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 : ??**). 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.

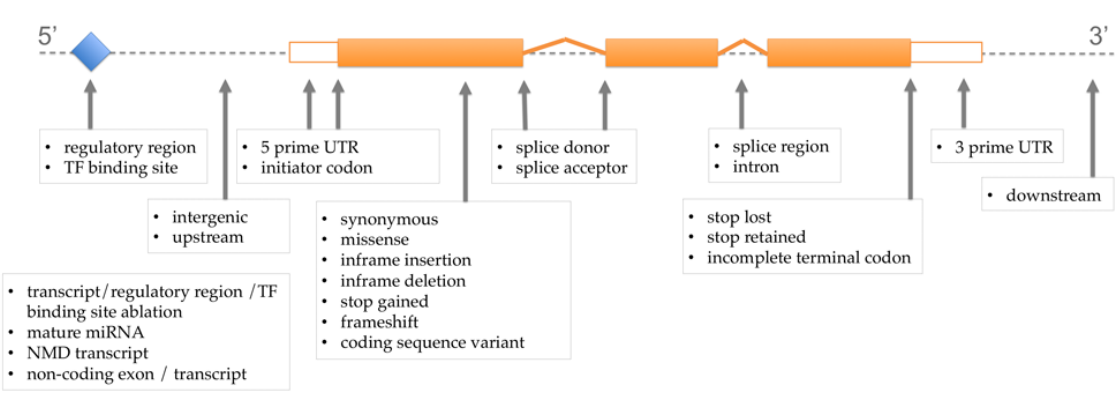

Figure 4.1 – Listes des différentes conséquences prédites par VEP et leurs positionement sur le transcrit d'après [VEP site](http ://www.ensembl.org/info/genome/variation/consequences.jpg)

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 4.1.

Table 4.1 – 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 les Pour rappel, l'alignement consiste à repositionner-l'ensemble des *reads* générés au cours de l'étape de séquençage le long d'un génome de référence. La quantité de *reads* peut varier en fonction de plusieurs paramètres et n'est donc pas éguale pour chaque patient bien que l'ordre de grandeur reste le même exceptés pour les deux frères AZ1 et AZ2 pour lesquels on a près de 3 fois plus de *reads* que pour les autres patients (**Figure :** 4.2 - **A**). Ceci peut être expliqué car ces deux patients sont les deux seuls à voir été séquencé au Mount Sinaï Institut or leur protocol d'amplification contient un nombre de cycles de PCR superieur à ceux appliqué au Génopole d'Evry où ont été séquencé les autres patients (**Table :** 4.1), L'ensemble de nos exomes ayant été réalisés en *paired-end*, les deux extrémités de chaques fragments sont séquencés chaque *end* d'un même *read* peut donc être considéré comme un *read* à part entière. Celle-ci sont ensuite alignées **indépendemment** le long du génome de référence, l'information fournit par le *paired-end* n'est utilisé qu'à *posteriri* en tant que critère qualité. Ainsi, après avoir filtré les *reads* ne s'étant pas aligné sur le génome et le *reads* orphelines (une seule des deux *ends* s'est alignée sur le génome), la "compatibilité entre les deux *ends* d'un même *reads* est analysée. Un *read* est dit"compatible" lorsque les deux *ends* qui le composent 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 la taille médiane de l'insert. Ici encore, seul les *reads* ayant des *ends* "compatibles" seront conservés. Pour l'ensemble de nos patients, les *reads* compatibles sont environs 10 fois plus important que la somme des *reads* non compatible, orphelins ou non mappés (**Figure :** 4.2 - **B**). Suite à cela, le nombre de site auxquels se sont alignés les *reads* est analysé. En effet, certaine zone du génome étant dupliqué, l'une des problématique des *short-reads* est qu'il est possible que ceux-ci s'alignent à plusieurs endroit du génome. Afin d'éviter toute ambiguité, seul ceux s'étant aligné sur un site unique sont conservés pour la suite des analyse. ces *reads* représente entre ... et ... % des *reads* ayant passé les précédents filtres (**Figure :** 4.2 - **C**). Les *reads* ayant passé l'ensemble des critères qualité seront ensuite utilisé pour effectué l'appel des variants.

