

Université de Montréal

Séquençage d'exomes d'une cohorte de familles caucasiennes simplex dont les patients sont atteints du syndrome d'interruption de la tige hypophysaire

Par

Martineau Jean-Louis

Programme de Bio-Informatique, Département de Biochimie
Faculté de Médecine

Mémoire présenté à la faculté de Médecine en vue de l'obtention
du grade de Maitrise en sciences Bio-Informatiques

Avril 2017

© Martineau Jean-Louis, 2017

Résumé

Le syndrome d'interruption de la tige hypophysaire (PSIS) est un désordre rare qui affecte la fonction du système endocrinien. Jusqu'à nos jours, l'imagerie par la résonance magnétique (IRM) demeure la méthode la plus entreprise afin d'évaluer in vivo l'anomalie d'organogenèse de la tige hypophysaire chez les patients. L'absence de la tige caractérise une déficience permanente en hormone de croissance (GH) pendant que l'étiologie du syndrome demeure inconnue. PSIS se définit comme l'hypopituitarisme congénital et il se caractérise soit par une ectopique post-hypophysaire, soit par une hypoplasie antéhypophysaire, ou encore par une hypoplasie de la tige hypophysaire. Notre objectif consiste à déterminer les mutations génétiques partagées entre les sujets affectés de l'étude et qui pourraient expliquer les causes du syndrome. Pour y parvenir, nous avons analysé les données de séquençage d'exomes (WES) provenant de sept familles caucasiennes simplex, une famille d'origine arabe et cinq autres dont la généalogie est incomplète. Ces données ont été précédemment analysées, pour le même but, par d'autres membres de l'équipe en utilisant le pipeline bio-informatique standard basé sur l'utilisation du logiciel GATK. Nous avons préférentiellement opté pour une nouvelle analyse en utilisant deux différents pipelines bio-informatiques indépendants, pour ensuite comparer conjointement les résultats obtenus. Notre protocole consiste à assembler : d'abord, deux pipelines alternatifs de détection de mutations génétiques ponctuelles (SNV). Ils sont composés d'un logiciel d'alignement de séquence (Bowtie2) et deux logiciels d'appel de variantes (Freebayes et SAMtools). Ensuite, nous avons assemblé trois pipelines de détection de variations du nombre de copies (CNV) génomiques composés communément d'un logiciel d'alignement de séquence (BWA) et trois logiciels d'appel de CNV (CoNIFER, fishingCNV, xHmm). Nos résultats nous ont permis d'identifier des mutations candidates additionnelles qui n'ont jamais été identifiées. De plus, notre méthodologie nous a permis de caractériser certains résultats faux positifs, par conséquent elle pourra nous aider à améliorer la performance des pipelines de détection de variations génomiques existantes.

Mots-clés : tige, hypophysaire, PSIS, IRM, WES, GH, SNP, CNV, NGS, séquençage, exome, GATK, Bowtie2, Freebayes, SAMtools, BWA, CoNIFER, fishingCNV, xHmm, génomique, pipeline, syndrome, désordre, rare.

Abstract

Pituitary stalk interruption syndrome (PSIS) is a rare disorder that affects the function of the endocrine system of the affected individuals. The absence of the pituitary stalk, assessed by MRI, characterizes patients with permanent growth hormone deficiency while the etiology of the syndrome remains unknown. PSIS is defined by clinical hypopituitarism together with anatomical findings including a hypoplastic anterior pituitary, ectopic posterior pituitary and reduced or hypoplastic pituitary stalk. We aim to find shared variations (SNP, CNV) among affected patients in coding regions which could explain the origin of the syndrome. We analyzed the exome NGS data from 8 affected French Canadian trio families, with one additional consanguineous Arabic trio family and 5 families with incomplete pedigree. These data were previously analyzed, for the same objective, by other members of the team using a standardize GATK based bioinformatics pipeline. It was desired to reanalyze the complete data set with two other independent pipelines, followed by a comparison of the SNP discovery results. In the present aspect of the study, we built two SNP discovery pipelines, both composed of a different NGS data aligners (Bowtie2) and each composed a different variant caller (Freebayes, SAMtools), then a CNV discovery pipeline which is composed of three different CNV callers (CoNIFER, fishingCNV, xHmm). In addition to the candidate mutations identified in the previous analysis, we identified additional candidate mutations which had not been detected and never been reported. Furthermore, our method helps to discover the sources of variation false discovery which could help to improve existing genomic mutation discovery pipelines.

Keywords: Pituitary, stalk, PSIS, rare, disorder, endocrine, MRI, magnetic, resonance, growth, hormone, SNP, CNV, exome, NGS, GATK, pipeline, Bowtie2, Freebayes, SAMtools, CoNIFER, fishingCNV, xHmm, genomic, mutation, WES.

Table des matières

Résumé	i
Abstract	ii
Liste des figures	vi
Liste des tableaux	viii
Sigles et abréviations	xi
Dédicace	xii
Remerciements	xiii
Contexte	1
Chapitre 1 - Introduction à la génétique moléculaire.....	1
1.1 - Concepts clés de la génétique moléculaire	1
1.1.1 - Caractéristiques des acides nucléiques.....	2
1.1.2 - Propriétés des gènes codants et non codants	4
1.1.3 - La ploïdie et les allèles	7
1.1.4 - Variant génétique et expression du génome.....	7
1.1.5 - Les variations et les mécanismes d'expression génique	8
1.1.6 - Impact des variations du nombre de copies.....	9
Chapitre 2 - Survole du système endocrinien	11
2.1 - Composantes et fonctions du système endocrinien.....	11
2.1.1 - Le système endocrinien	11
2.1.2 - Composantes fonctionnelles du système endocrinien.....	11
2.1.3 - L'hypothalamus et l'hypophyse.....	12
2.1.4 - Anatomie de l'hypophyse.....	12
2.1.5 - Organogenèse de l'hypophyse	13
2.1.6 - La formation de la poche de « Rathke ».....	15
2.1.7 - Origine des cellules spécialisées de l'antéhypophyse	17
2.2 - Populations et fonctions des cellules de l'antéhypophyse	20
2.2.1 - Les populations cellulaires de l'antéhypophyse	20
2.2.2 - Variation de la fonction de sécrétion hormonale	23
Chapitre 3 - Analyses automatisées des données NGS.....	27
3.1 - L'analyse et l'étiologie du PSIS.....	27

3.1.1 - Historiques du séquençage de molécules d'acides nucléiques	28
3.1.2 - Évolution de la technologie de séquençage	29
3.1.3 - Les diverses générations de séquençages	30
3.1.4 - Catégories séquençage d'ADN	32
3.1.5 - Flux de données massives.....	33
3.1.6 - Les données Fasta et FastQ	34
3.2 - Analyse bio-informatique de données NGS.....	35
3.2.1 - Outils d'alignement de séquences.....	36
3.2.2 - Alignement local et global.....	38
Chapitre 4 - Aspects expérimentaux du projet	42
4.1 - Hypothèses et objectifs du projet	42
4.2 - Matériels et méthodologies	43
4.2.1 - Provenance des données et évaluation des sujets	43
4.2.2 - Analyse des données de séquençage NGS	45
4.2.3 - Correction de la qualité des lectures	47
4.2.4 - Choix des outils d'alignement des séquences.....	51
4.2.5 - Procédure de détection de variation	57
4.2.6 - Méthodes de sélection des variations	61
4.2.7 - Annotation fonctionnelle des variations.....	62
4.2.8 - Recherche des variations du nombre de copies	71
4.2.9 - Expression spécifique de gènes	77
Chapitre 5 - Résultats	78
5.1 - Analyses et sélection des mutations	78
5.1.1 - Mutations partagées à travers les patients	78
5.1.2 - Analyse des mutations récessives rares.....	79
5.1.3 - Recherche des mutations <i>de novo</i> , non-sens et dominantes	80
5.1.4 - Mutations hemizygotés délétères liées au chromosome X.....	82
5.1.5 - Comparaison de résultats de SNV rares.....	85
5.1.6 - Vérification des INDEL avec les données de dbSNP	87
5.1.7 - Variations du nombre de copies.....	88
Chapitre 6 - Discussion et conclusion	92
6.1 - Méthodologie et mutations partagées	92

6.1.1 - Mutations ponctuelles candidates	93
6.1.2 - Variations du nombre de copies.....	95
6.1.3 - Analyse comparative des pipelines	96
6.2 - Conclusion	99
6.3 - Perspectives.....	99
Indexes	100
Annexes	104
Références	106

Liste des figures

Figure 1-1 - Représentation des composantes de la cellule eucaryote et d'un modèle simpliste d'expression de l'ADN.....	3
Figure 1-2 - Représentation d'une séquence d'ADN double brin formé de quatre nucléotides appariés.	4
Figure 1-3 - Table du code génétique de vertébrés et représentation moléculaire des bases azotées de l'ADN.....	5
Figure 1-4 - Représentation de la structure d'un gène eucaryote et ses régulateurs de transcription	6
Figure 1-5 - Variation fonctionnelle et moléculaire des gènes affectés par des SNV et des INDEL.....	10
Figure 2-1 - Localisation physiologique et représentation schématique de l'hypophyse. 14	
Figure 2-2 - Développement embryonnaire de la glande hypophysaire et sa composition en cellules spécialisées.....	16
Figure 2-3 - Modèle de développement embryonnaire de l'hypophyse dans des modèles murins.	19
Figure 2-4 - Différenciation séquentielle des cellules de l'antéhypophyse sous l'effet d'activation de différents facteurs génétiques.....	21
Figure 3-1 - Représentation graphique de la décroissance du coût de séquençage du génome humain à travers les ans.	29
Figure 3-2 - Procédure de détection des nucléotides incorporés durant le pyro-séquençage de l'ADN.....	31
Figure 3-3 - Représentation d'une lecture d'ADN séquencée sous un format Sanger FastQ.	34
Figure 3-4 - Représentation graphique du modèle d'alignement local et global d'une séquence mutante contre sa référence.....	41
Figure 4-1 - Illustration de différentes étapes d'enrichissement d'exons de l'ADN génomique.....	45
Figure 4-2 – Plan d'analyse des pipelines de détection de variants ponctuels spécifiques aux outils GATK, FREEBAYES, SAMTOOLS.	46
Figure 4-3 - Image comparative des résultats d'analyse de séquences d'un échantillon avant et après le traitement de la qualité via Trimmomatic PE par l'outil FastQC.	50

Figure 4-4 - Représentation de la couverture des SNV en fonction de la sensibilité des cinq meilleurs pipelines.	53
Figure 4-5 - Représentation du niveau de faux alignements générés par les logiciels BOWTIE2, segemehl, BWA et STAR.	54
Figure 4-6 - Évaluation du temps d'exécution de la recherche d'un site d'alignement par séquence des logiciels BOWTIE2, SEGEMEHL, BWA et STAR.....	54
Figure 4-7 - Évaluation de la consommation en mémoire lors de la recherche d'un site d'alignement par séquence des logiciels BOWTIE2, SEGEMEHL, BWA et STAR.	55
Figure 4-8 - Représentation des substitutions nucléotidiques du type transition ou traversions.	61
Figure 4-9 - Modèles de variations génétiques retenue dans nos protocoles de sélection de variants candidates	64
Figure 4-10 - Représentation de l'analyse de la performance comparative des algorithmes de prédiction des scores délétères des mutations codantes.....	70
Figure 4-11 - Représentation des scores d'impacts délétères des variations codantes prédits par les algorithmes de prédiction SIFT, PolyPHEN2 et MutationTaster.....	71
Figure 4-12 - xHmm, FISHINGCNV, CONIFER combined CNV discovery pipelines..	75
Figure 5-1 - Analyse de partage de mutations rares délétères entre les sujets affectés. ...	79
Figure 5-2 - Représentation du diagramme de Venn des SNV générés par les protocoles de détection de variants génomiques de GATK, SAMTOOLS et FREEBAYES.	86
Figure 5-3 - Représentation par diagramme de Venn de la comparaison des résultats d'INDEL présents et absents dans la base de données de dbSNP (snp138).....	88
Figure 5-4 - Représentation graphique de la délétion du nombre de copie homozygote du gène MGAM indiquée au tableau 5.6.	91

Liste des tableaux

Tableau 2-1 - Déficiences héréditaires en hormone hypophysaire causées par des facteurs de transcription mutants.....	25
Tableau 3-1 - Relation des scores phred avec la précision de la détection des bases.....	35
Tableau 4-1 - Disponibilité des différents logiciels utilisés pour déroulement des deux pipelines de détection de mutations ponctuelles.....	49
Tableau 4-2 - Analyse des ratios Ti/Tv des résultats de détection de variants par les logiciels FREEBAYES et SAMTOOLS.....	60
Tableau 4-3 - Paramètres de filtrage de résultats de détection des variations ponctuelles.....	62
Tableau 4-4 - Disponibilité des différents logiciels utilisés pour déroulement des trois pipelines de détection de CNV.	76
Tableau 5-1 - Mutation récessives rares retenues selon nos critères de sélections indiquées au Tableau 4-3.	81
Tableau 5-2 - État de la vérification des mutations délétères de novo par la méthode de séquençage Sanger PCR.	82
Tableau 5-3 - Annotation de la pathogénicité et du score d'effets délétères des SNV <i>de novo</i>	83
Tableau 5-4 - Candidats de mutations récessives rares, non-sens liées à l'X.....	84
Tableau 5-5 - CNV de novo retenus après les procédures de filtrage et de sélection.....	90
Tableau S 1 - Description clinique des attributs phénotypiques de la cohorte de 13 patients	104
Tableau S 2 - Résultats détaillés d'analyse de proportion de variants partagées entre les outils de détections de SNV.....	105

Sigles et abréviations

ACTH:	Hormone adrenocorticotropique
ACTHR:	Récepteur d'hormones adrenocorticotropique (MC2R)
ATP:	Adénosine triphosphate
ADN:	Acide désoxyribonucléique
ARN:	Acides ribonucléique
ARNm:	ARN messenger
ARNpri:	ARN primaire
BAM:	Alignement de séquences en format binaire (file)
BMP4:	Protéine de la morphogenèse des os (Famille 4)
BWA:	Aligneur de séquences « Burrow wheeler »
CLINVAR:	Variation clinique
CNV:	Variation du nombre de copie
CRH(CRF):	Facteur de libération de la corticotrophine
CRHR:	Récepteur d'hormones de libération de la corticotrophine
EPH:	Ectopique postérieure de l'hypophyse
FGF10:	Facteur de croissance des fibroblastes (Famille 10)
FGF8:	Facteur de croissance des fibroblastes (Famille 8)
FSHR:	Récepteur d'hormones stimulantes folliculaires.
FSHb:	hormones stimulantes folliculaire du type bêta
GATK:	Genome analysis toolkit
GH:	Hormone de croissance
GH1:	Hormone de croissance type 1
GHRH:	Facteur de libération d'hormones de croissance
GHRHR:	Récepteur d'hormones de libération d'hormones de croissance
GHR:	Récepteur d'hormones de croissance
GNHRH:	Facteur de libération de la gonadotrophine

GNRH:	Facteur de synthèse de la gonadotrophine
HPRL:	Hyperprolactinémie
HESX1	Homeobox expressed in Embryonic Stem cells 1
hGH-N:	Hormone de croissance humaine (Normal) 1
hGH-V:	Hormone de croissance humaine (Variant) 2
HAH :	Hypoplasie antéhypophysaire
HTH :	Hypoplasie de la tige hypophysaire
INDEL:	Petites insertions et délétions
ISL1:	insulin gene enhancer protein ISL-1
Kb:	kilo paires de bases
LH:	Hormone lutéinisante
LH3:	lysine hydroxylase type 3
LHR:	Récepteur d'hormones lutéinisantes
LHX3:	LIM homeobox protéine type 3
LHX4:	LIM homeobox protéine type 4
LHb:	Hormone lutéinisante type beta
MAF:	Fréquence d'allèles mineurs
MC2R:	Récepteur d'hormones melanocortin type 2
IRM:	Imagerie par résonance magnétique
PCR:	Réaction en chaine par l'ARN polymérase
POMC:	Pro-opiomelanocortin
PRL:	Prolactine
PRLR:	Récepteur de la prolactine
PSIS:	Syndrome d'interruption de la tige hypophysaire
SHH:	Sonic hedgehog
SNV:	Variation ponctuelle d'un nucléotide
THRH:	Facteur de libérations d'hormones Thyroïdiennes
TSH:	Hormone stimulant de la thyroïde
TSHB:	Hormone stimulant de la thyroïde du type beta

TSHR: Récepteur d'hormones stimulant de la thyroïde

WES : Séquençage du génome complet

Dédicaces

A ma courageuse femme Murielle.

Remerciements

Le projet a été soutenu par le Réseau de la Médecine Génétique Appliquée (RMGA), Génome Canada et Génome Québec, et le Centre de Recherche du CHUSJ. Nous remercions l'équipe dans le service d'endocrinologie du CHUSJ, surtout Profs. Cheri Deal, Johnny Deladoey et Guy Van Vliet. La cohorte de patients a été définie avec l'aide du Dr. Caroline Hasselmann et celui du Dr. Despoina Manousaki. Les analyses préliminaires bio-informatiques ont été effectuées par M. Roman Serpa. Les données génomiques de sujets ont été fournies avec l'aide de Mr. Dan Spiegelmann et Alexandre Lapointe sous la supervision du Prof. Guy Rouleau.

La totalité des manipulations, telles que l'acquisition, les traitements et les analyses des données génomiques ont été effectuées par moi-même sous la supervision intégrale du professeur Mark Samuels et à quelques reprises de Mr. Dan Spiegelmann. Par contre, j'ai utilisé le pronom « nous » à travers tout ce document afin de mettre en valeur le travail d'équipe effectué au sein de notre laboratoire.

Je remercie ma femme et tout le reste de ma famille, incluant mes enfants Arielle et William, pour toute la patience qu'ils m'ont accordée dans le cheminement vers la réalisation de mon rêve.

Je remercie énormément mon directeur de recherche Dr. Mark E. Samuels, d'avoir cru en moi et de m'avoir permis d'acquérir des connaissances qui n'ont pas de prix à travers nos discussions et ses projets.

Je remercie M. Jean-Leroy Amori St-Paul, Guillaume Huguet et Mme. Catherine Schramm pour les conseils et les supports inestimables qu'ils m'ont offerts durant la rédaction de ce mémoire.

Contexte

Le syndrome d'interruption de la tige hypophysaire (PSIS) est un désordre très rare qui affecte la fonction du système endocrinien. Ce syndrome se caractérise par plusieurs déformations physiques de la glande hypophyse: soit une absence ou un amincissement prononcé de la tige, soit par la présence d'une antéhypophyse hypoplasique ou un postérieur ectopique de l'hypophyse. De plus, il faut souligner que les trois caractéristiques précédemment mentionnées peuvent être simultanément présentes chez le sujet atteint. En se basant sur plusieurs données cliniques, il a été montré que le PSIS est un syndrome qui appartient à un vaste spectre phénotypique comme des anomalies de la ligne médiane [1], des anomalies ophtalmologiques et endocrinologiques [2-3]. De plus, plusieurs études cliniques ont montré que les sujets atteints du syndrome PSIS présentent des traits continus qui varient d'une déficience isolée en hormone de croissance à une déficience combinée en hormone pituitaire. À cause de la nature rare de ce syndrome, il est très difficile d'établir sa prévalence dans la population mondiale, d'autant plus que l'étiologie du syndrome demeure inconnue.

Chapitre 1 - Introduction à la génétique moléculaire

1.1 - Concepts clés de la génétique moléculaire

Les travaux de ce mémoire ont pour but de caractériser les impacts génétiques du syndrome d'interruption de la tige hypophysaire. Afin de comprendre nos démarches et nos résultats, nous allons impérativement développer certains mécanismes de fonctionnement biochimique et moléculaire des gènes. De ce fait, je vais brièvement

détailler dans ce chapitre, certaines propriétés génétiques fondamentales nécessaires à la survie d'un organisme.

1.1.1 - Caractéristiques des acides nucléiques

Les informations génétiques des organismes eucaryotes sont encodées dans une macromolécule appelée acide désoxyribonucléique (ADN). Cette dernière se localise dans le noyau de la cellule, entourée d'une membrane lipidique appelée membrane nucléaire (Figure 1-1). Cette membrane restreint les accès en contrôlant les transports de molécules à travers les canaux et les pores nucléaires. L'ADN est constitué de monomères de nucléotides qui sont représentés par quatre molécules différentes que sont l'adénine (A), la guanine (G), la cytosine (C) et la thymine (T). Les nucléotides sont composés d'une base azotée qui peut être soit une purine ou une pyrimidine et un pentose (glucide) qui se lie aux autres monomères par un lien phosphodiester (Figure 1-2).

Les nucléotides sont des molécules fonctionnelles, comme l'ATP (adénosine triphosphate) qui joue un rôle énergétique dans la cellule et le GTP (guanosine triphosphate) qui peut s'associer parfois aux protéines des récepteurs membranaires. Un simple nucléotide ne code pour aucune protéine. Pour avoir cette propriété, ce dernier doit être un polymère de nucléotides, arrangé en plusieurs triplets de monomères appelés codon. Un polymère de nucléotides représente un gène, si et seulement si celui-ci est capable d'être transcrit en ARN codant ou non codant et traduit en protéine dans le cas d'un gène codant (Figure 1-3).

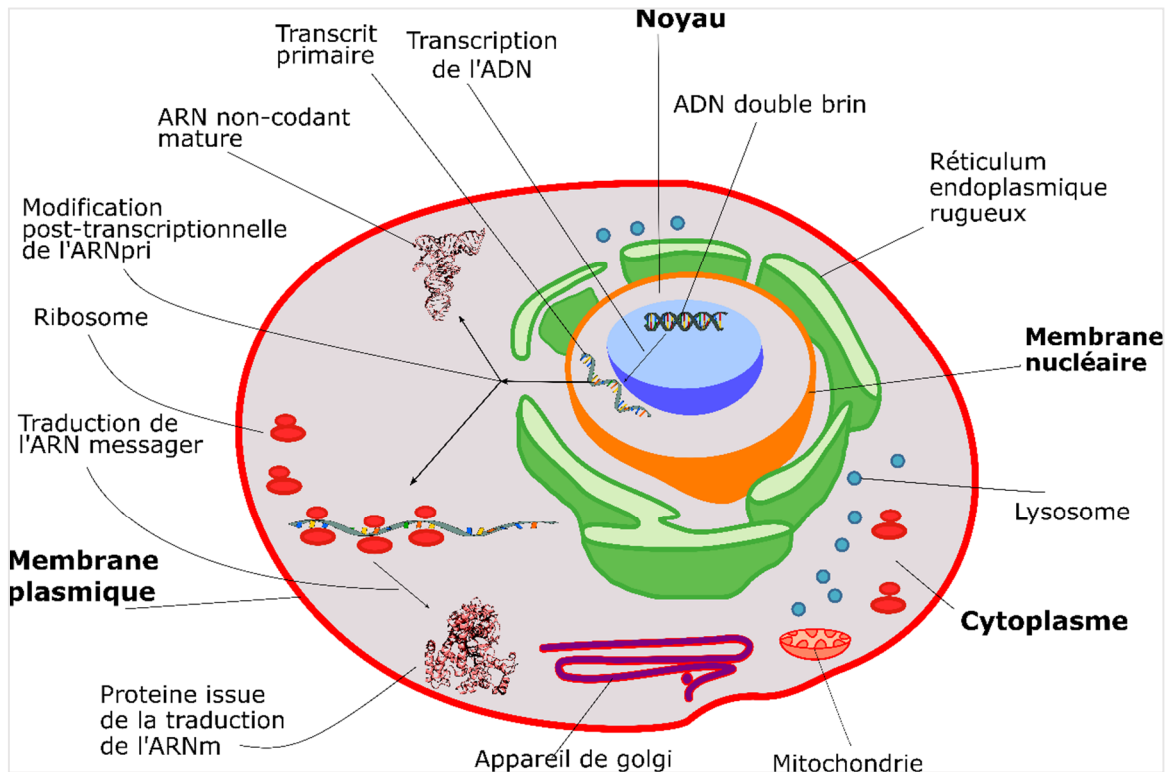


Figure 1-1 - Représentation des composantes de la cellule eucaryote et d'un modèle simpliste d'expression de l'ADN.

Nous montrons sur cette figure, les diverses organelles d'une cellule eucaryote, parmi lesquelles nous indiquons graphiquement un résumé du mécanisme d'expression de l'ADN à partir du noyau. Nous indiquons par les flèches unidirectionnelles, l'ADN qui se transcrit en ARN primaire (ARNpri). Ce dernier, selon la fonction génétique, se mature en ARN message suivie de la traduction en protéine. Dans certains cas, celui-ci demeure sous la forme d'ARN non codant afin de procéder à d'autres fonctions géniques.

Afin de décoder ou en d'autres termes d'exprimer le code génétique nucléaire d'un organisme, la séquence d'ADN doit détenir certains prérequis, tel que des éléments qui permettent de reconnaître et de recruter l'ARN polymérase appelés promoteurs. De plus, elle doit posséder au moins un exon ainsi qu'un facteur de terminaison de la transcription. Lorsqu'une séquence de nucléotides détient toutes ces propriétés qui lui permettent d'exercer une fonction dans la cellule, celui-ci devient alors un gène.

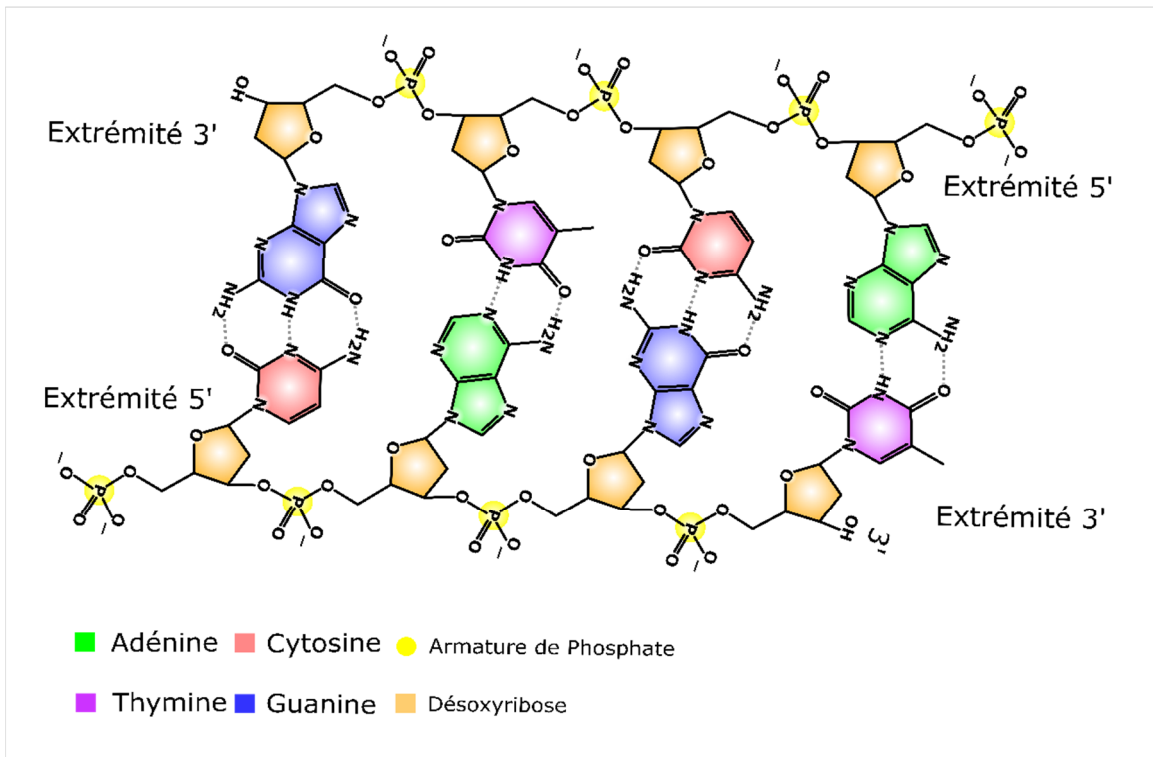


Figure 1-2 - Représentation d'une séquence d'ADN double brin formé de quatre nucléotides appariés.

Sur cette figure, nous représentons une proportion d'ADN double brin sur laquelle nous indiquons l'appariement vis-à-vis des nucléotides via des liens d'hydrogène et la liaison des nucléotides de l'ADN simple brin via des liens phosphodiesters. Les nucléotides sont formés d'une base azotée (l'adénine, la guanine, la cytosine ou la thymine), d'un désoxyribose en conformation cyclo-pentose et une armature de phosphates.

1.1.2 - Propriétés des gènes codants et non codants

Un gène n'est pas nécessairement une séquence d'ADN codant pour une protéine. Celui-ci doit détenir des propriétés additionnelles que ceux dont nous avons indiquées dans la section 1.1.1 (page 2). Contrairement à un gène non codant, un gène codant est constitué d'une coiffe (7-méthylguanosine), un codon d'initiation et un codon d'arrêt de la traduction de la séquence du gène. La coiffe est une guanine située au début de la séquence de l'ARNm. Celle-ci est modifiée à la position sept par la procédure de la méthylation durant la transcription du gène. Cette dernière permet de protéger l'ARNm contre la dégradation par les exonucléases et participe à l'initiation de la traduction du

gène [4]. La taille d'un gène, d'un exon ou le nombre d'exons d'un gène ne sont pas des éléments caractéristiques spécifiques d'un gène. Prenons l'exemple du gène interféron A6 (IFNA6), celui-ci mesure 0.57 kbps, constitue d'un seul exon et la traduction de celui-ci est basée sur la proportion totale du gène. L'exemple précédent montre qu'un gène peut exister même avec un nombre minimal d'exon et une taille raisonnablement petite comparativement à d'autres gènes comme la dystrophin (DMD) (2090 kbps, 79 exons) [5].

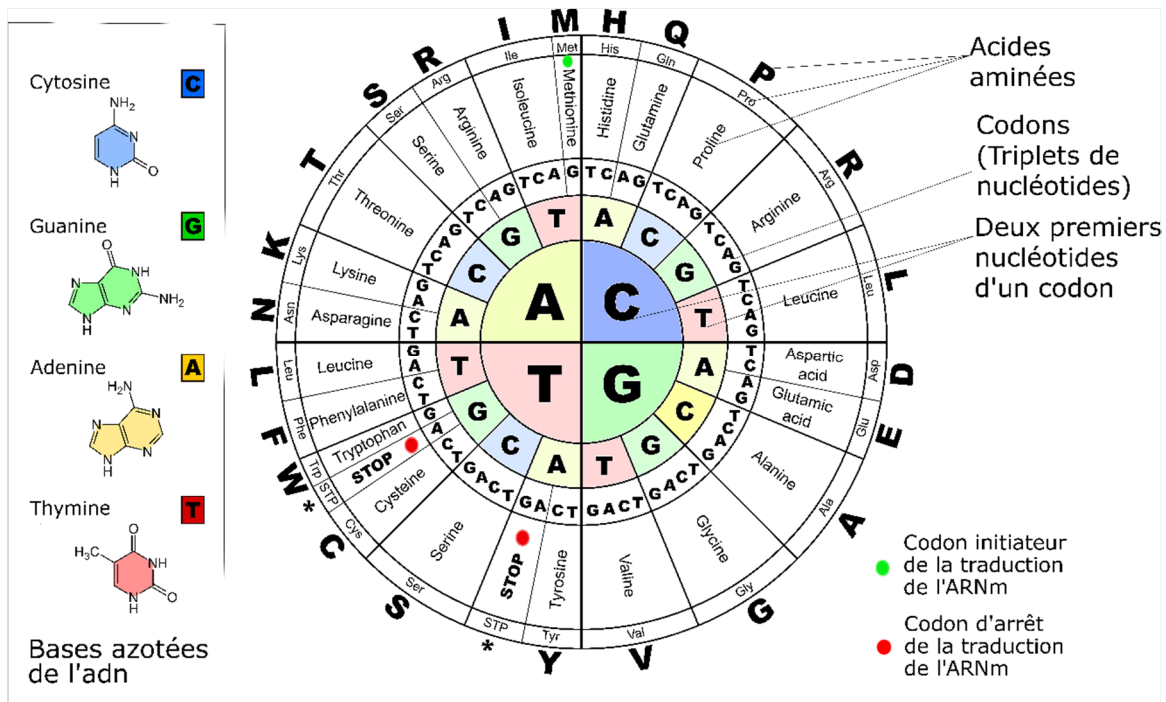


Figure 1-3 - Table du code génétique de vertébrés et représentation moléculaire des bases azotées de l'ADN.

Sur cette figure nous indiquons la méthode de décryptage du code génétique des vertébrés. Celle-ci est utilisée de façon procédurale par la machinerie de transcription et de traduction afin d'exprimer les éléments codants de l'ADN. À gauche nous indiquons à titre d'information la conformation chimique de bases azotées des nucléotides. À partir des codes de couleurs, nous indiquons les nucléotides représentés par des lettres dont la combinaison en triplet for des codons qui se traduisent ensuite en acides aminés. Nous indiquons par un point vert, le codon qui se traduit en méthionine et en rouge, ceux qui se traduisent en signaux d'arrêt d'élongation de la protéine.

La région promotrice d'un gène se situe souvent à la position 5' en amont de la séquence de celui-ci. Cette dernière favorise l'interaction du gène avec des facteurs de transcription qui recrutent d'autres éléments comme des répresseurs, amplificateurs, etc (Figure 1-4). L'expression d'un gène est un mécanisme qui peut se varier sous l'effet de plusieurs facteurs, que sont : soit des mutations génomiques ou des effets environnementaux. Ces variations peuvent avoir des conséquences néfastes sur les fonctions des gènes et des organes de l'organisme. Par contre, chez les organismes eucaryotes diploïdes, il existe certaines manières de compenser ces variations, celles-ci consistent à exprimer l'allèle qui n'est pas affecté par la variation.

