquement identifiables : Beckwith-Wiedeman, Silver-Russell, par exemple. Les cliniciens orientent les analyses des patients de façon adaptée. Si un syndrome provoqué par une disomie uniparentale est suspectée à posteriori, nous pourrions proposer l'analyse ROH DRAGEN sur les données d'exome, avec des réserves quant à résolution. D'autre part, l'analyse n'a pas été programmée en trio, la détection d'hétérodisomie n'est donc pas possible (27).

Sur une large cohorte de 29 723 individus séquencés en exome, Yaou et al., avaient identifié 14 patients porteurs d'isodisomie, soit un taux diagnostique de 0,047%. Les patients avec plus de deux ROH sur des chromosomes différents ont été exclues de l'analyse de disomie en raison de leur probable consanguinité ou identité par descendance (84).

L'analyse en trio de 4 912 patients, détectant les iso et les hétérodisomies est plus fructueuse avec un taux diagnostique à 0,2% (n=22).

E- Perspective de la réanalyse d'exome et échanges internationaux

L'exome ou le génome offrent la possibilité de réanalyser les données génétiques des patients avec l'évolution des connaissances. Liu et al., ont réanalysé en décembre 2017 des exomes complets séquencés entre 2011 et 2012 ; avec une approche semi-automatisée de priorisation des variants, ils doublent leur taux de diagnostic, passant de 24,8% à 46,8% (85). La grande majorité des diagnostics réalisés sont par la découverte de nouveaux gènes de DI, quelques reclassifications de variants, CNV et études familiales (85). L'évolution des outils bio-informatiques permet aussi la reclassification de variants (28).

Les patients de l'UF n'ayant pas reçus de résultats concluants pourront bénéficier d'une réanalyse d'exome dans un projet à court terme de l'Institut de génétique médicale.

D'autre part, des plateformes d'échange internationales existent, comme Genematcher, PhenoDB et Variantmatcher, où des cliniciens, biologistes et scientifiques du monde entier échangent sur des variants/gènes, proposent des regroupements de cohortes pour clarifier un phénotype ou imputer la causalité d'un variant (86,87). Ces plateformes permettent des collaborations sur des pathologies excessivement rares.

F- Autres pipelines et perspectives

1) Repérer les expansions de séquences répétées

Les expansions de séquences répétées sont responsables de plus de 40 pathologies monogéniques (expansion de Polyglutamine des gènes « SCA » dans les ataxies cérébelleuses dominantes, l'ataxie de Friedreich, le syndrome de l'X-fragile, la maladie de Huntington), certaines d'entre elles font des centaines voire des milliers de paires de bases (88). Ces « Short tandem Repeat » (STR) sont ubiquitaires dans le génome et jouent d'importants rôles dans les procédés cellulaires (89).

Les techniques pour détecter ces expansions sont les « Repeat-Primed PCR » RP-PCR et les PCR longs fragments analysées en taille de fragment. L'interprétation de ces analyses est délicate : on ne peut toujours avoir la taille du fragment, plus le fragment est long plus il est difficile à détecter, des INDELs ou des SNP dans les régions flanquantes peuvent amener des faux positifs ou des faux négatifs (88). Le Southern-Blot, gold standard de ces analyses, est difficile à mettre en place dans les laboratoires, il requiert de grande quantité d'ADN et est sensible à l'hétérogénéité somatique (88).

Des outils informatiques ont été développés, tel ExpansionHunter, LobSTR, GangSTR ou HipSTR pour repérer des expansions pathogènes de STR à travers le génome (88,90–92). Ils ont la capacité d'identifier des expansions plus longues que le read lui-même (89). ExpansionHunter montre des résultats satisfaisants avec *C9ORF72*, impliqué dans différentes pathologies neurodégénératives, avec 95% [98-100%] de sensibilité et 99% [99,7-100%] de spécificité (88). Il est applicable sur génome short read ou long read (89).

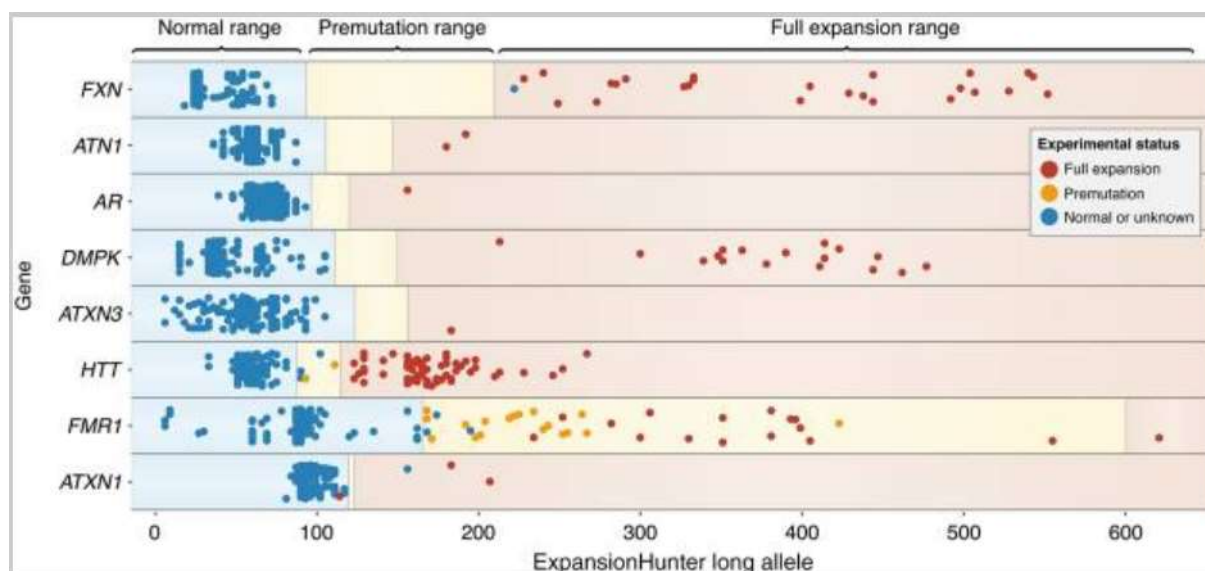


Figure 38 : *Présentation des capacités d'analyses des STR d'ExpansionHunter sur différentes pathologies d'après Dolzhenko E et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. nov 2017. (88)*

L'algorithme, pour quantifier la longueur de la répétition via les reads de SHD, utilise les reads flanquants, ceux au sein et ceux recouvrant la totalité de l'expansion. Une estimation de la longueur est réalisée. ExpansionHunter est programmé pour être utilisé sur les données de WGS car nombre de ces expansions ne sont pas exoniques et l'estimation de la taille de l'expansion nécessite une profondeur moyenne homogène (88). Les séquences STR ne sont pas spécifiques, cependant un allèle pathogène a le plus souvent une longueur bien supérieure à la limite, le off-target gêne l'algorithme dans son calcul de longueur de fragment mais pas dans la détection d'une expansion pathogène. Cet outil permettrait de réaliser l'analyse d'expansion en une fois, sur différentes régions du génome et non par des RP-PCR ciblées (88). Une évaluation des différents algorithmes de détection de STR a été réalisée, ils démontrent que certains comme GangSTR sont utilisables en exome et permettent la détection de STR non connus, malgré la couverture irrégulière et le biais de GC (93) (Figure 38).

