UNIVERSITÉ GRENOBLE-ALPES

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE-ALPES

Spécialité : **Modèles, méthodes et algorithmes en biologie, santé et environnement**

Arrêté ministériel : ?

Présentée par

Thomas Karaouzene

Thèse dirigée par **Pierre Ray**

Thèse co-dirigée par **Nicolas Thierry-Mieg**

préparée au sein du laboratoire
et de l'école doctorale **"Ingénierie de la Santé, de la Cognition et Environnement" (EDISCE)**

Écrire le titre de la thèse ici

Thèse soutenue publiquement le 31 octobre 2017,
devant le jury composé de :

Université Grenoble Alpes

Préface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Delete line 6 if you only have one advisor

Remerciements

Résumé

Chapitre 2

Introduction

Chapitre 3

Investigation génétique et physiologique de la globozoospermie

Chapitre 4

Mise en place d'une stratégie pour l'analyse des données exomiques – application en recherche clinique

4.1 Intro

Comme vu précédemment, l'émergence du séquençage haut débit, avec notamment le WGS et le WES, a révolutionné les méthodes de recherche dans le cadre d'étude phénotype-génotype en permettant de manière rapide et à moindre coup le séquençage de la quasi totalité des gènes humains. Les causes de plusieurs centaines de pathologies ont pu être identifiées grâce à ces technique depuis leur premier succès pubilié en 2010 (Ng et al., n.d.). Dès lors, l'analyse des données issues du séquençage est devenu la clef dans la réussite de ces études.

Il existe de nombreux logiciels qui à partir des variants appelés effectuent les étapes d'annotation et de filtrage. C'est par exemple le cas d'Exomiser [TODO : insert ref and Exomiser describtion] ou encore de [TODO : insert at least one other soft]. La plupart de ces logiciels fonctionnent très bien, cependant tous prennent pour point de départ des variants appelés en amont. Ils ne contrôlent donc en aucune manière les étapes d'alignement et d'appel des variants. Or, comme il a été dit plus tôt, ces deux étapes constituent la bases de l'analyse [TODO insert ref] et les résultats

Dans ce chapitre, je détaillerai les résultats de 4 articles dont je suis coauteur :

1. **Whole-exome sequencing of familial cases of multiple morphological abnormalities of the sperm flagella (MMAF) reveals new DNAH1 mutations** : [todo]
2. **Homozygous mutation of PLCZ1 leads to defective human oocyte activation and infertility that is not rescued by the WW-binding protein PAWP** : Dans cet article j'ai, comme précédemment, effectué

l'integralité des analyses bioinformatiques des données d'exomes effectués sur deux frères infertiles présentant des échecs de fécondation.

3. **SPINK2 deficiency causes infertility by inducing sperm defects in heterozygotes and azoospermia in homozygotes** : Dans cet article j'ai effectuer non seulement l'intégralité des analyses bioinformatiques des données d'exomes de deux frères infertiles présentant un phénotype d'azoospermie mais aussi séquencer en Sanger les séquences codantes du gène *SPINK2* pour une parie des 611 individus analyser ainsi que contribué à l'extraction de l'ARN testiculaire des souris pour l'analyse fonctionelle du gène *Spink2* sur le modèle murin.

4. **** : [todo]

4.2 Résultats

4.2.1 Description de la pipeline

Notre pipeline d'analyse effectue l'ensemble des étapes allant de l'alignement des données jusqu'au filtrage des variants

1. **L'alignement** : L'alignement des *reads* le long du génome de référence est effectué par le logiciel MAGIC (Su et al., 2014). Celui-ci l'intégralité pour l'ensemble des analyses en aval l'ensemble des *reads* dupliqués et / ou s'alignant à plusieurs zone du génome. Au cours de cette étape, MAGIC va produire également quatre comptages pour chaque position couverte du génome : R+, V+, R- et V- :

   a. **R+ et R-** : Ces deux comptages correspondent au nombres de *reads forward* (+) et *reverse* (-) sur lesquels est observé l'allere de **référence** (R) à une position donnée.

   b. **V+ et V-** : À l'inverse de R+ et R-, ces comptages correspondent au nombres de *reads forward* et *reverse* sur lesquels est observé un allele de **variant** (V) à une position donnée.

2. **L'appel des variants** : Comme nous l'avons vu plus tôt, il est fortement conseillé d'effectuer l'appel des variants en tenant compte de l'aligneur choisi (Nielsen, Paul, Albrechtsen, & Song, 2011, M. A. DePristo et al. (2011), Lunter & Goodson (2011)). C'est pourquoi, nous avons conçu notre propre algorithme d'appel des variants spécialement conçu pour l'analyse des données de MAGIC. Ainsi, l'appel des variants sera directement basé sur les quatre comptages vu précédement. Tout d'abord, les positions ayant une

couverture < 10 sur l'un des deux *strands* sera considérée comme de faible qualité, celles aynant une couverture < 10 sur les deux *strands* seront exclus. Ensuite pour chaque variant, des appels indépendant seront effectués pour chaque *strand*. L'appel final sera une synthèse de ces deux appels où seul les cas où ces deux appels sont concordants seront considérés comme de bone qualité.

3. **L'annotation** : Chaque variant retenu sera ensuite annoté tout d'abord par le logiciel *variant effect predictor* (VEP) (W. McLaren et al., 2016) qui nous indiquera pour chaque variant la conséquence que celui-ci aura sur la séquence codante de l'ensemble des transcrits Ensembl qu'il chevauche (**Figure :** 4.1) (**Table :** 4.1). Suite à cela nous ajoutons, lorsque celle-ci est disponible, la fréquence du variant dans les bases de données ExAC (Lek et al., 2016), ESP600 [TODO] et 1000Genomes [TODO] donnant ainsi une estimation de sa fréquence dans la population générale. De même, la particularité de cette pipeline est qu'elle conserve l'ensemble des variants identifiés dans les études effectués précédement permettant d'ajouter aux annotations la fréquences d'un variant chez les individus déjà séquencé et donc la fréquence d'un variant dans chaque phénotype étudié créant ainsi une base de données interne qui pourra servir de contrôle dans les études ulterieur.

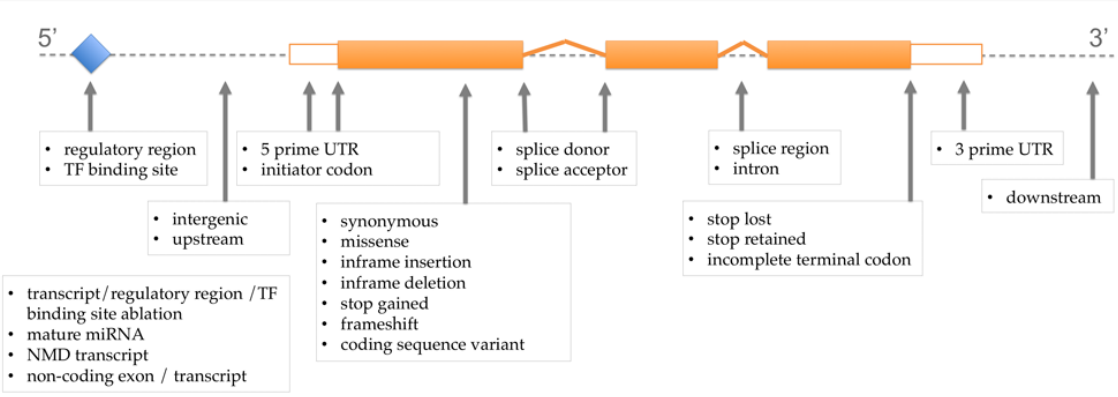


Figure 4.1 – Listes des différentes conséquences prédites par VEP et leurs positionement sur le transcrit d'après [VEP site](http ://www.ensembl.org/info/genome/variation/consequences.jpg)

4. **Le filtrage des variants** : L'étape de filtrage est extremement importante si l'on souhaite analyser de manière efficace les données provenant de WES. C'est pourquoi elle occupe une place importante dans notre pipeline. L'intégralité des paramètres de cette étape peuvent être modifier par l'utilisateur de sorte à faire correspondre les critères de filtre aux bsoins de l'étude. Afin de rendre son utilisation le plus efficace possibe, nous avons souhaité définir des paramètres par défauts pertinent dans la plupart des étude de séquençage exomique de sorte que à moins que le contraire ne soit spécifié, seul les variants impactant les transcrits codant pour une protéine sont conservés. De même les variants synonymes ou affectant les séquences UTRs sont filtrés ainsi que les variants

ayant une fréquence $\geq 1\%$ dans les bases dans l'une des bases données (ExAC, ESP6500 ou 1KH). Aussi, pour un phénotype donné, l'ensemble des variants observés chez les individus étudiés présentant un phénotype différent sont de même enlevés de la liste finale.