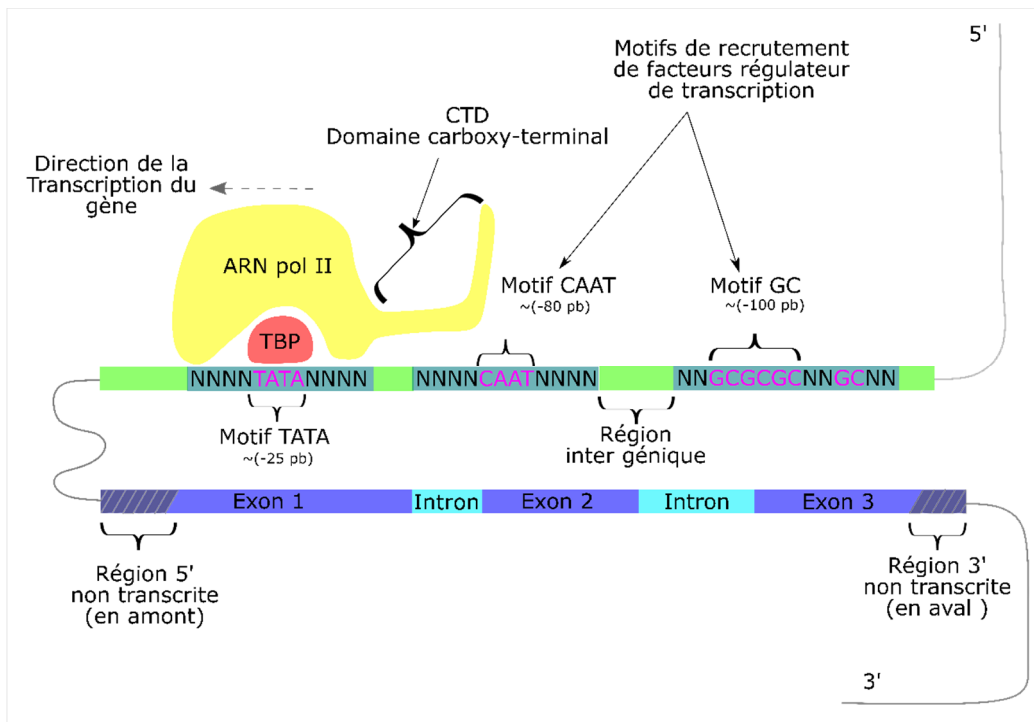


Figure 1-4 - Représentation de la structure d'un gène eucaryote et ses régulateurs de transcription

Cette figure représente la structure d'un gène eucaryote nucléaire, c'est-à-dire des gènes qui se localisent dans le noyau d'une cellule eucaryote et qui ne sont pas présents dans le génome mitochondriaux. Celui-ci se divise en deux sections : la séquence codante du gène et la séquence en amont où se situent les éléments de régulation de la transcription du gène. Nous représentons sur cette figure, en amont de la séquence codante du gène, les éléments régulateurs, tels que : la boîte « TATA » qui se situe à environ 25 paires de bases de la séquence codante, la boîte « CAAT » qui se situe à ~80 paires de bases et les motifs « GC » qui se situent à ~ 100 paires de bases. Nous représentons par « N », les nucléotides flanquant en amont et en

aval des éléments régulateurs. La structure en rouge représente la protéine de liaison TATA (TBP : TATA binding protein) et nous constatons qu'elle se lie au motif « TATA ». La liaison de TBP au motif « TATA » se fait par une interaction spécifique. La protéine « TATA » a pour but de recruter l'enzyme ARN polymérase II (ARN pol II) par une cascade d'interactions protéiques incluant les complexes de protéines de celle-ci. Nous indiquons la direction de la transcription par une flèche unidirectionnelle en haut de l'ARN pol II (5' vers 3'). Nous indiquons la queue carboxy-terminal de l'ARN pol II qui se constitue de motifs de polypeptide : Tyrosine, Serine, Proline, Thréonine, Serine, Proline, Serine. Les sections en vert représentent des régions inter géniques non fonctionnelles. Dans la séquence codante du gène, nous indiquons deux régions non transcrites qui se trouvent en amont et en aval. La région non transcrites en amont contient une région (5'-UTR) dont le but consiste à recruter des facteurs de transcription. Le 5'-UTR et les éléments de liaison des facteurs de transcription forment la structure du promoteur du gène. Celle en aval contient des éléments ayant pour but de terminer la transcription du gène.

1.1.3 - La ploïdie et les allèles

La ploïdie est un terme qui désigne le nombre d'exemplaire, dans une ou dans les cellules d'un organisme, de jeux complets des chromosomes du génome de ce type d'organisme. De ce fait, un organisme haploïde comprend un seul jeu complet de chromosomes par cellule, celui diploïde possède deux jeux complets de chromosomes et ainsi de suite. De plus, un génome diploïde comprend deux copies du même allèle dont chacune provient d'un parent différent.

Cependant, il parvient que deux allèles du même locus d'un génome diploïde soient différents l'une de l'autre. Dans ce cas, ces derniers sont considérés comme des allèles hétérozygotes. Dans le cas contraire, ils sont considérés comme des allèles homozygotes.

1.1.4 - Variant génétique et expression du génome

Afin de comprendre les effets des maladies liées à des facteurs génétiques, il faut d'abord comprendre les mécanismes de la régulation des gènes par des variations génomiques. Tel que nous avons déjà mentionné dans les sections précédentes, il existe différents facteurs qui sont capable de réguler l'expression génique, parmi lesquelles les mutations génomiques ponctuelles, structurales, épigénétiques, etc. Les variations

génomiques sont des altérations de la molécule d'ADN qui ne sont pas observées dans le génome de références. Elles peuvent se présenter sous plusieurs aspects, que sont : les SNV, les INDEL, les CNV et les réarrangements chromosomiques [6-8]. Ces dernières peuvent avoir plusieurs impacts sur la séquence d'un gène, elles peuvent supprimer ou dupliquer un locus, changer le cadre de lecture et même modifier les fonctions des transcrits. Les conséquences d'une mutation varient selon le type et l'emplacement de celle-ci, comme : les mutations localisées dans des promoteurs de gènes, des sites d'épissage alternatif, des exons codants et des régions de régulation quantitative de l'expression des gènes.

1.1.5 - Les variations et les mécanismes d'expression génique

Afin de bien étudier les mécanismes d'expression des gènes sous l'effet des variations génomiques, il faut d'abord comprendre les effets de ces dernières sur la régulation moléculaire de ces gènes. Nous savons qu'il existe différentes caractéristiques de mutations génomiques, qui sont : les mutations synonymes appelées encore mutation silencieuses, les mutations faux-sens, nonsenses, celles qui changent les cadres de lecture des gènes codants et finalement les petites et grandes variations structurales. Ces dernières peuvent avoir différents impacts sur la transcription des gènes et les interactions moléculaires.

Les SNV sont des variations qui sont capables de modifier la séquence codante d'un gène (SNV faux-sens) ou de modifier les fonctions associées à celle-ci dans le cas d'une séquence non codante comme les SNV qui affectent les régions d'épissage alternatif et qui provoquent les sauts d'exon (SNV silencieux). D'autres variations, comme les INDEL, sont en mesure de modifier les cadres de lecture des gènes codants en

ajoutant ou en diminuant des paires de bases dans la séquence du gène. Celles-ci engendrent un décalage dans l'ordre d'arrangements successifs des nucléotides de la séquence du gène lorsque la variation n'est pas un triplet. Cette dernière peut être responsable de la synthèse de nouveaux acides aminés dans la protéine mutante ou même générer un codon d'arrêt prématuré durant la traduction de l'ARNm (Figure 1-5).

1.1.6 - Impact des variations du nombre de copies

Les CNV sont des variations structurelles qui affectent des locus de très grandes tailles du génome. Comparativement à certains INDEL qui modifient les cadres de lecture des gènes codants, ces dernières peuvent affecter des locus qui s'étendent sur la taille de plusieurs gènes et provoquent des impacts qui s'évaluent selon le nombre et les fonctions des gènes affectés. La conséquence des CNV est plus facilement observable dans le cas des gènes sensibles au dosage. Par conséquent, il est plus probable d'observer un changement phénotypique chez un organisme dont la copie du gène est partiellement ou totalement affectée par la variation. Prenons le cas d'une délétion hétérozygote d'un gène sensible au dosage chez un organisme eucaryote. Malgré l'expression de la copie non mutante du gène affecté, le phénotype sera aussi bien altéré car l'expression normale des deux allèles est requise afin d'observer un phénotype normal. Cette caractéristique définit l'haplo-insuffisance du gène. D'autre part, il est connu que les duplications affectent l'expression phénotypique de certains gènes [9-10]. Par contre le nombre de copie dupliquée varie d'un gène à un autre. En conclusion, dans une analyse de détection et d'annotation d'impacts fonctionnels de CNV, la notion d'haplo-insuffisance demeure un aspect important car elle reflète la conséquence du CNV sur l'expression du génome

haploïde.

Conséquences des mutation dans des séquences d'ADN codants	
Types de mutations au niveau de l'ADN	Impact au niveau moléculaire
Absence de mutation	<p>Séquence non mutante</p> <p>Séquence codant pour la protéine non mutante</p>
Mutations du type transition ou transversion	<p>Mutation synonyme</p> <p>Codon mutant spécifie le même acide aminé</p>
	<p>Mutation faux-sens (conservative)</p> <p>Codon mutant spécifie un acide aminé chimiquement similaire</p>
	<p>Mutation faux-sens (nonconservative)</p> <p>Codon mutant spécifie un acide aminé chimiquement différent</p>
	<p>Mutation non-sens</p> <p>Signal d'arrêt d'élongation de la protéine</p>
INDEL	
Insertion d'une ou de plusieurs nucléotides	<p>Modificatrice du cadre de lecture</p> <p>Modification au niveau des acides aminés en élongation</p>
Deletion d'une ou de plusieurs nucléotides	<p>Modificatrice du cadre de lecture</p>

Figure 1-5 - Variation fonctionnelle et moléculaire des gènes affectés par des SNV et des INDEL.

Dans cette figure nous indiquons les différents modèles par lesquels un SNV affecte la traduction d'un codon en acides aminés. Cette figure est un hybride de tableau et d'image. Celle-ci est représentée en deux colonnes, la première indique le type de variation (SNV), pendant que la deuxième indique l'impact de celle-ci sur la traduction de l'acide aminé codé par le codon de la région affectée. Nous indiquons aussi des sous-sections de la première colonne. Cette dernière est divisée en deux catégories, que sont : famille de SNV (sous colonne de gauche) et type de SNV (sous colonne de droite).

Chapitre 2 - Survole du système endocrinien

2.1 - Composantes et fonctions du système endocrinien

Afin de mieux préparer nos idées en rapport au sujet de l'étude, il est nécessaire de comprendre ou de rafraîchir nos connaissances sur les caractéristiques et les composantes du système endocrinien, dans le but de mieux cibler les diverses propriétés embryologiques et fonctionnelles de la glande hypophyse. Pour ce faire, je vais d'abord vous élaborer dans les lignes qui suivent, les caractéristiques fonctionnelles du système endocrinien.

2.1.1 - Le système endocrinien

Le système endocrinien est un réseau formé de glandes sécrétrices d'hormones ou un réseau de signalisations via des hormones. Ce système est étendu sur divers compartiments du corps et détient plusieurs rôles dans l'organisme : d'abord il permet au sujet de maintenir l'homéostasie en équilibrant le dosage d'hormones nécessaires à la survie cellulaire et ensuite de réagir aux événements environnementaux par le biais des *stimuli*, enfin il permet de maintenir une croissance chronologique équilibrée des organes humains. Le mécanisme d'action du système est efficace, car il utilise les vaisseaux sanguins comme voie de transport des facteurs d'activation de signaux à distance.

2.1.2 - Composantes fonctionnelles du système endocrinien

Afin de mieux comprendre la fonction du système endocrinien, il est nécessaire de connaître ses composantes et les fonctions qui y sont associées. D'abord, nous signalons que le réseau du système est formé de six glandes sécrétrices d'hormones qui sont : les glandes surrénales, thyroïdes, sexuelles (testicules, ovaires), les glandes pancréatiques,

mammaires et les plus importantes de toutes, l'hypothalamus et l'hypophyse. Chacune des glandes du système endocrinien détient des fonctions et des propriétés particulières. Certaines glandes sont de nature exocrine, c'est-à-dire, elles sont constituées de cellules ayant la propriété de sécréter des substances destinées vers l'extérieur de l'organisme, d'autres sont de nature endocrine car elles ont la propriété de sécréter des facteurs d'activation de récepteurs d'hormones, via les vaisseaux sanguins, et ceci, afin de réguler les activités des autres glandes sous une forme de régulation distante.

2.1.3 - L'hypothalamus et l'hypophyse

L'hypothalamus est la structure mère du système endocrinien, plus précisément le pilier central de celui-ci, elle régule la glande hypophyse en sécrétant des facteurs de libération d'hormones qui sont reconnues par des récepteurs situés à la surface des cellules de la glande hypophyse. La fonction principale de l'hypothalamus est de réguler le niveau de sécrétion d'hormones qui permet en retour de maintenir le bon fonctionnement des organes.

Les facteurs de libération d'hormones sont des petits polypeptides synthétisés par les cellules spécialisées de l'hypothalamus et qui sont sécrétés à travers des axones qui forment une connexion directe entre l'hypothalamus et l'hypophyse. Nous notons quatre facteurs : le facteur de sécrétion d'hormone de croissance, corticotropique, gonadotropique et d'hormone thyroïdienne.

2.1.4 - Anatomie de l'hypophyse

Nous attirons l'attention sur la glande « hypophysaire » (Figure 2-1), car c'est l'organe principal sur lequel se base notre étude. Celle-ci mesure environ 4.78 cm³ [11], elle se situe à la base de la boîte crânienne plus précisément dans la cavité de l'os

sphénoïdale appelée sella turcica [12]. Elle est en contact direct avec l'hypothalamus par l'intermédiaire d'une tige appelée la tige hypophysaire (pituitary stalk) à travers laquelle sont transmis les signaux de sécrétion d'hormones (Figure 2-1).

L'hypophyse est divisée en deux lobes fonctionnellement différents : un lobe antérieur connu sous le nom d'adénohypophyse et un lobe postérieur appelé neurohypophyse [12-13]. Le lobe antérieur représente plus de la moitié du volume de la glande, environ 70% de la taille de l'hypophyse et elle est constituée d'environ cinq types cellulaires qui se sont différenciés pour avoir chacun des fonctions de sécrétion spécifiques [12-14] (Figure 2-2).

Nous voulons attirer l'attention sur les différents types cellulaires de l'adénohypophyse car notre protocole a été majoritairement consacré à la caractérisation des impacts des variations génétiques qui affectent l'expression des gènes associés au développement des tissus spécialisés de la glande antéhypophysaire.

2.1.5 - Organogenèse de l'hypophyse

La détermination des causes associées au PSIS nécessite préalablement une connaissance enrichie sur le développement embryonnaire de la glande hypophysaire. Certaines études ont montré que ce dernier est à l'origine d'une interaction physique entre deux structures dermatologiques, qui sont : la surface antérieure de l'ectoderme olfactif et l'ectoderme neural [15-17]. C'est une interaction qui requiert l'expression combinée et adéquate de plusieurs facteurs de transcriptions : Sonic hedgehog (*SHH*), protéine morphogénétique de l'os 4 (*BMP4*) et les facteurs de croissance des cellules fibroblastes 8 et 10 (*FGF8* et *FGF10*).

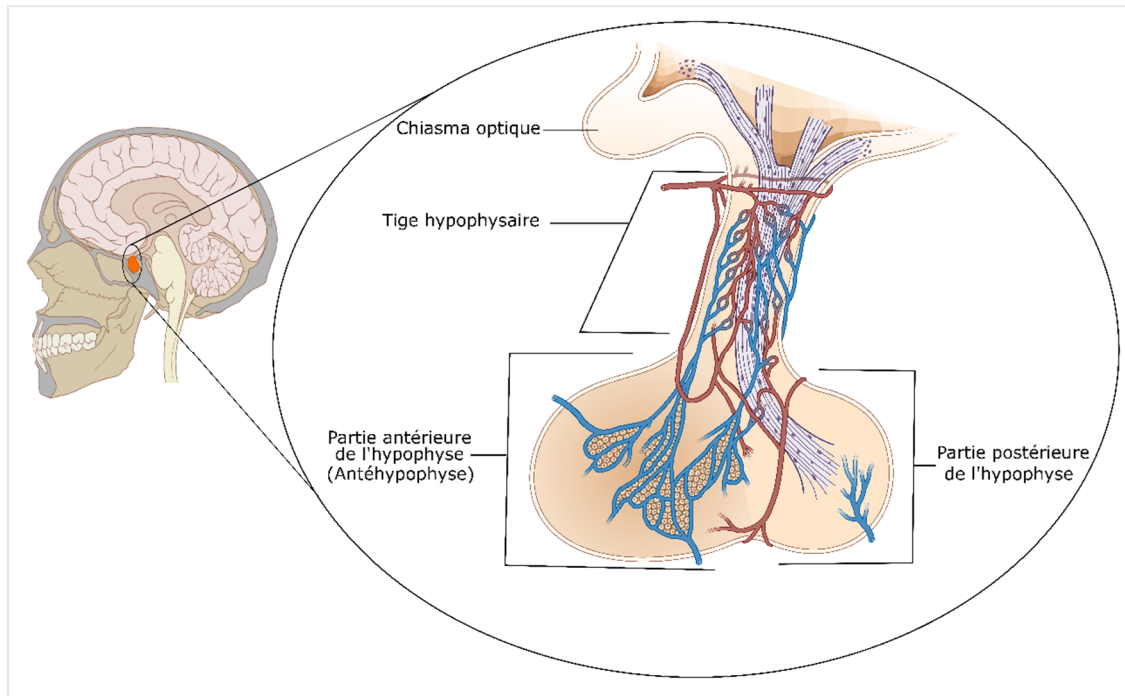


Figure 2-1 - Localisation physiologique et représentation schématique de l'hypophyse.

Figure modifiée et simplifiée issue de Scheithauer, B.W. et al.,1991 et Kaiser, U. et al.,2016 [12]

Dans cette image, nous représentons l'hypophyse et ses divers compartiments. La glande est divisée en trois sections : un lobe antérieur appelé antéhypophyse, un lobe postérieur et la tige hypophysaire. De plus, la glande est constituée d'une région intermédiaire située entre les deux lobes, par contre, celle-ci n'a pas été indiquée sur la figure par les auteurs. La coupe transversale du crâne représentée sur la figure montre simplement la localisation de l'hypophyse (structure colorée en rouge) dans la boîte crânienne.

De plus, au cours de l'embryogenèse, l'ectoderme neural se modifie morphologiquement pour devenir le lobe postérieur de l'hypophyse, pendant que la surface de l'ectoderme olfactif génère une structure appelée la poche de « Rathke »[14] qui est un précurseur de l'antéhypophyse et de la partie intermédiaire de la glande pituitaire (Figure 2-2). La formation de cette dernière est un processus très complexe et

très bien synchronisé. Il est divisé en trois grandes étapes qui débutent par l'expression des gènes Lim-homeobox 3 et 4 (LH3, LHX4), gamma-glutamylcystéine synthase (GSH4) et ISL1, en réponse au signal d'induction de la formation de la structure provenant de la région du diencephale. Plus de détails seront fournis à la section 1.1.6.

Par ailleurs, Lim-homeobox est une famille de gènes qui codent pour des facteurs de transcription. Ceux-ci jouent un rôle fondamental dans le développement de l'organisme et ils couvrent des régions génomiques bien conservées à travers plusieurs espèces. Les variations délétères qui affectent ces régions peuvent avoir des effets néfastes sur le développement embryonnaire de certains organes [18-20].

2.1.6 - La formation de la poche de « Rathke »

La formation de la poche de « Rathke » est une étape obligatoire de la formation de la glande hypophysaire, la première étape consiste à un épaissement et un durcissement de la paroi de l'ectoderme oral vers le onzième jour de vie embryonnaire [14] (Figure 2-2, Figure 2-3). Ce dernier génère une structure appelée « placode » dont la formation nécessite l'expression du gène BMP4. D'ailleurs, certaines études [21] ont montré qu'en éliminant la fonction du gène BMP4, la formation de la poche de « Rathke » n'est pas induite et le gène n'est pas exprimé. Contrairement à d'autres études, cette dernière suggère que la formation de la poche ne requiert pas l'expression du gène ISL1 [21-22].

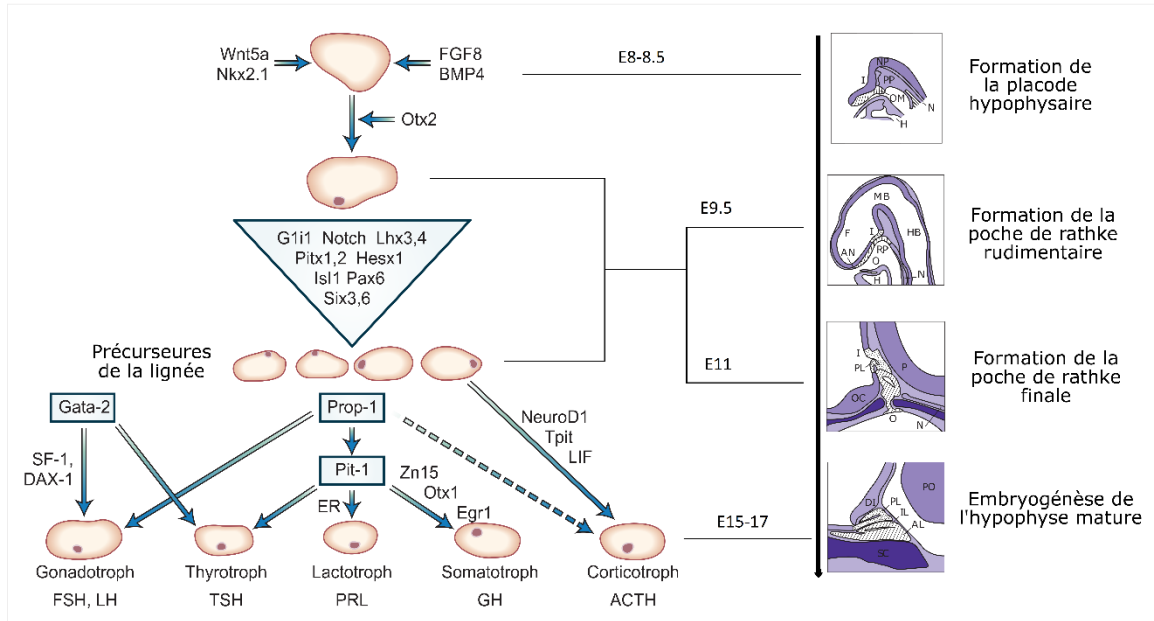


Figure 2-2 - Développement embryonnaire de la glande hypophysaire et sa composition en cellules spécialisées.

Figure modifiée et traduite de Shimon, L. et al. 1996

Cette figure nous montre un modèle de développement embryonnaire de la glande hypophyse, basé sur des études effectuées sur des rats et des souris. Nous indiquons le temps de développement embryonnaire par la lettre E suivie du nombre de jour. Du côté droit de la flèche noire descendante, on indique les différentes étapes de la formation de la « Rathke's pouch ». La figure de gauche représente une cascade d'expressions géniques, spécifiquement des facteurs de transcription, qui aboutit à la formation de la « Rathke's pouch », mais aussi à la formation de l'hypophyse mature. Nous observons, par l'intermédiaire des flèches, la direction de la voie de sécrétion ou elle indique la glande en aval qui est stimulée par des facteurs d'activation de sécrétion d'hormones TSH, ACTH, GH, PRL, LH et FSH. Cette image provient du manuel « *Endocrine Physiology, 4e; 2013* »

La deuxième étape consiste à débiter la formation définitive de la poche. La voie de signalisation métabolique de la synthèse de la protéine BMP4 qui permet d'induire la formation de la structure n'est malheureusement pas bien connue. D'autre part, il est connu que le gène FGF8 joue un rôle important dans la formation embryonnaire de cette dernière [23-26]. Le développement de la poche requière l'expression de deux facteurs

importants, LHX3 et LHX4. De plus, il est requis qu'au moins un de ces deux facteurs soit fonctionnel pour que la formation de la poche soit définitive.

La dernière étape consiste à la différenciation des cellules de l'ectoderme oral à la 3^e semaine du stade embryonnaire. Celle-ci se réalise grâce à l'expression des différents facteurs de transcription dans la partie dorsale de la glande en développement [26-28]. La poche de « Rathke » est la structure de base de l'antéhypophyse. De plus, les cellules embryonnaires qui sont à l'origine de cette dernière proviennent de la différenciation des cellules souches de la poche [29-30].

En conclusion, la formation de la poche de « Rathke » est un évènement complexe qui dépend d'une cascade d'interactions et d'activations de plusieurs facteurs de transcription [14] (Figure 2-3). Connaissant le rôle de cette structure dans l'embryogenèse de l'hypophyse, l'annotation fonctionnelle des gènes dans la formation de celle-ci pourrait mieux nous guider à caractériser les variations génétiques associées à la malformation de la tige hypophysaire.

2.1.7 - Origine des cellules spécialisées de l'antéhypophyse

Les cellules spécialisées de l'antéhypophyse proviennent des cellules souches progénitrices qui sont activées durant la formation de la poche de « Rathke ». Au cours de celle-ci, ces cellules sont marquées par l'expression du gène SOX2 et subissent une série d'activations via l'expression de plusieurs facteurs de transcription [14], parmi lesquelles on trouve: TPIT, ISL1, PROP1, PIT1, POMC, GATA2, PITX1/2, LHX3/4, HESX1 et SF1 [31-40]. Ces cellules ont été soumises à différentes étapes de différenciation cellulaire qui résultent en la formation de cinq types de cellules spécialisées qui sont: les cellules somatotropes, les lactotropes, les gonadotropes, les

thyroïdiques et enfin les corticotropes que nous allons détailler dans les paragraphes qui suivent (Figure 2-4).

En se basant sur des modèles d'étude de développement embryonnaire de l'hypophyse chez la souris [14], Les corticotropes ont été les premières cellules à se différencier, et ceci grâce à l'expression connue du gène SF1 vers la 13^e jour de développement embryonnaire. Les gonadotropes proviennent de la même cellule progénitrice que les corticotropes, mais celles-ci ne se sont différenciées que plus tard vers le 17^e jour de développement embryonnaire par l'activation de TPIT.

Les thyroïdiques, les lactotropes et les somatotropes sont des cellules provenant d'une même cellule progénitrice qui nécessite l'expression de PROP1 [41-42] afin de générer un précurseur intermédiaire qui permettra la différenciation des thyroïdiques via l'activation de GATA2, la différenciation des somatotropes et des lactotropes via l'expression de PIT1 [42-43]. Enfin, les mélanotropes sont les seules cellules qui proviennent d'une cellule progénitrice unique qui se localise dans le lobe intermédiaire de la glande en développement. Sachant que les mélanotropes partagent le même facteur d'activation (TPIT) que les corticotropes, certaines études [43-44] avancent que les deux types cellulaires seraient issus d'une même cellule progénitrice (Figure 2-4).

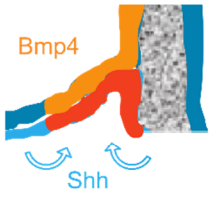
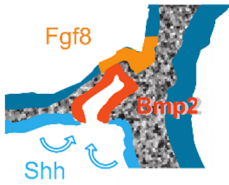
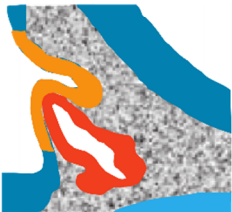


		Facteurs de transcription exprimés		Arrêt développemental / Mutagenèse chez la souris	
	e9.5	Hesx1 Pitx1/2 Lhx3/4 Isl1			
	e10.5	Pitx1/2 Lhx3/4 Hesx1 Isl1	Prop1 Gata2	<i>Pitx1^{-/-}; Pitx2^{-/-} Lhx3^{-/-}; Lhx4^{-/-}</i>	
	e12.5	Pitx1/2 Lhx3/4	Prop1 Gata2	<i>Pitx2^{-/-} Lhx3^{-/-}</i>	
	e13.5	○ Cellules progénitrices cycliques ● Cellules progénitrices non-cycliques ● Cellules différenciées	Pitx1/2 Lhx3/4	Pit1 Gata2 Sf1 Tpit	<i>Prop1^{-/-}</i>
	e17.5	M = melanotrope L = lactotrope T = thyrotrope S = somatotrope G = gonadotrope C = corticotrope			

Figure 2-3 - Modèle de développement embryonnaire de l'hypophyse dans des modèles murins.

Figure modifiée et traduite de Drouin, J. et al., 2011 [14]

Sur cette figure, nous montrons un modèle d'expression de facteurs de transcription spécifiques au développement embryonnaire de l'hypophyse. Cette figure est séparée en quatre compartiments. Dans le premier compartiment, nous montrons des images modélisées du développement embryonnaire de l'hypophyse à travers le temps. Dans le deuxième, nous indiquons les facteurs de transcription qui sont exprimés à l'étape de développement approximatif de la glande. Dans le troisième compartiment, nous indiquons les facteurs de transcription mutants qui engendrent l'arrêt de développement embryonnaire de l'hypophyse dans des modèles murins. Finalement, les lettres dans la partie inférieure de la figure représentent les types cellulaires de l'antéhypophyse.

Les connaissances du mécanisme de développement embryonnaire de l'hypophyse, basées sur l'expression génique spécifique à celle-ci, sont des atouts importants qui permettent de comprendre les impacts des variations génétiques sur l'expression des gènes durant la formation de celle-ci. Par contre, nous n'écartons pas l'hypothèse que d'autres facteurs non génétiques pourraient altérer le développement embryonnaire de l'hypophyse, comme des impacts environnementaux et la toxicité aux produits chimiques.

2.2 - Populations et fonctions des cellules de l'antéhypophyse

Le lobe antérieur de l'hypophyse (Figure 2-1) est formé de cinq populations de cellules qui se sont différenciées durant le développement embryonnaire initiale de celle-ci afin d'exécuter des fonctions endocriniennes différentes. Ces cellules sont : les somatotropes, les lactotropes, les gonadotropes, les thyrotropes et finalement les corticotropes dont nous allons brièvement détailler les aspects physiologiques et fonctionnels dans les paragraphes qui suivent.

2.2.1 - Les populations cellulaires de l'antéhypophyse

Les somatotropes représentent environ 50% de la proportion cellulaire de l'antéhypophyse [13], elles ont la propriété de synthétiser des hormones de croissance (GH) appelées encore somatotrophines qui sont d'une importance capitale pour le développement et la survie de l'organisme. La sécrétion d'hormones de croissance est un mécanisme complexe qui requiert un contact sélectif et physique (ligand-récepteur) entre le facteur d'activation (GHRHR) et des récepteurs de la membrane plasmique des somatotropes (GHRH). Le mécanisme de la sécrétion d'hormones précédemment mentionnées n'est pas valable que pour les somatotropes, mais aussi valable pour tous les

autres types cellulaires, c'est une interaction de base nécessaire à toute activité endocrinienne.

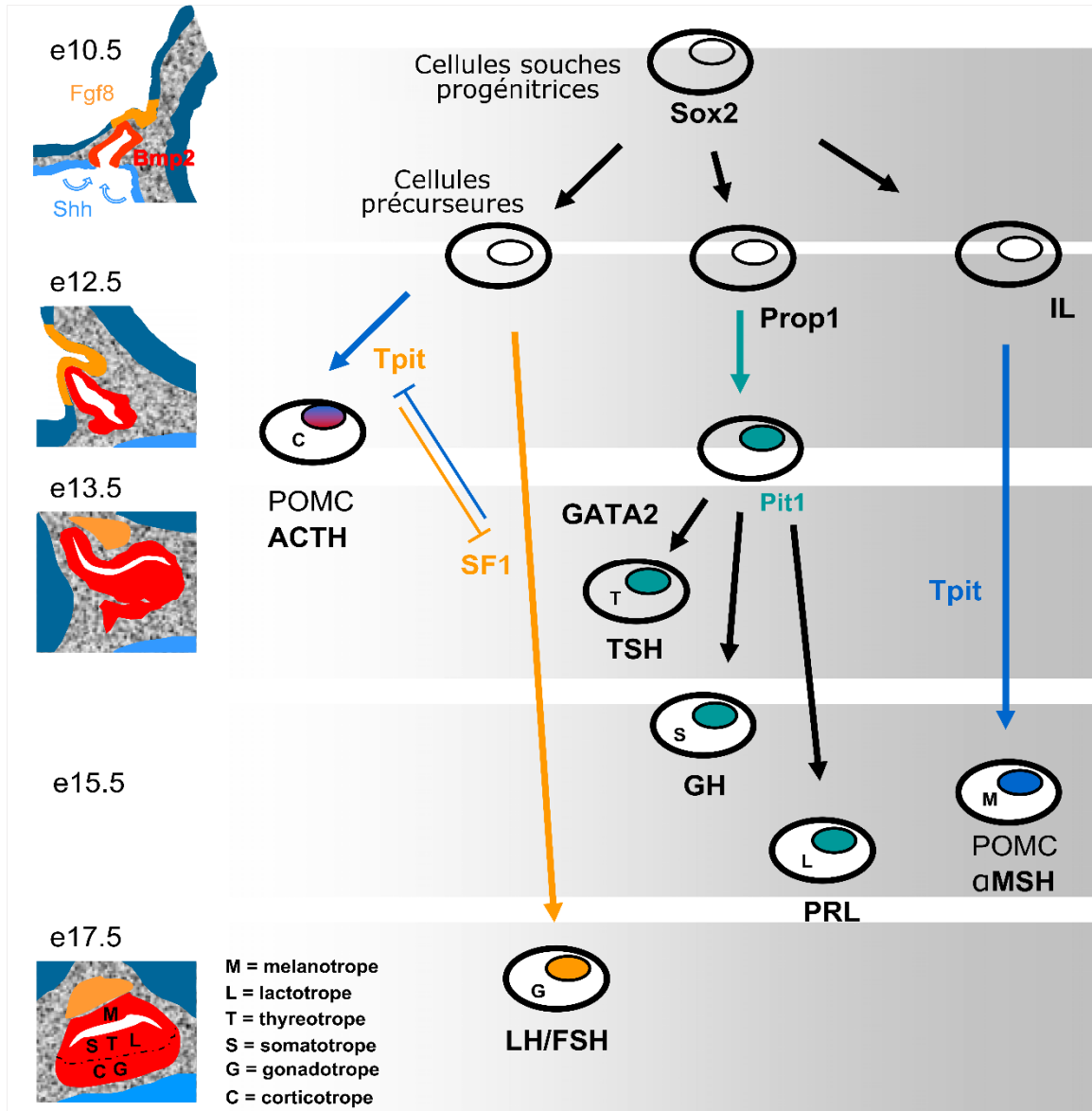


Figure 2-4 - Différenciation séquentielle des cellules de l'antéhypophyse sous l'effet d'activation de différents facteurs génétiques.