2) Localiser l'insertion d'éléments mobiles

Les Insertions d'Eléments Mobiles (MEI) représentent environ 25% des SV d'un génome humain. Elles sont déjà bien connues en pathologie car elles peuvent

interrompre des gènes, en modifier l'expression, comme l'interruption *de novo* du FIX de la coagulation par une séquence *Alu* chez un patient hémophile B décrit en 1993 par Vidaud et al. (94).

MELT (Mobile Element Locator Tool) est un outil développé lors du projet 1000 Genomes car les MEI ne sont pas détectées en routine sur génome (16,95). Il travaille sur les fichiers .bam, il recherche les soft-clipping et les paires discordantes pour repérer les MEI, il évalue l'impact de l'insertion sur le gène et/ou sur les gènes autour (95).

Localiser les MEI avec l'analyse SV aurait pu être intéressant. Cependant, les filtres appliqués sur les fichiers de SV DRAGEN se concentraient sur les gènes d'intérêt dans les troubles du neurodéveloppement, et manuellement nous recherchions les zones soft-clippées. Les MEI ne sont pas particulièrement rapportées dans la littérature scientifique en pathologies du neurodéveloppement, à l'inverse de l'oncologie où ce mécanisme est bien décrit (96).

3) Génome « *short read* »

Les plateformes nationales de séquençage, avec SeqOIA et AURAGEN, donnent aujourd'hui un accès plus simple au génome short read (11). Le génome complet permet d'avoir à la fois les SV par la continuité de sa couverture et les SNV introniques et exoniques (18). Le taux diagnostic du génome en première intention en DI (DEFIDIAG) espère atteindre les 60% (12). La diminution des coûts de séquençage tend à généraliser cette solution en première intention (12,13).

Des algorithmes de recherche spécifique tel que MELT, ExpansionHunter, SV DRAGEN, ROH sont particulièrement adaptés en génome. Ils profitent de la continuité de couverture et de la richesse des données pour extraire et analyser divers aspects du génome comme les STR, les MEI et les SV (26,27,88,95).

4) Génome « *Long read* »

Le génome long read apporte de nombreux avantages, ceux du génome complet mais aussi de la grande taille des reads, qui permettent un meilleur alignement des

séquences répétées. Des technologies comme Oxford Nanopore Technologies et PacBio Sequencing peuvent séquencer des fragments de 10kb à plus de 1Mb directement sur de l'ADN natif (97). Cette évolution du séquençage a permis d'améliorer les connaissances que nous avons sur le génome humain et d'obtenir la séquence de régions encore inexplorées dans les régions télomériques et subtelomériques pour améliorer le génome de référence (22). Audano et al. recensent 22755 SV repartis sur 11Mb en moyenne sur chaque personne séquencée, principalement dans les 5 derniers mégabases des bras des chromosomes (régions subtelomériques avec plus de STR). Il y a significativement plus de SV à proximité des SNV *de novo* (22). Cependant, le séquençage *long read* a lui aussi des limites : un excès de petites insertions et délétions (1-2pb), les reads ne sont pas assez long pour accéder à certaines régions du génome (centrosomes, régions acrocentriques) (22). Parce que ces technologies *Long read* séquencent des brins d'ADN natif, elles peuvent retrouver les modifications de bases telle la méthylation et donnent accès au méthylome sur brins natifs (97).

« These advances will dramatically improve our understanding of human heritability, population diversity, mutational processes, and the genetic basis of disease. Notably, adoption of long-read technology will also change how we discover and catalog human variation [...] » d'après Logsdon et al (97).

5) La cartographie optique

Optical mapping ou la cartographie optique, développée dans les années 90, est une technique proche de la cytogénétique permettant d'assembler *de novo* un génome à partir de molécules d'ADN uniques colorées avec « code barre » selon des motifs nucléotidiques (98). Cette technologie a une étape de fragmentation en très long read, cela conserve les structures de l'ADN. Elle est particulièrement intéressante pour la détection des SV, car n'est pas sensible aux répétitions mais nécessite une grande quantité d'ADN de très bonne qualité, est coûteuse, n'a pas un débit rapide et ne détecte pas les SNV (99,100).

V-Conclusion

Les outils de détection des SV sont complémentaires. CNV DRAGEN est performant pour la détection des CNV à partir de 500pb. La comparaison ACPA/Exome montre des résultats satisfaisant pour la détection des CNV en exome excepté pour les mosaïques, et évidemment dans des régions pauvres en exon. Sa sensibilité et sa finesse de résolution permettent de faire des diagnostics, non vu en ACPA, mais l'exome offre aussi la possibilité de la réanalyse des données avec l'évolution des connaissances.

SV DRAGEN est complémentaire de CNV DRAGEN pour la détection de plus petit SV, intra-exoniques comme l'atteste la détection de délétions dans *MECP2* (Figure 31 & Figure 32). Cet outil a un seuil de détection de 50pb et met en évidence des duplications en tandem et des inversions, événements que ne voit pas CNV DRAGEN. ROH DRAGEN n'a pas permis de faire des diagnostics supplémentaires et n'a pas montré son efficacité en exome.

Les troubles du neurodéveloppement sont une indication privilégiée du SHD en exome ou en génome complet par la grande hétérogénéité génique qui cause ces troubles. Le succès de ces techniques n'est que plus grand lors de collaboration entre biologistes et cliniciens, mais aussi, entre cytogénéticiens et biologistes moléculaires. Les outils bio-informatiques comme MELT, ExpansionHunter, SV DRAGEN ou différents pipelines d'analyses des données génétiques vont rendre accessibles des diagnostics via des mécanismes différents que ceux explorés jusqu'alors en SHD.