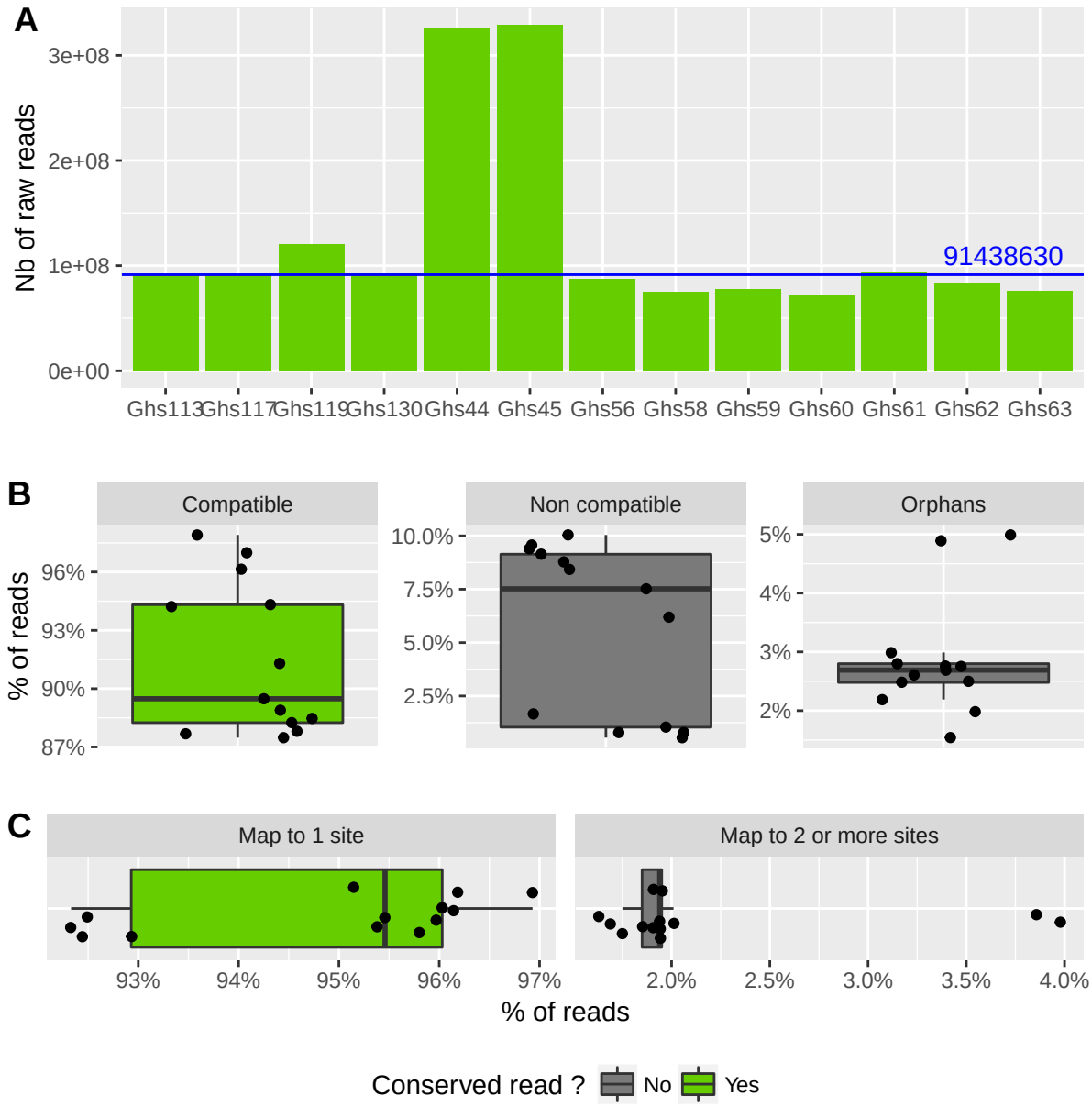

Figure 4.2 – Processus simplifié du contrôle qualité des *reads* : **A** : Quantité de *reads* bruts générés pour chaque patients au cours de l'étape de séquençage. On constate que ce nombre reste pour chaque patient dans le même ordre de grandeur sauf pour les frères AZ1 et AZ2 qui contienent presque 3 fois plus de *reads* que la mediane représentée en bleue.. **B** : Distribution pour chaque patient des *reads* compatibles, incompatibles, orphelins et non mappés. Comme attendu, les reads compatible sont les plus important. Ils sont les seuls à être utilisé dans le reste de l'analyse.. **C** : Présentation pour chaque *reads* du nombre de site auxquels ils s'alignent. Seuls les reads s'alignant sur un site unique sont conservés