Figure modifiée et traduite de Drouin, J. et al., 2011 [14]

Cette figure indique une cascade de différenciations des cellules souches embryonnaires de l'hypophyse au courant du développement de celle-ci. A chaque étape de différenciation cellulaire, il est indiqué le nom du gène qui a favorisé cette différenciation, pendant que les flèches indiquent la direction de chaque évènement. À gauche de la figure, il est indiqué l'image de développement embryonnaire temporelle de l'hypophyse. Enfin, les lettres M, L, T, S, G et C représentent les abréviations des différentes familles de cellules.

Les lactotropes occupent à peu près 15% [18] du volume de l'antéhypophyse. Elles sont responsables de la sécrétion des hormones prolactines (PRL). Cette dernière se réalise via l'activation des récepteurs PRLR par les œstrogènes et les TRH qui stimulent non seulement les glandes mammaires au courant de l'allaitement [12-13]. Il est aussi connu que cette hormone joue un rôle dans la survie cellulaire par l'interruption des processus d'apoptose.

Les gonadotropes représentent environ 10% de la proportion de l'antéhypophyse. Elles secrètent des hormones qui appartiennent à la classe des glycoprotéines à sous unités alpha et bêta, qui sont très importantes pour le développement des organes sexuels. Les gonadotropes sécrètent de la gonadotrophine (LH, FSH) qui est un duo d'hormones dont la sous unité alpha est encodé par les mêmes séquences d'acide nucléique, mais elles se différencient par la séquence en acides aminés de leur sous unité bêta. Les gonadotropes sont activées par les GnRH et de l'œstrogène via le récepteur GnRHR, et celui-ci est dans le but de produire des stéroïdes et des hormones (FSH, LH) nécessaires au développement des organes.

Les thyrotropes sont des cellules spécialisées qui représentent environ 5% de la population des cellules de l'antéhypophyse. Elles sécrètent l'hormone stimulante thyroïdienne (TSH) qui appartient à la famille des glycoprotéines à sous-unités alpha et bêta. Ces dernières permettent de réguler la fonction sécrétrice de la glande thyroïde qui se situe dans la façade antérieure du cou. Le TSH active les cellules de la thyroïde afin de synthétiser les composés (T3) triiodothyronine, (T4) thyroxine et de la calcitonine par l'intermédiaire d'activation des récepteurs membranaires (TSHR) localisés à la surface des cellules thyroïdiennes.

Les corticotropes forment une population cellulaire qui occupe environ 15% de la taille de l'antéhypophyse. Elles sécrètent le précurseur pro-Opiomelanocorticothines (POMC) qui est clivé par des protéases intracellulaires afin de relâcher de l'adrénocorticotrophine (ACTH), bêta-endorphine, hormone stimulant mélanocyte alpha, bêta et gamma. À partir de l'activation des récepteurs couplés aux protéines G (GPCR), dans les corticotropes, les molécules d'AMPc déclenchent les processus d'activation des protéines kinases via la libération des ions calcium qui ont pour but de promouvoir la transcription du gène POMC.

2.2.2 - Variation de la fonction de sécrétion hormonale

La surproduction du GH, dans les somatotropes, est habituellement à l'origine de la tumeur de l'hypophyse qui s'exprime par un excès de croissance de la glande. Nous reconnaissons plusieurs types de surcroissance de l'hypophyse comme le craniopharyngiome, l'adénome hypophysaire, la masse stellaire invasive non adénomateuse [45-46], etc. La surcroissance de l'hypophyse est souvent associée à la maladie du « gigantisme » chez les enfants [46]. D'autres maladies sont associées à la surproduction du GH, comme l'acromégalie qui s'exprime par un élargissement et un épaissement des membres en périphérie du corps (les mains, les pieds et la tête). D'autre part, certaines études [46-48] sur des souris mutantes ont montré que l'altération de la fonction du gène PROP1 pourrait engendrer la déficience en GH qui est un marqueur associé à l'insuffisance de croissance de l'organisme, par conséquent, elle serait associée à la maladie du « dwarfism » hypophyse dépendante [48]. La maladie du « dwarfism » est un syndrome qui se caractérise par un défaut de croissance globale extrême d'un individu [49].

L'expression altérée de la fonction des gonadotrophines est connue pour être associée à certaines pathologies, comme l'hypoprolactinémie et l'hyperprolactinémie. L'hyperprolactinémie est la plus fréquente parmi les maladies associées aux dysfonctionnements des lactotrobes. Cette dernière pourrait être à l'origine de plusieurs facteurs: physiologique, pathologique ou l'exposition aux produits toxiques [50-52]. Certaines études ont montré qu'une carence en prolactine est associée à une augmentation de la concentration d'œstrogène ovarien et du niveau de sécrétion de la progestérone. Par contre, le contraire de cette dernière inverserait la biosynthèse de la progestérone et de l'œstrogène ovarien [52].

En ce qui concerne la sécrétion des gonadotrophines, certaines études ont montré que les variabilités génétiques du gène LHR qui code pour le récepteur de LH sont connues pour être responsables de la puberté précoce familiale chez l'homme (MIM # 176410) et la résistance à l'hormone lutéinisante (MIM# 238320) [53]. D'autre part, les variants qui affectent le gène FSH-R sont reconnus pour leurs associations à la maladie de la dégénérescence et de l'hyperstimulation ovarienne. La dégénérescence ovarienne est une maladie à transmission autosomique récessive qui se caractérise par une aménorrhée principalement chez les femmes, et par un développement variable des caractéristiques sexuelles secondaires (MIM# 233300) [12, 53].

En se basant sur des études antérieures, nous avons retenu dans ce chapitre, qu'il existe des liens génétiques qui seraient associés aux dysfonctionnements des cellules différenciées de l'antéhypophyse. Évidemment, nous nous sommes référés aux différents résultats déjà reportés dans la littérature pour être associés à la malformation congénitale de la tige hypophysaire. En conséquence, ces derniers nous ont permis de comprendre

que l'altération fonctionnelle des gènes, spécifiquement des facteurs de transcription (Tableau 2-1) qui sont primordiales au développement embryonnaire de l'hypophyse, pourraient nous aider à mieux caractériser le syndrome PSIS dans notre étude.

Tableau 2-1 - Déficiences héréditaires en hormone hypophysaire causées par des facteurs de transcription mutants.

Gene	Band cytogénétique	Déficience en hormone hypophysaire	Observation par IRM	Malformations associées	Modèle de transmission génétique
POU1F1	3q11	GH, PRL, ±TSH	Normal ou hypoplasie antéhypophysaire		Récessif; Dominant
PROP1	5q35	GH, PRL, TSH, LH, FSH, ± ACTH	Normal, hyperplasie, hypoplasie hypophysaire ou hypophyse Cystique		Récessif
HESX1	3p21	GH, PRL, TSH, LH, FSH, ACTH, disfonctionnement post hypophysaire	Normal ou hypoplasie antéhypophysaire; normal ou ectopique post hypophysaire	dysplasie septo optique	Récessif
PITX2	4q25	GH, PRL, TSH, LH, FSH		Syndrome de Reiger	Dominant
LHX3	9q34	GH, PRL, TSH, LH, FSH	hyperplasie, hypoplasie antéhypophysaire	cou court; colonne cervicale rigide	Récessif
LHX4	1q25	GH, TSH, ACTH	hypoplasie antéhypophysaire; ectopie post hypophysaire		Dominant
TBX19	1q23	ACTH	Hypophyse normale		Récessif
OTX2		GH, TSH, ACTH, PRL, LH, FSH	hypoplasie antéhypophysaire; ectopie post hypophysaire	Malformation des yeux	Dominant négatif
SIX6	14q22		hypoplasie hypophysaire; absence du chiasme	Syndrome brachio-otorenal; syndrome oculoauriculo vertébral	Insuffisance haplotypique
SOX2	3q26	GH, FSH, LH	hypoplasie hypophysaire; déféctuosité hem encéphalique	Anophtalmie; Atrésie œsophagique	
SOX3	Xq27	GH, TSH, ACTH, FSH, LH	hypoplasie antéhypophysaire; ectopie post hypophysaire		Récessif lié au X
IGSF1	Xq25	GH, PRL, TSH		Elargissement testiculaire	Récessif lié au X
NR5A1	9q33	FSH, LH		insuffisance adrénérergique, déféctuosité des gonades, inversion des chromosomes sexuels	Dominant; Récessif
NROB1	Xp21.3	FSH, LH		Hypoplasie congénitale des glandes adrénérergiques; déféctuosité des gonades; inversion des chromosomes sexuels	Dominant lié au X

Dans ce tableau, nous indiquons les caractéristiques des facteurs de transcription connus pour être associés aux déficiences hormonales de l'hypophyse. Dans la première colonne, nous indiquons le nom des gènes qui codent pour des facteurs de transcription. Dans la deuxième, nous indiquons la localisation

cytogénétiques des mutations connues pour avoir engendrée la perte de fonction des gènes. Dans la troisième colonne, nous indiquons les déficiences en hormone hypophysaire lorsque le facteur de transcription indiqué est défectueux. Dans la quatrième colonne, nous indiquons la structure de l'hypophyse observée par IRM en présence du facteur mutant. Dans la cinquième colonne, nous indiquons la malformation hypophysaire connue pour être associée au facteur de transcription indiqué dans la colonne 1. Finalement, la sixième colonne indique les modèles de transmission de la variation génétique affectant le gène du facteur de transcription.

Chapitre 3 - Analyses automatisées des données NGS

3.1 - L'analyse et l'étiologie du PSIS

Dans le premier chapitre de ce mémoire, nous avons défini différents concepts de la génétique humaine qui nous permet de comprendre les mécanismes d'expression des gènes sous l'effet des différents types de variation génomiques. Ces informations nous ont permis de comprendre comment les variations génomiques peuvent altérer les fonctions des gènes clés dans l'embryogenèse de l'hypophyse et les syndromes qui ont été découverts à partir de celles-ci (chapitre 2). En continuité, nous allons montrer dans ce chapitre, comment les nouvelles connaissances et l'évolution de la technologie informatique biomédicale pourraient influencer les méthodes d'analyse des données de séquences d'ADN de nos jours. De plus, nous allons détailler l'implication expérimentale des nouvelles technologies dans les analyses génétiques moléculaires ainsi que leurs contributions en termes de rendement et de performance.

Cependant, notre but principal dans ce projet était bien la recherche des causes génétiques associées à la malformation congénitale de la tige hypophysaire. Plusieurs méthodologies ont déjà été mises en application pour caractériser les causes associées au PSIS, que sont la méthode biochimique qui se base sur la mesure de la concentration d'hormone dans les glandes en périphérique [54-55] et celle d'analyse d'imagerie par la résonance magnétique (IRM) [56-58]. Cette dernière exploite la capacité des protons à émettre des signaux d'énergie capables d'être captés par l'appareil d'imagerie afin de permettre la visualisation de ces régions sur une image monochrome.

L'analyse des données de séquençage WES [59] est la plus récente méthodologie permettant de caractériser le syndrome de la malformation congénitale de la tige hypophysaire. Celle-ci est la méthode dont nous avons fait usage dans ce projet et elle se trouve parmi les plus populaires et les plus appliquées de nos jours.

Les démarches précédemment mentionnées n'ont malheureusement pas toutes les mêmes sensibilités ; leurs caractéristiques distinctes peuvent engendrer des impacts sur les couts et les délais d'analyse.

3.1.1 - Historiques du séquençage de molécules d'acides nucléiques

Le séquençage génomique est une approche qui permet de lire et d'extraire les informations contenues dans le génome d'un organisme. Cette dernière est apparue pour la première fois, en 1977, par les découvertes d'Allan Maxam et Walter Gilbert [60-62]. Leurs approches consistaient à utiliser des molécules chimiques (le diméthyle sulfate et l'hydrazine) qui provoquent des cassures dans l'ADN, afin d'obtenir des plus petites molécules qui sont ensuite marquées, à l'extrémité 5', par des isotopes radioactifs (Phosphore ^{32}P , Soufre ^{35}S). Le marquage radioactif n'avait que pour but de favoriser la visualisation de leurs migrations dans un gel d'électrophorèse. Cette méthode est spécifiquement fondée sur la dégradation chimique de la molécule d'ADN.

Vers les années 1977, Frederick Sanger a introduit une nouvelle méthode de séquençage d'ADN, qui se caractérise par l'élongation de la molécule par une procédure de synthèse enzymatique de la séquence poly-nucléotidique. Cette méthode consiste à utiliser des molécules fluorescentes de didéoxynucléotide triphosphate, qui sont des molécules ayant la propriété de prévenir l'élongation d'un polymère nucléotidique [62].

Leurs marquages fluorescents facilitent la visualisation et la détection de l'incorporation des nucléotides par l'enzyme à la fin de chaque séquence.

3.1.2 - Évolution de la technologie de séquençage

L'évolution de la technologie de séquençage d'ADN est un évènement très favorable pour le développement de la recherche génomique. Celui-ci a permis d'améliorer la qualité des résultats, de diminuer le coût et le temps associés au service de séquençage et d'analyse d'ADN. Le coût du séquençage génomique de nos jours est très abordable, celui-ci est passé de plusieurs millions de dollars, dans les années 2001, à quelque milliers des dollars (Figure 3-1). Cette diminution de coût a beaucoup contribué à l'avancement des projets d'analyses de données génomiques car le séquençage de cohortes de grande taille est de plus en plus faisable.

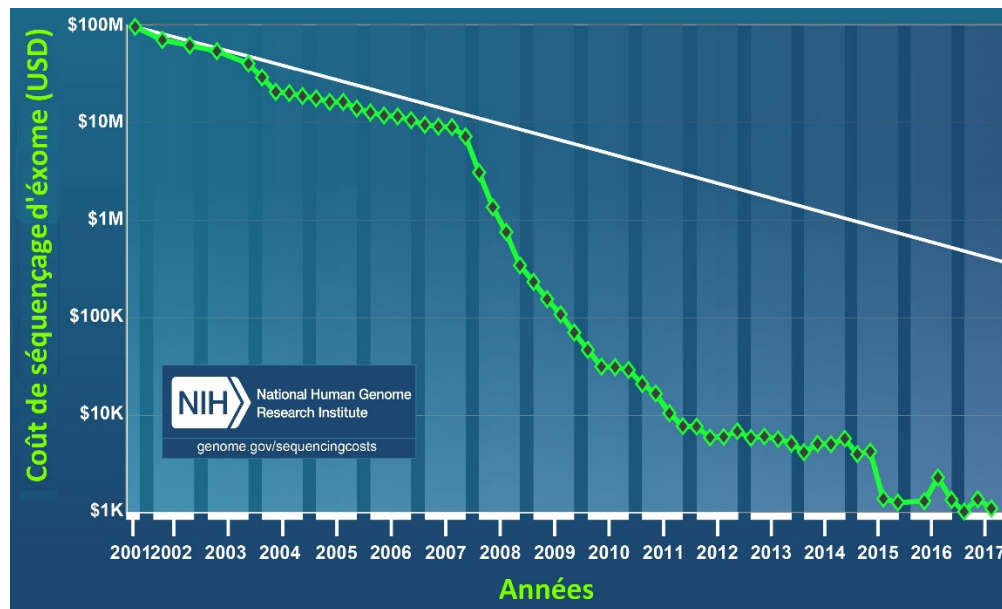


Figure 3-1 - Représentation graphique de la décroissance du coût de séquençage du génome humain à travers les ans.

Cette figure montre la décroissance du coût de séquençage par génome humain à travers le temps. Sur l'axe des abscisses, nous représentons l'intervalle de temps (en années) sur lequel les coûts de séquençage (dollars US) par génome humaine ont été évalués. Les abréviations M et K représentent respectivement 10^6 et 10^3 .

3.1.3 - Les diverses générations de séquençages

Il existe plusieurs protocoles de séquençage d'ADN qui sont de plus en plus évolués à travers le temps appelés séquençage prochaine génération (NGS). La première génération de séquençage d'ADN est apparue avec les méthodes enzymatiques de Frederick Sanger qui représentent la méthodologie de base de plusieurs autres technologies de séquençage qui dépendent de la synthèse et de l'élongation de la chaîne poly-nucléotidique. Ces technologies sont identifiées sous le nom de séquençage prochaine génération, deuxième génération et troisième génération qui sont chacune à la base d'une méthode de détection différente.

En 2005, le fondateur de la société « 454 Life Sciences » Jonathan Rothberg a introduit la toute première plateforme de séquençage prochaine génération appelée « GS20 » qui a permis de séquencer le génome complet de la bactérie *Mycoplasma Genitalium* ayant une taille génomique d'à peu près 0.58 Mbp [63]. En 2006, la compagnie « Roche/454 Life Sciences » a introduit une nouvelle machine de séquençage appelée « GS FLX Titanium », cette dernière a été la première à adopter la méthodologie du pyro-séquençage [64-66]. Le principe de ce dernier consiste à associer l'incorporation des nucléotides à des émissions de signaux lumineux durant l'élongation de la molécule d'ADN. Ces signaux sont détectés grâce au photorécepteur de la machine de séquençage (Figure 3.2) [67-71].

En 2007, la méthode de séquençage à haut débit « High throughput sequencing » a été nommée la méthode de choix de l'année [72] car celle-ci se repose sur le séquençage en parallèle, à haut débit, d'immenses quantités d'ADN. Le principe général de cette dernière consiste à hybrider les polymères d'acide nucléique simple brin avec des sondes

préfabriquées, fixées dans des puits de la plaque de séquençage afin de permettre la synthèse d'ADN complémentaire simple brin par un enzyme d'ADN polymérase à haute performance. De plus, la technique de séquençage par synthèse a demeuré le principe de base de toutes les plateformes de séquençage de nos jours (Illumina, ABI, Hélicos, Roche, etc) [73-75].

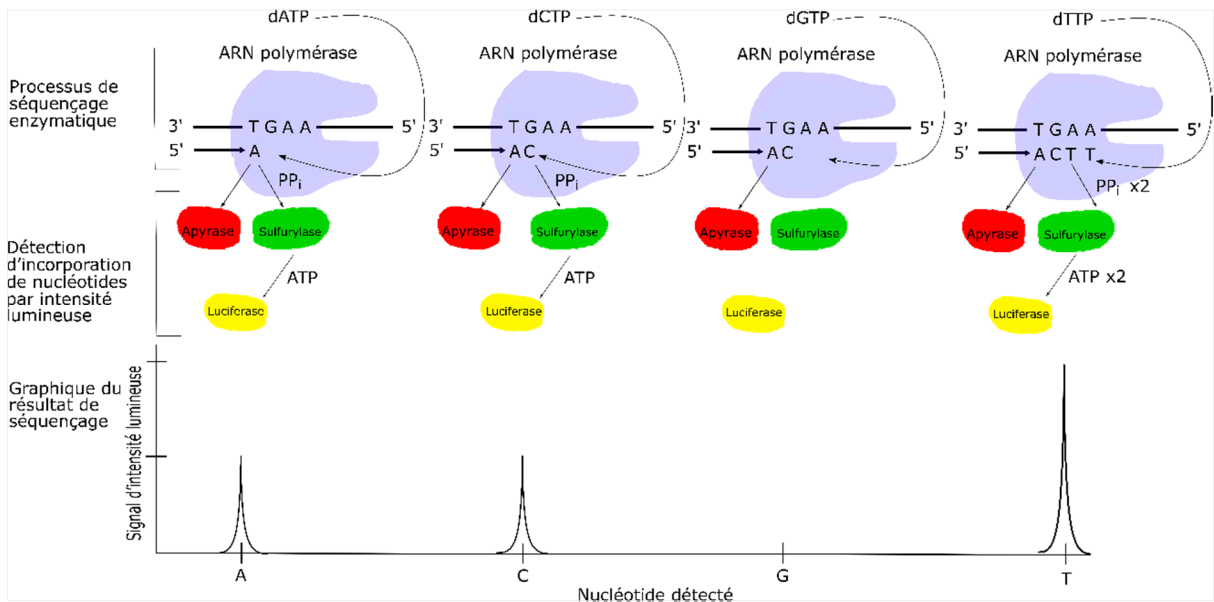


Figure 3-2 - Procédure de détection des nucléotides incorporés durant le pyro-séquençage de l'ADN.

Selon d'Ahmadian, H. et al., 2006 [64]

Cette figure montre un modèle de séquençage d'ADN par la méthode de pyro-séquençage. Cette dernière consiste à émettre un signal lumineux à chaque incorporation de nucléotide dans l'amorce durant l'élongation de l'ADN. Le séquençage de l'ADN se fait dans la direction de 5' vers 3'. Nous avons indiqué en bleu l'ARN polymérase attaché sur l'amorce de l'ADN, en rouge l'enzyme apyrase, l'enzyme sulfurylase est représentée en vert et celle-ci permet de métaboliser l'ATP en recyclant le groupement di-phosphate libéré par le nucléotide incorporé. L'ATP est ensuite utilisée comme carburant par l'enzyme luciférase afin de produire un signal lumineux qui est ensuite détecté par le pyrogramme et génère enfin le résultat graphique de l'ADN séquençé.

3.1.4 - Catégories séquençage d'ADN

Il existe plusieurs façons de séquencer les matérielles génomiques, qui sont tous efficaces. Par contre, nous allons nous intéresser sur les méthodes de séquençage les plus populaires que sont : le séquençage du génome complet (WGS), celui des régions codantes (WES), et celui du génotype (génotypage). Le séquençage WGS permet d'extraire la totalité de l'information génétique du génome. Ces informations sont obtenues à partir de l'extraction d'ADN génomique et mitochondrial. Elles doivent ensuite être séquencées afin de digitaliser les données. Enfin, les données digitalisées sont assemblées par des protocoles d'alignement de séquences afin de reconstituer intégralement le génome. Dans d'autres cas, elles sont cartographiées sur un génome de référence afin d'étudier particulièrement les données d'un génome via les informations génomiques ancestrales. D'autre part, le séquençage WES permet d'extraire seulement les informations génétiques des régions codantes du génome. Par conséquent, l'alignement des données du WES permet de reconstruire seulement le génome codant qui représente environ 2% de la totalité de l'information génétique.

Les méthodes de séquençage WGS et WES ne sont pas appropriées à toutes les analyses, en plus que les avantages et les inconvénients varient selon ces dernières. Grâce au séquençage WGS, il est possible d'analyser non seulement les régions codantes du génome, mais aussi d'autres éléments régulateurs qui pourraient potentiellement être localisés dans les régions non codant de celui-ci. Ce dernier est considéré comme la méthode de séquençage la plus couteuse à ce jour [76-77]. Contrairement au séquençage WGS, le séquençage du WES permet d'analyser les régions codantes du génome avec une bonne couverture des informations génétiques des locus et celle-ci se base sur la

profondeur de lecture spécifique à la technologie de séquençage d'ADN utilisée. De plus, le séquençage WES est la seule technique, à ce jour, qui fournit autant d'information du génome codant à un coût raisonnablement faible par rapport aux autres techniques.

En conclusion, nous savons que certaines maladies sont caractérisées par au moins une altération génomique qui affecte des éléments régulateurs importants dans l'expression des gènes [78-79]. Sachant que l'analyse du WES ne permet pas de détecter des variants qui affectent certains éléments régulateurs localisés dans des introns, à l'exception des UTR et des sites d'épissages alternatifs, celle-ci permet d'exploiter uniquement les informations des régions codantes. De ce fait, ce protocole n'est pas recommandé pour caractériser toutes les causes génétiques associées à la malformation congénitale de la tige hypophysaire, surtout lorsqu'elles impliquent des éléments régulateurs localisés dans les introns du génome.

3.1.5 - Flux de données massives

Il existe diverses technologies de séquençages d'ADN très évoluées et très performantes sur le marché. Ce sont des technologies qui offrent beaucoup d'avantages, sans oublier certains inconvénients. Sachant que le choix d'une plateforme de séquençage ne devrait pas être arbitraire, cela implique qu'il faut déjà s'attendre à maîtriser le format des fichiers qui sont fournis par le service de celle-ci. Sur ce point, nous allons introduire les deux formats de représentation numérique des résultats de séquençages d'ADN les plus connus à ce jour.

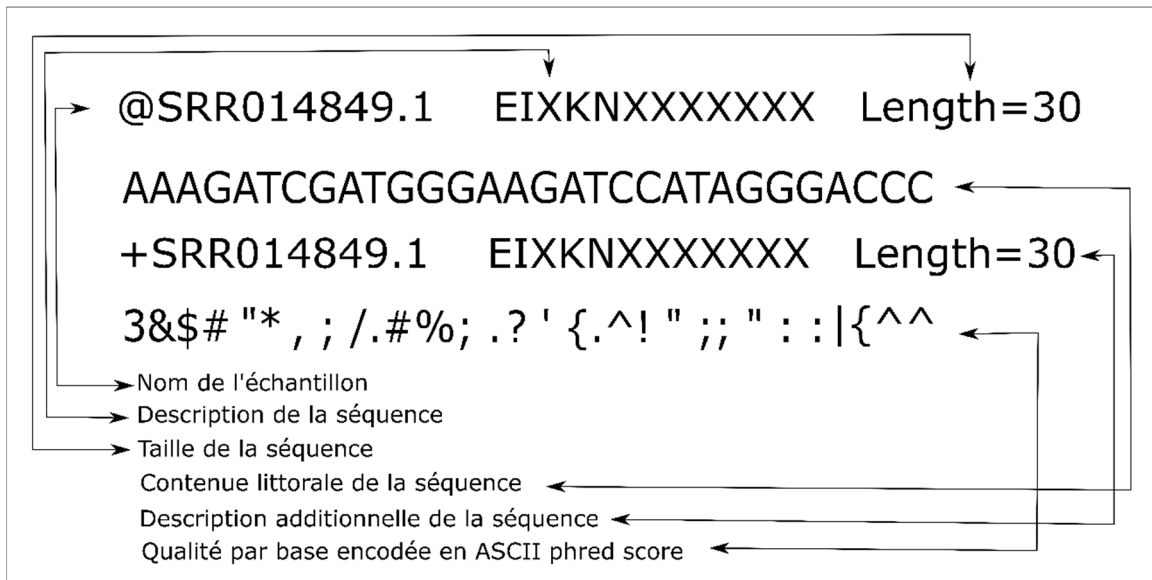


Figure 3-3 - Représentation d'une lecture d'ADN séquencée sous un format Sanger FastQ.

Cette Figure indique une représentation littorale d'une séquence d'ADN issue de séquençage via la technologie d'Illumina. Nous constatons sur la première ligne, le nom de l'échantillon d'ADN séquencée débutant obligatoirement par un préfixe « @ », suivie d'une description de celle-ci et finalement la taille de la séquence. La deuxième ligne indique la séquence encodée par les lettres A, G, C et T et celle-ci est suivie par une troisième ligne qui indique la qualité par base encode en ASCII phred 33 ou phred 64.

3.1.6 - Les données Fasta et FastQ

Afin de représenter les informations de l'ADN séquencée sous un format de données digitalisées qui soient lisible par des humains, chaque nucléotide séquencé y est représenté par une lettre faisant partie de l'alphabet A, G, C et T. De plus, ces dernières sont accompagnées de leurs qualités de détection par base, encodées en ascii phred33 ou phred64 [80]. C'est deux types d'encryptions n'influencent pas la représentation de la qualité des bases, elles sont plutôt spécifiques à des technologies de séquençage. Nous distinguons deux de ces types de formats : le format FastQ et le format Fasta qui ne rapporte pas la qualité Phred des bases séquencées (Figure 3-3).

Le concept du score de qualité phred est une mesure qui se réfère à la probabilité de la détection d'une base séquencé. Cette mesure est automatiquement compilée par la

machine de séquençage et elle se mesure sur une échelle allant de 10 à 50. Le score phred se calcule par la formule suivante : $Q = -10 \text{Log}_{10} P$ où Q est le score de qualité phred et P la probabilité de détecter la mauvaise base [81] (Tableau 3-1).

Les formats Fasta et FastQ ne sont pas les seuls formats qui représentent les données génomiques, il existe d'autres types de données comme celles issues des biopuces de génotypage par exemple, mais celles-ci répondent à des attentes différentes.

Tableau 3-1 - Relation des scores phred avec la précision de la détection des bases.

Q (Score de qualité phred)	P (Probabilité fractionnaire d'identification incorrecte de base)	Précision
10	1 / 10	0.9
20	1 / 100	0.99
30	1 / 1000	0.999
40	1 / 10000	0.9999
50	1 / 100000	0.99999

Ce tableau indique la signification de la qualité phred score par rapport à la probabilité d'erreur de séquençage et de la précision de détection d'une incorporation de base durant le séquençage. Dans la première colonne, nous indiquons la qualité phred score, la deuxième indique la probabilité d'erreur de séquençage et la troisième indique la précision de la détection de la base.

3.2 - Analyse bio-informatique de données NGS

L'analyse des données NGS consiste à regrouper un ensemble d'outils bio-informatique selon des critères d'efficacité bien établis afin que ceux-ci forment une chaîne d'analyse séquentielle des données, appelé « Pipeline ». Il n'existe pas de règle formelle sur la constitution d'un pipeline, il suffit que celui-ci soit capable de traiter des données soumises en entrée et de générer des résultats avec le plus faible taux d'erreur. Les pipelines se diffèrent l'un à l'autre par leurs natures et leurs constituants.

Il existe une grande quantité d'outils bio-informatiques disponibles en ligne qui permet d'analyser les données NGS sur lesquels nous n'allons pas discuter dans ce mémoire, par contre nous allons nous concentrer sur ceux qui sont pertinents pour les procédures de détection de variants dans ce projet.

3.2.1 - Outils d'alignement de séquences

L'alignement de séquences génomiques est une méthode d'analyse en bio-informatique qui permet de rechercher l'emplacement, la plus vraisemblable, d'une séquence dans un génome de référence fourni. Dans certains contextes, cette approche a pour but de faire ressortir les différences et les ressemblances par base ou par sous séquences à travers des données comparées. Les outils d'alignement de séquences sont indispensables dans le domaine d'analyse génomique.

L'alignement de séquences est un concept polyvalent qui est utilisé dans plusieurs domaines autres que la génomique, par exemples la recherche de mot clé ou d'expression dans un dictionnaire par le moteur de recherche « Google ». Dans le domaine d'analyse de données génomique, l'alignement de séquence permet de réaliser des fonctions comme suivantes : l'assemblage *de novo* d'un génome inconnu, l'alignement de séquence contre un génome de référence et enfin l'alignement comparatif de séquences. Ces derniers ne répondent pas aux mêmes requêtes expérimentales, chacune de ces méthodes répond à un but d'analyse spécifique. Nous allons préférentiellement nous intéresser aux méthodologies d'alignement de séquence ayant pour but de favoriser la détection de variations génomiques par rapport au génome de référence humain dans notre étude. La version du génome de référence utilisé est bien celle d'hg19 de la base de données de séquences de référence d'NCBI (RefSeq).

Nous notons plusieurs types de logiciels d'alignement de séquences qui sont majoritairement disponibles gratuitement en ligne. Les plus populaires sont : BOWTIE, BWA, Blast, Trinity et finalement Spades. L'alignement de séquences dépend énormément d'une bonne application des paramètres d'alignement, ces derniers peuvent avoir des impacts à plusieurs niveaux sur les résultats, bien qu'aucune étude n'ait pu encore rapporter l'existence de paramètres universels qui soient capables de fournir des résultats précis. Les conséquences d'un alignement efficace ou non peuvent être perçues tant au niveau quantitatif que qualitatif. En fonction des paramètres d'alignement, il est possible d'imposer certaines conditions de base telles que : l'alignement complet ou partiel de la séquence contre la référence, la proportion d'incompatibilité requise ou d'échec de comparaison par base et le nombre des essais de la recherche de la séquence. Les paramètres d'exécution des outils d'alignement de séquences ont beaucoup d'impact sur leurs performances en termes de temps d'exécution. Dans certaines analyses d'alignement de séquences, plus les paramètres d'alignement sont sévères, plus l'exécution de la tâche durera plus longtemps et vice versa. Un paramètre est dit sévère lorsque celui-ci impose une analyse hautement précise à un outil tout en se limitant à un ensemble restreint de résultats. Les logiciels d'alignement ont des avantages et des inconvénients spécifiques en termes d'efficacité et de sensibilité qui se définissent par les modalités d'utilisation des paramètres. Les algorithmes d'alignement sont implémentés à partir des approches heuristiques, ce qui signifie qu'ils sont des algorithmes qui recherchent les meilleures solutions d'alignement d'une séquence possibles, dans un temps raisonnablement court, à travers l'univers des solutions de la problématique [82]. En conclusion, les méthodologies d'alignement de séquences heuristiques sont des

algorithmes qui ont la propriété de rechercher des solutions d'alignement en sacrifiant la précision, l'optimisation afin de fournir les résultats dans un temps d'exécution raisonnable [83]. Dans notre projet, ainsi que dans d'autres catégories de recherche, nous sommes limités à faire usage aux algorithmes heuristiques afin de respecter les échéances des analyses. De ce fait, nous allons détailler, dans la section suivante, la méthodologie d'alignement local et global des algorithmes d'alignement des séquences assez connues, qui sont : BOWTIE2 et BWA.

3.2.2 - Alignement local et global

L'alignement global appelé encore alignement « bout-à-bout », est une méthode d'alignement qui requiert que toutes les caractères de la séquence recherchée soient séquentiellement retrouvées, dans le même ordre, dans la séquence de référence. Il parvient que la séquence recherchée ne soit pas complètement semblable à la séquence retrouvée dans la référence, en conséquent, celui-ci recherche un ensemble de caractères issues de la référence, ordonnées de la même façon et semblable à celle recherchée mais avec certaines incohérences attendues. La méthodologie d'alignement globale impose parfois qu'une taille différente à celle de la séquence recherchée soit considérée afin que les comparaisons soient acceptées dans l'ensemble des solutions (Figure 3-4). L'alignement global assume méthodologiquement la présence d'INDEL dans la séquence recherchée, de ce fait, certaines méthodes de renforcement d'alignement sont fortement recommandées suite à ce choix d'analyse, par exemple l'alignement à bout appariés des séquences (Paired-End).

L'alignement global de séquences est une méthode qui consiste à aligner l'intégralité des bases des séquences contre la séquence de référence. De plus, les

processus de cette méthode appliquent certaines contraintes d'incompatibilité par base spécifiées en paramètres. Lorsqu'une faible proportion de bases incompatibles est requise, plus la taille de la séquence recherchée est grande, plus la probabilité de succès devient très faible. Ce dernier cause souvent l'exclusion de la séquence dans l'alignement. En conclusion, l'alignement global est une méthodologie déconseillée lorsque couverture de séquence représente un critère importante pour l'analyse.