Au-delà des perspectives bio-informatiques, des perspectives technologiques disponibles en routine comme le séquençage d'ARN ou le profilage de méthylation, en combinaison du SHD permettront probablement de valider de nouveaux gènes de troubles du neurodéveloppement et de faciliter l'interprétation de certains variants.

On pourrait envisager grâce à ces projets technologiques, et sur le modèle de DEFIDIAG, de réaliser une analyse unique en appliquant plusieurs méthodes d'analyse à la recherche de SV, d'expansion de triplets et SNV.

D'autres techniques comme le méthylome et le transcriptome donnent un aspect fonctionnel aux analyses de génétique : elles permettent de mettre en évidence des

anomalies de méthylation ou des conséquences de variations sur l'épissage des transcrits ou la méthylation de l'ADN.

Bibliographie

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders [Internet]. Fifth Edition. American Psychiatric Association; 2013. Disponible sur: <https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596>
2. Hanly C, Shah H, Au PYB, Murias K. Description of neurodevelopmental phenotypes associated with 10 genetic neurodevelopmental disorders: A scoping review. *Clin Genet.* mars 2021;99(3):335-46.
3. Wilfert AB, Sulovari A, Turner TN, Coe BP, Eichler EE. Recurrent de novo mutations in neurodevelopmental disorders: properties and clinical implications. *Genome Med.* 27 nov 2017;9(1):101.
4. Jensen M, Girirajan S. Mapping a shared genetic basis for neurodevelopmental disorders. *Genome Med.* 14 déc 2017;9(1):109.
5. World Health Organization; 2022. International Classification of Diseases Eleventh Revision (ICD-11) [Internet]. Disponible sur: <https://icd.who.int/browse11/l-m/fr#/http%3a%2f%2fid.who.int%2fcd%2fentity%2f334423054>
6. Maulik PK, Mascarenhas MN, Mathers CD, Dua T, Saxena S. Prevalence of intellectual disability: a meta-analysis of population-based studies. *Res Dev Disabil.* avr 2011;32(2):419-36.
7. Vissers LELM, Gilissen C, Veltman JA. Genetic studies in intellectual disability and related disorders. *Nat Rev Genet.* janv 2016;17(1):9-18.
8. Wechsler, D. Wechsler Adult Intelligence Scale--Fourth Edition (WAIS-IV). APA PsycTests.
9. Saleem M, Beail N, Roache S. Relationship between the Vineland Adaptive Behaviour Scales and the Wechsler Adult Intelligence Scale IV in adults with intellectual disabilities. *J Intellect Disabil Res.* sept 2019;63(9):1158-62.
10. van Ool JS, Hurks PPM, Snoeijen-Schouwenaars FM, Tan IY, Schelhaas HJ, Klinkenberg S, et al. Accuracy of WISC-III and WAIS-IV short forms in patients with neurological disorders. *Dev Neurorehabil.* févr 2018;21(2):101-7.
11. Plan France Médecine Génomique 2025 [Internet]. PFMG 2025. Disponible sur: <https://pfmg2025.aviesan.fr/>
12. Binquet C, Lejeune C, Faivre L, Bouctot M, Asensio ML, Simon A, et al. Genome Sequencing for Genetics Diagnosis of Patients With Intellectual Disability: The DEFIDIAG Study. *Front Genet.* 1 févr 2022;12:766964.
13. Lejeune C, Robert-Viard C, Meunier-Beillard N, Borel MA, Gourvès L, Staraci S, et al. The Economic, Medical and Psychosocial Consequences of Whole Genome Sequencing for the Genetic Diagnosis of Patients With Intellectual Disability: The DEFIDIAG Study Protocol. *Front Genet.* 4 avr 2022;13:852472.

14. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet.* janv 2010;11(1):31-46.
15. Collins FS, Morgan M, Patrinos A. The Human Genome Project: Lessons from Large-Scale Biology. *Science.* 11 avr 2003;300(5617):286-90.
16. Siva N. 1000 Genomes project. *Nat Biotechnol.* mars 2008;26(3):256.
17. Awadalla P, Boileau C, Payette Y, Idaghdour Y, Goulet JP, Knoppers B, et al. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. *International Journal of Epidemiology.* oct 2013;42(5):1285-99.
18. The 100,000 Genomes Project Pilot Investigators, Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, et al. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. *N Engl J Med.* 11 nov 2021;385(20):1868-80.
19. Sherman RM, Salzberg SL. Pan-genomics in the human genome era. *Nat Rev Genet.* avr 2020;21(4):243-54.
20. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nat Rev Genet.* mars 2020;21(3):171-89.
21. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature.* 1 oct 2015;526(7571):68-74.
22. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell.* 24 janv 2019;176(3):663-675.e19.
23. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet.* mai 2011;12(5):363-76.
24. Illumina DRAGEN Bio-IT Platform v3.10. CNV Pipeline Input [Internet]. Disponible sur: https://support-docs.illumina.com/SW/DRAGEN_v310/Content/SW/DRAGEN/CNVInputs.htm
25. Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, et al. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics.* 15 oct 2018;34(20):3572-4.
26. Illumina DRAGEN Bio-IT Platform v3.10. DRAGEN SV Caller Overview [Internet]. Disponible sur: https://support-docs.illumina.com/SW/DRAGEN_v39/Content/SW/DRAGEN/SVOverview.htm
27. Illumina DRAGEN Bio-IT Platform v3.10. ROH Caller [Internet]. Disponible sur: https://support.illumina.com/content/dam/illumina-support/help/Illumina_DRAGEN_Bio_IT_Platform_v3_7_1000000141465/Content/SW/RegionsOfHomozygosity_fDG.htm

28. Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. *Brief Funct Genomics*. sept 2015;14(5):305-14.
29. Dharmadhikari AV, Ghosh R, Yuan B, Liu P, Dai H, Al Masri S, et al. Copy number variant and runs of homozygosity detection by microarrays enabled more precise molecular diagnoses in 11,020 clinical exome cases. *Genome Med*. 17 mai 2019;11(1):30.
30. Karampetsou E, Morrogh D, Chitty L. Microarray Technology for the Diagnosis of Fetal Chromosomal Aberrations: Which Platform Should We Use? *JCM*. 20 juin 2014;3(2):663-78.
31. Riggs ER, Andersen EF, Cherry AM, Kantarci S, Kearney H, Patel A, et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet Med*. févr 2020;22(2):245-57.
32. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. mai 2020;581(7809):434-43.
33. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. janv 2014;42(Database issue):D986-992.
34. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen--the Clinical Genome Resource. *N Engl J Med*. 4 juin 2015;372(23):2235-42.
35. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet*. avr 2009;84(4):524-33.
36. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BWM, Vulto-van Silfhout AT, Bosco P, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet*. oct 2014;46(10):1063-71.
37. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. *Nat Genet*. 14 août 2011;43(9):838-46.
38. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res*. 8 janv 2019;47(D1):D1038-43.
39. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. mai 2015;17(5):405-24.