4.2.2 Utilisation de la pipeline dans des cas familiaux :

Description des familles

Dans cette partie, je me concentre sur l'analyse bioinformatique des résultats des séquençages exomiques effectués entre 2012 et 2014 de 13 individus infertiles provenant de 6 familles différentes. Parmi celles-ci, 3 phénotypes différents ont été observés :

1. **L'Azoospermie :** Comme nous avons pu le voir, l'azoospermie est un phénotype d'infertilité masculine caractérisé par l'absence de spermatozoïde dans l'éjaculat.

2. **Echec de fécondation :** Ce phénotype d'infertilité se caractérise par l'incapacité des spermatozoïdes à féconder l'ovocyte.

3. **MMAF** : Le syndrome MMAF (*multiple morphological abnormalities of the sperm flagella*) caractérise comme son nom l'indique les patients présentant une majorité de spermatozoïdes atteins par une mosaïque d'anomalie morphologique du flagelle.

Un récapitulatif des familles et de leur phénotype est disponible dans la table 4.2.

Table 4.2 – Tableau recapitulatif des familles séquencées et de leur phénotype

Familly	Individuals	Phenotype	Year	Plateform	Place
Az	2	Azoospermia	2012	Illumina HiSeq2000	Mount Sinai Institut
FF	2	Fertilization failure	2014	Illumina HiSeq2000	Genoscope (Evry)
MMAF1	2	MMAF	2014	Illumina HiSeq2000	Genoscope (Evry)
MMAF2	2	MMAF	2014	Illumina HiSeq2000	Genoscope (Evry)
MMAF3	2	MMAF	2014	Illumina HiSeq2000	Genoscope (Evry)
MMAF4	3	MMAF	2014	Illumina HiSeq2000	Genoscope (Evry)

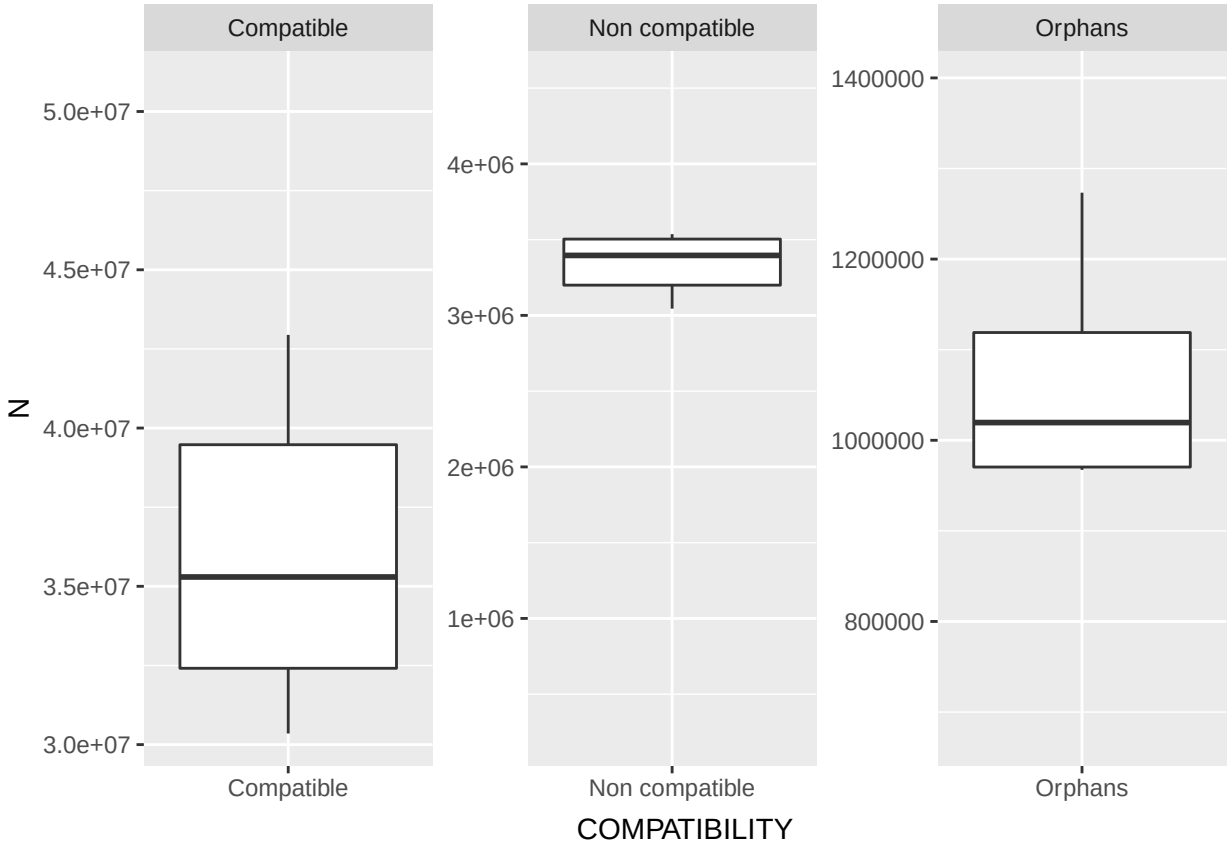
Resultats des exomes

Résultat de l'alignement L'ensemble de nos exomes ayant été réalisés en *paired-end*, les deux extrémités de chaques fragments sont séquencés créant ainsi deux *reads*. Après avoir aligné indépendemment les deux *ends* d'un même *read*, seul ceux pour lesquels les deux *ends* présentent un alignement "compatible" sont conservés. Un alignement est dit "compatible" lorsque les deux *ends* s'alignent face à face (une sur le *strand* + et l'autre sur le *strand* - et couvrent une zone ne faisant pas plus de 3 fois

Table 4.1 – Liste des conséquences prédites par VEP avec leur description et impact associée (trouver comment la faire tenir

Consequence	Description
Transcript ablation	A feature ablation whereby the deleted region includes a transcript
Splice acceptor variant	A splice variant that changes the 2 base region at the 3' end of an
Splice donor variant	A splice variant that changes the 2 base region at the 5' end of an
Stop gained	A sequence variant whereby at least one base of a codon is changed
Frameshift variant	A sequence variant which causes a disruption of the translational r
Stop lost	A sequence variant where at least one base of the terminator codon
Start lost	A codon variant that changes at least one base of the canonical sta
Transcript amplification	A feature amplification of a region containing a transcript
Inframe insertion	An inframe non synonymous variant that inserts bases into in the c
Inframe deletion	An inframe non synonymous variant that deletes bases from the co
Missense variant	A sequence variant, that changes one or more bases, resulting in a
Protein altering variant	A sequence_variant which is predicted to change the protein encod
Splice region variant	A sequence variant in which a change has occurred within the regio
Incomplete terminal codon variant	A sequence variant where at least one base of the final codon of an
Stop retained variant	A sequence variant where at least one base in the terminator codon
Synonymous variant	A sequence variant where there is no resulting change to the encod
Coding sequence variant	A sequence variant that changes the coding sequence
Mature miRNA variant	A transcript variant located with the sequence of the mature miRN
5 prime UTR variant	A UTR variant of the 5' UTR
3 prime UTR variant	A UTR variant of the 3' UTR
Non coding transcript exon variant	A sequence variant that changes non-coding exon sequence in a non
Intron variant	A transcript variant occurring within an intron
NMD transcript variant	A variant in a transcript that is the target of NMD
Non coding transcript variant	A transcript variant of a non coding RNA gene
Upstream gene variant	A sequence variant located 5' of a gene
Downstream gene variant	A sequence variant located 3' of a gene
TFBS ablation	A feature ablation whereby the deleted region includes a transcript
TFBS amplification	A feature amplification of a region containing a transcription factor
TF binding site variant	A sequence variant located within a transcription factor binding sit
Regulatory region ablation	A feature ablation whereby the deleted region includes a regulatory
Regulatory region amplification	A feature amplification of a region containing a regulatory region
Feature elongation	A sequence variant located within a regulatory region
Regulatory region variant	A sequence variant located within a regulatory region
Feature truncation	A sequence variant that causes the reduction of a genomic feature,
Intergenic variant	A sequence variant located in the intergenic region, between genes

la taille médiane de l'insert.