Contrairement à l'alignement global, l'alignement local consiste à rechercher l'intégralité de la séquence d'intérêt ou une proportion de celle-ci à travers la séquence de référence. Cette proportion n'est pas une mesure automatiquement déterminée par l'algorithme, mais plutôt elle est fournie en paramètre comme critère d'inclusion de la séquence dans les résultats d'alignement. L'algorithme d'alignement local s'assure de rechercher la séquence afin que celle-ci y soit parfaitement retrouvée, toujours sous les contraintes d'inclusions fournies, comme le nombre d'incompatibilités par base ou la taille des coupures de séquences en amont et en aval à considérer (Leading-end and Trailing-end soft clipping).

« Leading-end and Trailing-end soft clipping » appelé encore coupure en amont et en aval des séquences, est une méthodologie de base de l'alignement local qui permet d'augmenter la probabilité de retrouver la séquence recherchée en ignorant les caractères incompatibles qui se trouvent en amont et en aval de celle-ci. Dans certains cas, cette méthodologie engendre la formation de séquences non spécifiques, car plus la séquence devient courte, plus il y aura des possibilités de régions de cartographie alternative disponible dans le génome pour aligner cette séquence.

Le concept de « soft-clipping » et de « hard-clipping » est une technique de manipulation de séquences qui diffère selon la nature de l'analyse. Au cours de l'alignement des séquences, les proportions de lectures qui ne sont pas alignées, selon les critères d'exclusion fournies en paramètres, subissent une manipulation appelée soft-clipping. Ce dernier n'apporte pas de modification aux séquences mais identifie par des flags que certaines régions de celles-ci ne font pas partie de l'alignement.

D'autre part, il parvient que le score de qualité Phred d'une lecture soit partiellement ou globalement basse selon le seuil standard. Dans ce cas, la séquence de la lecture est imprécise et l'alignement de celle-ci pourrait conduire à des interprétations erronées des résultats. De ce fait, sachant qu'il n'existe aucune alternative connue qui permettent de réparer des lectures de mauvaise qualité, la modification ou l'élimination de celle-ci demeure la meilleure solution.

En conclusion, l'application des paramètres de coupure de fin de séquence devrait se faire avec beaucoup de réserve afin d'éviter l'alignement des séquences aux mauvais locus de la référence.



Figure 3-4 - Représentation graphique du modèle d'alignement local et global d'une séquence mutante contre sa référence.

Dans cette figure, nous montrons un modèle d'apparition de mutations ponctuelles par rapport aux méthodes d'alignement de séquences. Les séquences figurant sur les deux premières lignes indiquent respectivement la séquence à aligner et la séquence de référence. De la troisième à la neuvième ligne, nous représentons l'alignement global des séquences issues du locus de la séquence à aligner. Dans ce dernier, nous montrons un exemple où l'alignement global des séquences pourrait induire des insertions (GCT en vert) et des délétions (AGG en rouge). La mutation ponctuelle (A>T) est indiquée dans l'alignement à titre exemplaire mais toutefois elle n'est pas engendrée par la méthode d'alignement. A partir de la dixième à la seizième ligne, nous montrons que la méthode d'alignement local génère des résultats différents comparativement à la méthode globale. Les séquences en amont et en aval qui sont colorées en bleues indiquent des régions dont la séquence alignée a subi le « soft clipping ». Ces régions ne font pas partie de l'alignement car elles ont été exclues en fonction d'un signal par l'algorithme d'alignement. Le sens des séquences de l'alignement est de 5' vers 3'.

Chapitre 4 - Aspects expérimentaux du projet

4.1 - Hypothèses et objectifs du projet

Dans ce projet, nous nous intéressons à déterminer les divers facteurs génétiques qui expliqueraient l'origine de la malformation embryonnaire de la tige hypophysaire. Cette dernière requiert l'expression de plusieurs facteurs de transcription, comme : TBX19, PROP1, POU1F1, HESX1, LHX3 et LHX4 [84], ainsi que des facteurs de croissance cellulaire, afin de favoriser le développement normal et la différenciation des cellules de tissus spécifiques à l'hypophyse. D'autres membres de notre laboratoire ont déjà reporté, dans une première étape de notre étude, des variants candidats qui pourraient être associés au PSIS. En utilisant des protocoles alternatifs, nous allons pouvoir déterminer les variants qui n'ont pas pu être détectés par le protocole primaire. Par conséquent, nous serons en mesure de déterminer les variants rares et *de novo* qui nous permettront de caractériser le défaut de développement embryonnaire de la tige hypophysaire. De plus, plusieurs études ont caractérisé l'altération fonctionnelle des gènes HESX1, LHX4, PROP1, SOX2, SOX3, OTX2 et GLI2 qui sont associés aux malformations congénitales de l'hypophyse [85-90]. Ces dernières sont classées en trois grandes catégories, qui sont : l'hypoplasie antéhypophysaire (HAH/HAP), l'hypoplasie de la tige hypophysaire (HTH/HPS), finalement l'ectopie postérieure de l'hypophyse (EPH/EPP).

Afin de déterminer les variants délétères rares et *de novo* qui pourraient affecter les gènes associés à la malformation congénitale de l'hypophyse, nous avons visé d'analyser les données WES de huit familles trio, en implémentant deux pipelines de détection de variants génomiques. Ultérieurement, les données WES des sujets ont été

analysées par un pipeline de détection de variants qui se basait principalement sur GATK. Aussi bien que les protocoles de notre étude, ce pipeline avait pour but de caractériser les facteurs génétiques associés au PSIS. Malheureusement, aucun jeu de données de variants majeurs significativement associés au syndrome n'a été détecté (travaux de Roman Serpa, étudiant du Professeur Samuels, M).

A présent, nous avons implémenté deux pipelines de détection de variants supplémentaires dans le but de déterminer des variants candidats qui n'avaient pas été détectés dans la première phase du projet. De plus, nous avons visé de comparer les mutations détectées entre les pipelines afin de caractériser l'efficacité de détection de variants de chacun.

Le concept d'efficacité de détection de variants dans ce projet se réfère à la capacité de nos pipelines à détecter une quantité attendue de variations codantes. Selon certaines études, il est attendu que les régions codantes du génome humain soient composées de ~10413 SNV synonymes et de ~10389 SNV non synonymes [91]. Nous avons aussi créé un pipeline de détection de variations structurales afin d'explorer différentes possibilités d'altération des gènes associés à la malformation de l'hypophyse. Enfin, nous comptons exploiter nos observations critiques des résultats afin d'exposer les failles de notre méthodologie et les possibilités d'amélioration.

4.2 - Matériels et méthodologies

4.2.1 - Provenance des données et évaluation des sujets

Les participants de notre étude ont été évalués au département du service d'endocrinologie du Centre de Recherche du CHU Ste-Justine sur une période de 5 ans.

La sélection des patients se basent sur l'observation de la déficience d'au moins une hormone pituitaire (Tableau S1). Les patients, malgré leurs jeunes âges, ont subi une évaluation structurelle de l'hypophyse via l'imagerie par résonance magnétique (IRM). Cette technique utilise un composé chimique appelé « gadolinium » afin d'obtenir un meilleur contraste visuel de la glande. Coût de séquençage.

L'étude a été approuvée par le comité d'éthique du Centre de Recherche et les consentements éclairés ont été obtenus auprès des patients ayant l'âge légal à consentir aux directives de l'étude. Les échantillons d'ADN ont été extraits à partir de cellules sanguines et l'extraction de celles-ci a été effectuée selon le protocole standard de notre laboratoire biochimique. Il n'est pas été normalement possible d'obtenir des biopsies des tissus de la glande hypophysaire elle-même.

Notre cohorte se compose de 13 familles majoritairement canadiennes-françaises (12/13), qui ne sont pas toutes des trios, parmi lesquelles 8 des sujets affectés sont présents avec les deux parents (7 Canadiennes françaises, 1 venant du Moyen-Orient). Les échantillons d'ADN des sujets ont été séquencés selon le protocole de capture par hybridation des exons de la compagnie d'Agilent SureSelect. Elles ont été séquencées ensuite par la technologie d'Illumina HiSeq 2000, en utilisant la technique de séquençage à bouts appariés. De plus, la longueur des séquences requises dans nos analyses est sensiblement 100 paires de bases et ces dernières ont été séquencées dans les deux directions de la lecture de l'ADN afin de générer deux séquences à bouts appariés, chacune d'une longueur environ 100 paires de bases (pbs) (2 x (~100 pbs)). Pour plus de détails qui résument les grandes étapes du mécanisme de capture d'exon via la méthode d'Agilent SureSelect, veuillez consulter la Figure 4-1.

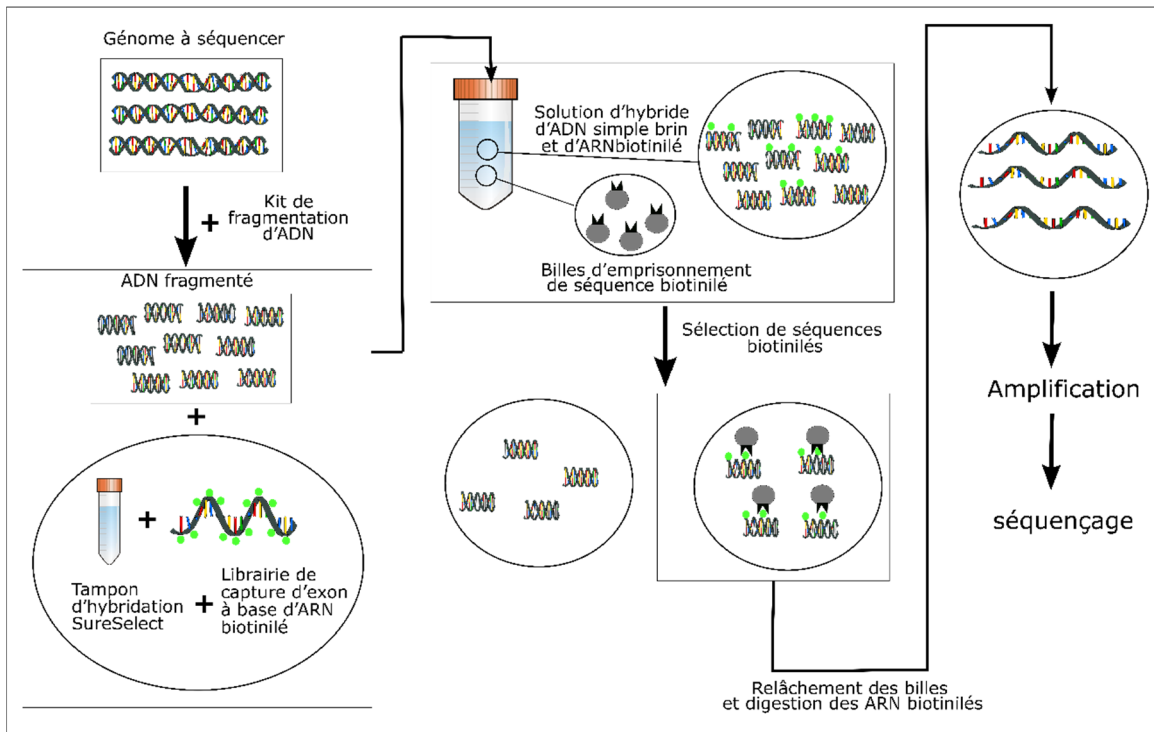


Figure 4-1 - Illustration de différentes étapes d'enrichissement d'exons de l'ADN génomique.

Cette figure a été générée par l'auteur de ce mémoire.

Dans cette figure, nous représentons les procédures de capture d'exon des séquences d'ADN codantes via des pré-mélanges de la compagnie SureSelect d'Agilent. Cette procédure consiste à fragmenter l'ADN génomique en petites séquences (à l'ordre de kilo paires de base), puis les hybridées avec des ARN biotinilés via un tampon d'hybridation issue de la d'Agilent. Ces derniers sont ensuite sélectionnés via des billes spécifiquement utilisées pour emprisonner des séquences d'acide nucléique biotinilées. Les séquences captées par les billes sont ensuite isolées de ces dernières afin de procéder à leur amplification par PCR et leur séquençage via la méthode désirée.

4.2.2 - Analyse des données de séquençage NGS

Les données WES ont été analysées par les deux pipelines d'analyse de variations génomiques qui permettent de détecter les SNV et les INDEL. Ces pipelines ont ensuite servi pour analyser les jeux donnés de variants détectés afin de sélectionner celles qui sont associés au syndrome de la malformation congénitale de l'hypophyse. Chacun de nos pipelines sont caractérisés par un outil de détection de variants différents, qui sont : SAMTOOLS et FREEBAYES. Ceux-ci avaient pour but de découvrir d'autres variations validées qui n'ont pas été retrouvées par l'ancien pipeline (GATK-UK). Les pipelines

sont constitués de logiciels et d'analyses computationnelles différemment implémentés (Figure 4-2). Il est attendu que ces derniers génèrent des résultats qui reflètent leurs particularités.

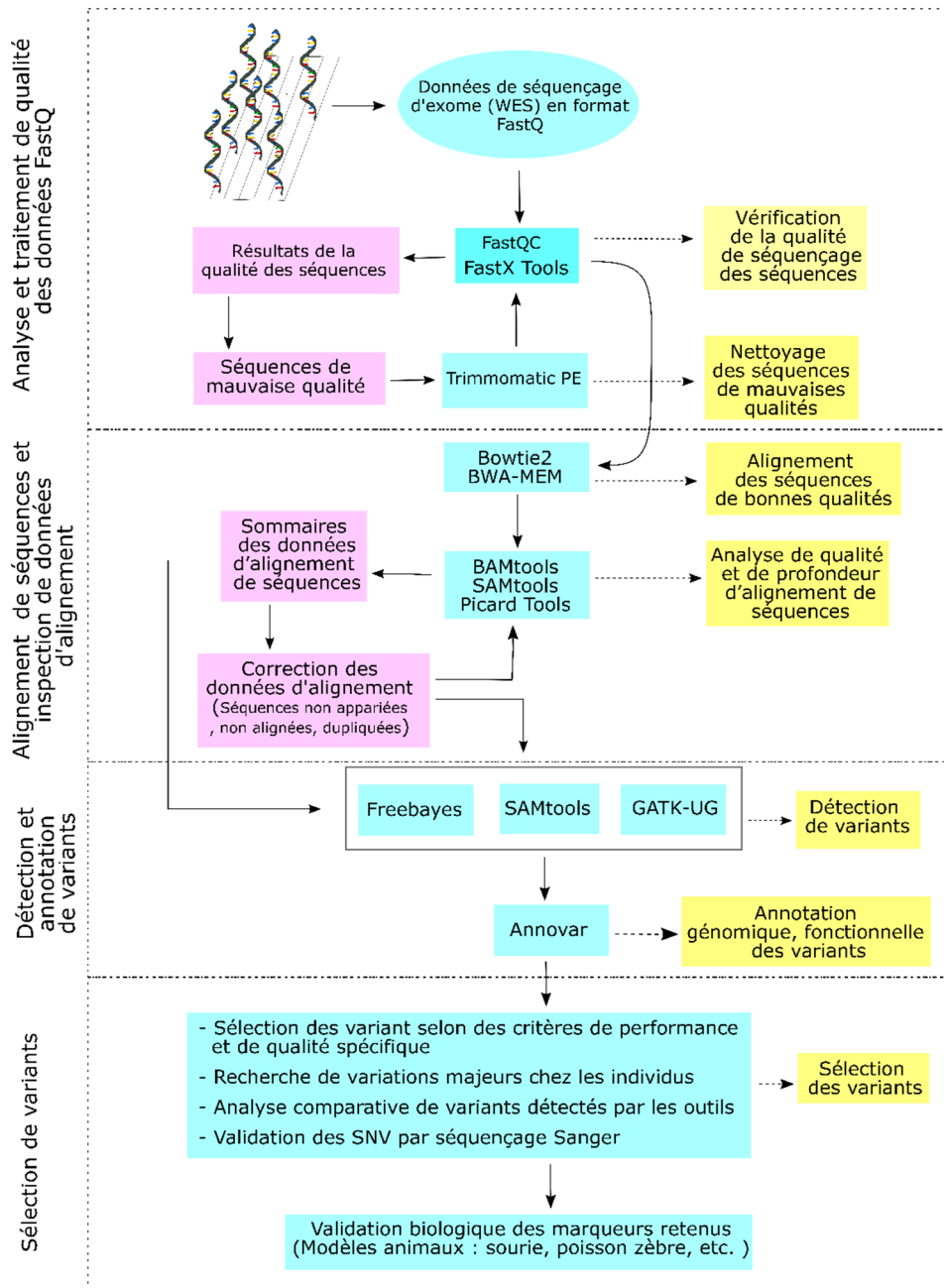


Figure 4-2 – Plan d’analyse des pipelines de détection de variants ponctuels spécifiques aux outils GATK, FREEBAYES, SAMTOOLS.

La figure ci-dessus indique le plan des trois pipelines de détection de variantes utilisées dans notre projet. Nous indiquons en vert les outils et les protocoles appliqués dans nos analyses, nous indiquons en rose les

étapes de traitement de qualité et de modification de données et en jaune, nous définissons par une brève explication, le but de chaque outil à chaque étape. Nous constatons que le pipeline est divisé en quatre grandes étapes : L'analyse et traitement de qualité des données de FastQ fournies en entrée, l'alignement des séquences FastQ et le traitement de qualité des données d'alignement de séquence, la détection de SNV à partir des données d'alignement et l'annotation fonctionnelle des SNV et finalement la sélection spécifique des SNV et la validation biologique fonctionnelle des SNV sur des modèles animaux.

Afin de limiter les biais de mauvaise qualité associé aux données, nous avons inspecté la qualité de ces dernières en utilisant l'outil « FASTQC » [92] (Tableau 4-1) qui a été implémenté par les bibliothèques du langage de programmation JAVA et qui est disponible en ligne de commande et en interface graphique. Ce dernier est en mesure de procéder à une inspection approfondie de la qualité des lectures, pour ensuite générer un sommaire d'évaluation statistique globale. Dépendamment de la qualité des lectures, il est parfois nécessaire de corriger celles-ci en éliminant certaines proportions (*hard clipping*) qui ont un score de qualité Phred [80] inférieur au seuil standard reporté dans la littérature (section 3.2.2).

4.2.3 - Correction de la qualité des lectures

Suivant l'inspection de la qualité des lectures (voir l'exemple des résultats dans la Figure 4.3), nous avons tenté de remanier celles qui ont échoué les critères d'inclusion de la qualité, et ce afin de retenir des lectures dont la qualité moyenne des bases totalise le seuil minimum standard du score de qualité phred. Pour ce faire, nous avons utilisé l'utilitaire d'analyse de traitement de séquences NGS appelé « TRIMMOMATIC PE ». De même que le logiciel FASTQC, celui-ci est implémenté par les bibliothèques du langage de JAVA et il est disponible qu'en ligne de commande.

Le traitement des séquences avec TRIMMOMATIC PE (Paired end) consiste à rechercher à travers les lectures, certaines proportions de celles-ci qui totalisent une

qualité moyenne inférieure au seuil minimum standard reporté dans les littératures. Le but de rechercher ces lectures est d'éliminer les proportions celles-ci ayant une faible qualité et de retenir celles qui dépassent le seuil standard. Les proportions de lectures analysées sont retenues selon les critères de taille minimum de la séquence (MINLEN) et de la qualité moyenne requise. Ces deux critères sont imposés par les paramètres utilisateurs, que sont : « SLIDINGWINDOW » et « MINLEN ». Il existe plusieurs autres paramètres qui sont impliqués dans l'analyse des séquences dont nous n'allons pas développer dans cette section par manque de pertinence, veuillez consulter la documentation en ligne (Tableau 4-1) pour les détails de chacun. Le mécanisme de fonctionnement de l'utilitaire TRIMMOMATIC PE consiste à traverser séquentiellement chaque intervalle de la séquence sous forme de fenêtre d'analyse (SLIDINGWINDOW). Cette dernière est fournie en paramètre comme critère d'analyse et de performance et elle est définie par une taille précise et constante durant l'analyse.

Enfin, certains paramètres de traitement de séquences par TRIMMOMATIC peuvent avoir des conséquences néfastes sur les analyses en aval, par exemple l'alignement des séquences et la détection de mutations à partir de celles-ci. De façon plus détaillée, il parvient que certaines séquences issues de l'analyse par TRIMMOMATIC dépassent le seuil de qualité standard, tout en ayant une taille qui rend leur alignement commun dans le génome de référence. De ce fait, elles auront tendance à mapper à plusieurs locus du génome et génèrent des observations de variations non valides, tel est souvent le cas des séquences qui sont issues des pseudo-gènes.

Tableau 4-1 - Disponibilité des différents logiciels utilisés pour déroulement des deux pipelines de détection de mutations ponctuelles.

Outils et bases de données en ligne	Liens de téléchargement
Human Genome Hg19	ftp.illumina.com/Homo_sapiens/UCSC/hg19/Homo_sapiens_UCSC_hg19.tar.gz
FASTQC v0.11.5	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.5.zip
TRIMMOMATIC v0.36	http://www.usadellab.org/cms/uploads/supplementary/TRIMMOMATIC/TRIMMOMATIC-Src-0.36.zip
BOWTIE2 v2.3.1	https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.3.1
SAMTOOLS v1.3.1	https://github.com/samtools/samtools/releases/download/1.3.1/samtools-1.3.1.tar.bz2
BCFtools v1.3.1	https://github.com/samtools/bcftools/releases/download/1.3.1/bcftools-1.3.1.tar.bz2
Tabix v0.2.6	https://sourceforge.net/projects/samtools/files/tabix/tabix-0.2.6.tar.bz2/download
HTSlib v1.3.2	https://github.com/samtools/htslib/releases/download/1.3.2/htslib-1.3.2.tar.bz2
FREEBAYES v0.9.10	https://github.com/ekg/freebayes
Annovar v2016Feb01	http://www.openbioinformatics.org/annovar/annovar_download_form.php
1000Genome project frequency	http://www.internationalgenome.org/
OMIM	https://www.omim.org/
SIFT	http://sift.jcvi.org/www/SIFT_chr_coords_submit.html
PolyPhen2	http://genetics.bwh.harvard.edu/pph2/
MutationTaster	http://www.mutationtaster.org/
BAMtools v2.4.1	https://github.com/pezmaster31/bamtools
VCFTools v0.1.13	https://sourceforge.net/projects/vcftools/files/vcftools_0.1.13.tar.gz/download

Dans la première colonne du tableau, nous indiquons le nom des logiciels utilisés dans les pipelines et la deuxième indique le lien web à partir duquel ces derniers sont disponibles.

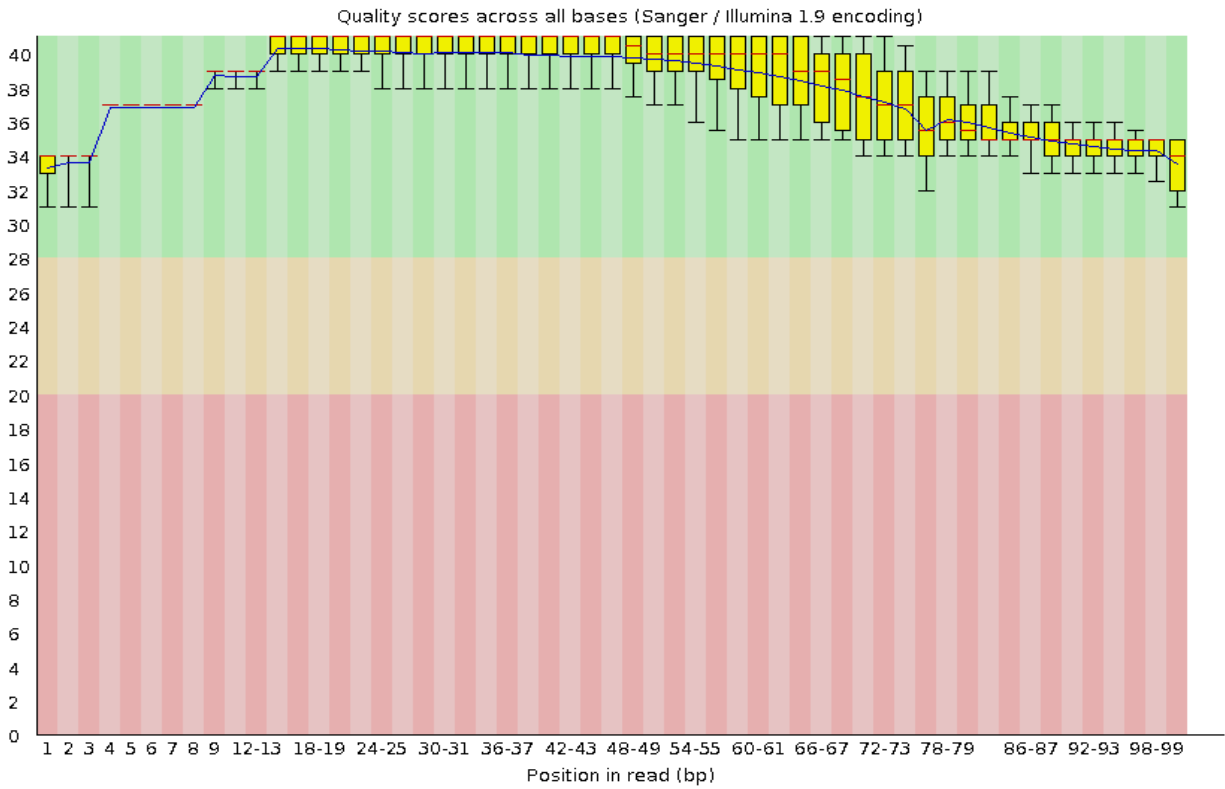
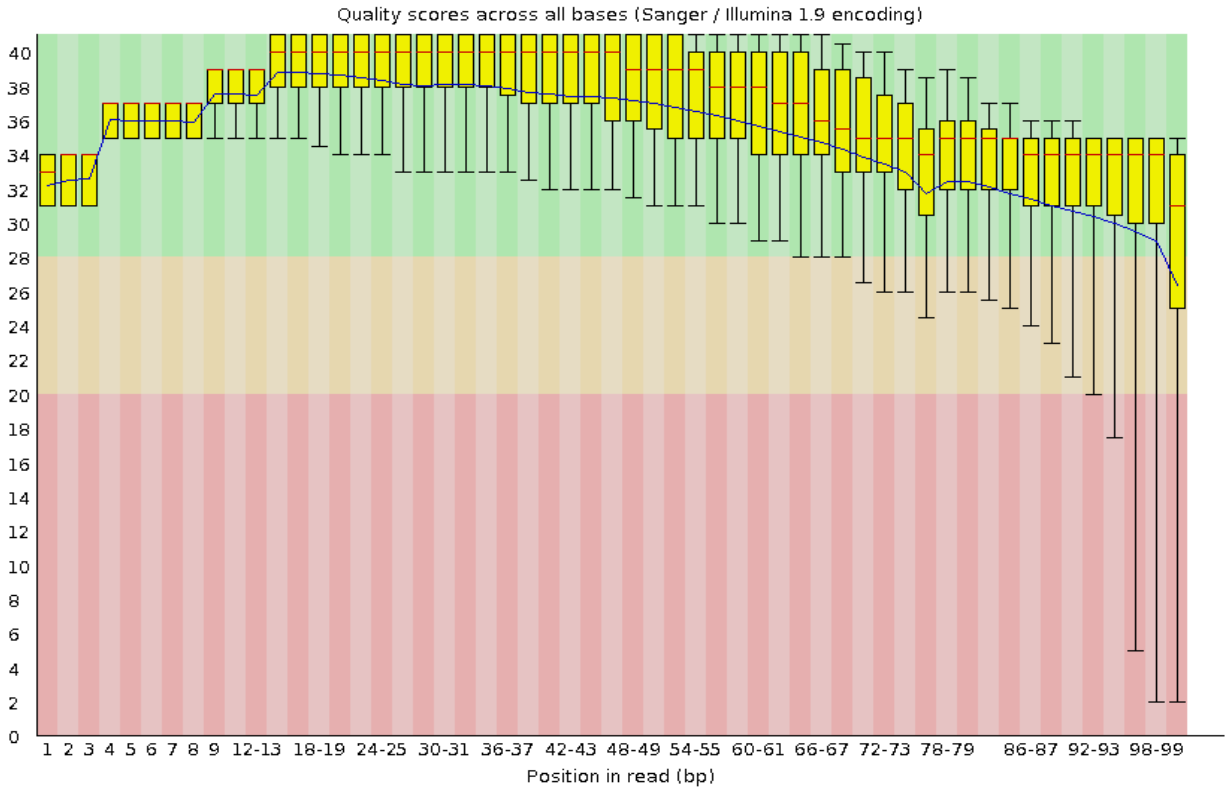


Figure 4-3 - Image comparative des résultats d'analyse de séquences d'un échantillon avant et après le traitement de la qualité via Trimmomatic PE par l'outil FastQC.

Cette figure indique la qualité des séquences mesurées à l'aide du score de qualité phred en ordonnée, en abscisse nous observons la taille en paire de bases des séquences examinées. Le graphe est séparé en trois couleurs : rouge, orange et verte, qui représente un indicateur de gradient de qualité. La zone rouge représente la section où les séquences sont de mauvaise qualité (phred score allant de 0 à 20), celle en orange indique une qualité acceptable des séquences (phred score allant de 21 à 28) et finalement la région verte qui indique une bonne qualité des séquences (phred score allant de 29 et plus). Les boîtes jaune représente le nombre de lectures d'une taille indiquée en abscisse ayant une qualité moyenne de phred score indiquée en ordonnées. La taille des boîtes jaunes reflète une estimation de la proportion des séquences associées aux valeurs indiquées et les lignes verticales noires indiquent l'écart type standard de la moyenne de la qualité des séquences. Finalement, la qualité moyenne des séquences est indiquée par une courte barre horizontale rouge à l'intérieur des boîtes jaunes. Nous avons représenté à travers l'image en haut de la figure, l'analyse de la qualité des données brutes, et l'image en bas, l'analyse de la qualité des données analysées par TRIMMOMATIC PE. Nous constatons dans cette dernière que la qualité montre une amélioration de la qualité moyenne des séquences ainsi que l'écart types standard. Nous constatons l'analyse a reporté très peu des résultats pour les séquences de petites tailles car ces dernières représentaient majoritairement des adaptateurs de séquences exclus par l'analyse de TRIMMOMATIC.

4.2.4 - Choix des outils d'alignement des séquences

Les lectures qui ont été retenues suivant l'analyse du traitement de qualités fait par TRIMMOMATIC, ont été ensuite mappées sur le génome de référence hg19 (human genome version 19) de l'humain selon le protocole d'alignement de séquences du logiciel « BOWTIE2 ». Bien entendu, il existe plusieurs autres logiciels d'alignement de séquences, que sont : BWA, CUSHAW, MOSAIK et NOVOALIGN. Ces outils procèdent tous à l'alignement de séquences selon des méthodologies, des performances et des sensibilités différentes.

Afin de choisir un logiciel d'alignement de séquences capable de minimiser le taux d'erreur d'alignement dans nos analyses, nous nous sommes référés aux études de Cornish, A. et al., 2015. Les auteurs ont analysé la performance des logiciels d'alignement de séquences ainsi que ceux de la détection des variants [93]. Ils ont aussi analysé conjointement les logiciels d'alignement de séquence et de détection de variants dans le but de déterminer le pipeline qui fournisse la meilleure performance de résultat.

Selon leurs résultats, nous avons jugé que les combinaisons « BOWTIE2/FREEBAYES » et « BOWTIE2/SAMTOOLS » représentent les meilleures méthodes d'analyse de variations (Figure 4-4).

La performance des pipelines ont été évaluée selon les deux critères suivants : la précision (PPV) et la sensibilité (TPR). La précision et la sensibilité des résultats sont calculées en fonction des formules représentées ci-dessous.

$$\text{Précision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Sensitivité} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

TP (vrai positif) représente le nombre de vrais résultats validés, FP (faux positif), le nombre de vrais résultats non validés et finalement FN (Faux négatif) qui représente le nombre de faux résultats non validés.

Les auteurs ont comparé leurs résultats à partir d'une base de données de variations pré-validées (Gold standard) appelée NIST-GiaB. Cette dernière est disponible en accès libre via la base de données de GIAB (Genome in a bottle) à l'adresse suivant : <http://jimb.stanford.edu/giab/>.

D'autre part, nous avons renforcé notre choix d'aligneur de séquences grâce aux études d'évaluation des performances des aligneurs de séquences [94]. Les auteurs de ces dernières ont construit leur propre base de données de validation d'alignement de séquences (Gold standard) basée sur un nombre limité de sites de cartographie multiple par lecture (MML, Multiple Mapping Loci) qui est de 10. Toutes autres séquences qui dépassent une valeur MML de 10 sont automatiquement considérées comme des faux

alignements. Les auteurs ont évalué le taux de faux positif (Figure 4-5), le temps de recherche d'un site d'alignement (Figure 4-6) et la consommation en mémoire RAM de l'exécution de la recherche (Figure 4-7). Leurs résultats nous ont permis de constater qu'entre autre BWA qui générer un niveau optimal de faux positif, BOWTIE2 représente le logiciel ayant la meilleur performance algorithmique basée sur les critères d'évaluation de l'étude.

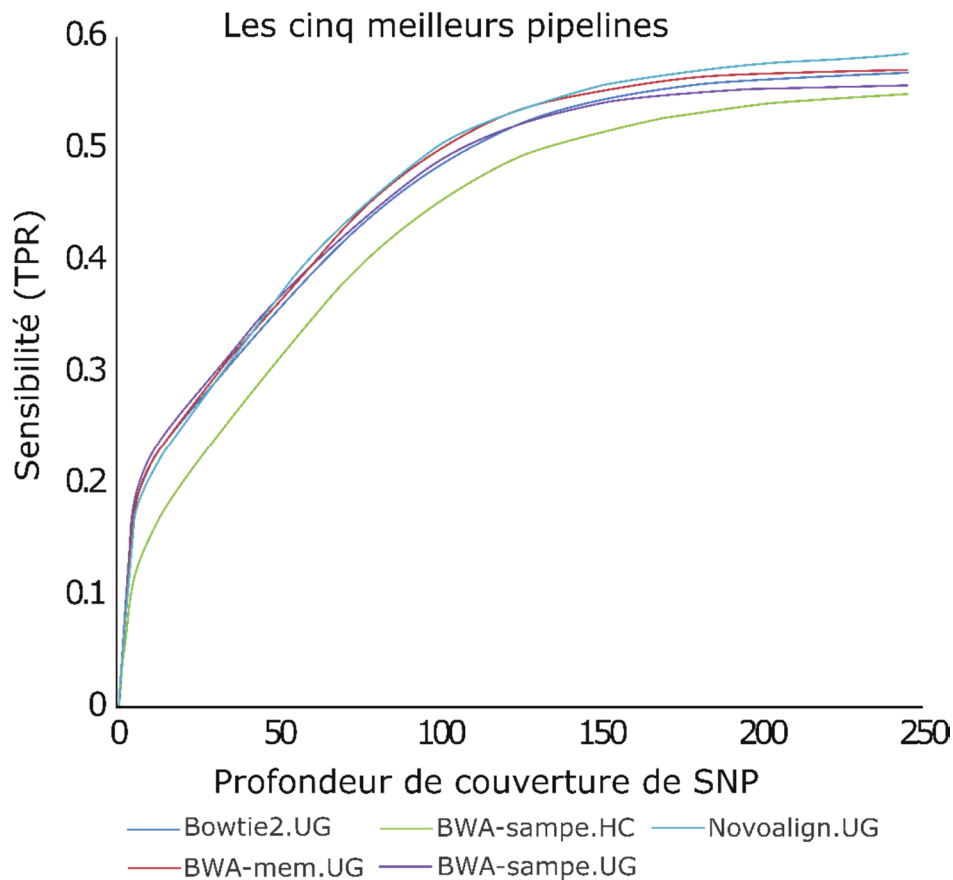


Figure 4-4 - Représentation de la couverture des SNV en fonction de la sensibilité des cinq meilleurs pipelines.