40. O'Leary NA, Wright MW, Brister JR, Ciuffo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 4 janv 2016;44(D1):D733-745.
41. Sagoo GS, Butterworth AS, Sanderson S, Shaw-Smith C, Higgins JPT, Burton H. Array CGH in patients with learning disability (mental retardation) and congenital anomalies: updated systematic review and meta-analysis of 19 studies and 13,926 subjects. *Genet Med.* mars 2009;11(3):139-46.
42. Monies D, Abouelhoda M, Assoum M, Moghrabi N, Rafiullah R, Almontashiri N, et al. Lessons Learned from Large-Scale, First-Tier Clinical Exome Sequencing in a Highly Consanguineous Population. *Am J Hum Genet.* 6 juin 2019;104(6):1182-201.
43. Adey AC. Tagmentation-based single-cell genomics. *Genome Res.* oct 2021;31(10):1693-705.
44. Auwera G van der, O'Connor BD. Genomics in the cloud: using Docker, GATK, and WDL in Terra. First edition. Sebastopol, CA: O'Reilly Media; 2020. 467 p.
45. Hunt SE, Moore B, Amode RM, Armean IM, Lemos D, Mushtaq A, et al. Annotating and prioritizing genomic variants using the Ensembl Variant Effect Predictor-A tutorial. *Hum Mutat.* 24 nov 2021;
46. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* mars 2014;46(3):310-5.
47. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* janv 2014;42(Database issue):D980-985.
48. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature.* 1 oct 2015;526(7571):75-81.
49. Quenez O, Cassinari K, Coutant S, Lecoquierre F, Le Guennec K, Rousseau S, et al. Detection of copy-number variations from NGS data using read depth information: a diagnostic performance evaluation. *Eur J Hum Genet.* janv 2021;29(1):99-109.
50. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* juin 2011;21(6):974-84.
51. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol.* avr 2016;12(4):e1004873.
52. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods.* sept 2009;6(9):677-81.

53. Chen K, Chen L, Fan X, Wallis J, Ding L, Weinstock G. TIGRA: A targeted iterative graph routing assembler for breakpoint assembly. *Genome Res.* févr 2014;24(2):310-7.
54. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol.* janv 2011;29(1):24-6.
55. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics.* 1 juin 2016;32(11):1749-51.
56. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet.* 6 oct 2016;99(4):877-85.
57. Chennen K, Weber T, Lornage X, Kress A, Böhm J, Thompson J, et al. MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants. *PLoS One.* 2020;15(7):e0236962.
58. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell.* 24 janv 2019;176(3):535-548.e24.
59. Leman R, Parfait B, Vidaud D, Girodon E, Pacot L, Gac GL, et al. SPiP: Splicing Prediction Pipeline, a machine learning tool for massive detection of exonic and intronic variant effect on mRNA splicing. *Preprints;* 2022 févr.
60. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* août 2021;596(7873):583-9.
61. Khanna T, Hanna G, Sternberg MJE, David A. Missense3D-DB web catalogue: an atom-based analysis and repository of 4M human protein-coding genetic variants. *Hum Genet.* mai 2021;140(5):805-12.
62. Köhler S, Gargano M, Matentzoglou N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Research.* 8 janv 2021;49(D1):D1207-17.
63. Fraga D, Meulia T, Fenster S. Real-Time PCR. *Current Protocols Essential Laboratory Techniques.* févr 2014;8(1).
64. Peluso F, Caraffi SG, Zuntini R, Trimarchi G, Ivanovski I, Valeri L, et al. Whole Exome Sequencing Is the Minimal Technological Approach in Proband Born to Consanguineous Couples. *Genes (Basel).* 24 juin 2021;12(7):962.
65. Yigit G, Wieczorek D, Bögershausen N, Beleggia F, Möller-Hartmann C, Altmüller J, et al. A syndrome of microcephaly, short stature, polysyndactyly, and dental anomalies caused by a homozygous KATNB1 mutation. *Am J Med Genet A.* mars 2016;170(3):728-33.
66. Bodea GO, McKelvey EGZ, Faulkner GJ. Retrotransposon-induced mosaicism in the neural genome. *Open Biol.* juill 2018;8(7):180074.

67. Scuffins J, Keller-Ramey J, Dyer L, Douglas G, Torene R, Gainullin V, et al. Uniparental disomy in a population of 32,067 clinical exome trios. *Genet Med.* juin 2021;23(6):1101-7.
68. Stefanski A, Calle-López Y, Leu C, Pérez-Palma E, Pestana-Knight E, Lal D. Clinical sequencing yield in epilepsy, autism spectrum disorder, and intellectual disability: A systematic review and meta-analysis. *Epilepsia.* janv 2021;62(1):143-51.
69. Srivastava S, Love-Nichols JA, Dies KA, Ledbetter DH, Martin CL, Chung WK, et al. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet Med.* nov 2019;21(11):2413-21.
70. Manickam K, McClain MR, Demmer LA, Biswas S, Kearney HM, Malinowski J, et al. Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the American College of Medical Genetics and Genomics (ACMG). *Genet Med.* nov 2021;23(11):2029-37.
71. McDonald-McGinn DM, Hain HS, Emanuel BS, Zackai EH. 22q11.2 Deletion Syndrome. In: Adam MP, Everman DB, Mirzaa GM, Pagon RA, Wallace SE, Bean LJ, et al., éditeurs. *GeneReviews®* [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cité 15 août 2022]. Disponible sur: <http://www.ncbi.nlm.nih.gov/books/NBK1523/>
72. Cerruti Mainardi P. Cri du Chat syndrome. *Orphanet J Rare Dis.* 5 sept 2006;1:33.
73. Deardorff MA, Noon SE, Krantz ID. Cornelia de Lange Syndrome. In: Adam MP, Everman DB, Mirzaa GM, Pagon RA, Wallace SE, Bean LJ, et al., éditeurs. *GeneReviews®* [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cité 15 août 2022]. Disponible sur: <http://www.ncbi.nlm.nih.gov/books/NBK1104/>
74. Abdelgadir D, Nowaczyk MJM, Li C. Trisomy 22 mosaicism and normal developmental outcome: report of two patients and review of the literature. *Am J Med Genet A.* mai 2013;161A(5):1126-31.
75. Slavotinek AM. The genetics of common disorders - congenital diaphragmatic hernia. *Eur J Med Genet.* août 2014;57(8):418-23.
76. Ballif BC, Rorem EA, Sundin K, Lincicum M, Gaskin S, Coppinger J, et al. Detection of low-level mosaicism by array CGH in routine diagnostic specimens. *Am J Med Genet A.* 15 déc 2006;140(24):2757-67.
77. Wang F, Zhao J, Zhang M, Yang J, Zeng G. Genome-wide analysis of the mouse LIM gene family reveals its roles in regulating pathological cardiac hypertrophy. *FEBS Lett.* sept 2021;595(17):2271-89.
78. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell.* 21 mai 2015;161(5):1012-25.