Résultat de l'appel des variants Comme dit précédement, l'appel des variants fait suite à l'alignement et consiste à comparer la séquence d'un individus avec celle d'un génome de référence afin d'en relever les différences. La particularité de notre algorithme d'appel est d'effectuer pour chaque position deux appels indépendents. Le premier sera effectué en utilisant uniquement les *reads forward* et le second le *reads reverse*. Les positions ayant une couverture ≤ 10 sur **les deux** *strands* seront filtrés (NS). Les autres seront conservés bien que ceux ayant une couverture ≤ 10 sur **un des deux** *strands* (SS) seront considéré comme de faible qualité et seront leurs interprétations sera plus précotionneuse. Ainsi, chez nos `r` [TODO nb of patients] entre ... et ... variants sont filtrés car leur couverture est < 10 sur les deux *ends* du *reads* (**Figure :** 4.3 - **A**).

Pour les position ayant une couverture ≥ 10 sur **les deux** *strands* (DS) les résultats des deux appels sont comparés et seul les appel concordant seront conservés, c'est à dire environs ...% des variants DS. Les appels ambigus et discordants seront filtrés et non considérés dans les analyses en aval (**Figure :** 4.3 - **B**).

Dès lors il est intéréssant de noter que bien que les variants *single strand* (SS) soient conservés, on peut s'attendre à ce qu'également ... % de ces variants soient abérents, ceux-ci n'ayant put subir le même contrôle que les SS. Pour l'ensemble des variants ayant passé les premiers filtres, c'est à dire les variants SS et les variants DS avec appels concordants, le génotype est déterminé en fonction du pourcentage de *reads* portant le variant à cette position. Par exemple, si à une position donnée, 0% des *reads* portent un variant, l'individu sera appelé "Homozygote référence", si 50% des *reads* sont portent un variant, l'appel sera "hétérozygote" et si 100% des *reads* portent un variant, l'appel sera "Homozygote variant". Ainsi, pour chaque individu nous avons pu établire une liste de variants avec leur génotype associé (**Figure :** 4.3 - **C**).