Figure inspirée et traduite, issue de l'étude Cornish et al., 2015

Sur cette figure nous indiquons en abscisse la profondeur de couverture des SNV. Les valeurs en ordonnée indiquent l'évaluation de la sensibilité du top 5 des pipelines retenus dans l'analyse. Les courbes sur le graphe représentent les valeurs de la sensibilité des tops cinq pipelines en fonction de la couverture des SNV. Les pipelines dans le graphe sont constitués de cinq aligneurs, chacun combiné avec le logiciel de détection de variants appelé GATK Unified Genotyper. Nous remarquons, selon l'analyse des auteurs, que

le pipeline NOVOALIGN.UG détient la meilleure performance, pendant que celle de BOWTIE2.UG et BWA-MEM.UG montre une égalité approximative.

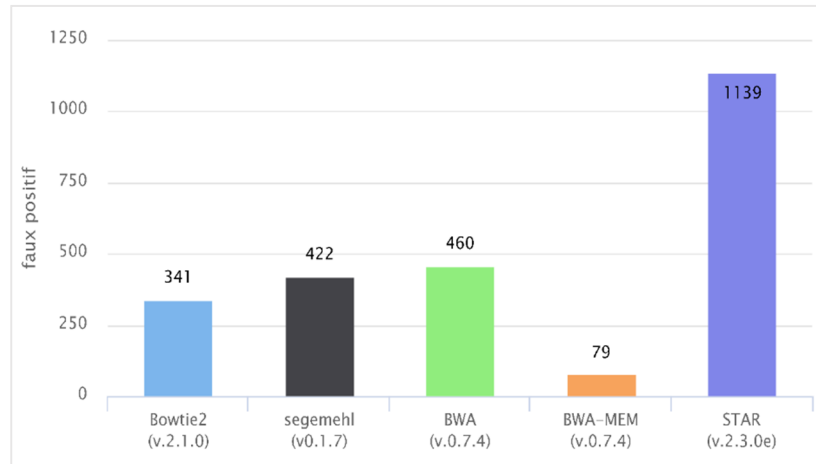


Figure 4-5 - Représentation du niveau de faux alignement générés par les logiciels BOWTIE2, segemehl, BWA et STAR.

Figure inspirée et traduite issue de l'étude Otto, C. et al., 2014 [94]

Cette figure représente l'analyse de la proportion de résultats faux positifs que génèrent les logiciels d'alignement de séquences suivants : BOWTIE2, SEGEMEHL, BWA, BWA-MEM et STAR. Les logiciels d'alignement de séquences sont représentés sur l'axe des abscisses, et les valeurs du nombre de faux alignements sont représentées en ordonnées. Chaque boîte de couleur différente représente un aligneur unique sur le graphe. Nous constatons qu'à l'exception de BWA-MEM, BOWTIE2 est l'aligneur qui génère moins de résultats faux positifs.

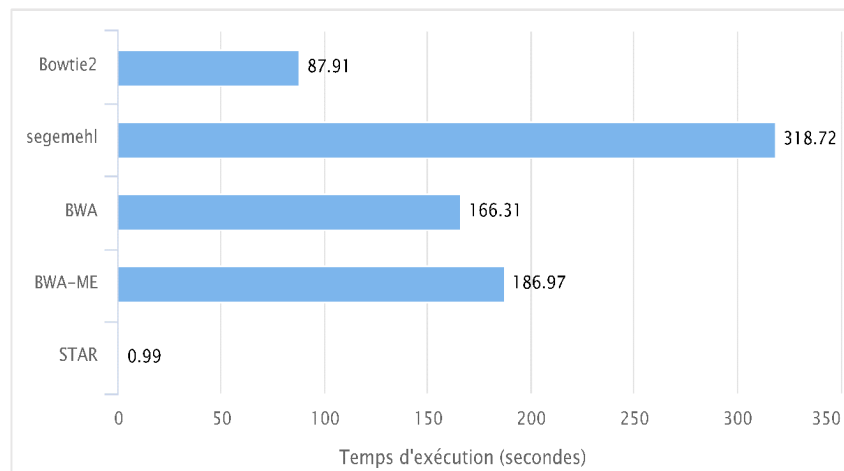


Figure 4-6 - Évaluation du temps d'exécution de la recherche d'un site d'alignement par séquence des logiciels BOWTIE2, SEGEMEHL, BWA et STAR.

Figure inspirée et traduite issue de l'étude Otto, C. et al., 2014 [94]

Cette figure indique les résultats d'analyse de la performance des logiciels d'alignement de séquences en temps d'exécution. La liste des logiciels est indiquée en ordonnée et ces derniers sont représentés par des barres bleues orientées horizontalement. Les valeurs du temps d'exécution des logiciels sont indiquées en abscisse et l'unité de mesure du temps d'exécution est en seconde. Le niveau des barres indique le temps consommé par les logiciels afin de retrouver un site d'alignement par séquence. En comparant les résultats de la figure, nous avons constaté qu'à l'exception de l'aligneur STAR, respectivement les deux logiciels ayant mieux performés en temps d'exécution sont BOWTIE2 et BWA.

En conclusion, nous avons été prudents en combinant les logiciels des pipelines, en plus d'éviter que la performance soit un critère de refus d'utilisation de ceux-ci. Nos observations nous ont permis de nous informer sur les forces et les faiblesses des logiciels d'alignement de séquences, par conséquent nous comptons générer plusieurs résultats qui seront comparés afin d'extraire les meilleurs de tous.

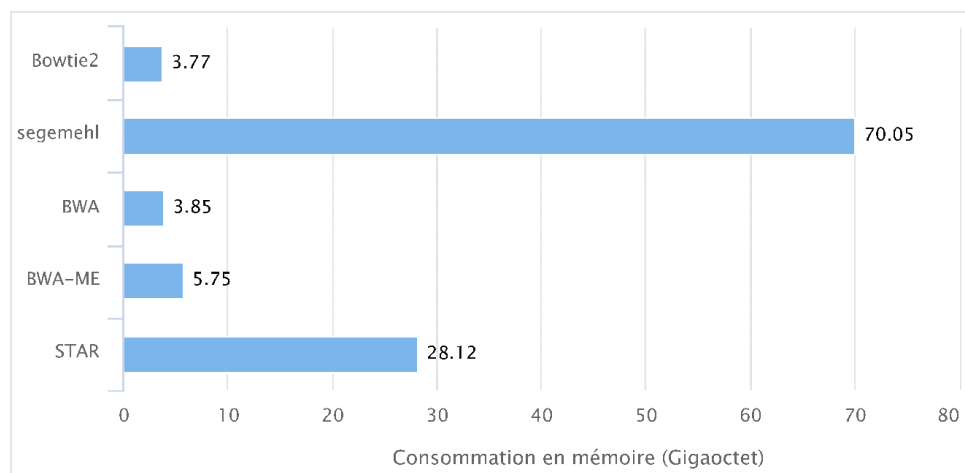


Figure 4-7 - Évaluation de la consommation en mémoire lors de la recherche d'un site d'alignement par séquence des logiciels BOWTIE2, SEGEMEHL, BWA et STAR.

Figure inspirée et traduite issue de l'étude Otto, C. et al., 2014 [94]

Cette figure indique les résultats d'analyse de la performance des logiciels en consommation de mémoire « ram ». La liste des logiciels est indiquée en ordonnée et ces derniers sont représentés par des barres bleues orientées horizontalement. Les valeurs du niveau de mémoire consommée sont indiquées en abscisse et l'unité de mesure de la mémoire est en Gigaoctet (Go). Le niveau des barres indique la capacité de mémoire RAM requise par les logiciels pour la recherche d'un site d'alignement par séquence. Finalement,

les résultats de la figure nous ont permis de constater que BOWTIE2 requiert la plus faible quantité de mémoire RAM pour la recherche des sites d'alignement.

De plus, nous avons considéré que l'aligneur BOWTIE2 a été un bon choix, car il est parmi l'un des aligneurs de séquences génomiques des plus rapides, celui-ci construit des indexes de navigation génomique du type BTW (Burrows Wheeler Transformation) et cela permet d'utiliser moins des mémoires pour la navigation à travers les intervalles génomiques [95].

Les séquences de nos jeux de données WES ont été alignées contre le génome de référence humain (hg19) selon les protocoles par défaut de l'algorithme BOWTIE2. Nous avons opté pour l'alignement local des séquences plutôt qu'un alignement global car cette dernière pourrait potentiellement générer des fausses insertions « GAP » dans les alignements lorsque la cartographie optimale des séquences n'est pas retrouvée. Par conséquent, cette dernière engendre la mésinterprétation des résultats dans nos protocoles de détection de variants. D'autre part, il est attendu que l'alignement local pourrait générer certains alignements issues de séquences de petites tailles, de ce fait, celui-ci pourrait avoir un impact sur l'augmentation des sites potentiels de cartographie des séquences. Afin d'éviter cette dernière, nous avons corrigé notre procédure dans le but de diminuer le pourcentage d'erreurs. De ce fait, nous avons appliqué le paramètre de sensibilité maximale d'alignement local par défaut du logiciel, celui-ci est identifié par l'option « --very-sensitive-local ». Ce dernier se constitue de plusieurs autres options, que sont : le nombre de tentative de recherche (20 tentatives) d'un site d'alignement optimal par fenêtre de recherche, le nombre de reprise (3 reprises) en cas d'échec des tentatives, le nombre d'incompatibilité de base égale à zéro et finalement la taille des sous

séquences recherchées (20). Ces paramètres définissent les efforts que l'algorithme doit fournir afin de retrouver les meilleurs sites d'alignement de séquences. D'autant plus, il est possible d'augmenter encore plus la sensibilité de l'analyse d'alignement en modifiant les paramètres précédemment indiqués. Par exemple, plus le paramètre (--L) d'ajustement de la taille des sous-séquences recherchées est grand, moins il est probable de retrouver séquentiellement les séquences recherchées dans le génome de référence, car l'ensemble de solutions sera trop élevé. Par conséquent, l'algorithme consommera plus de temps et plus de mémoire vive pour trouver la solution optimale. Le besoin en optimisation d'alignement de séquences est une nécessité qui se présente souvent lorsque la taille des échantillons est grande ou la complexité des données est élevée, comme le séquençage génome complet ou d'ARN.

4.2.5 - Procédure de détection de variations

À la suite de l'alignement des séquences de notre jeu de données par BOWTIE2, nous avons utilisé ce dernier afin de rechercher des variations génomiques par les protocoles de détections de variants de SAMTOOLS et de FREEBAYES (Tableau 4-1). La méthodologie de détection de variants génomiques consiste à analyser des données d'alignement de séquences, de les comparer avec une séquence de référence et enfin déduire l'existence de mutations lorsqu'il y a une incompatibilité entre une ou plusieurs bases avec celles de la référence. Afin de déduire qu'un locus soit altéré, plusieurs paramètres sont pris en considération, que sont : le score de qualité phred du locus mutant, la couverture de lectures des variants et la proportion de variants par séquence.

Nous avons initialisé les paramètres de détection de variants aux valeurs par défaut du protocole de SAMTOOLS et de FREEBAYES. Ces derniers ont permis de

retenir une proportion raisonnable de variations qui n'est pas limitante pour notre analyse de sélections personnalisées. L'analyse de détection de variants génère un score de génotypage par allèle mutant détecté qui se mesure sur une échelle de 0 à 100. Ce dernier est calculé en fonction de la valeur des paramètres précédemment indiqués, en plus d'être utilisé comme critère de filtrage des variants.

Avant de commencer à filtrer la liste des variants retenues dans notre jeu de données, nous avons vérifié la proportion de transitions (Ti) par rapport aux transversions (Tv) des allèles. La transition de base est une altération qui substitue le nucléotide de départ vers une autre de la même catégorie, par exemple le changement d'une purine pour une autre purine. Contrairement à la transition, la transversion consiste à substituer le nucléotide de départ à une autre d'une catégorie différente (Figure 4-8, Tableau 4-2). La méthode de vérification de la qualité des données par l'analyse du ratio Ti/Tv permet de constater s'il n'y pas un excès anormal d'une famille d'allèles mutants (Ti) et vice versa. Nous entendons par excès anormal d'une famille d'allèle l'accumulation non attendue de purine ou de pyrimidine dans nos jeux de données. Selon Liu, Q. et al. 2012 [96], nous devrions nous attendre à un ratio de Ti/Tv approximative 3.0 pour une analyse de données WES.

Nous avons retenu des variants ayant un score de génotypage minimum de 90, où chaque variant requiert une couverture minimum de 15 lectures et qu'elle aille un score d'alignement d'au moins 40. (Tableau 4-3). La couverture de lecture des variants est le nombre de séquences du locus analysé étant affectées par la mutation. Il est connu que celle-ci est fortement corrélée le score de qualité du génotype de la variation [76]. Supposons qu'un locus soit couvert par 300 lectures parmi lesquelles 5 sont altérés par

une mutation. Dans ce cas, il est très difficile de déterminer la vraisemblance de celle-ci, en plus du fait qu'elle pourrait être engendrée par des erreurs d'alignement ou de séquençage. D'autre part, supposons qu'une variante soit couverte en moyenne à plus de 50% ou de 90% de profondeur de lecture au niveau du locus concerné. Dans ce cas, il est plus facile de déterminer si le génotype de celle-ci soit homozygote (couverture moyenne de ~95%) ou hétérozygote (~50%) (Section 1.1.3 et Figure 4-9).

D'autant plus, en se basant sur le modèle de transmission des allèles représenté dans la Figure 4-9, nous avons priorisé les mutations *de novo* qui consiste à retenir les allèles mutants chez les sujets atteints et qui ne sont pas observés chez aucun parent. Il ne nous a pas été possible de rechercher des variations *de novo* chez certains sujets (cinq) due à l'absence de données parentales.

En conclusion, les paramètres de sélection de variants ne permettent pas de faire un choix absolu d'une liste de vrais variants dans nos analyses mais plutôt de diminuer le taux de variants faussement détecté. D'autant plus, même avec des paramètres optimaux, il arrive de ne pas être en mesure de différencier les vraies parmi les fausses variations, ce qui est souvent le cas des variations localisées dans des régions hautement répétitives. Les seuls protocoles qui permettent de valider des variations génomiques sont : le séquençage Sanger, la PCR en temps réel (RQ-PCR) et la validation sur des résultats de référence standard. Jusqu'à ces jours, il n'existe pas de résultats de référence standard connus pour la validation des SNV *de novo* et rares (MAF<1%). De ce fait, il ne nous a pas été possible d'automatiser la validation de nos résultats autrement que de procéder à la visualisation individuelle des SNV. Enfin, notre meilleure solution pour diminuer les

choix de fausses variations dans notre analyse est d'effectuer une inspection visuelle du locus des variants en utilisant le logiciel de visualisation génomique IGV.

Tableau 4-2 - Analyse des ratios Ti/Tv des résultats de détection de variants par les logiciels FREEBAYES et SAMTOOLS.

Logiciel de détection de variants	Substitution de nucléotides	Données brutes	Régions codantes	MAF > 0.01	MAF ≤ 0.01	Exon, MAF > 0.01	Exon, MAF ≤ 0.01	
FREEBAYES v.0.9.10	A	C	7992	1117	6398	760	597	512
		G	36571	6625	29223	4138	3459	3126
		T	6186	684	4758	726	367	302
	C	A	12497	2095	7305	4044	1167	902
		G	11427	1967	8732	1424	1077	859
		T	46182	10048	33666	6223	5395	4496
	G	A	47065	10514	34056	6627	5556	4802
		C	11264	2061	8618	1488	1081	955
		T	12045	1810	7166	3720	1021	766
	T	A	6120	745	4711	723	385	342
		C	36288	6569	29034	4049	3528	2991
		G	8207	1155	6484	825	635	508
	transversion		75738	11634	54172	13710	6330	5146
	transition		166106	33756	125979	21037	17938	15415
Ratio Ti/Tv		2.19	2.90	2.33	1.53	2.83	3.00	
SAMTOOLS v.1.3.1	A	C	8012	1238	6335	961	641	591
		G	37098	6887	29393	4730	3508	3336
		T	5816	775	4335	934	391	369
	C	A	12174	2132	7218	3920	1127	974
		G	11542	2026	8770	1611	1072	922
		T	46393	10101	33846	6757	5317	4630
	G	A	47299	10581	34129	7180	5503	4921
		C	11593	2144	8739	1759	1090	1029
		T	11630	1789	6968	3607	959	804
	T	A	5842	856	4280	994	409	430
		C	36919	6895	29267	4691	3596	3252
		G	8184	1259	6427	977	668	581
	transversion		74793	12219	53072	14763	6357	5700
	transition		167709	34464	126635	23358	17924	16139
Ratio Ti/Tv		2.24	2.82	2.39	1.58	2.82	2.83	

Dans ce tableau, nous montrons les résultats d'analyses de ratio Ti/Tv des données de variants détectées par les logiciels SAMTOOLS et FREEBAYES. Dans la première colonne, nous indiquons le nom du logiciel de détection de variants concernées par l'analyse. Dans la 2^e et 3^e colonnes, nous indiquons les possibilités d'altérations de nucléotides, suivi du nombre total de variants détecté par le logiciel (colonne 4). Nous avons restreint la liste brute des variants à celles localisées dans des régions codantes (exons, UTR, site d'épissage alternative, etc.) et les valeurs du nombre de variants retenus sont reportées dans la

colonne 5. Nous avons ensuite dichotomisé la liste des variants en deux catégories, la colonne 6 qui représente les variants codants communs et la colonne 7 qui représente ceux qui sont rares. Nous avons effectué la même action précédente mais en restreignant les variants codants en variants localisés dans des exons. Ce dernier nous a permis de représenter dans la colonne 8, les variants communs localisés exclusivement dans des exons et dans la colonne 9, les variants rares localisés dans les exons. Au bas du tableau, nous représentons le nombre total de transitions, de transversions et leur ratio (Ti/Tv) pour chaque étape de filtrage de données (colonne 4 à 9). Le but de cet exercice est de parvenir à obtenir des données traitées de détection de variants localisés dans les régions codantes afin de comparer le calcul de ratio Ti/Tv de notre jeu de données avec celui de la littérature. Nous avons constaté que les résultats du ratio Ti/Tv de nos données (colonne 8 et 9) coïncident avec ceux attendus. Par conséquent, nos données de variants ne sont pas biaisées par des excès de variants issues de la transition ou de la transversion.

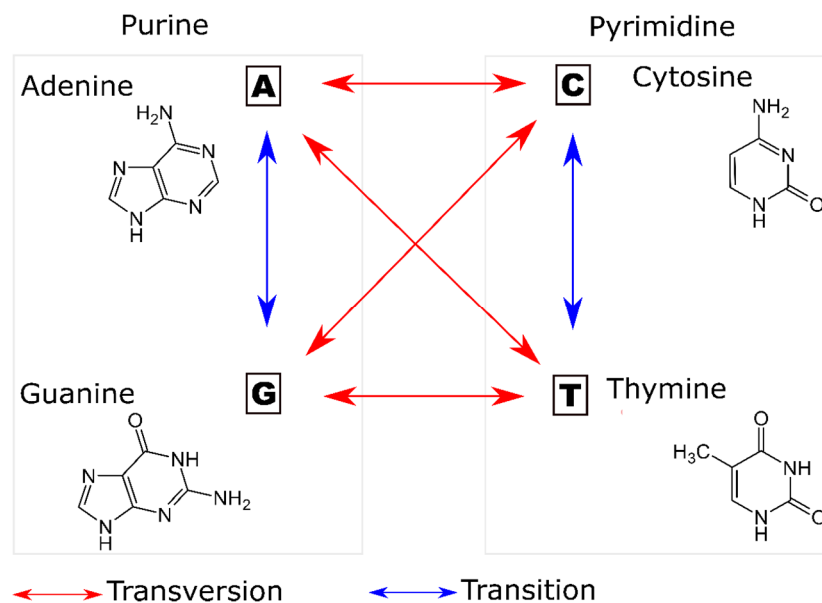


Figure 4-8 - Représentation des substitutions nucléotidiques du type transition ou traversions.

Cette figure montre un modèle des événements de transition et transversion des allèles. À gauche, nous indiquons les nucléotides de la famille des purines et à droites celles de la famille des pyrimidines. Les flèches rouges à double sens représentent les substitutions des nucléotides purines en pyrimidines et vice versa. Les flèches bleues indiquent les substitutions d'un nucléotide purine vers une autre purine, ainsi qu'une pyrimidine vers une autre.

4.2.6 - Méthodes de sélection des variations

Nous savons qu'il existe une très faible incidence d'individus atteints du syndrome PSIS dans la population canadien-française, de ce fait nous avons pensé que le profil génétique de ces derniers, basé sur des SNV et des CNV rares, permettrait de caractériser les aspects génétiques du syndrome. En conséquent, nous devons limiter nos résultats à un

nombre minimal attendu de SNV ≤ 5 mutations *de novo* par patient ou ≤ 400 variations rares délétères potentiellement transmises d'un parent. Alors, à partir des connaissances préétablies, nous avons bâti des scripts bio-informatiques qui sont capables de filtrer et trier sélectivement nos résultats de mutations à partir des contraintes bien définies. Ces contraintes sont : les modèles de transmission des variants, le niveau de partage de celles-ci, l'appartenance à des patrons d'expression génique spécifique connus, etc.

4.2.7 - Annotation fonctionnelle des variations

Nous avons annoté les variations retenues selon les méthodologies du logiciel ANNOVAR. Ce dernier est très performant et largement utilisé par la communauté scientifique actuelle. Le logiciel d'ANNOVAR permet d'annoter le génome de plusieurs organismes autres que celui de l'humain, par contre les options d'annotation disponibles demeurent majoritairement adaptées au génome de l'humain. Le logiciel annote plusieurs caractéristiques d'une variation, comme les suivantes : le type et les impacts dommageables des variations sur les régions codantes, les fréquences génétiques populationnelles, etc.

Tableau 4-3 - Paramètres de filtrage de résultats de détection des variations ponctuelles.

Localisation des variants	Exon, site d'épissage alternatif
Types de variations codantes	Non synonymes, codon stop, insertion et délétion de cadre de lecture
Fréquence d'allèles mineurs 1000Genome	≤ 0.01 ou 1%
Profondeur de lectures des variants	≥ 15
Qualité du génotype des variants	≥ 90
Niveau de partage des variants	Entre 4 à 13 sujets affectés
Types de transmission de variants	<i>de novo</i> , autosomique récessive, dominante, récessive liée à l'X, hétérozygote composée

Nous indiquons dans la première colonne, les paramètres de filtrage utilisés dans notre méthode de sélection de SNV et dans la deuxième colonne nous indiquons la valeur des paramètres.

Afin d'annoter les locus du génome, plusieurs prérequis doivent avoir déjà été mis en place pour favoriser l'annotation via ANNOVAR. Régulièrement, les données d'annotations ne sont pas disponible localement par défaut. Avant de procéder à l'analyse du jeu de données, le logiciel doit enregistrer les prérequis nécessaire aux processus d'annotation de façon permanente via le protocole de transfert ftp sur le disque local. Ces données sont : les fréquences d'allèles mineurs issues des bases de données en ligne, les scores d'impact délétère et d'haplo-insuffisance des variants, etc. Il est possible de procéder à trois types d'analyse d'annotation génomique avec ANNOVAR: l'annotation basée des gènes, des régions fonctionnelles et finalement des critères de filtrage des variants. L'annotation basée sur des gènes consiste à comprendre les effets des variations sur la fonction altérée des gènes et des transcrits de gènes codants ou non codants. D'autre part, l'annotation basée sur des régions fonctionnelles s'intéresse plutôt aux impacts des variants sur l'expression des éléments régulateurs et à leurs effets prédits sur l'expression quantitative des gènes. Finalement, l'annotation des critères de filtrage consiste à caractériser les variants en fonction des données de score d'impact délétère et de fréquences populationnelles. Les critères de filtrage sont majoritairement utilisés à des fins de stratification du jeu de données dépendamment de la maladie génétique de l'étude.

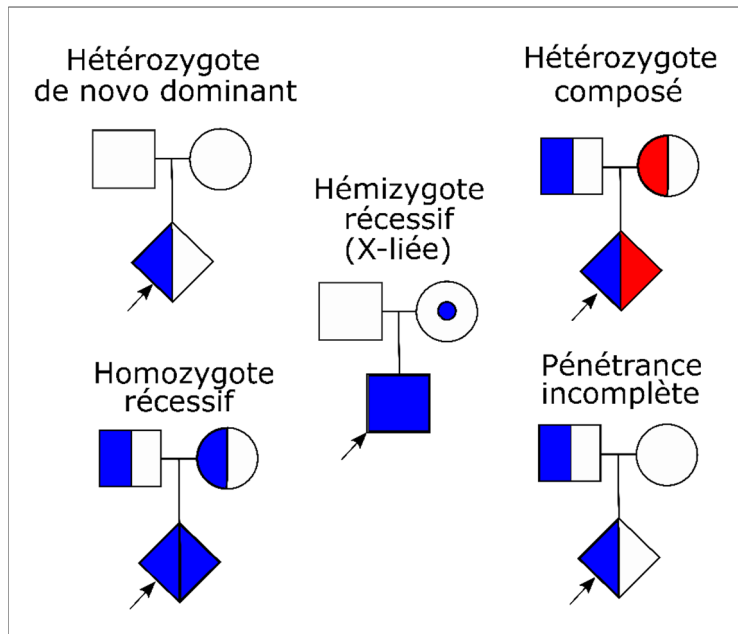


Figure 4-9 - Modèles de variations génétiques retenue dans nos protocoles de sélection de variants candidates

Cette figure représente les différents modèles de transmission d'allèle mutants dont nous avons interrogé dans nos analyses. Sur cette figure, nous constatons plusieurs formes géométriques, que sont : des carrés, des losanges et des cercles. Chacune représente un sexe différent. La forme carrée représente les hommes, le cercle représente les femmes et le losange représente un individu de sexe non déterminé. La couleur bleue et la couleur rouge représentent deux allèles mutants différents. La couleur blanche indique des allèles sauvages. Lorsqu'une forme est remplie de moitié, cela indique une transmission d'allèle hétérozygote, pendant que la forme complètement remplie indique une transmission d'allèle homozygote. Le point localisé au centre du cercle indique que seulement le gonosome de l'individu est affecté. Nous avons représenté cinq modèles de transmission de variants sur cette figure. D'abord, le modèle *de novo* dominant où seul l'enfant est affecté par une mutation hétérozygote. Ensuite, le modèle homozygote récessive où l'enfant porte la même variation sur ces deux locus parentaux. Le modèle hémizygone récessive (transmission liée au chromosome X) indique que seuls les garçons sont affectés car ils n'ont qu'une copie d'allèle parental du chromosome X. Le modèle hétérozygote composé indique que les mêmes locus parentaux sont affectés par deux types d'altérations différentes. Finalement le modèle de pénétrance incomplète indique que l'enfant et le parent portent le même allèle hétérozygote délétère, par contre ils ne partagent pas le même niveau d'affectation par le syndrome.

La majorité des bases de données utilisées par le logiciel d'ANNOVAR durant l'annotation ne sont pas installées sur le disque local. Bien entendu, seulement les utilitaires indispensables sont téléchargés, comme : les données du génome de références (format fasta), les données des séquences d'ARN (format fasta), les données de fréquence allélique spécifique à certaines populations (format vcf ou autres), etc. La majorité des

informations des données d'annotation sont obtenues via des requêtes en ligne codées dans les algorithmes du logiciel. Ces requêtes ne sont pas toutes codées de façon similaire, car chaque base de données propose leur propre API web. Les API web sont des langages de programmation qui permettent aux utilisateurs d'interagir avec les bases de données en ligne via des lignes de commande. Cette méthode est préférablement utilisée lorsque l'utilisateur nécessite de faire des lots de requêtes. Il est connu que NCBI détient une base de données riche en données d'annotations fonctionnelles du génome de plusieurs espèces. Pour interroger celle-ci, les administrateurs du site ont mis en place un API de requêtes en ligne de commande appelée « eUtils », composé de plusieurs fonctions (esearch, efetch, esummary, epost, etc.) qui permettent d'accéder aux données.

Certaines études ont constaté que plus les variations fonctionnelles codantes sont rares plus l'impact délétère sur la fonction de la protéine est sévère, ce qui augmente leurs degrés d'association aux phénotypes du désordre génétique étudié [97-98]. D'autant plus, ces variations ont souvent tendance à disparaître très rapidement dans la population car certains critères empêchent la ségrégation de celles-ci, par exemple : le taux élevé de mortalité ou une absence d'interaction sociale due à la variation. La sévérité de l'impact délétère des variants codants se définit sur une échelle allant de 0 à 1, où 0 représente un effet délétère négligeable sur la protéine et 1, l'effet le plus sévère. Nous expliquons cette sévérité par l'habilité de récupération partielle de la fonction des protéines. En d'autres termes, lorsqu'une protéine est incapable d'exécuter sa fonction altérée due à une mutation, il est attendu que le score d'impact délétère de celle-ci soit élevé.

Les scores d'impacts délétères des variants sont des données prédites par des algorithmes de prédiction en ligne, parmi lesquels nous avons fait usage (SIFT, PolyPHEN2, MutationTaster).

Afin de pouvoir mieux interpréter les résultats des scores de prédictions générés par les logiciels précédemment mentionnés, nous avons jugé qu'il est nécessaire de comprendre les détails de fonctionnement des algorithmes. D'abord, l'algorithme de prédiction de scores délétères SIFT (Sorting Tolerant From Intolerent) est un algorithme qui fonctionne par la méthodologie des pipelines, c'est-à-dire avec des analyses fragmentées en plusieurs étapes. Son approche principale se base sur l'analyse des protéines ou des polypeptides par similarité de séquences afin de classifier les substitutions des acides aminés. De plus, La méthode d'analyse de SIFT dépend principalement de la conservation des acides aminés à travers les familles de protéines au cours de l'évolution, où nous nous attendons à ce qu'un acide aminé localisé dans une région très conservée de la protéine soit très intolérant aux mutations non synonymes.

Le fonctionnement de SIFT consiste à recueillir des données fonctionnelles d'une famille de protéine en utilisant l'algorithme PSI-Blast, et de bâtir un alignement multiple à partir des séquences. Ensuite, le logiciel détermine la probabilité d'observer un acide aminé quelconque à une position de l'alignement qui est normalisée par la fréquence la plus élevée de l'acide aminé à la position concernée. Pour garantir un meilleur résultat, le model de prédiction de l'algorithme de SIFT prend en considération le niveau de conservation des acides aminés qui se détermine par la formule suivante : $\text{Log}_2 \alpha$ où α est le nombre d'observation de l'acide aminé à la position. Le score de conservation des acides aminés se mesure sur une échelle allant de 0 (où les 20 acides aminés distincts sont

observés au locus d'alignement multiple) à 4.32 (où rien qu'un seul acide aminé est présent à la position).

La performance de prédiction de l'algorithme de SIFT a été analysée selon les auteurs du logiciel SIFT et la précision de celui-ci s'évaluait à 69%. Enfin, le pipeline d'analyse de la prédiction de SIFT ne prend pas en considération la structure du repliement de la protéine, ainsi celui-ci ignore les impacts fonctionnels de cette dernière sur l'activité de la protéine.

D'autre part, la prédiction des scores délétères des variants par l'algorithme PolyPhen2 se fait sensiblement de la même manière que celui du SIFT. La méthode d'analyse de PolyPHEN2 s'effectue en trois étapes. D'abord, celui-ci recherche et analyse les annotations fonctionnelles de huit séquences de protéines homologues ainsi que leurs structures par un algorithme itératif vorace. Les alignements multiples des séquences homologues générés sont ensuite raffinés par un algorithme de classification pour ensuite être analysés par un modèle de classification bayésienne naïve [99].

Nous entendons par modèle de classification bayésienne un algorithme qui classe un évènement en fonction d'un ou des autres évènements dont la probabilité d'occurrence est connue. Supposons deux évènements X et Y où la probabilité de Y sachant celle de X est connue. Alors il est possible de déterminer la probabilité d'occurrence de X selon la loi de Bayes suivante :

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

$$P(X|Y) \cdot P(Y) = P(X \cap Y)$$

$$P(Y|X) = \frac{P(X \cap Y)}{P(X)}$$

$$P(Y|X) \cdot P(X) = P(X \cap Y)$$

Selon la loi de Bayes, la relation devient :

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)}$$

La méthode de prédiction de l'outil PolyPHEN2 permet de générer des scores probabilistes qui reflètent la distance proportionnelle entre une séquence protéine normale et celle d'une protéine altérée en se basant toujours sur des critères évolutives de celles-ci. De même que l'algorithme de prédiction SIFT, la performance de PolyPHEN2 a été analysée en utilisant deux jeux de données de variants non synonymes provenant des bases de données HumDiv et HumVar. Selon les auteurs de l'outil, celui-ci génère 20% de faux positif avec une sensibilité de 92% pour la prédiction des scores sur les jeux de données HumDiv et 73% pour celles de HumVar.

Enfin, nous avons fait usage de l'outil de prédiction des scores délétères appelé « MutationTaster ». Ce dernier est en mesure de prédire non seulement les impacts délétères des variants codants mais aussi certaines mutations régulatrices d'expression génique. De même que plusieurs prédicteurs de scores délétères, celui-ci utilise une approche de classification basée sur des modèles de prédiction bayésiennes. De plus, en réponse aux différents effets des variations délétères sur les protéines, trois types d'algorithme de classification bayésienne sont implémentés dans l'outil afin de prendre

en charge les trois types d'altérations suivantes : les substitutions d'un simple acide aminé, de plusieurs acides aminés et des substitutions synonymes et non codantes.