79. Roca I, González-Castro L, Fernández H, Couce ML, Fernández-Marmiesse A. Free-access copy-number variant detection tools for targeted next-generation sequencing data. *Mutat Res Rev Mutat Res.* mars 2019;779:114-25.
80. Moreno-Cabrera JM, Del Valle J, Castellanos E, Feliubadaló L, Pineda M, Brunet J, et al. Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. *Eur J Hum Genet.* déc 2020;28(12):1645-55.
81. Bauters M, Van Esch H, Friez MJ, Boespflug-Tanguy O, Zenker M, Vianna-Morgante AM, et al. Nonrecurrent MECP2 duplications mediated by genomic architecture-driven DNA breaks and break-induced replication repair. *Genome Res.* juin 2008;18(6):847-58.
82. Ulintz PJ, Wu W, Gates CM. Bioinformatics Analysis of Whole Exome Sequencing Data. *Methods Mol Biol.* 2019;1881:277-318.
83. Pippucci T, Magi A, Gialluisi A, Romeo G. Detection of runs of homozygosity from whole exome sequencing data: state of the art and perspectives for clinical, population and epidemiological studies. *Hum Hered.* 2014;77(1-4):63-72.
84. Yaury K, de Leeuw N, Yntema HG, Pfundt R, Gilissen C. Accurate detection of clinically relevant uniparental disomy from exome sequencing data. *Genet Med.* avr 2020;22(4):803-8.
85. Liu P, Meng L, Normand EA, Xia F, Song X, Ghazi A, et al. Reanalysis of Clinical Exome Sequencing Data. *N Engl J Med.* 20 juin 2019;380(25):2478-80.
86. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat.* oct 2015;36(10):928-30.
87. Wohler E, Martin R, Griffith S, Rodrigues E da S, Antonescu C, Posey JE, et al. PhenoDB, GeneMatcher and VariantMatcher, tools for analysis and sharing of sequence data. *Orphanet J Rare Dis.* déc 2021;16(1):365.
88. Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, van Blitterswijk M, Narzisi G, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* nov 2017;27(11):1895-903.
89. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics.* 1 nov 2019;35(22):4754-6.
90. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.* juin 2012;22(6):1154-62.
91. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* 5 sept 2019;47(15):e90.
92. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. *Nat Methods.* juin 2017;14(6):590-2.

93. Rajan-Babu IS, Peng JJ, Chiu R, IMAGINE Study, CAUSES Study, Li C, et al. Genome-wide sequencing as a first-tier screening test for short tandem repeat expansions. *Genome Med.* 9 août 2021;13(1):126.
94. Vidaud D, Vidaud M, Bahnak BR, Siguret V, Gispert Sanchez S, Laurian Y, et al. Haemophilia B due to a de novo insertion of a human-specific Alu subfamily member within the coding region of the factor IX gene. *Eur J Hum Genet.* 1993;1(1):30-6.
95. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* nov 2017;27(11):1916-29.
96. Eyries M, Ariste O, Legrand G, Basset N, Guillerm E, Perrier A, et al. Detection of a pathogenic Alu element insertion in PALB2 gene from targeted NGS diagnostic data. *Eur J Hum Genet.* 11 mars 2022;
97. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet.* oct 2020;21(10):597-614.
98. Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang YK. Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science.* 1 oct 1993;262(5130):110-4.
99. Sahajpal NS, Barseghyan H, Kolhe R, Hastie A, Chaubey A. Optical Genome Mapping as a Next-Generation Cytogenomic Tool for Detection of Structural and Copy Number Variations for Prenatal Genomic Analyses. *Genes (Basel).* 11 mars 2021;12(3):398.
100. Chan S, Lam E, Saghbini M, Bocklandt S, Hastie A, Cao H, et al. Structural Variation Detection and Analysis Using Bionano Optical Mapping. *Methods Mol Biol.* 2018;1833:193-203.