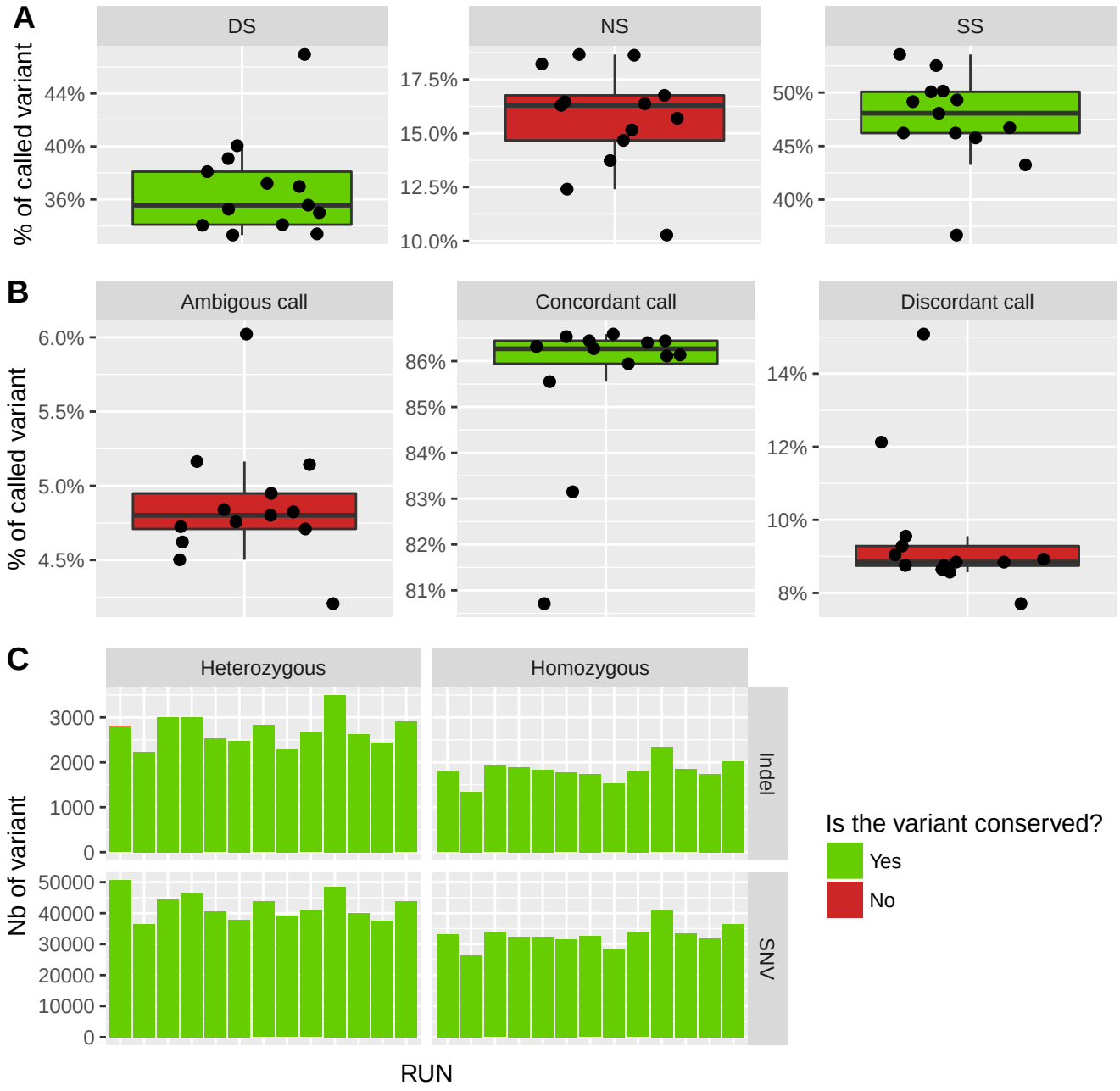

Figure 4.3 – Contrôle qualité des variants appelés : Pour l'ensemble des figures les variants verts sont conservés, les gris sont filtrés. **A** : Distribution du *stranding* des appels pour chaque patients. Environs 16 pourcents des vairants ont une couverture insufisante pour l'appel *forward* et l'appel *revers* et sont donc filtrés, les autres sont conservés. **B** : Comparaison des appels entre les deux *ends* des variants appelés DS. En fonction des individus, 80 à 90 pourcent des appels sont concordants. Les autres appels sont filtrés des analyses ulterieures. **C** : Distribution des SNVs et indels en fonction de leur génotype pour chaque patients (représentés par une barre).

Résultats de l'annotation Afin de connaitre l'effet qu'auront chacun des variants appelés sur les différents transcrits qu'ils chevauchent nous utilisons le logiciel VEP. Grâce à celà, nous pouvons constater que pour chaque patient ... gènes sont en moyenne affecté par au moins un variant tandis que ... sont impactés (soit environs `r` transcrits par gènes) (**Figure :** 4.4 - **A**). Chaque variant affectera l'ensemble des transcrits qu'il chevauche, ainsi un même variant pourra impacter plusieurs transcrits. Ces impacts sont ensuite classés par VEP en quatre catégories qui sont, de la plus délétère à la moins délétère : HIGH, MODERATE, LOW, MODIFIER. Comme attendu, les variants ayant un impact tronquant se retrouvent être les moins fréquent chez chacun de nos patients. Ceci est d'autant plus flagrant pour l'impact HIGH qui regroupe, entre autre, les variants créant un codon stop ou encore ceux causant un décalage du cadre de lécture, se retrouvent en quantité extrêment faible puisqu'ils ne représentent en moyenne que ... % des variants (soit environs ... par patient (en nb ici)) (**Figure :** 4.4 - **B**).