Les auteurs de l'outil ont procédé à des analyses de performance multiples en utilisant plusieurs jeux de données et en comparant les prédictions obtenues à celles des autres outils les plus connues. Ils ont montré que MutationTaster est en mesure de prédire l'effet délétères d'une variation codante avec une précision de 87.5% et une sensibilité de 88.7%. La performance de MutationTaster a été évaluée en se comparant à celles de SIFT, PolyPHEN2 et PROVEAN qui est une extension de SIFT. Nous constatons à partir de la figure de la courbe ROC des résultats de leurs analyses que celui-ci détient la meilleure performance de prédiction (AUC = 0.97, Figure 4-10).

Nous avons annoté les scores de prédiction délétères des mutations codantes rares en utilisant les trois algorithmes de prédiction de score délétères précédemment indiqués. Nous avons procédé ainsi afin de pouvoir comparer les résultats des trois logiciels et d'analyser la similarité approximative des valeurs des scores prédits. En se référant aux résultats de performance des outils de prédiction (Figure 4-10), il est attendu d'observer une forte similarité de scores d'effets délétères prédits par les analyses de nos jeux de données. Celui-ci n'a pas été le cas car les prédictions des effets délétères des variants ont révélé des valeurs majoritairement discordantes générées par ces outils (Figure 4-11).

En conclusion, malgré les démonstrations de la haute performance des algorithmes de prédiction d'impacts délétères de variants, il parvient que les effets délétères soient prédits différemment et proprement à un jeu de données quelconque. En d'autres termes, nous ne déconseillons pas l'utilisation de ces outils de prédictions d'effets délétères, mais plutôt d'utiliser ceux-ci avec extrême prudence.

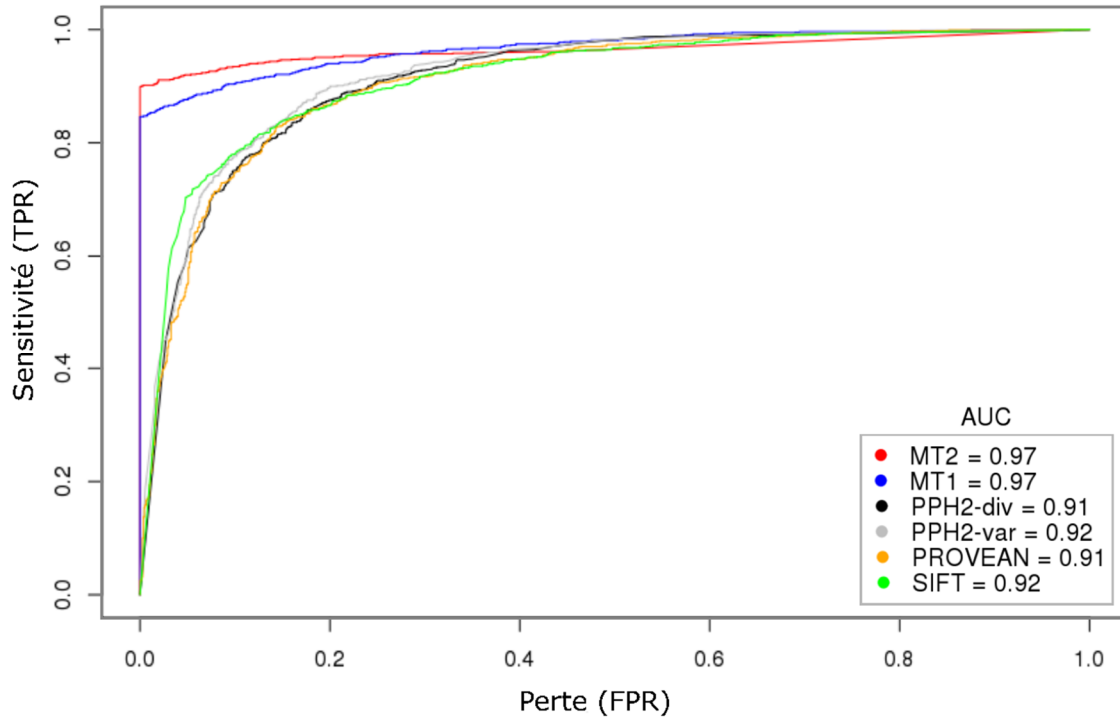


Figure 4-10 - Représentation de l'analyse de la performance comparative des algorithmes de prédiction des scores délétères des mutations codantes.

La figure 4.10 représente les résultats de la comparaison de la performance des algorithmes de prédiction suivant : MutationTaster2 (MT2, rouge), MutationTaster (MT1, bleu), PolyPHEN2-HumDiv (PPH2-div, noir), PolyPHEN2-HumVar (PPH2-var, gris), PROVEAN (orange), SIFT (vert). La performance de ces derniers a été évaluée selon le protocole de comparaison de données de la courbe ROC (Receiver Operating Characteristic). Le protocole de ce dernier consiste à représenter le taux de vraie prédiction (TPR) des outils en fonction du taux de fausse prédiction (FPR). Le score total de la performance de chaque algorithme est donné par le calcul de l'intégral en dessous de la courbe (AUC : Area Under the Curve) de la figure de celui-ci. Nous avons constaté que les résultats d'analyse de la performance de MT2 et MT1 équivalaient au score maximum d'AUC de 0.97.

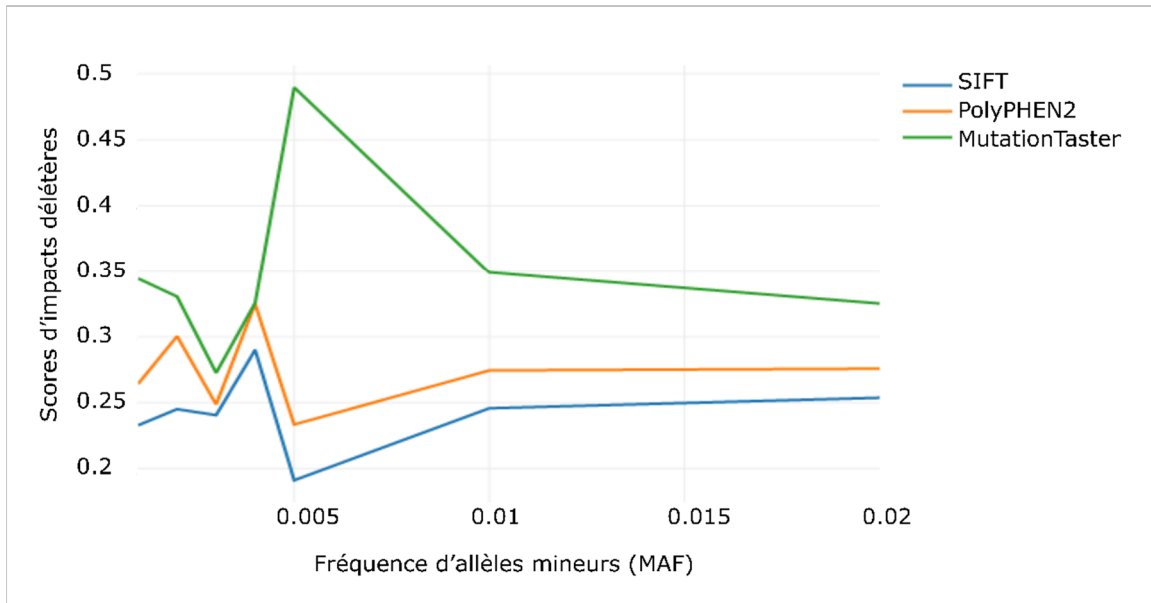


Figure 4-11 - Représentation des scores d'impacts délétères des variations codantes prédits par les algorithmes de prédiction SIFT, PolyPHEN2 et MutationTaster.

La figure 4.11 représente les valeurs des scores de prédiction des impacts délétères des variants codants rares. Nous avons défini la rareté d'une variante par leur fréquence d'allèles mineurs où celle-ci se trouve entre un intervalle de 0 à 0.02. Les variations analysées dans cet exercice sont tous localisés dans des éléments codants et celles-ci sont annotées comme des modificateurs d'acides aminés de la protéine (variation non synonyme, codon stop). Sur l'axe des ordonnées, nous avons représenté les scores de prédictions délétères de variants en fonction de leurs fréquences d'allèles mineurs que nous avons représentés sur l'axe des abscisses. Sur cette figure, la courbe en bleue représente la distribution des valeurs prédits par l'algorithme SIFT, celle de PolyPHEN2 en orange et finalement celle de MutationTaster en verte. En se basant sur des attentes théoriques, nous nous attendons à observer une distribution décroissante où les variants les plus rares ont un plus haut scores et celles moins rare, un plus faible score. Celui-ci n'a pas été le cas car nous avons observé des distributions sans aucune monotonie à travers celles-ci. Alors, nous pensons que les outils de prédictions de scores d'impacts délétères des variant ne seraient pas des outils adéquates pour prédire l'impact des variant codants dans nos jeux de données.

4.2.8 - Recherche des variations du nombre de copies

Les CNV ont déjà fait l'objet d'étude d'association à maladies endocriniennes et plusieurs marqueurs ont déjà été reportés pour être associés à la malformation congénitale chez des sujets. De ce fait, nous pensons que la détection de CNV dans notre étude pourrait nous aider à caractériser les contributions génétiques du PSIS. Nous avons

assemblé trois pipelines de détection de CNV (Figure 4-12, Tableau 4-4) afin de détecter les CNV candidats délétères associés au PSIS.

Notre protocole de détection de CNV consiste à recueillir les profondeurs de lecture des locus présents dans les données d'alignements. Afin de recueillir ces dernières, nous avons appliqué la fonction « DeepOfCoverage » de GATK qui se réfère au génome de référence humain afin de calculer la profondeur de lecture de chaque exon. Cette procédure consomme énormément de ressource computationnelle car celle-ci est basée sur un algorithme linéaire. Cette dernière améliore le temps d'exécution en tenant compte des indexes des données d'alignement préalablement définis.

Nous avons considéré ensuite trois logiciels de détection de CNV rares dans nos analyses, que sont : XHMM [100-101], FISHINGCNV [102] et CONIFER [103]. Ces trois algorithmes utilisent les données de profondeur de lecture issues des analyses de couvertures de séquences des exons afin de déterminer le niveau de variation du nombre de copie des régions du génome. Cependant, les logiciels sont caractérisés par des protocoles de détection de CNV différents les uns des autres.

D'abord, le protocole de XHMM se base sur la méthodologie des modèles cachés de Markov (HMM) pour déterminer les CNV. Ces derniers sont entraînés sur des jeux de données fournies en entrée. Durant la phase d'entraînement, les algorithmes Viterbi et Forward-Backward sont utilisés afin de générer les scores de vraisemblance des chaînes cachées de Markov (CCM) observées dans l'HMM [104]. Le calcul des scores de vraisemblance des CCM se base sur l'application de différentes méthodologies statistiques majoritairement bayésiennes. Enfin, l'HMM entraîné est utilisé afin de prédire les CNV de chaque individu. Le but d'utilisation d'un HMM est de normaliser la

profondeur de lecture des locus à travers la population analysée, ensuite exploiter la propriété de prédiction probabiliste de l'algorithme de Viterbi [105-108] et de Forward-Backward dans la prédiction de la vraisemblance des régions avec un nombre de copie variable.

Le concept des chaînes de Markov a été inventé dans les années 1902 par le mathématicien Andrei Andreievitch Markov [109], celles-ci est utilisé dans le but de construire des modèles de Markov qui est un protocole de prédiction des CCM. Ces dernières forment un system qui provient du même concept que les modèles de Markov mais la procédure de prédiction des chaînes de Markov est inconnue ou cachée à travers les séquences d'observation. Les HMM sont constituées de systèmes d'automate formés de machine à état non-déterministe. Ces systèmes sont régis par deux critères importants, qui sont : des états de transitions et des probabilités de transition d'un état à un autre [110-111].

Les HMM procèdent par des approches mathématiques pour résoudre plusieurs types de problèmes. Leurs approches consistent à prédire les probabilités de transition dans un modèle ou prédire la probabilité la plus vraisemblable des transitions lorsqu'un ensemble d'observation est fourni. Pour prédire la vraisemblance de transition des états, d'autres algorithmes statistiques sont requis par le HMM, par exemple : Viterbi, Forward-Backward, Baum-Welch [104], Segmental K-means etc.

Contrairement à XHMM, le protocole de CONIFER ne dépend pas d'algorithme HMM. Ce dernier consiste à normaliser par la médiane et par l'écart standard, la profondeur de lecture des séquences alignées à travers les échantillons analysées (RPKM : Read per kilo base per million). Le score Z de la profondeur de lecture

normalisée est ensuite transformé par une décomposition en valeurs singulières (SVD) afin de favoriser une meilleure visualisation des CNV détectés par l'algorithme. Finalement, comme XHMM, CONIFER nécessite obligatoirement la profondeur de lecture des exons en entrée afin de procéder à la détection des CNV.

Enfin, le protocole de FISHINGCNV procède généralement de la même manière que CoNIFER. Celui-ci consiste d'abord à normaliser les données de la profondeur de lecture de la population étudiée en RPKM, ensuite exclure les biais d'effet d'échantillonnage par une analyse par composante principale pour enfin détecter les CNV à partir des données RPKM, en comparant les données de chaque individu avec la distribution des données RPKM.

Cependant, il faut souligner que les données WES ne sont pas les meilleures pour effectuer des analyses de détection de CNV. Cette technique comporte beaucoup de limitations et génère beaucoup de faux résultats. Il est recommandé que des données de contrôle positif et négatif soient disponibles afin d'effectuer une analyse soigneuse. De plus, il est connu que les meilleures données à partir desquelles la recherche de CNV est mieux adaptée sont : le WGS, les données de génotypage et le CGH (Comparative genomic hybridization). Le CGH est une méthode semi quantitative, elle consiste à hybrider de l'ADN génomique avec des sondes marquées et fixées sur une micro puce à ADN et la détection des CNV se fait en fonction de la fluorescence d'hybridation de séquences observée [112].

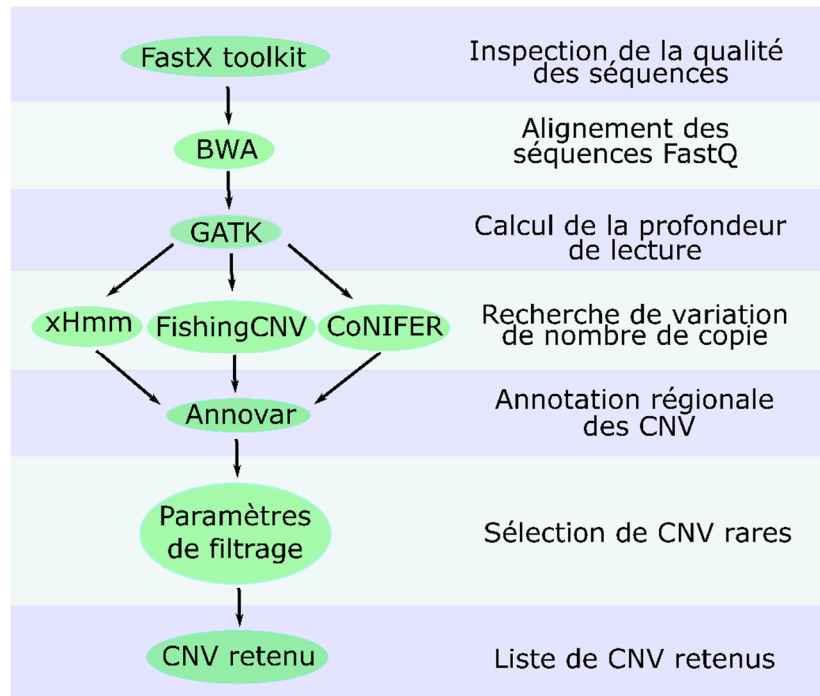


Figure 4-12 - xHmm, FISHINGCNV, CONIFER combined CNV discovery pipelines.

Nous montrons dans cette figure la séquence d’analyse des données WES afin de déterminer les variations du nombre de copies à partir de celles-ci. Nous indiquons le nom des logiciels utilisés à gauche de la figure et le but de leurs utilisations à droite de celle-ci. Nous avons aligné les séquences en utilisant le logiciel d’alignement de séquences BWA. Nous avons calculé la profondeur de couverture des régions codantes pour ensuite rechercher les variations du nombre de copies en utilisant les 3 logiciels de détection de CNV : xHmm, CONIFER et FISHINGCNV. Les résultats générés sont ensuite soumis à une étape de sélection de CNV par des méthodes de filtrage qui dépendent des paramètres spécifiques. Les CNV sont enfin soumis à la validation par Sanger PCR avant d’être retenu dans l’ensemble des résultats de CNV candidats.

Les analyses de détection de CNV à haute précision sur des données WES sont difficiles à obtenir et les résultats issus de celles-ci sont très difficiles à interpréter [113]. La difficulté d’interprétation des CNV est en partie à l’origine des biais spécifiques à la technologie de séquençage utilisée. Ces biais sont souvent reliés à l’incapacité de certaines sondes de capture d’exon à hybrider leurs cibles, ce qui a des conséquences néfastes sur la couverture de lectures [114-115]. De plus, il est difficile de comparer les CNV détectés entre différents logiciels [116], car aucune procédure computationnelle de

nos jours n'a encore permis de standardiser les protocoles de détection de CNV du WES. Nous avons constaté que les logiciels de détection des CNV performant avec une sensibilité approximative de moins de 25% dans nos analyses, ce qui concorde aux observations déjà reporté par d'autres études [114]. Ces observations nous permettent de questionner la performance des logiciels afin de réduire la proportion de CNV artefacts dans nos analyses.

Il ne serait économiquement pas avantageux de séquencer à nouveau le génome de nos sujets, simplement dans le but d'effectuer une analyse de détection de CNV plus précise via des méthodologies de séquençage alternatives. De ce fait, nous avons pensé qu'une analyse stricte des données et une méthode de sélection efficace des résultats consensus des CNV vont permettre de retenir des résultats de plus en plus vraisemblables [117].

Tableau 4-4 - Disponibilité des différents logiciels utilisés pour déroulement des trois pipelines de détection de CNV.

Outils	Adresse web de téléchargement
Fastx ToolKit v0.0.14	https://github.com/agordon/fastx_toolkit/releases/download/0.0.14/fastx_toolkit-0.0.14.tar.bz2
BWA v0.7.15	https://sourceforge.net/projects/bio-bwa/files/bwa-0.7.15.tar.bz2/download
GATK v3.7	https://software.broadinstitute.org/gatk/download/
XHMM	http://atgu.mgh.harvard.edu/xhmm/tutorial.shtml
FISHINGCNV v1.5.3	https://sourceforge.net/projects/fishingcnv/
CONIFER v0.2.2	http://conifer.sourceforge.net/download.html

4.2.9 - Expression spécifique de gènes

Nous avons considéré certains gènes spécifiquement associés au développement embryonnaire de l'hypophyse et nous avons analysé l'association de l'altération fonctionnelle de ceux-ci avec les variations non-synonymes. Pour y parvenir, nous avons récupéré, de la base de données PaGenBase (<http://bioinf.xmu.edu.cn/PaGenBase/>), une liste de gènes potentiellement exprimés spécifiquement dans l'embryogenèse de l'hypophyse et nous avons restreint notre analyse sur l'ensemble des gènes figurant dans celles-ci. À partir des analyses computationnelles proprement implémentées pour ce projet, nous avons recherché toutes les variations délétères localisées dans la liste des gènes présélectionnés issues de la base de données de PaGenBase et nous y avons appliqué des critères de sélection, telles qu'indiquées au Tableau 4-3. Les gènes retenus sont ensuite manuellement vérifiés et seront biochimiquement validés selon les protocoles de séquençage Sanger.

En conclusion, nous avons constaté dans ce chapitre que notre défi a été d'appliquer un ensemble de paramètres de référence absolue, défini par la communauté scientifique, qui permet de générer des résultats de détections de variation optimaux. Malheureusement, il n'a pas été possible de compiler un tel ensemble de paramètres, par conséquent, nous avons utilisé certains paramètres de filtres de données et validé biochimiquement les variations afin de garantir la qualité des résultats générés.

Chapitre 5 - Résultats

5.1 - Analyses et sélection des mutations

Je rappelle que le but de notre étude a été d'identifier des variations délétères rares et *de novo* qui n'ont jamais été identifiées par d'autres études et qui sont nécessairement associées au syndrome du PSIS. Nous nous concentrons sur la recherche des variations qui se localisent dans les exons, les UTR et les variations du site d'épissage. Ces dernières pourraient altérer négativement la fonction normale des gènes, néanmoins, les variations dans les introns, les régions intergéniques et les régions régulatrices non codantes ont aussi la propriété d'altérer la fonction des gènes dont elles régulent. Par contre, nous n'avons pas priorisé ces variations car nous n'avons pas établi les protocoles nécessaires qui les considèrent dans nos analyses. Pour y arriver, nous avons analysé les données WES des 13 familles simplex constituées d'un sujet affectés par famille. Nous avons construit des pipelines de détection de variants ponctuelles et de variations de nombre de copies afin d'analyser et de détecter, par des approches computationnelles, les meilleurs candidats de variations contenues dans nos données. Finalement, nous avons mis l'emphase sur la recherche des variations candidates qui sont partagées entre les sujets affectés afin de mieux caractériser les causes génétiques associées au phénotype de la triade classique.

5.1.1 - Mutations partagées à travers les patients

Nous avons recherché les mutations partagées entre les sujets de l'étude afin de vérifier la présence de celles délétères transmises par des effets fondateurs. Nous avons appliqué les critères de filtrage des mutations tel qu'indiquées au Tableau 4-3 de la section 4.2.4 et nous avons déterminé les divers niveaux de partage de mutation entre

les sujets. Nous n'avons malheureusement pas réussi à observer aucunes mutations délétères majeures dans notre analyse (Figure 5-1). Seulement un marqueur est observé avec la fréquence maximale de 61% chez les individus affectés. Cette observation pourrait être due soit par une absence des mutations candidates ou une hétérogénéité génétique très élevée du syndrome étudié.

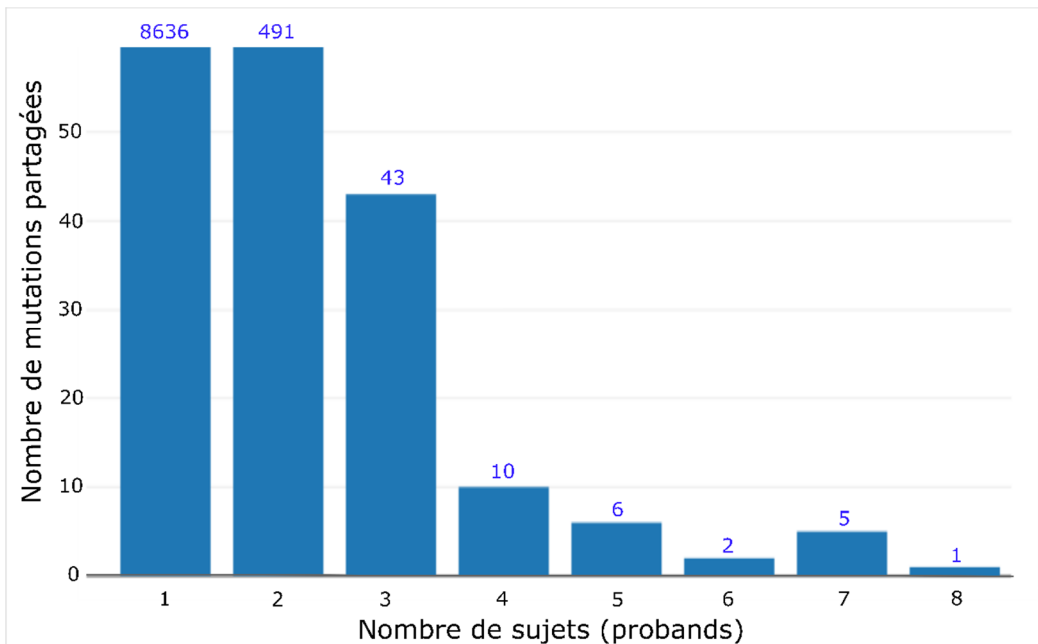


Figure 5-1 - Analyse de partage de mutations rares délétères entre les sujets affectés.

Cette figure représente les résultats d'analyse de mutations rares délétères partagées entre les sujets affectés de notre étude. A partir d'un consensus de mutations détectées par les trois outils de détection de variants génomiques, nous avons appliqué les paramètres de sélection des mutations du Tableau 4-3 afin de retenir un jeu de données de variations délétères rares. Nous avons analysé celui-ci en recherchant les mutations qui sont partagées par les sujets affectés. Sur l'axe des abscisses, nous représentons le nombre d'individu, puis en ordonnée, nous représentons le nombre de mutations partagées entre les sujets. De ce fait, nous constatons que pour 8636 mutations rare délétères, au plus un seul patient porte celles-ci. Pendant qu'au plus d'une seule variation est portée par au plus huit sujets affectés.

5.1.2 - Analyse des mutations récessives rares

Nous avons ensuite recherché des mutations récessives rares dans nos résultats de SNV sélectionnés selon les critères indiquées au Tableau 4-3. Ce dernier consiste à

retenir des sujets qui portent des variants homozygotes pour lesquels les parents sont hétérozygotes. La difficulté de cette démarche est de garantir la performance d'exécution de la méthode sur un si gros jeu de données. Pour y arriver, nous avons appliqué des méthodes computationnelles basées sur des algorithmes de table de hachage. Ces derniers nous ont permis de lire, de trier et de fournir les résultats dans un délai raisonnable. Nos scripts combinent les résultats des SNV rares retenus et sélectionnent ceux qui sont homozygotes, rares et non synonymes et qui ne sont observés qu'à l'état hétérozygote chez les parents. Nous avons mis l'emphase sur des gènes mutants dont les fonctions spécifiques sont associées au développement embryonnaire de l'hypophyse. Nous avons retenu une liste de variants ayant satisfait les contraintes de sélection, mais qui nécessite de la validation via la méthode de séquençage Sanger (Tableau 5-1). Par contre, cette dernière n'a été effectuée que pour les gènes RBM28 et FAM154B.

5.1.3 - Recherche des mutations *de novo*, non-sens et dominantes

Nous avons appliqué les mêmes approches algorithmiques que dans la section 5.1.2 pour rechercher les variants *de novo* délétères. Notre but consiste à déterminer les variants délétères qui ne sont pas portés par les parents des individus atteints. Nous avons réussi à générer une liste de variations candidates qui ont d'abord passées les critères de sélection, ensuite, respectées les contraintes du modèle de transmission génétique recherché. Nous avons retenu une fraction de la liste des mutations générées pour lesquelles l'inspection visuelle et la validation par la méthode de séquençage Sanger ont été effectuées (Tableau 5-2, Tableau 5-3). Nous avons constaté que le nombre de variants était particulièrement élevé par rapport au nombre de variants attendus. Cette

dernière pourrait représenter un enjeu économique important pour vérification des résultats via le séquençage Sanger PCR.

Tableau 5-1 - Mutation récessives rares retenues selon nos critères de sélections indiquées au Tableau 4-3.

Chromosome	Position	SNV	Gènes	Identifiant des SNV (snp132)	Fréquence d'allèles mineurs
1	5969225	G>C	NPHP4		
1	158624487	C>A	SPTA1		
2	209308236	G>T	PTH2R		0.0009
3	149048185	G>T	TM4SF18		0.0005
4	57687770	C>A	SPINK2		
4	57843607	C>G	NOA1	rs114787590	0.01
5	465305	G>A	EXOC3		
6	31083945	C>T	PSORS1C1		0.0014
6	43413359	C>G	ABCC10		
7	127953238	G>A	RBM28	rs73230638	0.0014
8	143413129	C>T	TSNARE1		
8	144941474	G>A	EPPK1		
8	145580182	G>C	FBXL6		
8	145583505	C>A	SLC52A2	rs117500243	0.0023
9	127684074	T>C	GOLGA1	rs35237091	0.02
11	94731606	A>G	KDM4D	rs76057256	0.0037
11	108032676	G>A	NPAT	rs117925274	0.01
11	108277861	C>T	C11orf65		0.01
11	121444958	A>G	SORL1	rs62617129	0.0023
11	124624147	G>C	ESAM		
11	64074743	C>T	ESRRA	rs117285599	0.0032
11	94759020	G>A	KDM4E	rs2020210	
12	133682839	A>C	ZNF140		
15	83519982	T>C	HOMER2	rs79448007	0.0037
15	82575282	C>T	FAM154B		
16	4649353	G>A	C16orf96		
16	82033600	C>T	SDR42E1		
16	82033626	T>C	SDR42E1	rs79313103	0.0041
16	85120287	G>A	KIAA0513		
19	53209064	C>T	ZNF611		0.0005
19	13993709	C>T	C19orf57	rs111386677	0.0023
22	18300470	G>T	MICAL3	rs45518631	0.01

Nous rapportons, dans le tableau ci-dessus, les résultats de la sélection des variants homozygotes rares localisées dans les données des individus affectés. Les colonnes 1 et 2 représentent les coordonnées génomiques des variants, la colonne 3 représente la substitution des allèles, par exemple, la substitution d'une adénine vers une guanine est représentée comme suivante (A>G). Dans la quatrième colonne, nous

représentons la version « RefSeq » du nom des gènes du locus mutants. Dans la cinquième colonne, nous indiquons les identifiants des SNV tel qu'ils sont reportés dans la version 138 de la base de données de dbSNP. Finalement, dans la sixième colonne, nous indiquons la fréquence d'allèles mineurs des variants reportés dans la base de données 1000Genome (la version du mois d'avril 2015).

Tableau 5-2 - État de la vérification des mutations délétères *de novo* par la méthode de séquençage Sanger PCR.

Chromosome	Position	SNV	Gènes	Validation
1	240370952	C>T	FMN2	-
3	41277315	C>G	CTNNB1	-
4	76711837	G>A	USO1	-
7	12269397	G>C	TMEM106B	-
11	134009766	A>C	JAM3	-
11	67219634	T>C	GPR152	+
13	41885681	T>C	NAA16	-
15	40268998	G>C	EIF2AK4	-
16	55362984	G>A	IRX6	+
22	46931308	A>C	CELSR1	-

Dans ce tableau nous rapportons l'état de vérification des variants *de novo* délétères par la méthode de séquençage Sanger. Ces variants ont été retenus selon les critères de sélection indiqués au Tableau 4-3. Dans la première et la deuxième colonne, nous représentons les coordonnées génomiques des variants. Dans la troisième colonne, nous indiquons la variation des allèles parentaux. Dans la quatrième colonne, nous représentons la version « RefSeq » du nom des gènes du locus mutants, dans la cinquième colonne, nous indiquons l'état de la vérification de la mutation par séquençage Sanger. Nous indiquons par plus (+), les SNV qui ont été positivement confirmés par séquençage Sanger et le contraire par un moins (-).

5.1.4 - Mutations hemizygotiques délétères liées au chromosome X

Certains marqueurs pathogéniques ont déjà été rapportés d'être transmis via les gonosomes maternelles, comme l'agammaglobulinémie [118] (Xp22, MIM#300310) ou le retard mental [119] (Xp22.3, MIM#300428). Celui-ci consiste au fait que la mère transmet son chromosome X à son enfant masculin où celui-ci est porteur d'au moins une mutation délétère. Ces variants sont du type hemizygotique et elles sont spécifiquement transmises via le chromosome X. L'importance de l'impact délétère d'une telle mutation s'explique par l'incapacité du locus mutant à récupérer la fonction de ses gènes car le sujet ne détient qu'un seul allèle. Afin de rechercher ces types de variations dans nos

résultats, nous avons d'abord restreint notre recherche sur le chromosome X, puis appliqué nos critères de sélection de variants indiqués au Tableau 4-3 et retenir seulement les variants rares statuéés en tant qu'homozygotes. De plus, nous avons requis que seules les mères doivent être porteuses d'allèles délétères.

Tableau 5-3 - Annotation de la pathogénicité et du score d'effets délétères des SNV novo.

Chr	Position	SNV	Gènes	OMIM phénotype	MIM#	Prédiction des scores SIFT	Prédiction d'impact SIFT
1	240370952	C>T	FMN2	Retard mental, autosomique récessive 47	616193	0.002	Délétères
3	41277315	C>G	CTNNB1	Retard mental, Médulloblastome	615075, 155255	0.81	Bénigne
4	76711837	G>A	USO1	Immunodéficience 49	617237	1	Bénigne
7	12269397	G>C	TMEM106B	dégénération du lobe front temporal, Aphasie	607485	0.084	Bénigne
9	96326681	G>C	FAM120A			0.137	Bénigne
10	75006782	T>G	DNAJC9			0.003	Délétères
11	134009766	A>C	JAM3	Destruction hémorragique du cerveau, syndrome de Jacobsen	613730, 147791	0.54	Bénigne
11	67219634	T>C	GPR152			0.044	Délétères
12	7045939	T>G	ATN1	Néoplasme ovarienne secondaire maligne	613730, 147791	0.104	Bénigne
13	41885681	T>C	NAA16			0	Délétères
15	40268998	G>C	EIF2AK4	Maladie veno-occlusive pulmonaire	234810, 265450	1	Bénigne
16	89346906	T>G	ANKRD11	syndrome de KBG, retard psychomoteur, caractéristiques distinctes du visage, syndrome de Williams-Beuren	148050, 616728, 194050	0.055	Bénigne
16	55362984	G>A	IRX6			0.21	Bénigne
19	879988	C>T	MED16			1	Bénigne
20	61443713	T>G	OGFR			0	Délétères
22	46931308	A>C	CELSR1	Défauts du tube neural, syndrome de régression caudale	182940, 600145	0.001	Délétères

Dans ce tableau, nous rapportons les résultats d'annotation des scores d'effets délétères des variants prédits par SIFT et des phénotypes OMIM associés à celle-ci. Les variants retenus sont celles qui ont passées les

critères de filtrage spécifiées au Tableau 4-3 et qui ont été manuellement vérifiées dans nos analyses. Les colonnes 1 et 2 représentent les coordonnées génomiques des variants, la colonne 3 représente la substitution des allèles. La colonne 4 indique le nom des gènes du locus mutants selon la base de données de « RefSeq ». Nous avons rapporté dans la colonne 5 le nom des maladies associées aux variants, telles qu'indiquées dans la base de données d'OMIM, et qui sont suivis de leurs identifiants rapportés dans la colonne 6. Dans la colonne 7, Nous indiquons les scores de prédiction d'effet délétères des allèles générés par SIFT pendant que la dernière colonne indique la prédiction d'effet dommageable des variants sur les gènes.