Annexes

Annexe 1 : Tableau de comparaison ACPA versus exome

No Patient	Taille minimale ACPA en kb	Bornes CGH	SV	Classification CGH	Score qualité	Bornes exome HGVs hg38	Taille exome en kb	CNV DRAGEN	ROH DRAGEN	SV DRAGEN
1	170	arr[GRCh37] 15q21.3(54280754_54451481)x1	del	VOUS	119	seq[GRCh38]del(15)(q21.3) NC_0000015.10 :g.[? 54012902] [54143701_ ?]del	131	oui	non	non
2	869	arr[GRCh37] 9q22.31q22.32(96353582_97223224)x3	dup	VOUS	148	seq[GRCh38]dup(9)(q22.31)(q22.32) NC_0000009.12 :g.[? 93543219] [94454072_ ?]dup	911	oui	non	non
3	395	arr[GRCh37] 15q11.2(22784523_23179948)x1	del	PIEV	88	seq[GRCh38]del(15)(q11.2) NC_0000015.10 :g.[? 22786605] [23037175_ ?]del	251	oui	non	non
4	624	arr[GRCh37] 3p26.3(1638453_2262692)x3	dup	PIEV		/		non	non	non
4	1500	arr[GRCh37] 20q12(39314617_40817950)x1	del	VOUS	131	seq[GRCh38]del(20)(q12.1) NC_0000020.11 :g.[? 40687876] [42199397_ ?]del	1512	oui	non	non
5	390	arr[GRCh37] 2q24.3(167602904_167992598)x3	dup	VOUS	58	seq[GRCh38]dup(2)(q22.4.3) NC_0000002.12 :g.[? 169923961] [169924178_ ?]dup	217	oui	non	non
6	390	arr[GRCh37] 2q24.3(167602904_167992598)x3	dup	VOUS		/		non	non	non
7	XXX	arr(X)x3	XXX	Découverte fortuite	PASS	47,XXX		oui	non	non
8	289	arr[GRCh37] Xp21.1(31526372_31824743)x1	del	Découverte fortuite	78	seq[GRCh38]del(X)(q21.1) NC_0000023.11 :g.[? 31496736] [31875424_ ?]del	379	oui	non	non
9	405	arr[GRCh37] Xq26.2(131817692_132222801)x3	dup	VOUS	139	seq[GRCh38]dup(X)(q26.2) NC_0000023.11:g.[? 132669105] [133028271_ ?]dup	359	oui	non	non
10	3000	arr[GRCh37] 22q11.21(19023824_21801661)x3 pat	dup	VS4	112	seq[GRCh38]dup(22)(q11.21) NC_0000022.11:g.[? 19038851] [21127645_ ?]dup	2088	oui	non	non
11	56	arr[GRCh37] Xp11.4(38491539_38547951)x2	dup	VOUS	114	seq[GRCh38]dup(X)(p11.4) NC_0000023.11:g.[? 38646236] [38687718_ ?]dup	41	oui	non	non
12	333	arr[GRCh37] 11q13.1(64535451_64868919)x3	dup	VOUS	118	seq[GRCh38]dup(11)(q13.1) NC_0000011.10:g.[? 64751626] [65117655_ ?]dup	366	oui	non	non
13	222	arr[hg19] 10p12.2p12.1(24580596_24803060)x3	dup	VOUS	43	seq[GRCh38]dup(10)(p12.2)(p12.1) NC_0000010.11:g.[? 24380866] [24513436_ ?]dup	133	oui	non	non
14	170	arr[GRCh37] 19q13.32(47488275_47658389)x3	dup	VOUS	102	seq[GRCh38]dup(19)(q13.32)(p12.1) NC_0000019.10:g.[? 46987977] [47169925_ ?]dup	182	oui	non	non
15	478	arr[GRCh37] 2p13.2(72857637_73336374)x1	del	VOUS	112	seq[GRCh38]del(2)(p13.2) NC_0000002.12 :g.[? 72718101] [73112562_ ?]del	394	oui	non	non
16	XXY	arr(X)x2,(Y)x1	XXY	Découverte fortuite	PASS	47,XXY		oui	non	non
17	6200	arr[GRCh37] 6q15q16.1(88569372_94783161)x1	del	VS4	123	seq[GRCh38]del(6)(q15)(q16.1) NC_0000006.12 :g.[? 88047903] [93419343_ ?]del	5371	oui	non	non
18	490	arr[GRCh37] 15q13.3(32021733_32510863)x3	dup	PIEV	57	seq[GRCh38]dup(15)(q13.3) NC_0000015.10 :g.[? 32424611] [32447084_ ?]dup	81	oui	non	non
19	862	arr[GRCh37] 2q14.2(120870778_121732990)x3	dup	VOUS	79	seq[GRCh38]dup(2)(q14.2) NC_0000002.12 :g.[? 120073095] [120990656_ ?]dup	917	oui	non	non
20	1600	arr[GRCh37] 7q11.23(72726578_74339044)x3	dup	VS5	135	seq[GRCh38]dup(7)(q11.23) NC_0000007.14 :g.[? 73182083] [74896511_ ?]dup	1714	oui	non	non
21	1300	arr[GRCh37] 22q11.21q11.22(21075575_22423120)x3	dup	VOUS	77	seq[GRCh38]dup(22)(q11.21) NC_0000022.11:g.[? 20690029] [21478694_ ?]dup	789	oui	non	non