qsswxfdcsdfgds dfsqfdqfg fdrgsfdgh fdgzrfg fgdg s

```
## <ggproto object: Class ScaleDiscretePosition, ScaleDiscrete, Scale>
##     aesthetics: x xmin xmax xend
##     axis_order: function
##     break_info: function
##     break_positions: function
##     breaks: waiver
##     call: call
##     clone: function
##     dimension: function
##     drop: TRUE
##     expand: waiver
##     get_breaks: function
##     get_breaks_minor: function
##     get_labels: function
##     get_limits: function
##     guide: none
##     is_discrete: function
##     is_empty: function
```

```
##     labels: AZ1 AZ2 FF1 FF2 MMAF1.1 MMAF1.2 MMAF2.1 MMAF2.2 MMAF3.1  ...
##     limits: NULL
##     make_sec_title: function
##     make_title: function
##     map: function
##     map_df: function
##     n.breaks.cache: NULL
##     na.translate: TRUE
##     na.value: NA
##     name: waiver
##     palette: function
##     palette.cache: NULL
##     position: bottom
##     range: <ggproto object: Class RangeDiscrete, Range>
##         range: NULL
##         reset: function
##         train: function
##         super:  <ggproto object: Class RangeDiscrete, Range>
##     range_c: <ggproto object: Class RangeContinuous, Range>
##         range: NULL
##         reset: function
##         train: function
##         super:  <ggproto object: Class RangeContinuous, Range>
##     reset: function
##     scale_name: position_d
##     train: function
##     train_df: function
##     transform: function
##     transform_df: function
##     super:  <ggproto object: Class ScaleDiscretePosition, ScaleDiscrete, Scale>
```

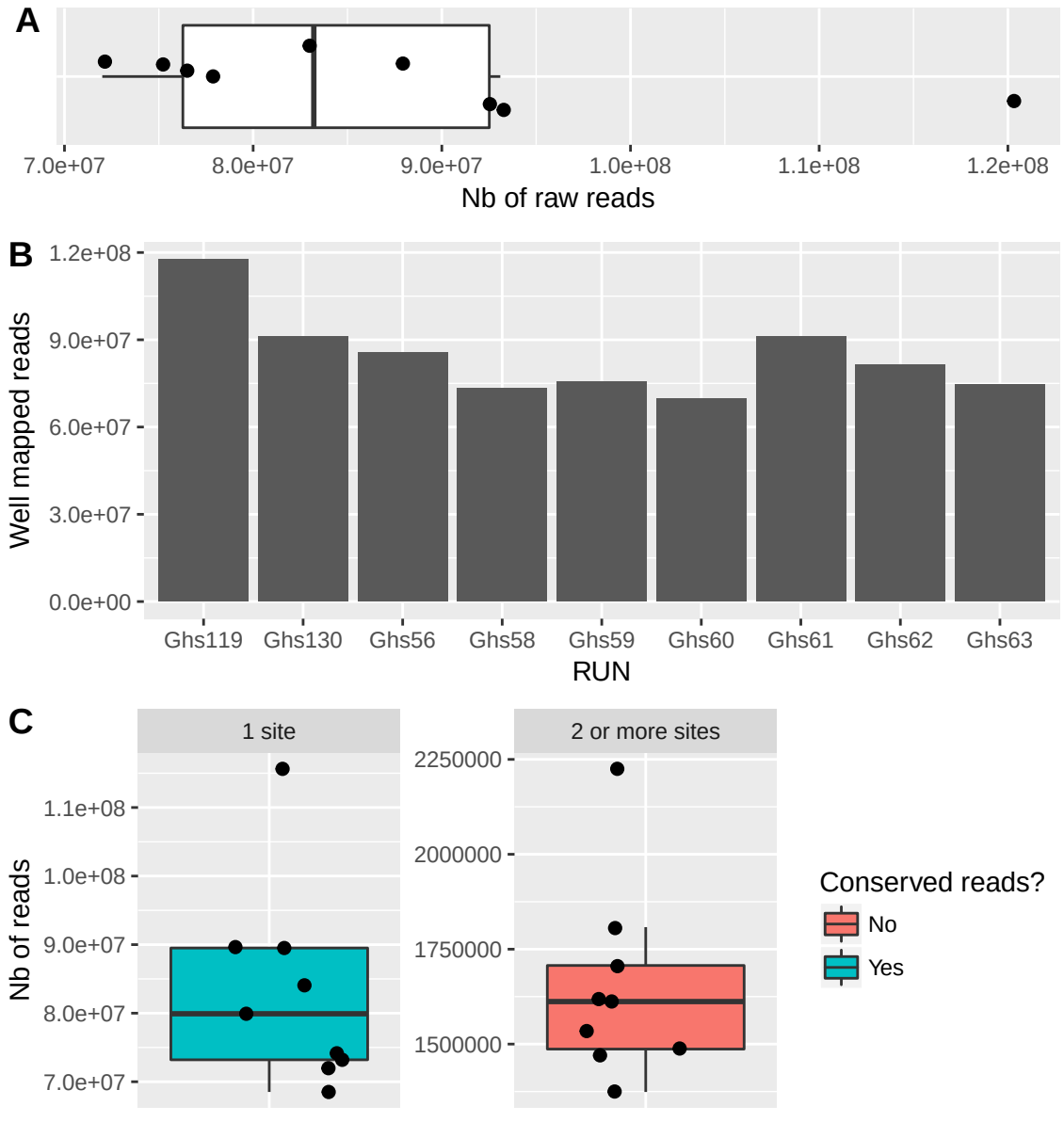


Figure 4.2 – Résultats du mapping des reads pour chaque Patients : **A** : Cette boite à moustache montre le nombre de *reads* bruts générés pour chaque patients (représentés par les points) au cours de l'étape de séquençage. On constate que ce nombre reste pour chaque patient dans le même ordre de grandeur sauf pour un des frères de la famille AZ qui contient presque 35 millions de *reads* en plus que la mediane. **B** : Ce diagramme en barre montre le nombre de *reads* correctement allignés pour chaque patients. On peut donc constater que la quasi totalité de *reads* générés ont été correctement allignés pour l'ensemble des patients. **C** : Ces deux boites à moustaches montrent .... On constate que la grande majorité des *reads* ne mappent qu'à un seul site (boite bleue), les autres (boite rouge) seront d'ailleur écartés des analyses à suivre

Résultat de l'appel des variants

Résultats de l'annotation plot du comptage des impact / patients

plot du nombre de variants ayant un match dans une des base de donnée (ExAC...)
plot de la distribution des MAF associées des MA

Résultats du filtrage Pour l'ensemble des individus de ces quatre familles nous avons appliqué notre pipeline d'analyse de sorte à obtenir pour chaque patient une liste de SNV et d'indel avec leur génotype associé (**Figure :** 4.3).