Nous avons réussi à localiser plusieurs variants qui nécessiteraient certaines vérifications moléculaires (Tableau 5-4), mais nous avons retenu particulièrement celles qui affectent les gènes IGSF1 et GPC4 car ces derniers seraient spécifiquement associés au développement embryonnaire de l'hypophyse.

Tableau 5-4 - Candidats de mutations récessives rares, non-sens liées à l'X.

Chromosome X				
Position	Mutations	Gènes	Identifiant de SNV	Fréquence d'allèles mineurs
130417091	T>C	IGSF1	s/o	s/o
20060681	G>C	MAP7D2	s/o	s/o
26235846	A>T	MAGEB5	s/o	s/o
47444408	G>A	SYN1 TIMP1	s/o	s/o
65417554	C>T	HEPH	s/o	0.0012
135308084	C>T	MAP7D3	s/o	0.0006
152612801	G>A	ZNF275	rs61999320	0.0024
7175525	G>A	STS	s/o	0.0006
13612963	G>A	EGFL6	rs16979010	0.01
132438817	C>T	GPC4	s/o	s/o
49047968	C>A	SYP	s/o	s/o
49079179	T>G	CACNA1F	s/o	0.0048
147026497	G>A	FMR1	s/o	s/o
69749000	C>T	TEX11	s/o	s/o
118377123	A>G	PGRMC1	s/o	s/o
7194089	G>A	STS	rs113416414	0.01
34150257	T>A	FAM47A	s/o	s/o

Les champs indiqués en gris représentent les gènes que nous avons sélectionnés à cause de leur association connue et spécifique au développement embryonnaire de la glande hypophysaire (selon la base de données de BioGPS et PAGENBASE).

5.1.5 - Comparaison de résultats de SNV rares

Nous avons comparé les résultats de détection de variants que nous avons représentés par des diagrammes de Venn (Figure 5-2, Tableau S2). Nous avons procédé ainsi afin d'analyser la proportion discordante de variants détectées entre les outils de détection appliqués dans ce projet. En conséquent, les résultats de celui-ci ont montré que les logiciels GATK, SAMTOOLS et FREEBAYES génèrent respectivement 38%, 89% et 89 % de l'ensemble des variants détectées. De plus, nous avons remarqué le protocole de détection de variants de GATK génère exclusivement près de la moitié (50.7%) des variants. Ce dernier nous a permis de constater que les critères de sélection de variants de celui-ci étaient moins sévères que celles de SAMTOOLS et de FREEBAYES.

D'autre part, nous pensons que les erreurs de détections de variants sont mieux contrôlées par le protocole de GATK, celui-ci inclut les variants dont la MAF est inférieure ou égale à 1% dans notre jeu de données de control positif locale. Les variants qui constituent cette dernière sont issues d'ADN génomiques qui ont été séquencés selon le même protocole de séquençage d'Illumina que celui appliqué dans ce projet. De ce fait, les variations dont la MAF ne corrèle pas à celui de la base de données de control local sont automatiquement exclues car elle représente un risque de fausse détection. Nous n'avons pas pu renforcer la précision du protocole de détection de variants de FREEBAYES et SAMTOOLS car nous ne détenions pas de contrôles positifs locaux. Les contrôles positifs locaux nous aident à détecter et éliminer des fausses détections de variants spécifiques à un protocole.

Les résultats de détection de variants ont été restreints exclusivement aux SNV, et ceci, dans le but de garder l'homogénéité des données. Toujours dans le même but, nous avons appliqué des paramètres par défaut aux analyses de chaque pipeline, plusieurs types d'analyse de comparaisons (scénarios) ont été considérés afin de vérifier la stabilité du ratio de partage de variants, ce sont : tous les SNV, ceux ayant un identifiant dbSNP, les SNV rares (MAF \leq 1%) et les SNV communs (MAF $>$ 1%). De plus, afin d'effectuer une comparaison sur une base équitable d'évaluation des données, les variants de chaque pipeline ont été évalués sur une même base d'intervalle génomique.

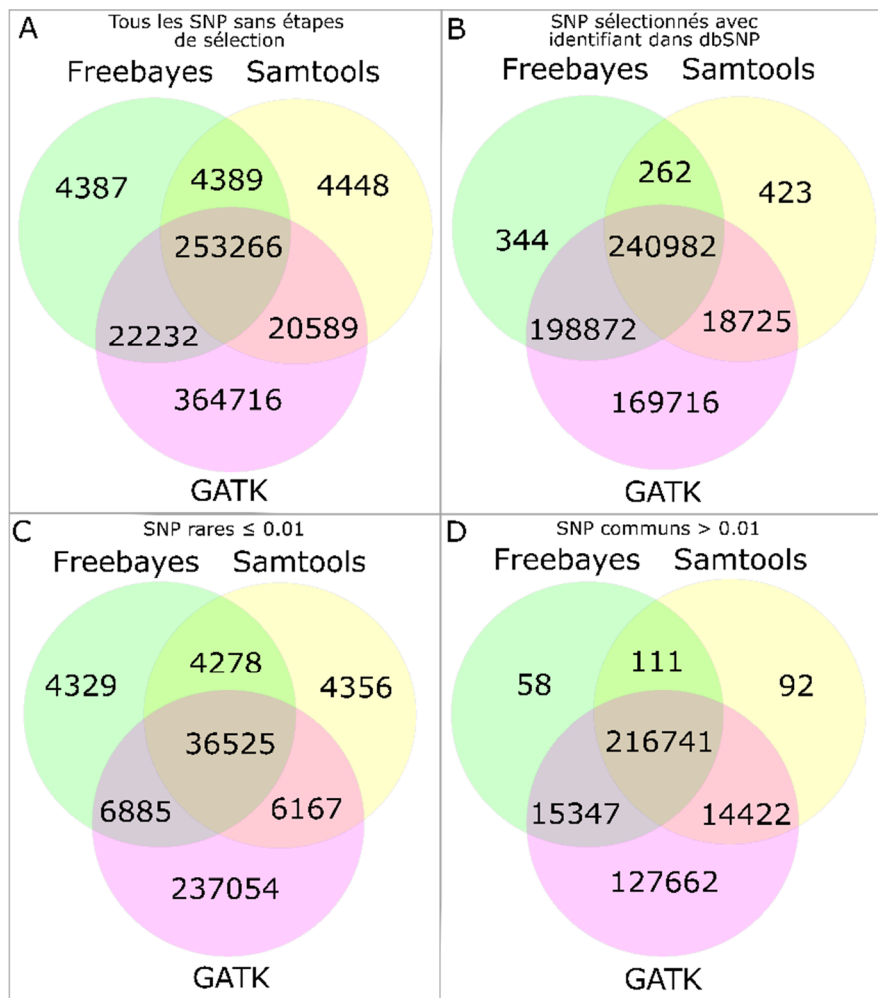


Figure 5-2 - Représentation du diagramme de Venn des SNV générés par les protocoles de détection de variants génomiques de GATK, SAMTOOLS et FREEBAYES.

Dans cette figure, nous représentons par un diagramme de Venn les résultats de l'analyse de comparaison de variants ponctuels générées par les outils GATK, SAMTOOLS et FREEBAYES. A montre le diagramme de Venn de la comparaison des résultats de tous les SNV indépendamment de leur MAF. B nous représentons les résultats de la comparaison des variants qui ont déjà reportés dans la base de données dbSNP (version snp138), C montre la comparaison des SNV rares ($MAF \leq 1\%$) et enfin montre le diagramme de Venn de la comparaison des résultats de SNV commun. La fréquence d'allèles mineurs des SNV est obtenue à partir de la base de données 1000Genomes. Nous avons comparé les résultats de SNV afin de pouvoir analyser, sous divers contraintes, les proportions de celui-ci partagées entre les outils de détection de variants. Nous avons constaté que SAMTOOLS et FREEBAYES génèrent moins de variants spécifiques ($\leq 10\%$) pendant que GATK génère au-delà de 86%. De plus, les résultats nous montrent que SAMTOOLS et FREEBAYES génèrent plus de 90% des variants partagés entre les trois outils pendant que GATK génère au-delà de 62% (Tableau S3).

5.1.6 - Vérification des INDEL avec les données de dbSNP

Les études d'O'Rawe et al. 2013 [116] ont montré qu'à travers la majorité des protocoles de détection des variants connus, il existe une faible concordance entre les variants du type INDEL détectés, c'est-à-dire une concordance de 43.3% entre les INDEL présents dans dbSNP et 4.7% entre les nouveaux INDEL. Afin d'étudier cette variabilité dans nos jeux de données, nous avons vérifié la présence des INDEL générés par nos analyses dans la base de données de dbSNP (version 138). Nous avons constaté que nos analyses génèrent une proportion insignifiante d'INDEL reportée dans cette dernière. Les résultats de vérification des INDEL dans la base de données dbSNP sont reportés sur la Figure 5-3. Nos résultats nous ont permis de constater que chaque outil de détection de variants génère moins d'1% de l'ensemble des INDEL disponibles dans dbSNP. Ces observations nous indiquent que les résultats de détection d'INDEL générés dans nos analyses, doivent être interprétés avec beaucoup réserves et de vérifications approfondies (Figure 5-3).

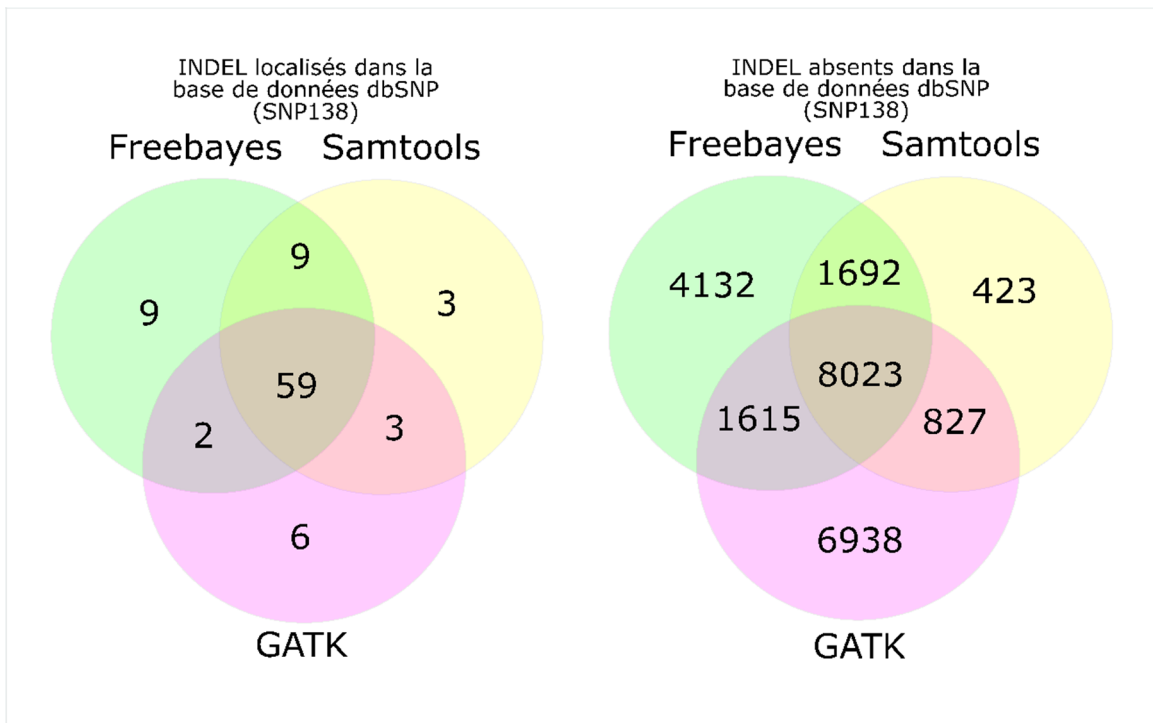


Figure 5-3 - Représentation par diagramme de Venn de la comparaison des résultats d'INDEL présents et absents dans la base de données de dbSNP (snp138).

dbSNP est une base de données qui rapporte des annotation génomiques et fonctionnelles des SNV qui ne sont pas nécessairement observé dans des études cliniques.

Nous représentons dans cette figure, les résultats des variants du type INDEL détectées par les trois outils de détections de variants appliquées dans nos protocoles. Nous avons d'abord comparé les résultats générés entre les outils afin d'analyser les proportions d'INDEL consensus retenues. Nous avons constaté qu'au-delà de 80% des INDEL sont partagés entre les outils de détection de variants, lorsqu'ils sont reportés dans la base de données dbSNP138 (l'image à gauche). Pour les INDEL qui sont absent dans dbSNP138, nos résultats indiquent qu'à peu près 50% d'INDEL consensus est observé dans les résultats de SAMTOOLS et de GATK et que 73% de ceux-ci est observé dans les données de FREEBAYES (l'image à droite).

5.1.7 - Variations du nombre de copies

Les études de Pfundt, R. et al. 2016 [120] ont montré que les CNV jouaient un rôle important dans l'attribution des diagnostics cliniques. Grace aux détections de CNV par la méthode de génotypage sur des puces à ADN, il est possible de diagnostiquer entre ~10% à ~20% de patients [121-122]. La conséquence des CNV peut avoir des

effets néfastes sur la fonction partielle ou intégrale d'un ou de plusieurs gènes simultanément. Selon les types de régions génomiques d'intérêt, l'analyse de ces variations peut présenter différentes complexités. Il est déconseillé de détecter des CNV avec des données WES car le protocole est limité par l'absence des informations des régions non codantes et la couverture biaisée des exons [114-115, 123]. Par contre, en limitant les résultats à ceux qui sont issus de la détection des CNV par plusieurs outils, il est possible d'améliorer la précision du protocole. Nous avons analysé les données de WES selon nos protocoles de détection de CNV détaillés dans la section 4.2.8 (page 71, Figure 4-12), ensuite, nous avons comparé les résultats de chaque outil de détection de CNV afin de mettre en évidence les résultats consensus de CNV détectés entre les outils. Au travers des résultats issus de chaque outil, nous avons recherché les CNV dont le pourcentage d'intervalle génomique commun soit identique à 90% et plus. De plus, nous avons retenu des CNV ayant des effets délétères sur les locus affectés et qui sont à caractère homozygote ou hétérozygote. Nos analyses nous ont permis de retenir certains CNV qui ont satisfaits les critères de sélection (Tableau 5-5). Parmi ces derniers, nous avons retenu une délétion homozygote du gène *MGAM* (Figure 5-4) qui se localise sur le bras q34 du chromosome 7 entre les intervalles génomiques : 141761987-141796113. Les CNV retenues seront vérifiées par la méthode de PCR en temps réel afin de confirmer nos observations.

En conclusion, nous avons constaté que nos protocoles de détection de CNV montrent que plus de 50% des CNV détectés dans nos résultats sont *de novo*. De plus, en considérant la faible probabilité d'occurrence d'un tel évènement, nous pensons que

certaines précautions sont à considérer afin de diminuer la proportion de faux CNV *de novo* dans nos résultats.

Tableau 5-5 - CNV *de novo* retenus après les procédures de filtrage et de sélection.

Locus (Chromosome:début-fin)	Taille (kbp)	Gènes atteints	Type
1 :17017083-17274004	256.92	ESPNP, MST1L, MIR3675, CROCC	Hétérozygote
6 :32604561-32633280	28.72	HLA-DQB1, HLA- DQA1	Hétérozygote
7 :141761987-141796113	34.13	MGAM	Homozygote
15 :22318389-22490518	172.13	OR4N4, OR4N3P, LOC727924, OR4M2	Hétérozygote
17 :36351287-36385555	34.27	LOC440434	Hétérozygote

Nous représentons dans ce tableau les résultats de CNV *de novo* détectés dans nos analyses. Dans la première colonne, nous représentons les coordonnées génomiques affectées par la variation du nombre de copies. Les coordonnées génomiques sont représentées sous le format de chromosome, position de début suivie de la position de fin. Nous rapportons dans la deuxième colonne, la taille des CNV où l'intervalle des régions non codantes n'y est pas compris. Cette dernière est rapportée sous une mesure de kilo pair de base. Nous rapportons dans la troisième colonne, l'ensemble des gènes qui sont atteints par le CNV. Finalement, la dernière colonne indique les types de CNV observé, que sont : hétérozygote lorsqu'une seule copie de l'allèle est affectée et homozygote lorsque deux copies de celui-ci sont affectées.

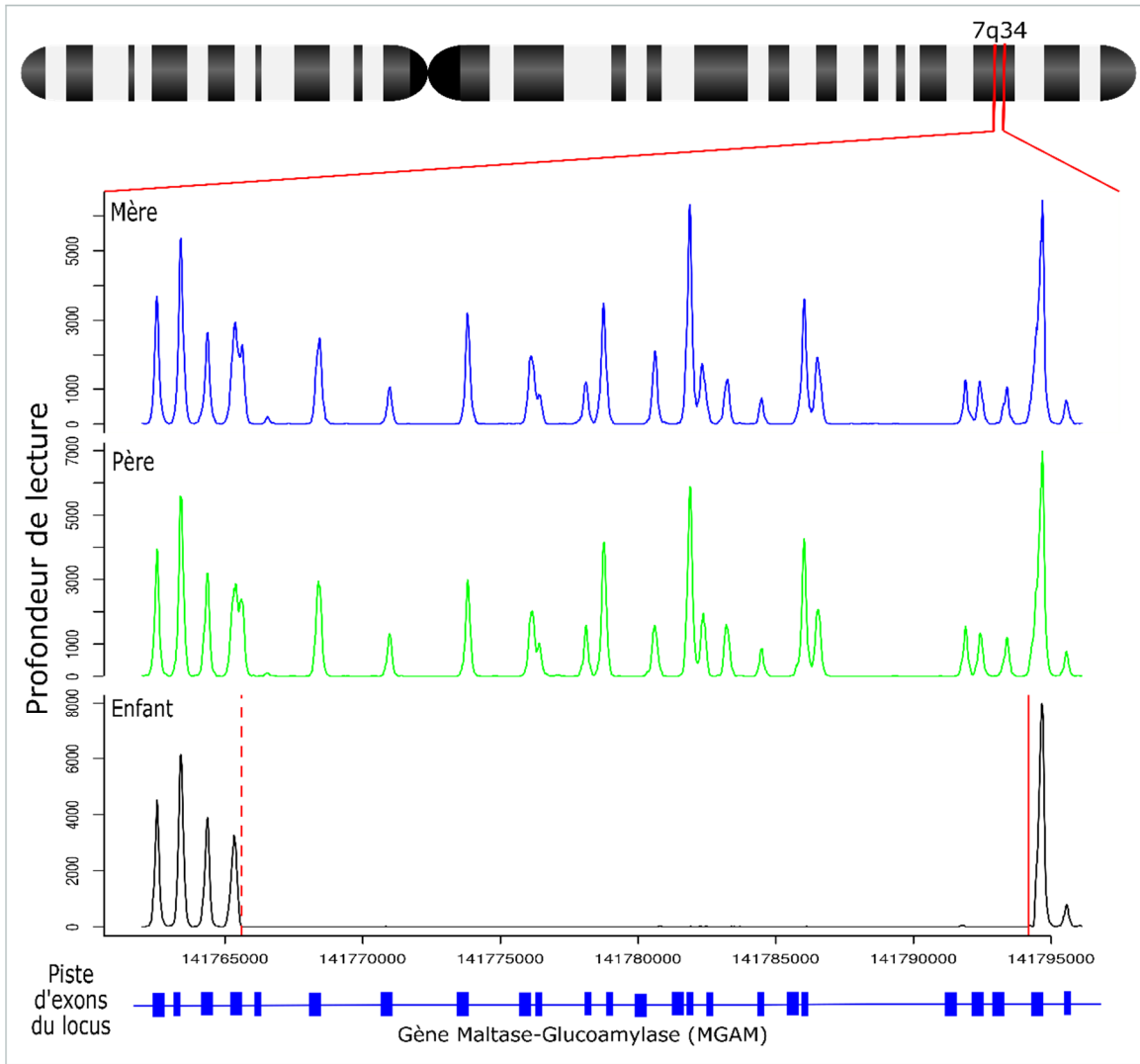


Figure 5-4 - Représentation graphique de la délétion du nombre de copie homozygote du gène MGAM indiquée au tableau 5.6.

Dans cette figure, nous montrons la délétion homozygote de copie d'allèles parentaux du gène MGAM. Dans la partie supérieure de la figure, nous indiquons la région cytogénétique où se localise le CNV. Ensuite, nous représentons respectivement, la distribution de la profondeur de lecture du locus affecté de la mère, du père et de l'enfant. Sur l'axe des abscisses, nous indiquons la profondeur de lectures brutes des locus et sur l'axe des ordonnées, nous indiquons les coordonnées génomiques des locus en pair de base. La représentation comparative des trois images de la figure est de nature quantitative et ne sert qu'à comparer le niveau de profondeur de lecture entre les données du trio. Finalement, au niveau inférieur de la figure, la piste codante du locus affectés du gène *MGAM*. Sur celle-ci, nous indiquons les introns par des lignes et les exons par des boites. Nous avons constaté que dans les données WES de l'enfant atteint, la couverture de lecture du locus représenté sur la figure est complètement absente. Nous avons indiqué ce dernier par un intervalle délimité par deux lignes verticales rouges.

Chapitre 6 - Discussion et conclusion

6.1 - Méthodologie et mutations partagées

Dans cette étude, notre but principal a été de déterminer les marqueurs génétiques qui sont partagés entre les patients afin de caractériser l'aspect génétique du syndrome. D'ailleurs, d'autres membres de l'équipe ont déjà tenté de découvrir les variants génétiques ponctuels partagés dans notre cohorte. Tel que nous l'avons précédemment indiqué dans la section 4.2.5, leurs approches de détection de variants s'opéraient selon le protocole de l'outil GATK. Ils n'ont malheureusement pas réussi à déterminer aucune mutation ponctuelle majeure qui affecte les fonctions des gènes associés aux développements embryonnaire de l'hypophyse. En conséquent, nous avons eu pour objectif d'analyser à nouveau les données du WES en utilisant d'autres approches bio-informatiques dans le but de retrouver, si possible, les variations génétiques qui ont pu systématiquement échapper à la précédente analyse. L'hétérogénéité génétique du PSIS n'est malheureusement pas un concept bien étudié dans la littérature. Par conséquent, nous pensons que celui-ci serait partiellement responsable de l'échec de détection de variants ponctuels majeures dans nos analyses.

Cependant, nos résultats nous ont clairement révélé que les trois approches de détection de SNV, appliqués dans cette étude, génèrent des résultats qui sont spécifiques à chacun. De ce fait, nous avons pensé qu'en combinant les résultats de toutes les approches, nous serions en mesure de déterminer certaines variations ponctuelles majeures causales. Alors, nous avons combiné les résultats de SNV détectés dans nos analyses, ceux-ci nous ont permis non seulement de confirmer certaines discordances entre les résultats générés par les trois protocoles, mais aussi de générer de nouveaux

résultats permettant de renforcer nos recherches de SNV causales majeurs dans nos jeux de données. Étonnamment, nous n'avons pas réussi à détecter de variations ponctuelles majeures dans nos analyses car aucune d'elle n'a été partagée entre les 13 individus affectés de l'étude (Figure 5-1).

Malgré qu'aucune mutation ponctuelle majeure n'ait pu être détectée dans nos analyses, nous avons pensé qu'il serait important d'analyser les SNV délétères individuellement détectés chez les sujets atteints. En procédant de la sorte, nous avons pensé que les SNV délétères candidats retenus pourraient partiellement aider à la caractérisation de certains aspects génétiques du syndrome.

En conclusion, notre échec d'identification de SNV majeurs candidats dans nos analyses ne permet pas nécessairement de nier l'évidence d'effet fondateur dans la ségrégation des allèles délétères chez nos patients. Il est possible de procéder à une analyse plus approfondie en se dotant d'un nombre suffisamment élevé de participants (~1000 et plus) afin de faire ressortir les variants rares d'intérêt qui manquent et qui sont partagés à travers les 13 sujets [124]. En conséquence, nous avons été obligés de mettre à l'écart l'évidence d'effet fondateur dans nos analyses, car nous n'avons pas pu démontrer, à travers celles-ci, l'existence de SNV délétères majoritairement partagés entre les patients.

6.1.1 - Mutations ponctuelles candidates

Notre protocole de détection de SNV se base sur des modèles de transmission d'allèles tel qu'indiquées sur la Figure 4-9. Celui-ci a permis de générer une liste de variations candidates considérablement dense et complexifie les procédures de vérification moléculaires des SNV par séquençage Sanger. Par contre, nous avons

procédé à certaines étapes supplémentaires qui consistent à sélectionner les SNV en fonction de leurs annotations génomiques fonctionnelles et leurs appartenances de voie métabolique. Nous sommes parvenus à sélectionner certains variants géniques qui pourraient être spécifiquement associées au PSIS. Ces dernières sont : IGSF1 (Xq26, c.A815G), JAM3 (11q25, c.97A>C) et RBM28 (7q32, c.2135C>T). Il est connu que la protéine mutantes du gène IGSF1, exprimée à partir du locus Xq26, serait associée au syndrome PSIS (MIM#:300388 [125-129]), ce dernier exprime une grande famille de protéine d'immunoglobuline agissant comme corécepteur dans la voie signalisation d'inhibine et qui régule les procédures de transduction des signaux extracellulaires [130-132].

Pour le gène JAM3, Il est connu que la sous-famille des protéines d'immunoglobuline agisse comme une protéine d'adhérence cellulaire et serait associé au développement du système nerveux [133-138]. De ce fait, sachant que le gène JAM3 (molécule de jonction cellulaire) partage aussi la même fonction d'adhésion cellulaire que les sous familles d'immunoglobuline (IGSF), nous pensons que JAM3 serait aussi associé au PSIS. De plus ce dernier est reconnue pour être associé à certaines maladies neuro-développementales (MIM#613730) comme la prosencéphale et les crises nerveuses [139-140]. De ce fait, nous avons pensé que ce dernier pourrait représenter une piste importante pour la caractérisation de la malformation congénitale de l'hypophyse depuis le développement embryonnaire de la poche de Rathke.

Quant au gène RBM28, nous y avons détecté un SNV rare délétère qui a déjà été reporté pour être associé au syndrome d'ANE [141-142] (Syndrome du déficit alopécique, neurologique et endocrinopathique MIM# 612079). Ce dernier engendre un

phénotype comparable à celui du DCHP (Déficiency Combinée en Hormone Pituitaire), selon Spiegel, R. et al. 2010. De plus, les sujets atteints du syndrome d'ANE montrent certaines caractéristiques qui sont partiellement comparables à celles du PSIS (Hypoplasie antéhypophysaire et posthypophyse ectopique). En outre, il a été montré que le gène RBM28 pourrait s'associer sélectivement aux complexes snRNA de la particule d'épissage alternatif [141]. De ce fait, nous avons pensé que certaines altérations fonctionnelles de celui-ci provoqueraient possiblement une déficience de l'épissage alternatif des ARN messenger auprès des sujets atteints.

6.1.2 - Variations du nombre de copies

Nous avons développé une méthodologie de détection de CNV qui s'opère selon les protocoles d'analyse des trois logiciels de détection de CNV différents indiqués dans la section 4.2.6 (page 71, Figure 4-12). Nous avons retenu certains CNV qui affectent le locus de plusieurs gènes et nous avons reporté les résultats détaillés au tableau 5.6. Nous avons priorisé le CNV homozygote *de novo* (récessive autosomique) affectant le gène MGAM à cause de la fonction connue du gène et le statut délétère du CNV. Celui-ci se localise au 7q34 et engendre une délétion des deux copies d'allèles parentales dont les tailles font 34.13 kilo pair de base (tableau 5.6, Figure 5-4). Le gène MGAM code pour l'enzyme Maltase-Glucoamylase dont la fonction connue de celui-ci consiste à métaboliser des nutriments dans des tissus intestinaux. Selon Younkin, S.G. et al. 2015 [143], la délétion d'une copie partielle du gène serait associée au syndrome de la « fente palatine » qui est un phénotype semblable à celui du défaut développemental de la ligne médiane du corps [144]. De ce fait, nous pensons que l'altération fonctionnelle du gène MGAM serait associée au syndrome du PSIS.

Nous savons que notre technique détient certaines limitations, par conséquent nous avons appliqué des paramètres strictes et retenu prioritairement les CNV consensus. D'ailleurs, il est connu que les données WES ne sont pas optimisées pour des analyses de détection de CNV [145], d'où la détection de CNV via des données de WGS ou de génotypage seraient des meilleures alternatives.

En conclusion, nous pensons qu'il devrait exister une méthodologie standard ayant été validée tant qu'au niveau de la performance que la vraisemblance des résultats. Cette dernière permettrait à la communauté de recherche clinique de garantir une meilleure observation des résultats d'analyse et d'annotation de CNV générés par les logiciels.

6.1.3 - Analyse comparative des pipelines

Je vous rappelle que notre but secondaire dans ce projet a été de comparer les résultats des trois outils de détection de variants afin de déterminer les proportions pertes systématiques de résultats spécifiques au pipeline. D'ailleurs, il existe plusieurs outils de détection de variants, disponibles gratuitement en ligne, desquels certains paramètres auraient été reportés optimaux. De plus, ces derniers ont été rapportés pour avoir la capacité d'améliorer la qualité et la performance des procédures de détection de variations.

Cependant, les résultats expérimentaux ne présentent pas toujours des observations attendues lorsque nous nous référons aux protocoles indiqués dans la littérature. De ce fait, nous pourrions nous demander à quelle mesure qu'un seul outil de détection de variants serait suffisamment optimal pour garantir les résultats les plus vraisemblables possible. De plus, nous savons que tout protocole d'analyse génère une

proportion d'erreur de type I et de type II. Par ailleurs, je vous rappelle que l'erreur de type I d'un protocole se réfère à la proportion de résultats faux positifs générés pendant que celui du type II se réfère plutôt à la proportion de faux négatifs. En conséquence, pour diminuer ces types d'erreurs dans nos analyses, notre règle d'or est de se doter d'au moins une méthode alternative à la procédure initiale afin de valider la cohérence des observations.

Pour évaluer la performance de nos démarches, nous avons décidé de comparer les résultats issus de ceux-ci, tant qu'au point de vue des SNV rares que des INDEL. Ceci nous a permis d'observer le niveau de partage des SNV détectés et procéder à une analyse approximative de la performance de ceux-ci. Notre procédure d'analyse de performance à des limites, celle-ci a été effectuée en absence de résultats de référence standard (gold standard data) qui aurait permis de mieux évaluer la précision et la sensibilité globale de nos méthodes. En conséquence, nous n'avons pas pu évaluer le niveau d'erreur de type II de notre méthodologie, par contre, celui du type I a été évalué en assumant que les résultats consensus des trois outils de détection de SNV soient vrais.

Nous avons retenu deux aspects des résultats représentés dans la Figure 5-2, que sont : la proportion de SNV consensus et privés. D'abord, nous avons constaté que GATK-UG détecte énormément de SNV rares (~86%) qui ne font pas parties des résultats consensus. Nous pensons que ces derniers pourraient représenter une proportion considérable de faux positives car ils sont majoritairement discordants entre les résultats consensus. Toutefois, GATK-UG génère un faible taux de faux négatif dans les résultats, ceci serait dû à l'utilisation de paramètres de détection de SNV trop relaxants.

Pour les outils de détection de SNV FREEBAYES et SAMTOOLS, nous avons trouvé que ces derniers génèrent beaucoup moins de SNV privés (~10%) comparativement aux résultats de GATK-UG. Ces observations nous permettent de constater que les paramètres d'analyse de base de SAMTOOLS et FREEBAYES sont plus stricts ou plus précis que ceux de GATK-UG. De plus, nos résultats révèlent que SAMTOOLS et FREEBAYES génèrent ~89% de SNV consensus, celui-ci montre qu'il existe une cohérence assez élevée dans les protocoles de ces deux outils.

Certaines études ont déjà exploré plusieurs sources de solutions permettant de résoudre certains défis associés aux analyses de détection des SNV du WES, par contre, pouvons-nous dire autant pour les analyses de détection des INDEL? Nous avons eu l'objectif d'analyser la proportion d'INDEL consensus détectés par les trois outils de détection de variants. Nous avons observé qu'une infime proportion d'INDEL, détectés par les trois outils, est reporté dans la base de données dbSNP138. De plus, nous avons effectué une comparaison entre les résultats d'INDEL détectés par entre les outils. Nous avons constaté que la proportion d'INDEL consensus détectés représente 50% à 73% (Figure 5-3), cette observation est différente de celle des SNV (62% à 90%). Ces résultats nous indiquent qu'il existe une incohérence considérable entre des résultats d'INDEL détectés par nos outils. De plus, cette incohérence est aussi observée à travers la comparaison de nos résultats et les données d'INDEL reportées dans dbSNP138.

Nous pensons que les protocoles de détection d'INDEL appliqués dans ce projet a été très instable, par conséquent, nos résultats doivent être traité avec beaucoup d'attention. De plus, nous avons constaté une discordance considérable entre les résultats d'analyse des données locales et ceux de la littérature (dbSNP138). Encore une fois, nous

pensons qu'un ensemble de résultats de référence standard nous serait nécessaire afin de mieux calibrer nos paramètres d'analyse.

6.2 - Conclusion

En conclusion, nous avons pu observer une proportion très élevée de SNV rares et *de novo* dans nos analyses. De plus nous ne sommes pas parvenus à détecter majoritairement aucun SNV rare dans nos analyses. Alors, nous pensons que ces observations seraient dues à une immense proportion d'hétérogénéité génétique du SNV causales dans notre cohorte. Concernant l'analyse comparative de résultats de détection de variants, nous avons constaté que la rigueur des paramètres de base diffère d'un outil à une autre. Cette dernière a permis de générer des résultats très hétérogènes qui nécessitent certaines vérifications via des données de référence standard.

6.3 - Perspectives

Nos analyses n'ont pas été concluantes sur la détection des marqueurs génétiques majeurs qui nous auraient permis de caractériser les phénotypes associés au syndrome du PSIS. Celui-ci pourrait être dû à plusieurs raisons, soit que les variants recherchés ne sont pas présents dans nos jeux de données ou la tailles de nos échantillons ne nous permet pas d'évaluer statistiquement les variants candidats détectés. Dans des travaux futures, nous envisagerons d'effectuer certaines collaborations avec d'autres équipes de recherche afin d'explorer d'autres analyses alternatives (des analyses qui impliquent des voies d'expressions de gènes connues). Nous allons aussi procéder à des analyses d'association « Gene Based Burden » qui nous permettront de caractériser les phénotypes associés au syndrome à l'échelle génique (gènes partagées) plutôt que moléculaire (SNV).