22	554	arr[GRCh37] Xq27.3(144703386_145257981)x2	dup	VOUS	148	seq[GRCh38]dup(X)(q27.3) NC_0000023.11:g.[?_145822423]_[146027883_ ?]dup	205	oui	non	non
23	XYX	arr(Y)x2	XYX	Découverte fortuite	PASS	47,XYX		oui	non	non
24	489	arr[GRCh37] 15q13.3(32021733_32510863)x3	dup	PIEV	28	seq[GRCh38]dup(15)(q13.3) NC_0000015.10 :g.[?_32030552]_[32112363_ ?]dup	22	oui	non	non
25	780	arr[GRCh37] 7q11.22(69293697_70070355)x1 mat	del	VOUS	53	seq[GRCh38]del(7)(q11.22) NC_0000007.1 :g.[?_68899283]_[68899500_ ?]del	0,2	oui	non	non
26	4300	arr[hg19] 5q23.2(121,842,978- 126,161,864)x1	del	VOUS	114	seq[GRCh38]del(5)(q23.2) NC_0000005.10 :g.[?_122775101]_[126836293_ ?]del	4061	oui	non	non
27	78	arr[GRCh37] 16q24.3(89429676_89507835)x1	del	VS2	28	seq[GRCh38]del(16)(q24.3) NC_0000016.10 :g.[?_89418245]_[89418368_ ?]del	0,1	oui	non	non
28	400	arr[GRCh37] Xp22.31(6705268_7106107)x2	dup	VOUS	143	seq[GRCh38]dup(X)(p22.31) NC_0000023.11:g.[?_7050294]_[7148012_ ?]dup	98	oui	non	non
29	X à 10%	(X)x1[0.1]	X 10%	VOUS		mos 45,X[1]/46,XX[18]		non	non	non
29	157	arr[GRCh37] 9q33.1(119419472_119576586)x1,	del	VOUS	123	seq[GRCh38]del(9)(q31.3) NC_0000009.12 :g.[?_116618321]_[116820785_ ?]del	202	oui	non	non
30	804	arr[GRCh37] 1q32.2(207780569_208585415)x3	dup	VOUS	130	seq[GRCh38]dup(1)(q32.2) NC_0000001.10:g.[?_207523609]_[208218084_ ?]dup	694	oui	non	non
31	236	arr[GRCh37] 8q22.2(100160197_100396502)x1	del	VOUS	105	seq[GRCh38]del(8)(q22.2) NC_0000009.12 :g.[?_99134580]_[99391755_ ?]del	257	oui	non	non
32	15150	arr[GRCh37] 13q21.1q21.33(56880053_72037960)x3	dup	VOUS	100	seq[GRCh38]dup(13)(q21.1)(q21.33) NC_0000023.11:g.[?_57141220]_[71479318_ ?]dup	14380	oui	non	non
33	15150	arr[GRCh37] 13q21.1q21.33(56880053_72037960)x3	dup	VOUS	100	seq[GRCh38]dup(13)(q21.1)(q21.33) NC_0000023.11:g.[?_57141220]_[71479318_ ?]dup	14380	oui	non	non
34	524	arr[GRCh37] 16p11.2(29673954_30198600)x3	dup	PIEV	125	seq[GRCh38]dup(16)(p13.11) NC_0000016.10 :g.[?_29565289]_[30207096_ ?]dup	642	oui	non	non
35	1570	arr[GRCh37] 3q29(195740357_197310451)x1	del	VOUS	118	seq[GRCh38]del(3)(q29) NC_0000003.12 :g.[?_196027171]_[197546452_ ?]del	1500	oui	non	non
35	628	arr[GRCh37] 9p24.3(534531_681480)x3	dup	VOUS	82	seq[GRCh38]dup(9)(p24.3) NC_0000009.12 :g.[?_676941]_[745254_ ?]dup	68	oui	non	non
36	6200	arr[GRCh37] 6q15q16.1(88569372_94783161)x1	del	VS4	123	seq[GRCh38]del(6)(q15)(q16.1) NC_0000006.12 :g.[?_88047903]_[93419343_ ?]del	5371	oui	non	non
37	XXY	arr(X)x2,(Y)x1	XXY	Découverte fortuite	PASS	47,XXY		oui	non	non
38	1200	arr[GRCh37] 1q21.1q21.2(146564743_147745610)x1 mat	del	PIEV	76	seq[GRCh38]del(1)(q21.1)(q21.2) NC_0000001.10 :g.[?_146960142]_[147936706_ ?]del	977	oui	non	non
39	306	arr[GRCh37] 15q11.2(22873688_23179948)x1	del	PIEV	78	seq[GRCh38]del(15)(q21.3) NC_0000015.10 :g.[?_22786605]_[23039545_ ?]del	253	oui	non	non
40	306	arr[GRCh37] 15q11.2(22873688_23179948)x1	del	PIEV	78	seq[GRCh38]del(15)(q21.3) NC_0000015.10 :g.[?_22786605]_[23039545_ ?]del	253	oui	non	non
41	139	arr[hg19] 4q28.3(134068963_134208021)x1	del	VOUS	82	seq[GRCh38]del(4)(q28.3) NC_0000004.12 :g.[?_133150138]_[133190200_ ?]del	40	oui	non	non
42	283	arr[GRCh37] 12q23.1(99844789_100127855)x1	del	VOUS	81	seq[GRCh38]del(12)(q23.1) NC_0000012.12 :g.[?_99443671]_[99655210_ ?]del	211	oui	non	non
43	164	arr[GRCh38] 16q24.3(89104234_89269079)x1	del	VOUS	106	seq[GRCh38]del(16)(q24.3) NC_0000016.10 :g.[?_89098589]_[89268714_ ?]del	170	oui	non	non
44	69	arr[hg19] 9p24.3(500525_569970)x1	del	VOUS		/		non	non	non

45	1600	arr[GRCh37] 18p11.32(597754_2190664)x1	del	VOUS	119	seq[GRCh38]del(18)(p11.32) NC_0000018.10 :g.[? 580406] [1359460 ?]del	779	oui	non	non
46	82	arr[GRCh37] Xp21.2(31333408_31415703)x0	del	VOUS	150	seq[GRCh38]del(X)(q21.1) NC_0000023.11 :g.[? 31323546] [31627913 ?]del	304	oui	non	non
47	191	arr[GRCh37] 7q35(146367349_146558801)x1	del	VOUS	58	seq[GRCh38]del(7)(q35) NC_0000007.14 :g.[? 146774219] [146839955 ?]del	66	oui	non	non
48	1100	arr[GRCh37] 3p24.1p23(301111101_31245911)x3	dup	VOUS	91	seq[GRCh38]dup(3)(p24.1)(p23) NC_0000003.12 :g.[? 30606832] [30894645 ?]dup	288	oui	non	non
49	93	arr[GRCh37] 14q32.2(100716802_100809813)x3	dup	VOUS	119	seq[GRCh38]dup14(q32.2) NC_0000014.9 :g.[? 100239193] [100343389 ?]dup	104	oui	non	non
50	429	arr[GRCh37] 16p12.2(21951379_22380197)x3	dup	VOUS	80	seq[GRCh38]dup(16)(p12.2) NC_0000016.10 :g.[? 21953330] [22374244 ?]dup	420	oui	non	non
51	729	arr[GRCh37] 5q31.1q31.2(135583508_136312689)x1	del	VOUS	143	seq[GRCh38]del(5)(q23.2) NC_0000005.10 :g.[? 136192212] [136985221 ?]del	793	oui	non	non
52	695	arr[GRCh37] 3p26.1(6274495_6969739)x3 pat	dup	VOUS	76	seq[GRCh38]dup(3)(p26.1) NC_0000003.12 :g.[? 6786589] [6863041 ?]dup	76	oui	non	non
53	329	arr[GRCh37] 16q24.2(87485102_87814273)x1	del	VOUS	123	seq[GRCh38]del(16)(q24.3) NC_0000016.10 :g.[? 87460005] [87765892 ?]del	305	oui	non	non
54	627	arr[GRCh37] 4q21.22q21.23(83900040_84526966)x3	dup	VOUS	105	seq[GRCh38]dup(4)(q21.22) NC_0000004.12 :g.[? 82978880] [83604769 ?]dup	626	oui	non	non
54	304	arr[GRCh37] 4q21.3(86937696_87277469)x3	dup	VOUS	93	seq[GRCh38]dup(4)(q21.3) NC_0000004.12 :g.[? 86017225] [86194418 ?]dup	626	oui	non	non
55	1800	arr[GRCh37] 8q21.13(81591520_83365015)x3	dup	VOUS	73	seq[GRCh38]dup(8)(q21.13) NC_0000008.11 :g.[? 80659582] [81839988 ?]dup	1180	oui	non	non
56	XYX	arr(Y)x2	XYX	Découverte fortuite	PASS	47,XYX		oui	non	non
57	13401	arr[GRCh37] Yp11.31q11.221(2656461_16057257)x2[0.2]	dup	Découverte fortuite		/		non	non	non
57	43	arr[GRCh37] Yq11.221q12(16188682_59335913)x0[0.2]	del	Découverte fortuite		/		non	non	non
58	1000	arr[GRCh37] 8p23.3(977218_1941913)x3	dup	VOUS	83	seq[GRCh38]dup(8)(q23.3) NC_0000008.11 :g.[? 1501347] [2059249 ?]dup	558	oui	non	non
59	1500	arr[GRCh37] Xp22.31(6552712_8115153)	del	VOUS	150	seq[GRCh38]del(X)(p22.31) NC_0000023.11 :g.[? 7057705] [8170351 ?]del	1113	oui	non	non
60	790	arr[GRCh37] 22q13.1q13.2(40338828_41128569)x3	dup	VOUS	78	seq[GRCh38]dup(22)(q13.1) NC_0000022.11:g.[? 39947095] [40774028 ?]dup	827	oui	non	non
60	252	arr[GRCh37] Xq21.1(79807232_80059046)x2	dup	VOUS	150	seq[GRCh38]dup(X)(q21.1) NC_0000023.11:g.[? 80676557] [80809567 ?]dup	371	oui	non	non
60	360	arr[GRCh37] Xq23(110047022_110406949)x2	dup	VOUS	87	seq[GRCh38]dup(X)(q23) NC_0000023.11:g.[? 154195878] [154252673 ?]dup	56	oui	non	non
61	177	arr[GRCh37] 10q22.1q22.2(74874146_75051097)x1 pat	del	VOUS	94	seq[GRCh38]del(10)(q22.1)(q22.2) NC_0000010.11 :g.[? 73114357] [73299683 ?]del	185	oui	non	non
62	176	arr[GRCh37] Xq26.1(128701259_128877928)x2	dup	VOUS	150	seq[GRCh38]dup(X)(q26.1) NC_0000023.11:g.[? 129540429] [129756569 ?]dup	216	oui	non	non
63	490	arr[GRCh37] 15q13.3(32021733_32510863)x3	dup	PIEV	48	seq[GRCh38]dup(15)(q13.3) NC_0000015.10 :g.[? 32030644] [32112363 ?]dup	82	oui	non	non
64	1500	arr[GRCh37] 15q13.2q13.3(31014508_32510863)x3	dup	PIEV	96	seq[GRCh38]dup(15)(q13.2) NC_0000015.10 :g.[? 30582890] [32112363 ?]dup	1529	oui	non	non