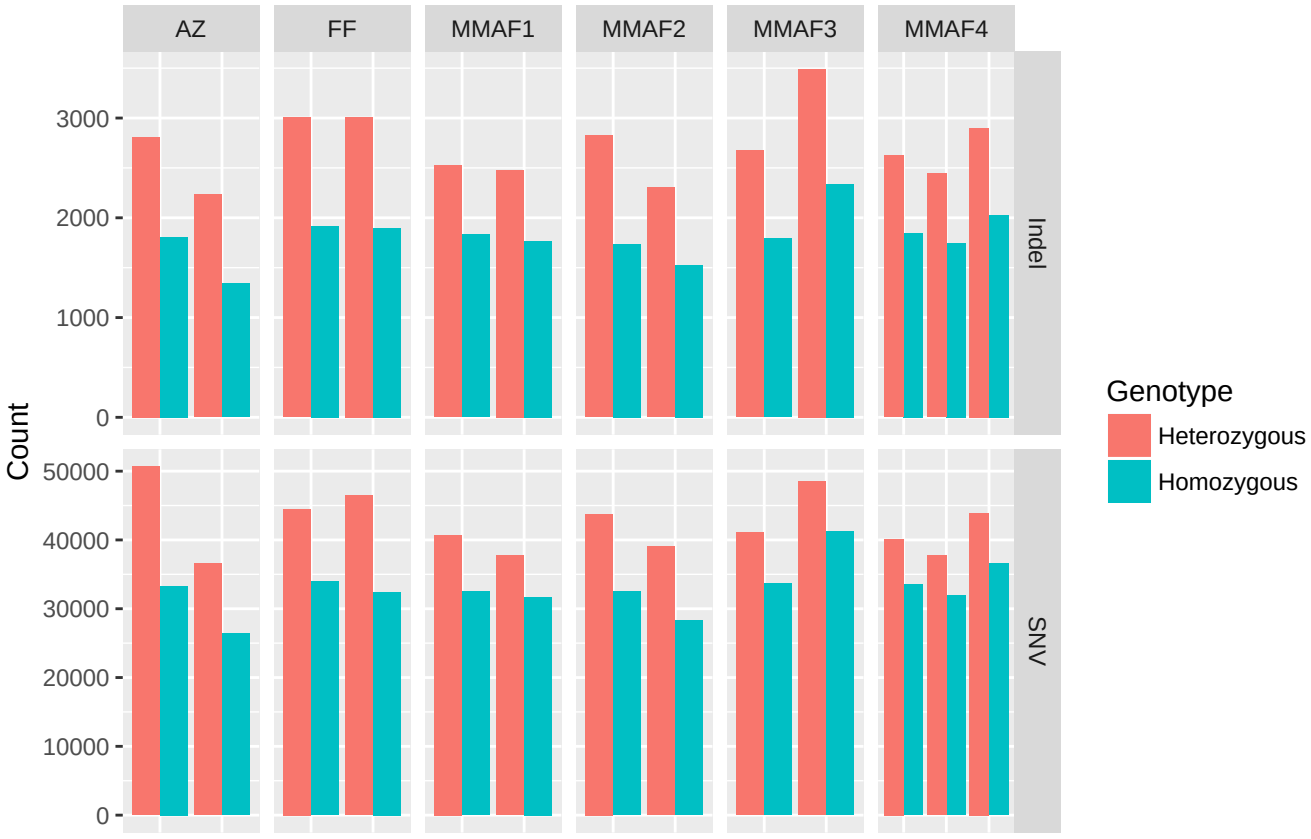


Figure 4.3 – Comptage des SNVs et indels retrouvés par patients avec leur génotypes associés

Ensuite, afin de ne conserver que les variants ayant la plus forte probabilité d'être responsable du phénotype nous avons appliqué succesivement six filtres :

1. **L'union des variants :** Dans ces différentes études, nos patients ont à chaque fois au moins un frère présentant le même phénotype. Ainsi nous avons pu formuler l'hypothèse d'une cause génétique commune entre les différents frères

d'une même famille et donc filtrer l'ensemble des variants qui ne sont pas partagés par l'ensemble des membre de la fraterie.

2. **Genotype des variants :** Dans ces études, nous avons emmis l'hypothèse d'une transmission recessive du phénotype. Ainsi, seul les variants homozygotes ont été conservés. Ce filtre est le plus efficace du pipeline en permettant de filtrer entre 38814 et 53448 variants par individus (**Figure :** 4.3, 4.6).

3. **Impact du variant :** Afin de ne conserver que les variants ayant un effet potentiellement tronquant sur la protéine, nous avons filtré les variants intonique et ceux tombant dans les sequences UTRs. De même les variants synonymes ne sont pas conservés (exeptés ceux se trouvant proches des régions d'épissage) car ceux-ci n'ont aucun effet sur séquences protéique. Pour les variants faux sens (changement d'un seul aa de la séquence protéique) il est plus difficile de se décider [TODO insert citation] nous avons donc utilisé les logiciels SIFT et Polyphen et filtré l'ensemble des fauxsens prédit comme *tolerated* par SIFT et *benign* par Polyphen.

4. **Transcrits NMD :** Le mécanisme NMD (*nonsense-mediated decay*) a pour but de controler la qualité des ARNm cellulaires chez les eucaryotes (Y.-F. Chang, Imam, & Wilkinson, 2007) en éliminant les ARNm qui comportent un codon stop prématuré (Baker & Parker, 2004), pouvant être le résultat d'une erreur de transcription, d'une mutation ou encore d'une erreur d'épissage. Il est donc peu probable que les variants présents sur transcrits annotés NMD soient responsables du phénotype. Nous avons donc filtré l'ensemble des variants chevauchant **uniquement** des transcrits annotés NMD. Cette étape de filtre permet à elle seule de filtrer systematiquement les variants de 2587 à 3212 transcrits (**Figure :** 4.4) en fonction des individus soit, entre 7261 et 10872 variants différents par individus (**Figure :** 4.6).

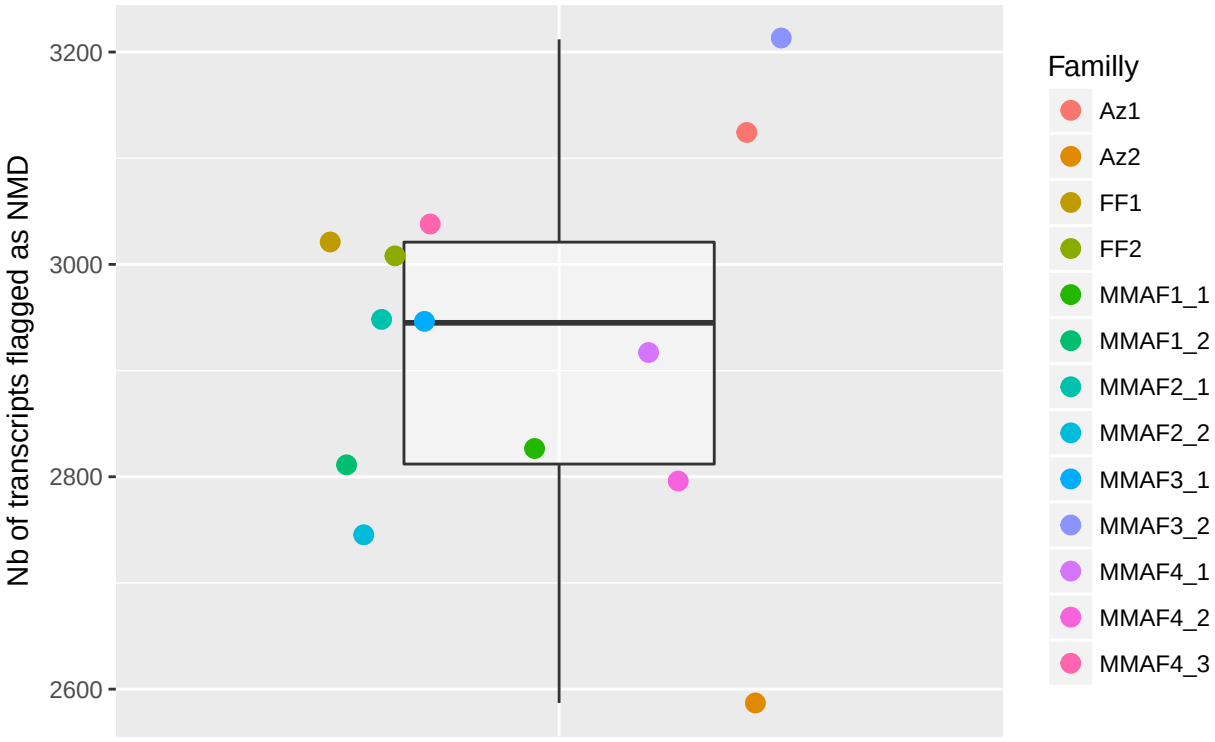


Figure 4.4 – Nombre de transcrits filtrés car ils sont annotés NMD : Chaque point représente un individu séquencé, la couleur et la forme du point dépend de la famille d'origine de l'individu