Indexes

A

ACTH	31
Adénohypophyse	20
ADN	8
Adrénocortico-trophine	31
Allan Maxam.....	36
Allèle	12
ANNOVAR.....	70
Antéhypophyse.....	7, 30
ARN polymérase	9
ATP.....	8

B

Baum-Welch	81
Bayes	75
BMP4.....	23
BOWTIE2	46
Burden	107
BWA.....	45

C

Calcitonine	30
CHRH	19
CNV.....	14, 104
Codon.....	8
Corticotropes.....	25, 26

D

DCHP.....	102
<i>de novo</i>	2, 6, 45, 98, 103, 107
DeepOfCoverage.....	80
Déficiéce	103
Délétion.....	97
Diploïde	13
Dwarfism	32

F

FASTQC	55
Faux-sens	14

Fibroblastes	21
Forward-Backward	81
Frederick Sanger	36
FREEBAYES	60

G

Gamma-glutamylcystéine synthase	22
GATK	50
GHRH	19
Gigantisme	31
Glandes	19
GnHRH	19
GS FLX Titanium	38
GS20	38
GSH4	22
GTP	8

H

Hard-clipping	48
Hemizygote	90
Hétérogénéité	100
Hétérozygotes	13
HiSeq	52
HMM	80
Homozygote	97
Homozygotes	13
Hormone	7
Hyperprolactinémie	32
Hypophyse	19
Hypoplasie	103
Hypothalamus	19

I

Inactivateurs	12
INDEL	14
Interruption	7
IRM	35
ISL1	23

J

Jonathan Rothberg	38
-------------------------	----

L

L'hypoprolactinémie	32
Lactotropes.....	26
LH3.....	22
LHX4.....	22
Ligne.....	7
Lim-homeobox.....	22

M

MAF	93
Mélanotropes.....	26
Mutation silencieuses	14
Mutations synonymes.....	14
MutationTaster	74

N

Nonsenses	14
-----------------	----

P

PaGenBase	85
Paired-End	47
Phred.....	42
Pituitary stalk	20
PolyPHEN2.....	74
POMC.....	31
Promoteurs.....	9, 14
Pro-Opiomelanocorticotrophines	31
PSIS.....	7, 20

R

Rathke.....	23, 24
RNAseq.....	43
RPKM.....	82

S

Sella turcica.....	20
Séquençage	37
SIFT.....	74
SNV.....	14, 53
Soft-clipping	48
Sonic hedgehog.....	21
SureSelect	52

Surrénales.....	19
Syndrome	7
Syndrome d'ANE.....	102
Système endocrinien.....	18

T

THRH	19
Thyroïdes	19
Thyroxine.....	30
Tige hypophysaire	7
Triiodothyronine	30
TSH.....	30

V

Viterbi.....	80, 81
--------------	--------

W

Walter Gilbert	36
WES.....	40
WGS	40

Annexes

Tableau S 1 - Description clinique des attributs phénotypiques de la cohorte de 13 patients

Identifiant familial du patient	Sexe	Origine ancestrale	Poids à la naissance (g)	Âge de recrutement	Diagnostique	Traitements hormonaux	Anomalie morphologique de la glande	Hauteur SDS
11-002	F	CF	3345	13	RC	GH, TSH, ACTH, FSH, LH	HAH, PHE	-4.23
11-006	M	CF	3800	~1	HoG	GH, ACTH, TSH, FSH*, LH*	PHE, THA, ANOg, Ase, Ape, ACC	na
11-011	M	CF	3360	78	RC	GH, TSH, ACTH	HAH, PHE, THA	-3.29
11-015	M	OAC	3330	78	RC	GH	PHE, THA	-3.84
11-016	M	CF	3140	78	RC	GH	HAH, PHE, ACC	-3.25
11-017	M	CF	2500	54	RC	GH, TSH, ACTH	PHE, THA, ACC	-4.74
11-019	M	CF	3030	24	RC	GH	HAH, PHE, THA	-3.19
11-021	M	CF	2975	~1	HoG	GH, TSH, ACTH, FSH, LH	PHE, THA	na
11-022	M	CF	3315	72	RC	GH	PHE, THA	na
11-025	M	CF	3350	30	RC	GH, TSH	PHE, THA	-5.75
11-026	M	CF	3690	42	RC	GH	normal	-3.76
11-028	F	CF	3070	24	RC	GH, TSH	PHE, THA	-3.34
11-031	F	CF	3380	36	RC	GH, TSH	HAH, PHE, THA	-4.81
Définitions des abréviations est des légendes du tableau								
CF = Canadien français			ANO = Atrophie du nerf optique (gauche ou droite)					
RC = Retard de croissance			ASe = Absence du septum					
HAH = Hypoplasie antéhypophysaire			Ape = Absence du pallidum					
PHE = Post hypophysaire ectopique			ACC = Absence du corpus callosum					
HoG = Hypoglycémie			CAO = Origine arabe consanguine					
THA = Tige hypophysaire AMINCIE/ABSCENT			~1= naissance					

Tableau S 2 - Résultats détaillés d'analyse de proportion de variants partagés entre les outils de détections de SNV.

Outil de détection de variants	Critères d'inclusions de variants	variants spécifiques	variants partagés	variants totaux générés	variants partagés (%)	variants spécifiques (%)
FREEBAYES	T1	4387	253566	257953	98.3	0.017007
	T2	344	240982	241326	99.9	0.001425
	Rare	4329	36525	40854	89.4	0.105963
	Commun	58	216741	216799	99.97	0.000268
SAMTOOLS	T1	4448	253566	258014	98.3	0.017239
	T2	423	240982	241405	99.8248	0.001752
	Rare	4356	36525	40881	0.893447	0.106553
	Commun	92	216741	216833	0.999576	0.000424
GATK	T1	364716	253266	617982	0.409827	0.590173
	T2	169716	240982	410698	0.586762	0.413238
	Rare	237054	36525	273579	0.133508	0.866492
	Commun	127662	216741	344403	0.629324	0.370676

Dans la première colonne du tableau, nous indiquons les outils de détection de variantes considérés dans cette analyse de comparaison de SNV. Dans la deuxième colonne, nous indiquons les types de variantes sélectionnés, soit: tous les SNV (T1), tous les SNV qui sont reportés dans dbSNP (T2), les SNV rares, les SNV communs. Nous entendons par variations spécifiques (troisième colonne), celles qui sont exclusivement détectés par un seul outil.

6.4 - Références

1. Triulzi F, Scotti G, di Natale B, Pellini C, Lukezic M, Scognamiglio M, et al. Evidence of a congenital midline brain anomaly in pituitary dwarfs: a magnetic resonance imaging study in 101 patients. *Pediatrics*. 1994;93(3):409-16.
2. Kamboj A, Lause M, Kumar P. Ophthalmic manifestations of endocrine disorders—endocrinology and the eye. *Translational Pediatrics*. 2017;6(4):286-99.
3. Tatsi C, Sertedaki A, Voutetakis A, Valavani E, Magiakou MA, Kanaka-Gantenbein C, et al. Pituitary stalk interruption syndrome and isolated pituitary hypoplasia may be caused by mutations in holoprosencephaly-related genes. *J Clin Endocrinol Metab*. 2013;98(4):E779-84.
4. Lynch M. The origins of eukaryotic gene structure. *Mol Biol Evol*. 2006;23(2):450-68.
5. Strachan T. *Génétique moléculaire humaine*. 4e éd.. ed. Read AP, editor. Paris: Paris : Médecine Sciences pub.; 2012.
6. Izumi K, Mine K, Inoue Y, Teshima M, Ogawa S, Kai Y, et al. Reduced Tyk2 gene expression in beta-cells due to natural mutation determines susceptibility to virus-induced diabetes. *Nat Commun*. 2015;6:6748.
7. Kaneko K. Proportionality between variances in gene expression induced by noise and mutation: consequence of evolutionary robustness. *BMC Evol Biol*. 2011;11:27.
8. Son MS, Kang MJ, Park HC, Chi SG, Kim YH. Expression and mutation analysis of TIG1 (tazarotene-induced gene 1) in human gastric cancer. *Oncol Res*. 2009;17(11-12):571-80.
9. Watkins-Chow DE, Pavan WJ. Genomic copy number and expression variation within the C57BL/6J inbred mouse strain. *Genome Res*. 2008;18(1):60-6.
10. Bai X, Huang Y, Hu Y, Liu H, Zhang B, Smaczniak C, et al. Duplication of an upstream silencer of FZP increases grain yield in rice. *Nat Plants*. 2017;3(11):885-93.
11. Soni BK, Joish UK, Sahni H, George RA, Sivasankar R, Aggarwal R. A Comparative Study of Pituitary Volume Variations in MRI in Acute Onset of Psychiatric Conditions. *J Clin Diagn Res*. 2017;11(2):TC01-TC4.
12. Kaiser U, Ho KKY. Chapter 8 - Pituitary Physiology and Diagnostic Evaluation A2 - Melmed, Shlomo. In: Polonsky KS, Larsen PR, Kronenberg HM, editors. *Williams Textbook of Endocrinology (Thirteenth Edition)*. Philadelphia: Content Repository Only!; 2016. p. 176-231.
13. Ben-Shlomo A, Melmed S. Chapter 2 - Hypothalamic Regulation of Anterior Pituitary Function. *The Pituitary (Third Edition)*. San Diego: Academic Press; 2011. p. 21-45.
14. Drouin J. Chapter 1 - Pituitary Development A2 - Melmed, Shlomo. *The Pituitary (Third Edition)*. San Diego: Academic Press; 2011. p. 3-19.
15. Pfaffle R, Blankenstein O, Wuller S, Heimann K, Heimann G. Idiopathic growth hormone deficiency: a vanishing diagnosis? *Horm Res*. 2000;53 Suppl 3:1-8.
16. Frisch H, Kim C, Hausler G, Pfaffle R. Combined pituitary hormone deficiency and pituitary hypoplasia due to a mutation of the Pit-1 gene. *Clin Endocrinol (Oxf)*. 2000;52(5):661-5.
17. Maintz D, Benz-Bohm G, Gindele A, Schonau E, Pfaffle R, Lackner K. Posterior pituitary ectopia: another hint toward a genetic etiology. *AJNR Am J Neuroradiol*. 2000;21(6):1116-8.
18. Zhu X, Lin CR, Prefontaine GG, Tollkuhn J, Rosenfeld MG. Genetic control of pituitary development and hypopituitarism. *Curr Opin Genet Dev*. 2005;15(3):332-40.
19. Sheng HZ, Zhadanov AB, Mosinger B, Jr., Fujii T, Bertuzzi S, Grinberg A, et al. Specification of pituitary cell lineages by the LIM homeobox gene Lhx3. *Science*. 1996;272(5264):1004-7.

20. Raetzman LT, Ward R, Camper SA. Lhx4 and Prop1 are required for cell survival and expansion of the pituitary primordia. *Development (Cambridge, England)*. 2002;129(18):4229-39.
21. Davis SW, Camper SA. Noggin Regulates Bmp4 Activity During Pituitary Induction. *Developmental biology*. 2007;305(1):145-60.
22. Takuma N, Sheng HZ, Furuta Y, Ward JM, Sharma K, Hogan BL, et al. Formation of Rathke's pouch requires dual induction from the diencephalon. *Development (Cambridge, England)*. 1998;125(23):4835-40.
23. Millán MIP, Camper SA. Chapter 4 - The Biology of Pituitary Stem Cells A2 - Ulloa-Aguirre, Alfredo. In: Conn PM, editor. *Cellular Endocrinology in Health and Disease*. Boston: Academic Press; 2014. p. 55-63.
24. Sheng HZ, Westphal H. Early steps in pituitary organogenesis. *Trends Genet*. 1999;15(6):236-40.
25. Zhao Y, Sheng HZ, Amini R, Grinberg A, Lee E, Huang S, et al. Control of hippocampal morphogenesis and neuronal differentiation by the LIM homeobox gene Lhx5. *Science*. 1999;284(5417):1155-8.
26. Treier M, Gleiberman AS, O'Connell SM, Szeto DP, McMahon JA, McMahon AP, et al. Multistep signaling requirements for pituitary organogenesis in vivo. *Genes Dev*. 1998;12(11):1691-704.
27. Carreno G, Apps JR, Lodge EJ, Panousopoulos L, Haston S, Gonzalez-Meljem JM, et al. Hypothalamic sonic hedgehog is required for cell specification and proliferation of LHX3/LHX4 pituitary embryonic precursors. *Development (Cambridge, England)*. 2017;144(18):3289-302.
28. Ericson J, Norlin S, Jessell TM, Edlund T. Integrated FGF and BMP signaling controls the progression of progenitor cell differentiation and the emergence of pattern in the embryonic anterior pituitary. *Development (Cambridge, England)*. 1998;125(6):1005-15.
29. Davis SW, Mortensen AH, Camper SA. Birthdating studies reshape models for pituitary gland cell specification. *Dev Biol*. 2011;352(2):215-27.
30. Castinetti F, Davis SW, Brue T, Camper SA. Pituitary stem cell update and potential implications for treating hypopituitarism. *Endocr Rev*. 2011;32(4):453-71.
31. Miletta MC, Flück CE, Mullis PE. Targeting GH-1 splicing as a novel pharmacological strategy for growth hormone deficiency type II. *Biochemical pharmacology*. 2017;124:1-9.
32. Giordano M. Genetic causes of isolated and combined pituitary hormone deficiency. *Best Practice and Research: Clinical Endocrinology and Metabolism*. 2016;30(6):679-91.
33. Di Iorgi N, Morana G, Allegri AEM, Napoli F, Gastaldi R, Calcagno A, et al. Classical and non-classical causes of GH deficiency in the paediatric age. *Best Practice and Research: Clinical Endocrinology and Metabolism*. 2016;30(6):705-36.
34. Birla S, Khadgawat R, Jyotsna VP, Jain V, Garg MK, Bhalla AS, et al. Identification of novel GHRHR and GH1 mutations in patients with isolated growth hormone deficiency. *Growth Hormone and IGF Research*. 2016;29:50-6.
35. Arnhold IJP, França MM, Carvalho LR, Mendonca BB, Jorge AAL. Role of GLI2 in hypopituitarism phenotype. *Journal of molecular endocrinology*. 2015;54(3):R141-R50.
36. Backeljauw PF, Dattani MT, Cohen P, Rosenfeld RG. Disorders of growth hormone/insulin-like growth factor secretion and action. *Pediatric Endocrinology: Fourth Edition* 2014.
37. Arman A, Dündar BN, Çetinkaya E, Erzaim N, Büyükgebiz A. Novel growth hormone-releasing hormone receptor gene mutations in Turkish children with isolated growth hormone deficiency. *JCRPE Journal of Clinical Research in Pediatric Endocrinology*. 2014;6(4):202-8.

38. Alatzoglou KS, Webb EA, Le Tissier P, Dattani MT. Isolated growth hormone deficiency (GHD) in childhood and adolescence: Recent advances. *Endocrine reviews*. 2014;35(3):376-432.
39. Tsubahara M, Hayashi Y, Nijima SI, Yamamoto M, Kamijo T, Murata Y, et al. Isolated growth hormone deficiency in two siblings because of paternal mosaicism for a mutation in the GH1 gene. *Clinical endocrinology*. 2012;76(3):420-4.
40. Stanley T. Diagnosis of growth hormone deficiency in childhood. *Current Opinion in Endocrinology, Diabetes and Obesity*. 2012;19(1):47-52.
41. Li S, Crenshaw EB, 3rd, Rawson EJ, Simmons DM, Swanson LW, Rosenfeld MG. Dwarf locus mutants lacking three pituitary cell types result from mutations in the POU-domain gene pit-1. *Nature*. 1990;347(6293):528-33.
42. Camper SA, Saunders TL, Katz RW, Reeves RH. The Pit-1 transcription factor gene is a candidate for the murine Snell dwarf mutation. *Genomics*. 1990;8(3):586-90.
43. Cohen LE, Radovick S. Molecular basis of combined pituitary hormone deficiencies. *Endocr Rev*. 2002;23(4):431-42.
44. Pulichino AM, Vallette-Kasic S, Tsai JP, Couture C, Gauthier Y, Drouin J. Tpit determines alternate fates during pituitary cell differentiation. *Genes Dev*. 2003;17(6):738-47.
45. Melmed S, Kleinberg D. Chapter 9 - Pituitary Masses and Tumors. *Williams Textbook of Endocrinology (Thirteenth Edition)*. Philadelphia: Content Repository Only!; 2016. p. 232-99.
46. Cooke DW, Divall SA, Radovick S. Chapter 24 - Normal and Aberrant Growth in Children A2 - Melmed, Shlomo. In: Polonsky KS, Larsen PR, Kronenberg HM, editors. *Williams Textbook of Endocrinology (Thirteenth Edition)*. Philadelphia: Content Repository Only!; 2016. p. 964-1073.
47. Puustinen L, Jalanko H, Holmberg C, Merenmies J. Recombinant human growth hormone treatment after liver transplantation in childhood: the 5-year outcome. *Transplantation*. 2005;79(9):1241-6.
48. Pfaffle R, Klammt J. Pituitary transcription factors in the aetiology of combined pituitary hormone deficiency. *Best Pract Res Clin Endocrinol Metab*. 2011;25(1):43-60.
49. Klingseisen A, Jackson AP. Mechanisms and pathways of growth failure in primordial dwarfism. *Genes Dev*. 2011;25(19):2011-24.
50. Noel GL, Suh HK, Stone JG, Frantz AG. Human prolactin and growth hormone release during surgery and other conditions of stress. *J Clin Endocrinol Metab*. 1972;35(6):840-51.
51. Chahal J, Schlechte J. Hyperprolactinemia. *Pituitary*. 2008;11(2):141-6.
52. Molitch ME. Pathologic hyperprolactinemia. *Endocrinol Metab Clin North Am*. 1992;21(4):877-901.
53. Lieblich JM, Rogol AD, White BJ, Rosen SW. Syndrome of anosmia with hypogonadotropic hypogonadism (Kallmann syndrome): clinical and laboratory studies in 23 cases. *Am J Med*. 1982;73(4):506-19.
54. Bas F, Uyguner ZO, Darendeliler F, Aycan Z, Cetinkaya E, Berberoglu M, et al. Molecular analysis of PROP1, POU1F1, LHX3, and HESX1 in Turkish patients with combined pituitary hormone deficiency: a multicenter study. *Endocrine*. 2015;49(2):479-91.
55. Vieira TC, Boldarine VT, Abucham J. Molecular analysis of PROP1, PIT1, HESX1, LHX3, and LHX4 shows high frequency of PROP1 mutations in patients with familial forms of combined pituitary hormone deficiency. *Arquivos brasileiros de endocrinologia e metabologia*. 2007;51(7):1097-103.
56. Crone J, Pfaffle R, Stobbe H, Prayer D, Gomez I, Frisch H. Familial combined pituitary hormone deficiency caused by PROP-1 gene mutation. Growth patterns and MRI studies in untreated subjects. *Hormone research*. 2002;57(3-4):120-6.

57. Bhattacharjee R, Chakraborty PP, Ghosh S, Mukhopadhyay P, Mukopadhyay S, Chowdhury S. 'Double bright spot' in T1 weighted non-contrast MRI of pituitary in multiple pituitary hormone deficiency. *Indian journal of pediatrics*. 2015;82(5):485-6.
58. Arslanoglu I, Kutlu H, Isguven P, Tokus F, Isik K. Diagnostic value of pituitary MRI in differentiation of children with normal growth hormone secretion, isolated growth hormone deficiency and multiple pituitary hormone deficiency. *Journal of pediatric endocrinology & metabolism : JPEM*. 2001;14(5):517-23.
59. Karaca E, Buyukkaya R, Pehlivan D, Charng WL, Yaykasli KO, Bayram Y, et al. Whole-exome sequencing identifies homozygous GPR161 mutation in a family with pituitary stalk interruption syndrome. *J Clin Endocrinol Metab*. 2015;100(1):E140-7.
60. Maxam AM, Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A*. 1977;74(2):560-4.
61. Maxam AM, Gilbert W. A new method for sequencing DNA. 1977. *Biotechnology*. 1992;24:99-103.
62. Kumar S, Fuller CW. Chapter 4 Advances in Dye-Nucleotide Conjugate Chemistry for DNA Sequencing. In: Mitchelson KR, editor. *Perspectives in Bioanalysis*. 2: Elsevier; 2007. p. 119-49.
63. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005;437(7057):376-80.
64. Ahmadian A, Ehn M, Hober S. Pyrosequencing: history, biochemistry and future. *Clin Chim Acta*. 2006;363(1-2):83-94.
65. Gharizadeh B, Herman ZS, Eason RG, Jejelowo O, Pourmand N. Large-scale pyrosequencing of synthetic DNA: a comparison with results from Sanger dideoxy sequencing. *Electrophoresis*. 2006;27(15):3042-7.
66. Gharizadeh B, Akhras M, Nourizad N, Ghaderi M, Yasuda K, Nyren P, et al. Methodological improvements of pyrosequencing technology. *J Biotechnol*. 2006;124(3):504-11.
67. Tawfik DS, Griffiths AD. Man-made cell-like compartments for molecular evolution. *Nature biotechnology*. 1998;16(7):652-6.
68. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry*. 1996;242(1):84-9.
69. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. *Journal of applied genetics*. 2011;52(4):413-35.
70. Nyren P. The history of pyrosequencing. *Methods in molecular biology*. 2007;373:1-14.
71. Ronaghi M, Uhlen M, Nyren P. A sequencing method based on real-time pyrophosphate. *Science*. 1998;281(5375):363, 5.
72. Markt P, McGoohan C, Walker B, Kirchmair J, Feldmann C, De Martino G, et al. Discovery of novel cathepsin S inhibitors by pharmacophore-based virtual high-throughput screening. *J Chem Inf Model*. 2008;48(8):1693-705.
73. Mardis ER. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*. 2008;9:387-402.
74. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG*. 2008;24(3):133-41.
75. Metzker ML. Sequencing technologies - the next generation. *Nature reviews Genetics*. 2010;11(1):31-46.
76. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences of the United States of America*. 2015;112(17):5473-8.

77. Schwarze K, Buchanan J, Taylor JC, Wordsworth S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genetics In Medicine*. 2018.
78. Elkon R, Agami R. Characterization of noncoding regulatory DNA in the human genome. *Nat Biotechnol*. 2017;35(8):732-46.
79. Kuklin EA, Alkins S, Bakthavachalu B, Genco MC, Sudhakaran I, Raghavan KV, et al. The Long 3'UTR mRNA of CaMKII Is Essential for Translation-Dependent Plasticity of Spontaneous Release in *Drosophila melanogaster*. *J Neurosci*. 2017;37(44):10554-66.
80. Liao P, Satten GA, Hu YJ. PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies. *Genet Epidemiol*. 2017;41(5):375-87.
81. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. 1998;8(3):175-85.
82. Pearl J. *Heuristics: Intelligent search strategies for computer problem solving*: Addison-Wesley Pub. Co., Inc., Reading, MA; None; 1984. Medium: X; Size: Pages: 382 p.
83. Kucharska E. Heuristic Method for Decision-Making in Common Scheduling Problems. *Applied Sciences*. 2017;7(10):1073.
84. Di Iorgi N, Morana G, Allegri AE, Napoli F, Gastaldi R, Calcagno A, et al. Classical and non-classical causes of GH deficiency in the paediatric age. *Best practice & research Clinical endocrinology & metabolism*. 2016;30(6):705-36.
85. Reynaud R, Castinetti F, Galon-Faure N, Albarel-Loy F, Saveanu A, Quentien MH, et al. Genetic aspects of growth hormone deficiency. *Archives de Pediatrie*. 2011;18(6):696-706.
86. Petkovic V, Mullis PE. Genetic defects causing functional and structural isolated growth hormone deficiency. *Translational Neuroscience*. 2011;2(2):152-62.
87. Mullis PE. Genetics of GHRH, GHRH-receptor, GH and GH-receptor: Its impact on pharmacogenetics. *Best Practice and Research: Clinical Endocrinology and Metabolism*. 2011;25(1):25-41.
88. Al-Hassnan ZN, Sakati N. Genetic disorders in Saudi Arabia. *Genetic Disorders Among Arab Populations 2010*. p. 531-73.
89. Alatzoglou KS, Dattani MT. Genetic causes and treatment of isolated growth hormone deficiency-an update. *Nature Reviews Endocrinology*. 2010;6(10):562-76.
90. Mullis PE. Genetic control of growth. *European Journal of Endocrinology*. 2005;152(1):11-31.
91. Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, et al. Genetic Variation in an Individual Human Exome. *PLOS Genetics*. 2008;4(8):e1000160.
92. Brown J, Pirrung M, McCue LA. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics*. 2017.
93. Cornish A, Guda C. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *BioMed Research International*. 2015;2015:11.
94. Otto C, Stadler PF, Hoffmann S. Lacking alignments? The next-generation sequencing mapper segemehl revisited. *Bioinformatics*. 2014;30(13):1837-43.
95. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.
96. Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC Genomics*. 2012;13 Suppl 8:S8.
97. Bhattacharya R, Rose PW, Burley SK, Prlić A. Impact of genetic variation on three dimensional structure and function of proteins. *PLoS ONE*. 2017;12(3):e0171355.
98. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biology*. 2017;18(1):77.

99. Sambo F, Trifoglio E, Di Camillo B, Toffolo GM, Cobelli C. Bag of Naïve Bayes: biomarker selection and classification from genome-wide SNP data. *BMC Bioinformatics*. 2012;13(14):S2.
100. Fromer M, Purcell SM. Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. *Curr Protoc Hum Genet*. 2014;81:7 23 1-1.
101. Yao R, Zhang C, Yu T, Li N, Hu X, Wang X, et al. Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Molecular Cytogenetics*. 2017;10(1):30.
102. Watson CM, Crinnion LA, Berry IR, Harrison SM, Lascelles C, Antanaviciute A, et al. Enhanced diagnostic yield in Meckel-Gruber and Joubert syndrome through exome sequencing supplemented with split-read mapping. *BMC Medical Genetics*. 2016;17(1):1.
103. Hong CS, Singh LN, Mullikin JC, Biesecker LG. Assessing the reproducibility of exome copy number variations predictions. *Genome Medicine*. 2016;8(1):82.
104. Miklós I, Meyer IM. A linear memory algorithm for Baum-Welch training. *BMC Bioinformatics*. 2005;6(1):231.
105. Xu Y, Peng B, Fu Y, Amos CI. Genome-wide algorithm for detecting CNV associations with diseases. *BMC Bioinformatics*. 2011;12:331.
106. Lam TY, Meyer IM. Efficient algorithms for training the parameters of hidden Markov models using stochastic expectation maximization (EM) training and Viterbi training. *Algorithms for Molecular Biology*. 2010;5(1):38.
107. Barrett N, Weber-Jahnke J. Building a biomedical tokenizer using the token lattice design pattern and the adapted Viterbi algorithm. *BMC Bioinformatics*. 2011;12(3):S1.
108. Churbanov A, Winters-Hilt S. Implementing EM and Viterbi algorithms for Hidden Markov Model in linear memory. *BMC Bioinformatics*. 2008;9(1):224.
109. Ista J. *Mathematical modeling for the life sciences*. Berlin: Springer; 2005. vi, 164 p. p.
110. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. In: Alex W, Kai-Fu L, editors. *Readings in speech recognition: Morgan Kaufmann Publishers Inc.*; 1990. p. 267-96.
111. Yoon BJ. *Hidden Markov Models and their Applications in Biological Sequence Analysis*. *Curr Genomics*. 2009;10(6):402-15.
112. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet*. 2007;39(7 Suppl):S16-21.
113. Mahamdallie S, Ruark E, Yost S, Ramsay E, Uddin I, Wylie H, et al. The ICR96 exon CNV validation series: a resource for orthogonal assessment of exon CNV calling in NGS data. *Wellcome Open Res*. 2017;2:35.
114. Yao R, Zhang C, Yu T, Li N, Hu X, Wang X, et al. Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Molecular Cytogenetics*. 2017;10:30.
115. Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics*. 2017;18(1):286.
116. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med*. 2013;5(3):28.
117. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*. 2013;14(11):S1.
118. Mensink EJ, Thompson A, Schot JD, van de Greef WM, Sandkuyl LA, Schuurman RK. Mapping of a gene for X-linked agammaglobulinemia and evidence for genetic heterogeneity. *Hum Genet*. 1986;73(4):327-32.

119. Arveiler B, Alembik Y, Hanauer A, Jacobs P, Tranebjaerg L, Mikkelsen M, et al. Linkage analysis suggests at least two loci for X-linked non-specific mental retardation. *Am J Med Genet.* 1988;30(1-2):473-83.
120. Pfundt R, del Rosario M, Vissers LELM, Kwint MP, Janssen IM, de Leeuw N, et al. Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genetics in Medicine.* 2017;19(6):667-75.
121. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet.* 2010;86(5):749-64.
122. 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.* 2011;43(9):838-46.
123. Wang Q, Shashikant CS, Jensen M, Altman NS, Girirajan S. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Scientific Reports.* 2017;7(1):885.
124. Timberlake AT, Choi J, Zaidi S, Lu Q, Nelson-Williams C, Brooks ED, et al. Two locus inheritance of non-syndromic midline craniosynostosis via rare SMAD6 and common BMP2 alleles. *eLife.* 2016;5.
125. Baker EM, Khorasgani MG, Gardner-Medwin D, Gholkar A, Griffiths PD. Arthrogryposis multiplex congenita and bilateral parietal polymicrogyria in association with the intrauterine death of a twin. *Neuropediatrics.* 1996;27(1):54-6.
126. Barkovich AJ, Lindan CE. Congenital cytomegalovirus infection of the brain: imaging analysis and embryologic considerations. *AJNR American journal of neuroradiology.* 1994;15(4):703-15.
127. Borgatti R, Triulzi F, Zucca C, Piccinelli P, Balottin U, Carozzo R, et al. Bilateral perisylvian polymicrogyria in three generations. *Neurology.* 1999;52(9):1910-3.
128. Brandao-Almeida IL, Hage SR, Oliveira EP, Guimaraes CA, Teixeira KC, Abramides DV, et al. Congenital bilateral perisylvian syndrome: familial occurrence, clinical and psycholinguistic aspects correlated with MRI. *Neuropediatrics.* 2008;39(3):139-45.
129. Clark M, Neville BG. Familial and genetic associations in Worster-Drought syndrome and perisylvian disorders. *American journal of medical genetics Part A.* 2008;146A(1):35-42.
130. Chapman SC, Woodruff TK. Modulation of activin signal transduction by inhibin B and inhibin-binding protein (INhBP). *Molecular endocrinology.* 2001;15(4):668-79.
131. Nishigaki S, Hamazaki T, Fujita K, Morikawa S, Tajima T, Shintaku H. A Japanese Family with Central Hypothyroidism Caused by a Novel IGSF1 Mutation. *Thyroid : official journal of the American Thyroid Association.* 2016;26(12):1701-5.
132. Schoenmakers N, Alatzoglou KS, Chatterjee VK, Dattani MT. Recent advances in central congenital hypothyroidism. *The Journal of endocrinology.* 2015;227(3):R51-71.
133. Leshchyn'ska I, Sytnyk V. Reciprocal Interactions between Cell Adhesion Molecules of the Immunoglobulin Superfamily and the Cytoskeleton in Neurons. *Front Cell Dev Biol.* 2016;4:9.
134. Atz ME, Rollins B, Vawter MP. NCAM1 association study of bipolar disorder and schizophrenia: polymorphisms and alternatively spliced isoforms lead to similarities and differences. *Psychiatric genetics.* 2007;17(2):55-67.
135. Barker TH, Grenett HE, MacEwen MW, Tilden SG, Fuller GM, Settleman J, et al. Thy-1 regulates fibroblast focal adhesions, cytoskeletal organization and migration through modulation of p190 RhoGAP and Rho GTPase activity. *Experimental cell research.* 2004;295(2):488-96.

136. Bodrikov V, Leshchyns'ka I, Sytnyk V, Overvoorde J, den Hertog J, Schachner M. RPTPalph is essential for NCAM-mediated p59fyn activation and neurite elongation. *The Journal of cell biology*. 2005;168(1):127-39.
137. Bodrikov V, Sytnyk V, Leshchyns'ka I, den Hertog J, Schachner M. NCAM induces CaMKIIalpha-mediated RPTPalph phosphorylation to enhance its catalytic activity and neurite outgrowth. *The Journal of cell biology*. 2008;182(6):1185-200.
138. Matthaus C, Langhorst H, Schutz L, Juttner R, Rathjen FG. Cell-cell communication mediated by the CAR subgroup of immunoglobulin cell adhesion molecules in health and disease. *Molecular and cellular neurosciences*. 2016.
139. Akawi NA, Canpolat FE, White SM, Quilis-Esquerria J, Morales Sanchez M, Gamundi MJ, et al. Delineation of the clinical, molecular and cellular aspects of novel JAM3 mutations underlying the autosomal recessive hemorrhagic destruction of the brain, subependymal calcification, and congenital cataracts. *Human mutation*. 2013;34(3):498-505.
140. Mochida GH, Ganesh VS, Felie JM, Gleason D, Hill RS, Clapham KR, et al. A homozygous mutation in the tight-junction protein JAM3 causes hemorrhagic destruction of the brain, subependymal calcification, and congenital cataracts. *American journal of human genetics*. 2010;87(6):882-9.
141. Damianov A, Kann M, Lane WS, Bindereif A. Human RBM28 protein is a specific nucleolar component of the spliceosomal snRNPs. *Biological chemistry*. 2006;387(10-11):1455-60.
142. Spiegel R, Shalev SA, Adawi A, Sprecher E, Tenenbaum-Rakover Y. ANE syndrome caused by mutated RBM28 gene: a novel etiology of combined pituitary hormone deficiency. *European journal of endocrinology / European Federation of Endocrine Societies*. 2010;162(6):1021-5.
143. Younkin SG, Scharpf RB, Schwender H, Parker MM, Scott AF, Marazita ML, et al. A genome-wide study of inherited deletions identified two regions associated with nonsyndromic isolated oral clefts. *Birth Defects Res A Clin Mol Teratol*. 2015;103(4):276-83.
144. Uslu VV, Petretich M, Ruf S, Langenfeld K, Fonseca NA, Marioni JC, et al. Long-range enhancers regulating Myc expression are required for normal facial morphogenesis. *Nature genetics*. 2014;46(7):753-8.
145. Wang X, Li X, Cheng Y, Sun X, Sun X, Self S, et al. Copy number alterations detected by whole-exome and whole-genome sequencing of esophageal adenocarcinoma. *Hum Genomics*. 2015;9:22.