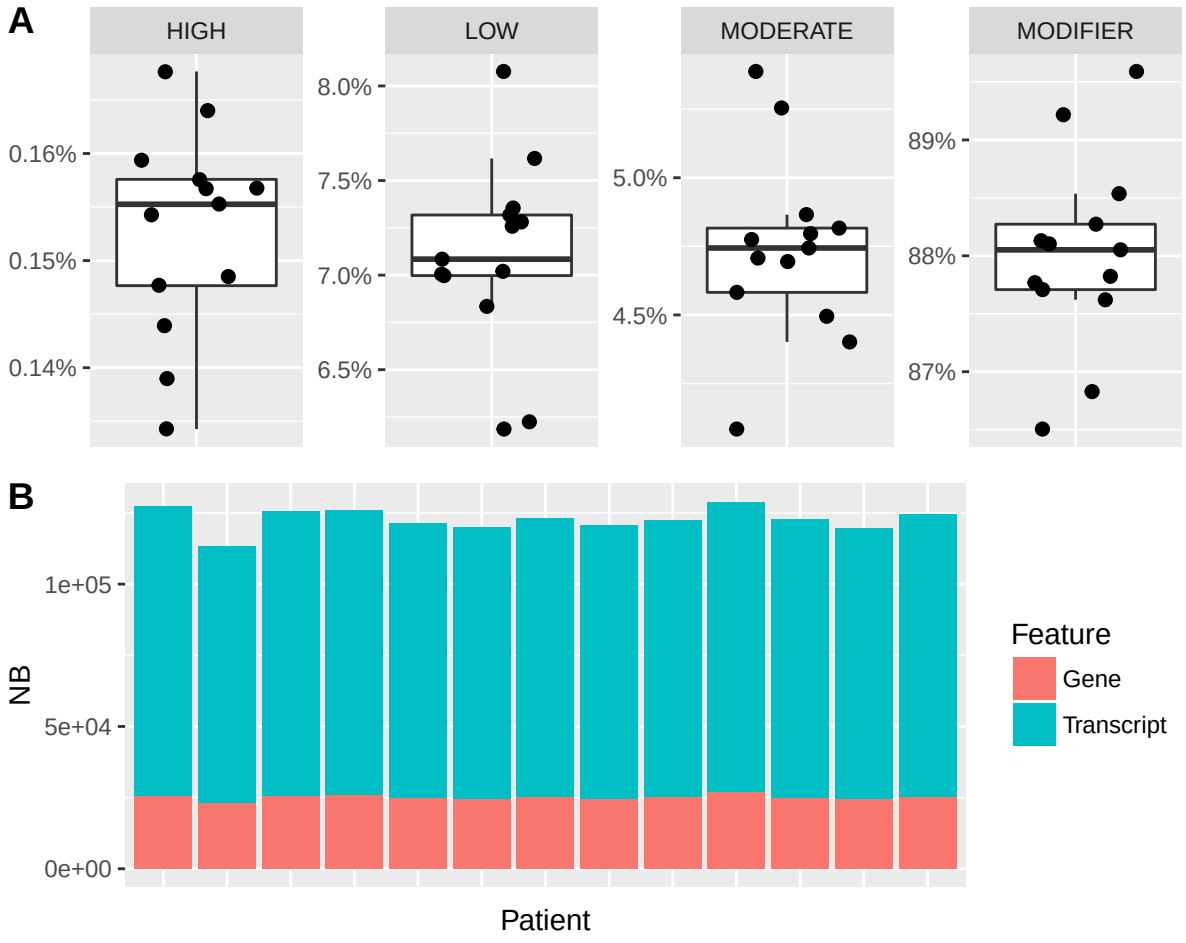

Figure 4.4 – TODO : TODOoooooooo oooooooooo oooooooooooooooooo ooooooooooooooooo oooooooooooo oooooooooooo oooo oooooooooo oooooooooooooooooooooooooooooo ooooooooooo ooooo oooooooooooooo ooooooooooo oooooooooo ooooooo.