5. **Frequence des variants :** La fréquence d'un variant dans la population générale est un moyen rapide d'avoir un avis sur l'effet délétère de celui-ci. En efft, il est peu probable qu'un retrouvé fréquement dans la population générale soit causal d'une pathologie sévère. Ainsi nous avons filtré pour l'ensemble de nos patients l'ensemble des variants ayant une fréquence ≥ 0.01 dans l'une des trois bases de données que sont ExAC, ESP et 1KG.

6. **Présence des variants dans la cohorte contrôle :** Au cours de nos différentes études, nous avons été ammené à séquencé 134. L'ensemble de ces individus peuvent être soit sains soit présenter l'un des 6 phénotypes étudié au cours de nos différentes études (**Table : ??**). Ces phénotypes étant très différent, il n'est pas abérant d'emmetre l'hypothèse qu'ils que leurs causes génétiques soient diffrentes. De même, les variants recherché étant rares, il est peu probable qu'un individu porte les variants de deux phénotypes différents. Ainsi, pour chacune des 6 familles, nous avons pu constituer une cohorte contrôle composée dans l'ensemble des patients précédemment analysés et ne présentant pas le même phénotype que celui étudié dans la famille (**Figure :** 4.5). Dès lors, nous avons put filtrer l'ensemble des variants retrouvés à la fois chez nos patients et observés à l'état homozygote dans la cohorte contrôle.

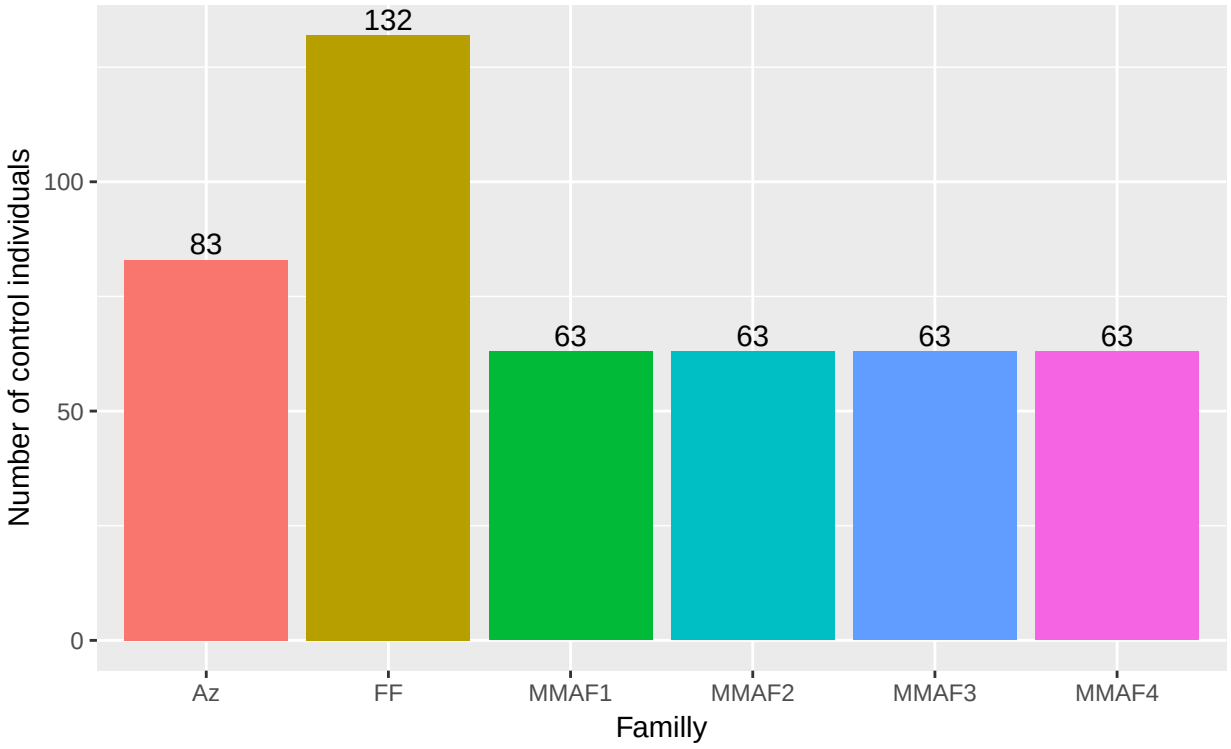


Figure 4.5 – Nombre d'individus la cohorte contrôle constituée pour chaque famille de l'analyse : Ici, chaque barre représente une famille et sa hauteur est déterminée par le nombre d'individus composant la cohorte contrôle à laquelle elle a été confronté. Chaque individus de la cohorte contrôle a été séquencés en WES par notre équipe. Afin d'être considérer comme "contrôle" un individus doit être sain ou présenter un phénotype d'infertilité différent de la famille étudiée. Par exemple, un individus MMAF pourra servir de contrôle aux familles AZ et FF mais pas aux familles MMAF1-4

Afin de comparer le pouvoir discriminant de chacun de ces filtres, nous avons compté le nombre de variant filtrés par chacun d'entre eux indépendamments des autres (**Figure :** 4.6).