65	1200	arr[GRCh37] 16p13.11(15048751_16276115)x3	dup	PIEV	67	seq[GRCh38]dup(16)(p13.11) NC_0000016.10 :g.[? 14708901] [16393923_?]dup	1685	oui	non	non
66	198	arr[GRCh37] 8q24.12(120743147_120942109)x3	dup	VOUS	88	seq[GRCh38]dup(8)(q24.12) NC_0000008.11 :g.[? 119731872] [119965412_?]dup	233	oui	non	non
67	278	arr[GRCh37] 16p13.3(154314_433278)x3	dup	VOUS	97	seq[GRCh38]dup(16)(p13.3) NC_0000016.10 :g.[? 58971] [397940_?]dup	339	oui	non	non
68	867	arr[GRCh37] 4q21.1(77828305_78695754)x3	dup	VOUS	103	seq[GRCh38]dup(4)(q21.1) NC_0000004.12 :g.[? 76895465] [77776392_?]dup	881	oui	non	non
68	391	arr[GRCh37] Xp11.22(53430723_53822101)x3	dup	VOUS	110	seq[GRCh38]dup(X)(p11.22) NC_0000023.11:g.[? 53394726] [53654158_?]dup	259	oui	non	non

FACULTE DE PHARMACIE DE LILLE

THESE EN VUE DU DIPLOME D'ETAT DE DOCTEUR EN PHARMACIE

Année Universitaire 2021/2022

Nom : PLANTE-BORDENEUVE

Prénom : Pauline

Titre de la thèse : Complémentarité des approches bio-informatiques en séquençage d'exome pour la détection des variations de structure : implications sur les taux de diagnostics dans une cohorte de patients porteurs d'anomalies du neurodéveloppement

Mots-clés : Variation de structure, exome, CNV, SV DRAGEN, CNV DRAGEN, ROH DRAGEN

Résumé :

Contexte : L'identification des variations de structures (SV) est essentielle pour l'interprétation des données de séquençages pangénomiques mais est historiquement un sujet complexe. Les SV sont extrêmement variés, >50pb à plusieurs mégabases. Ils sont présents dans le génome humain, affectent les fonctions cellulaires et moléculaires, les fonctions de régulation, les structures 3D et la machinerie transcriptionnelle. Leur détection est un enjeu majeur de la génétique actuelle, autant sur le plan de la connaissance du génome, de ses structures, qu'en pathologie humaine. De multiples stratégies ont émergé pour améliorer la détection des SV, notamment en exome : utilisation d'algorithmes couplés au séquençage haut débit, seul ou en association pour plus de performance, le séquençage troisième génération avec les Long Reads et la cartographie optique.

Méthode : Sur les données d'exome, nous avons évalué 3 approches bio-informatiques pour la détection de SV : CNV DRAGEN, SV DRAGEN et ROH DRAGEN. Nous avons mesuré leur impact sur l'apport diagnostique à l'Institut de Génétique Médicale du CHU de Lille sur une cohorte de patients porteurs de troubles du neurodéveloppement de Septembre 2020 à Mars 2022.

Résultats : La complémentarité de ces approches et leurs limites ont ainsi pu être identifiées : CNV DRAGEN trouve avec une grande efficacité les CNV/SV de plus de 500pb alors que SV DRAGEN à partir de 50pb. Cependant, la couverture fragmentée et partielle de l'exome ne permet pas une utilisation optimale de ces pipelines.

Conclusion : L'utilisation de CNV DRAGEN aura permis de faire 8 diagnostics supplémentaires et SV DRAGEN 2, sur 600 patients, soit un apport diagnostique de 1,7%.

Membres du jury :

Président :

Monsieur le Professeur Thierry BROUSSEAU

Faculté de Pharmacie, Université Lille

Directeur de Thèse :

Monsieur le Docteur Thomas SMOL

Faculté de Médecine, Université Lille

Assesseur(s) :

Madame la Professeure Claire-Marie DHAENENS

Faculté de Pharmacie, Université Lille

Madame la Professeure Catherine ROCHE-LESTIENNE

Faculté de Médecine, Université Lille

Monsieur le Docteur Jamal GHOUIM

Faculté de Médecine, Université Lille