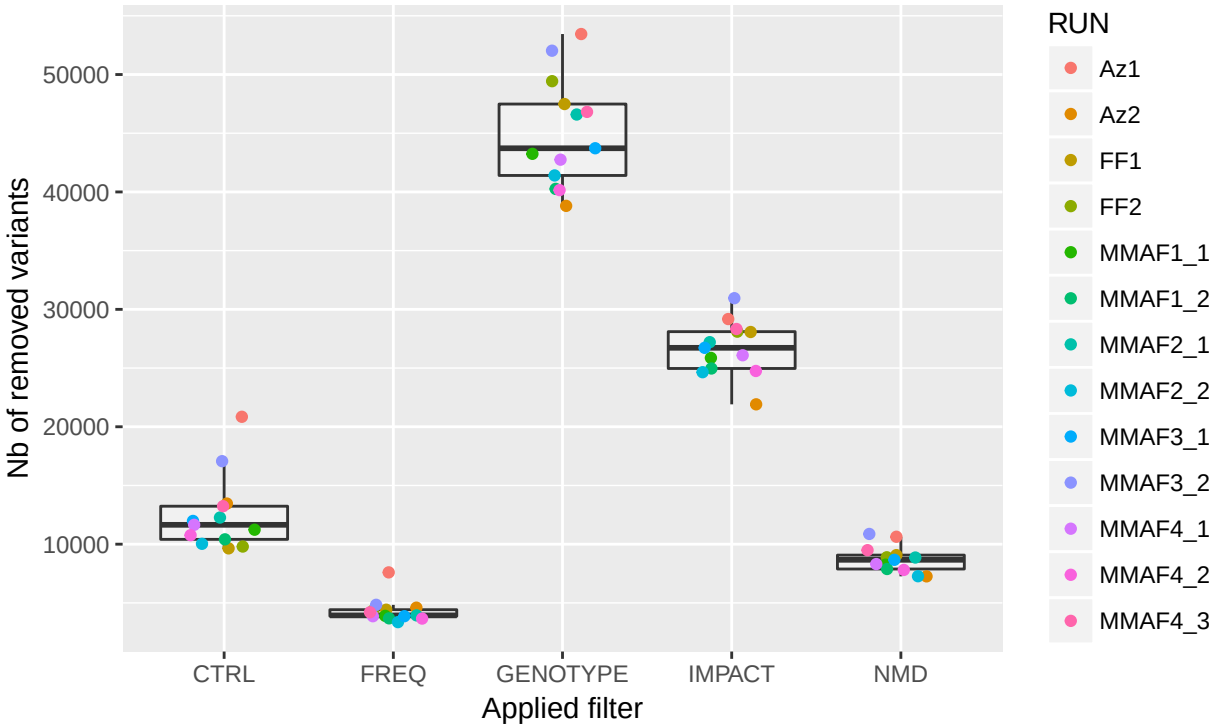


Figure 4.6 – Comparaison du pouvoir discriminant de chaque filtre employé TODO mettre union : Les boites à moustache représentent la quantité de variants filtrés par chacun des 6 filtres utilisés sur l'ensemble des patients. Chaque point représente un patient et sa hauteur informe du nombre exact de variant filtrés par uin filtre donné. Comme on peut le voir, le filtre le plus discriminant est celui consistant à filtrer l'ensemble des variants hétérozygotes ce qui n'est pas surprenant puisqu'ils représentent entre ... et ... pourcents des variants observés pour chaque patients

Après avoir effectuer l'ensemble de ces filtres, seuls quelques variants subsistent nous permettant d'obtenir unle liste de gènes restrainte pour chaque famille (**Table :** 4.3). Ainsi, la cause génétique expliquant le phénotype d'une famille a pu être mis en évidence dans ... familles sur ... [TODO] (**Figure :** 4.7). Il est a noté que l'ensemble des familles pour lesquelles la cause génétique a été identifiée présente un historique consanguin [figure arbre] ce qui n'était pas le cas pour les ... autres. Cette consanguinité observée dans une partie des famille nous a permi de justifier l'exclusion des variants hétérozygotes. En revanche pour les ... autres fa milles, rien ne justifiait un tel filtre. Ainsi, pour celles-ci il est probable que les variants responsables se soient vu exclus par ce filtre. C'est pourquoi, notre équipe se concentre actuellement sur les variants hétérozygotes de ces familles.

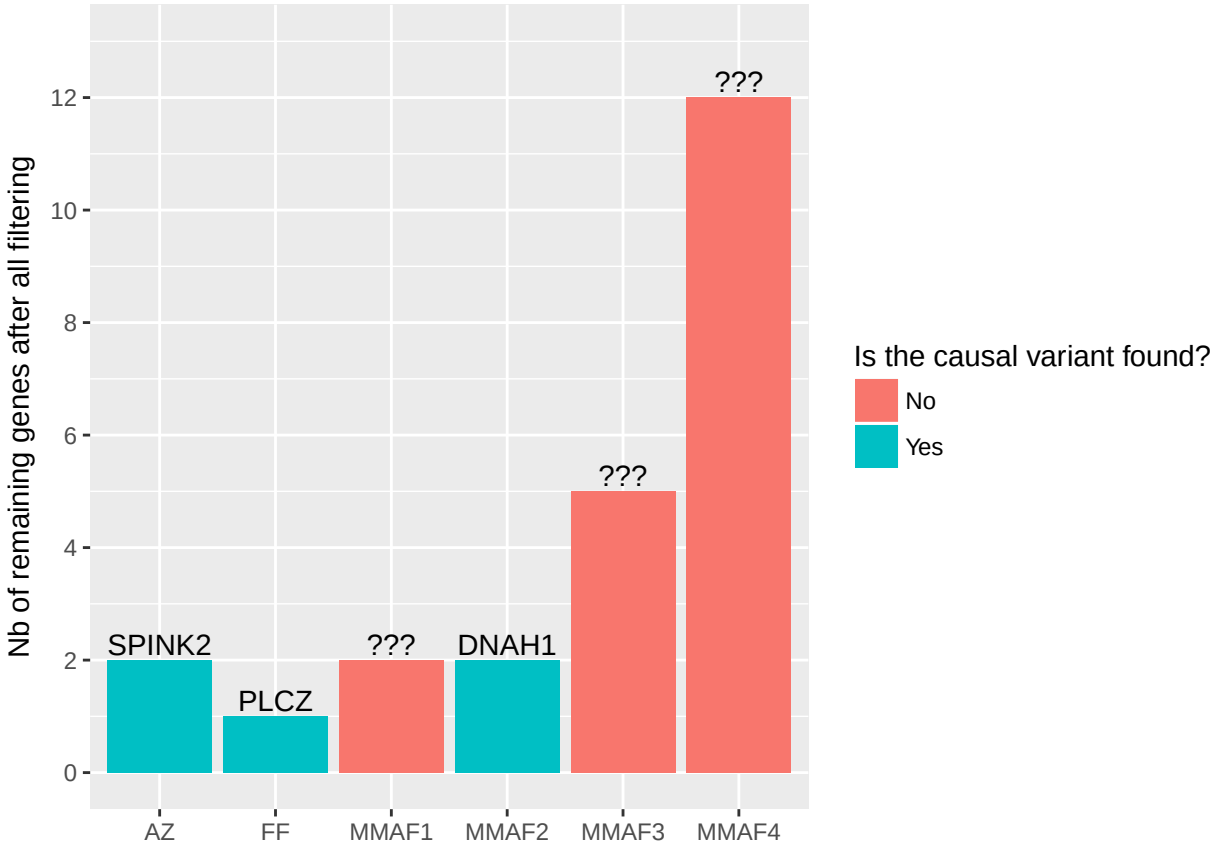


Figure 4.7 – Nombre de gènes passant l'ensemble des filtres par famille : Chaque barre représente une des familles analysées. La hauteure de cette barre correspond au nombre de gènes ayant passé l'ensemble des filtres pour chaque famille. Les barres bleues caractérisent les familles pour lesquelles le gène responsable de la pathologie a été identifié parmi la liste de gène (dans ce cas le symbole du gène est écrit au dessus de la barre). Les barres rouges indique qu'aucun des gènes ayant passé les filtres pour ne semble expliquer le phénotype (dans ce cas il est écrit "???" au dessus de la barre)

4.2.3 Etude d'une large cohorte de patients MMAF

Dans cette partie, nous allons détailler les analyses effectuées sur une cohorte de 62 individus présentant le phénotype MMAF pour lesquels nous avons effectués un séquençage WES. Nous avons ainsi pu appliquer notre pipeline d'analyse afin d'appeler et annoter les variants de ces 62 individus (**Figure :** 4.8).

Table 4.3 – Tableau des gènes ayant passé l'ensemble des filtres pour les fifférentes familles

AZ	FF	MMAF1	MMAF2	MMAF3	MMAF4
GUF1					
SPINK2					
	PLCZ1				
		PLA2G4B			
		JMJD7-PLA2G4B			
			MYH11		
			DNAH1		
				WEE2	
				PCSK5	
				ZFYVE28	
				GBP2	
				FCGR3A	
					MMP9
					TGIF2
					ZNF469
					HYDIN
					MTSS1L
					CDH23
					CCDC37
					DAPK1
					SEMA5B
					SLC13A3
					TMEM231
					ZNF276

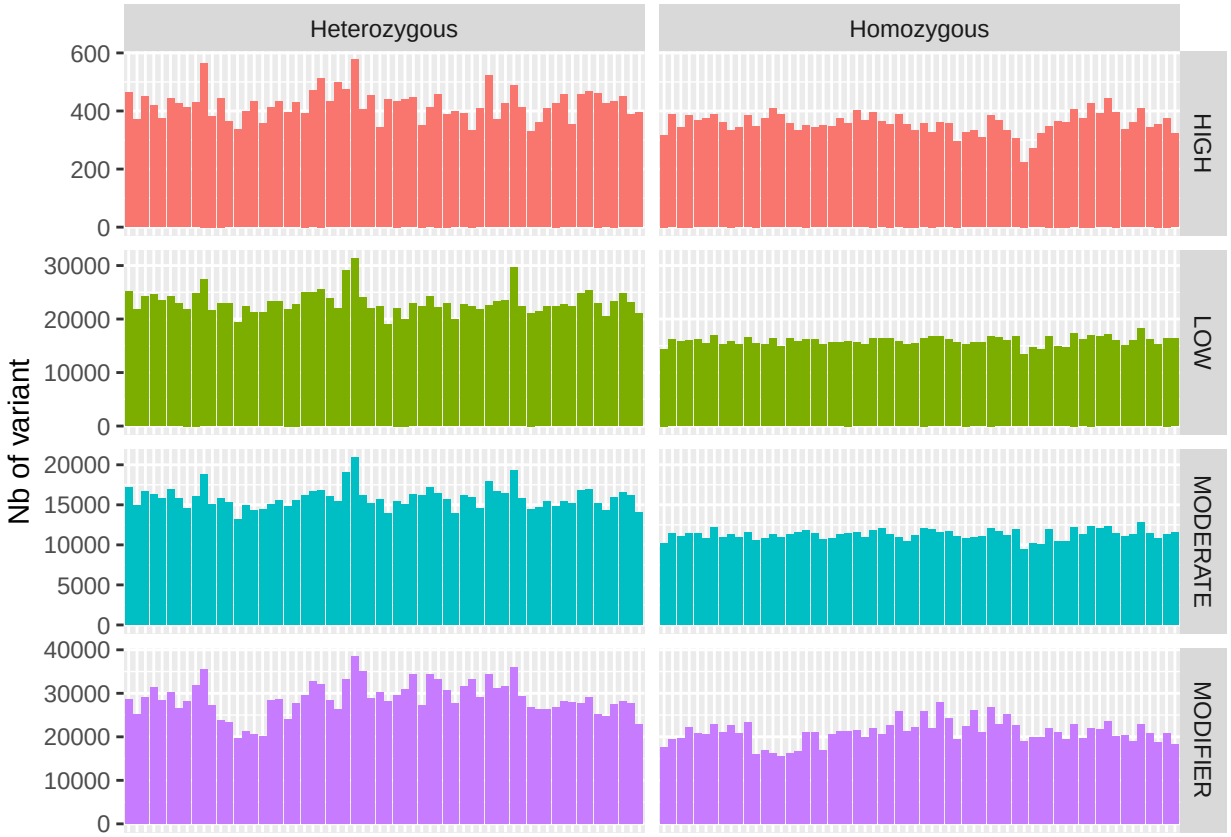


Figure 4.8 – Comptage des variants pour chaque individus avec leur génotype et l'impact prédite par VEP : VEP possède quatre niveaux d'impact pour ses variants : **HIGH** : variant ayant une forte probabilité de causer des dommages sévères à la protéine, **MODERATE** : Variants non-tronquant pouvant tout de même affecté la protéine, **LOW** : variant ayant peu de chance d'alterer la protéine, **MODIFIER** : Variants affectant les régions non codantes du transcrits et dont l'impact sur la protéine est difficile à prévoir. Chaque barre représente le comptage pour un individus

Les filtres utilisés ont été les mêmes que ceux détaillés dans l'études des cas familiaux, c'est à dire que seul les variants homozygotes ayant une fréquence ≤ 0.01 dans la population générales et n'étant pas observés dans la cohorte contrôle de 63 individus furent conservés. De même, les variants synonymes, impactant la séquences UTR ou chevauchant uniquement des transcrits annotés NMD par VEP on été filtrés. Ainsi, ces différents filtres nous ont permis d'obtenir une liste de 1369 SNVs (entre 1 et 77 différents par patients) et de 211 indels (entre 1 et 11 différents par patients) (**Figure :** 4.9). Cet ensemble de variant nous a alors permis d'obtenir des listes comprenant entre 1 et 74 gènes différents par patients (**Figure :** 4.10 - **A**) constituant ainsi un total de 1316 gènes différents pour l'ensemble des patients. Parmi ceux là, 1121 (85%) ont été retrouvés mutés chez un seul patients tandis que 195 (15%) ont été retrouvé

chez au moins 2 patients (**Figure :** 4.10 - **B**).

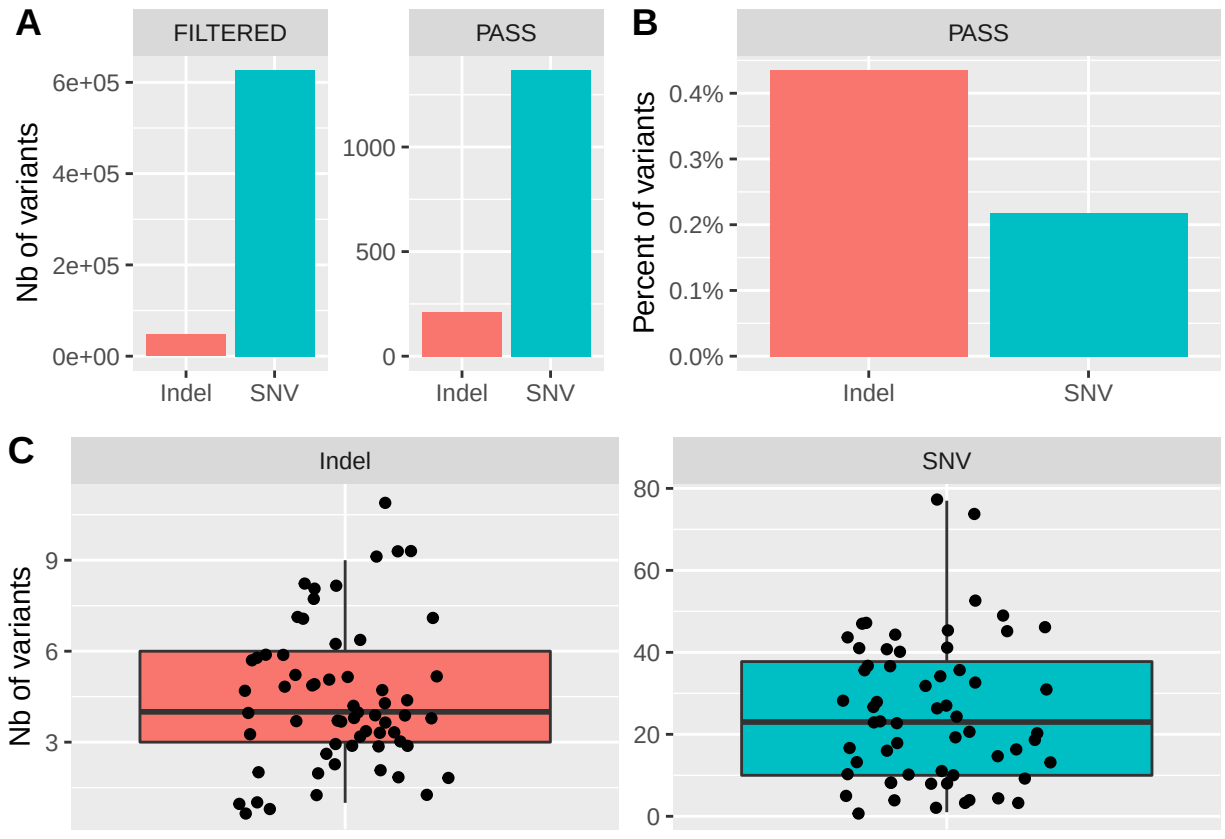


Figure 4.9 – Comptage des variants filtrés : **A** : Comptage des SNVs et Indels ayant été filtrés (FILTERED) et ayant passé les filtres (PASS), **B** : Pourcentage des SNVs et indels ayant passé les filtres, **C** : Comptage pour chaque individus du nombre de SNVs et d'indels ayant passé les filtres. Chaque point représente le comptage pour un individus

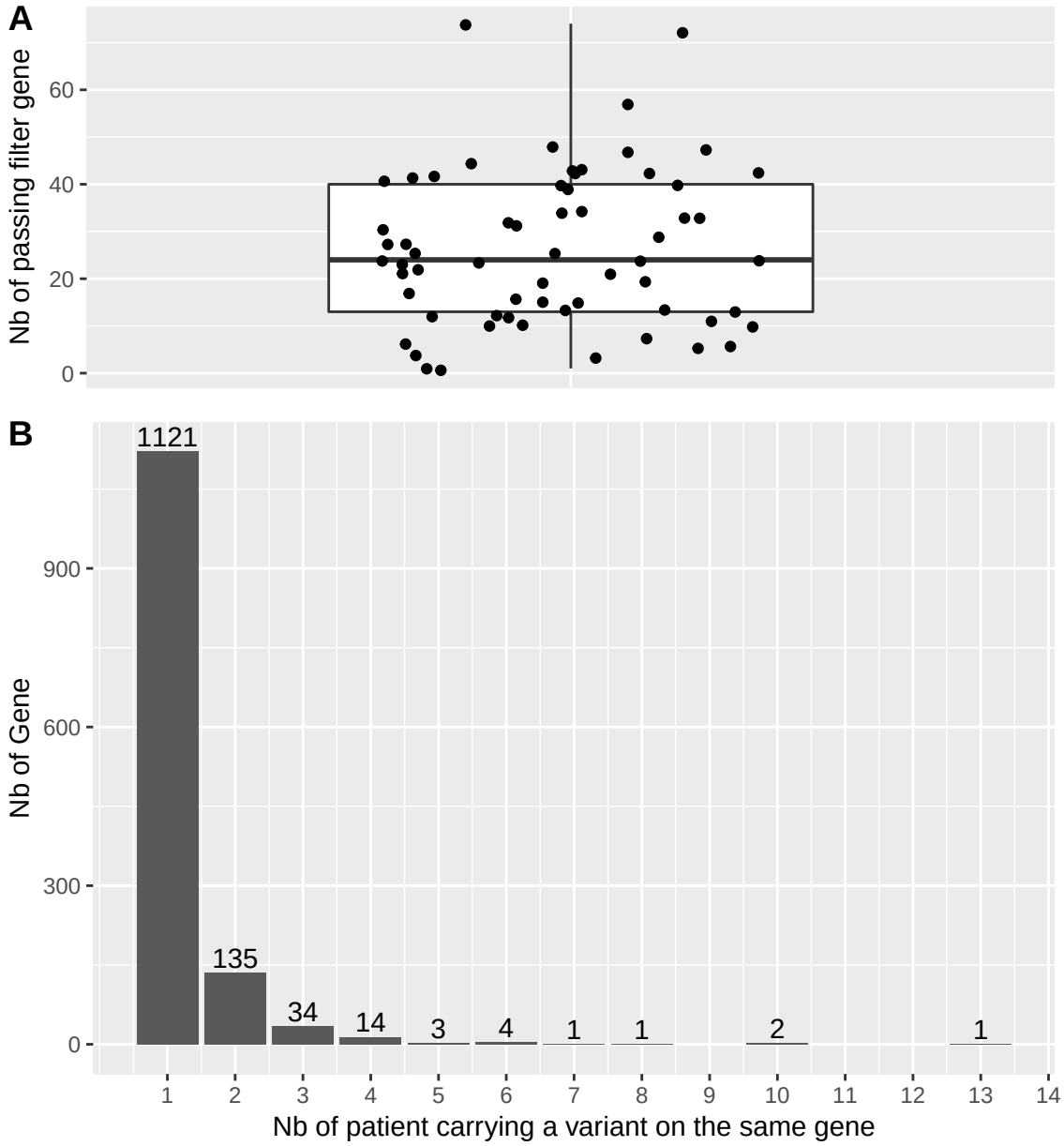


Figure 4.10 – Comptage des variants filtrés : **A** : Comptage des SNVs et Indels ayant été filtrés (FILTERED) et ayant passé les filtres (PASS), **B** : Pourcentage des SNVs et indels ayant passé les filtres, **C** : Comptage pour chaque individus du nombre de SNVs et d'indels ayant passé les filtres. Chaque point représente le comptage pour un individus

Le flagelle spermatique pouvant s'apparenté à un cil, nous avons ensuite comparés ces gènes avec une liste de 371 gènes prédits comme faisant parti du ciliome [TODO : insert ref]. Ainsi, sur notre ensemble de gène ayant passé les filtres 31 sont prédit comme faisant parti du ciliome (**Table :** 4.4).

Table 4.4 – Gènes ayant passé les filtres et annotés comme faisant partie du ciliome

Gene	Patient carrying a variant on the gene	Ciliome evidence
LRRC43	4	No evidence from previous studies
ARMC2	3	No evidence from previous studies
WDR52	3	Strong evidence from previous studies
AK7	2	Strong evidence from previous studies
EFCAB6	2	Strong evidence from previous studies
CCDC146	2	Strong evidence from previous studies
TTC29	2	Strong evidence from previous studies
KIAA0556	1	No evidence from previous studies
KIF9	1	Strong evidence from previous studies
FBXO15	1	Weak evidence from previous studies
C21orf59	1	Strong evidence from previous studies
FAM81B	1	Strong evidence from previous studies
WDR16	1	Strong evidence from previous studies
CCDC147	1	Strong evidence from previous studies
KIF6	1	Strong evidence from previous studies
SPAG17	1	Weak evidence from previous studies
C6orf118	1	Strong evidence from previous studies
RSPH9	1	Strong evidence from previous studies
KIAA0319	1	No evidence from previous studies
SPEF2	1	Strong evidence from previous studies
C6	1	Weak evidence from previous studies
ZMYND10	1	Strong evidence from previous studies
MIPEP	1	Weak evidence from previous studies
PROM1	1	Strong evidence from previous studies
DLEC1	1	Strong evidence from previous studies
CCDC65	1	Strong evidence from previous studies
HYDIN	1	Strong evidence from previous studies
C21orf58	1	No evidence from previous studies
SLFN13	1	Weak evidence from previous studies
ACYP1	1	No evidence from previous studies
STK33	1	Strong evidence from previous studies

Ces analyses nous ont permis de mettre en évidence certains candidats évidents (TODO : insert table avec candidats évidents) nous permettant ainsi d'identifier la cause génétique de ... patients soit ...% de notre cohorte (TODO).

Chapitre 5

MutaScript

Conclusion

Chapitre 6

The First Appendix

References

Baker, K. E., & Parker, R. (2004). Nonsense-mediated mRNA decay : terminating erroneous gene expression. *Current Opinion in Cell Biology*, *16*(3), 293–9. http://doi.org/10.1016/j.ceb.2004.03.003

Chang, Y.-F., Imam, J. S., & Wilkinson, M. F. (2007). The Nonsense-Mediated Decay RNA Surveillance Pathway. *Annual Review of Biochemistry*, *76*(1), 51–74. http://doi.org/10.1146/annurev.biochem.76.050106.093909

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... Pritchard, E. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. http://doi.org/10.1038/ng.806

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., ... Exome Aggregation Consortium, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–91. http://doi.org/10.1038/nature19057

Lunter, G., & Goodson, M. (2011). Stampy : A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, *21*(6), 936–939. http://doi.org/10.1101/gr.111120.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., ... Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. http://doi.org/10.1186/s13059-016-0974-4

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., ... Bamshad, M. J. (n.d.). Exome sequencing identifies the cause of a Mendelian disorder. http://doi.org/10.1038/ng.499

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, *12*(6), 443–51. http://doi.org/10.1038/nrg2986

Su, Z., Łabaj, P. P., Li, S. S., Thierry-Mieg, J., Thierry-Mieg, D., Shi, W., ... Shi, L. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology*, *32*(9), 903–14. http://doi.org/10.1038/nbt.2957