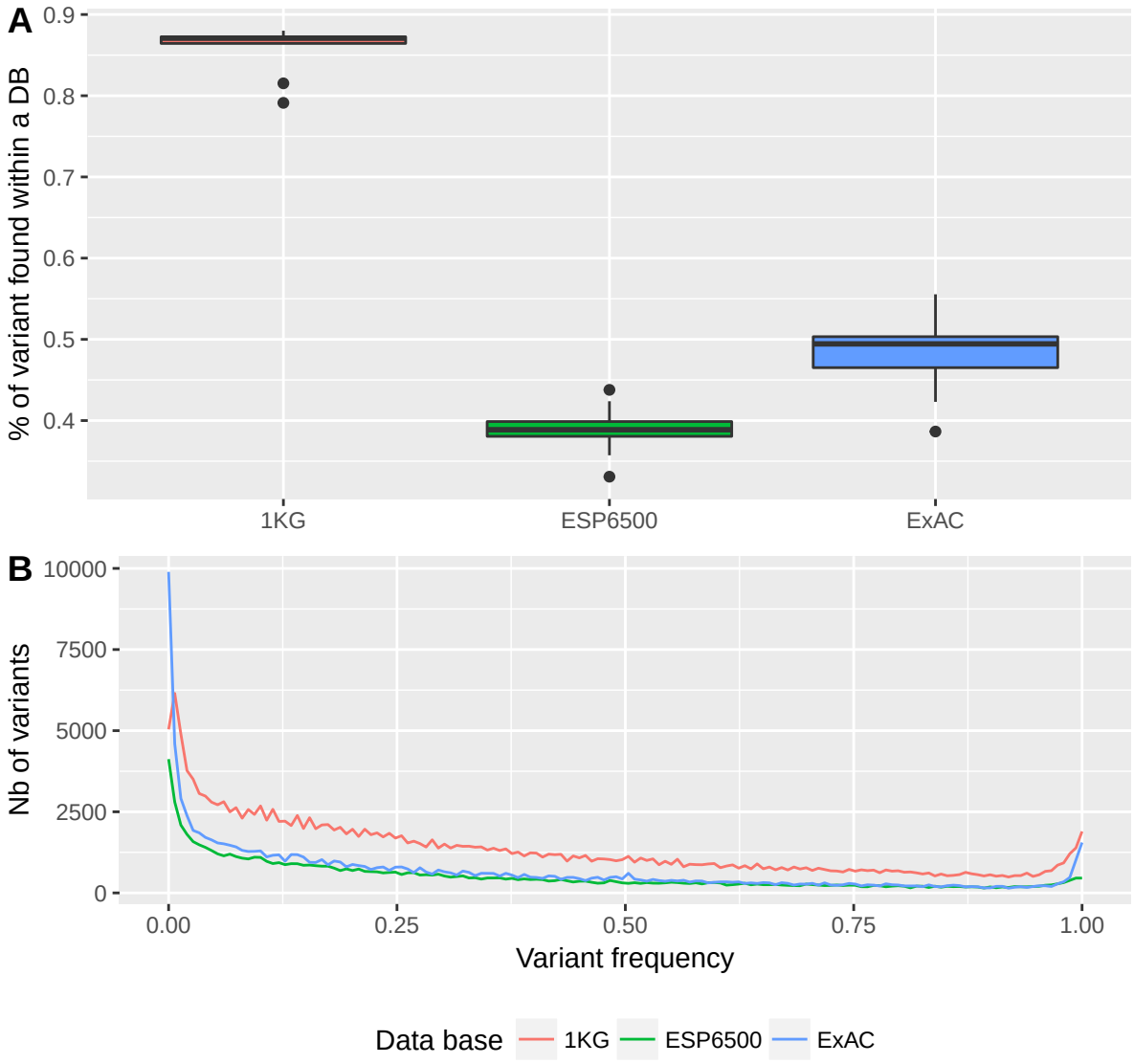

Figure 4.5 – TODO : TODOoooooooo oooooooooo oooooooooooooooooo ooooooooooooooooo oooooooooooo oooooooooooo oooo oooooooo ooooooooooooooooooooooooooooo ooooooooooo ooooo oooooooooooooo ooooooooooo ooooooooooo ooooooo.

Résultats du filtrage

```
##             used  (Mb) gc trigger   (Mb)  max used   (Mb)
## Ncells    754897  40.4    2564361  137.0   3205452  171.2
## Vcells 115550903 881.6  363498973 2773.3 454172341 3465.1
```

Les étapes précédentes nous ont permis de mettre en évidence pour chaque patient une liste de variants passant l'ensemble de nos critères qualités. Ces variants ont dès lors put être annotés nous permettant entre autre d'avoir connaissance de leurs l'impacts sur les différents transcrits qu'ils chevauchent ou encore leur fréquence dans la population générale. Desormai, afin de ne conserver que les variants ayant la plus forte probabilité d'être responsable du phénotype de ces patients, nous avons appliqué succesivement six filtres basés à la fois sur les différentes annotations que nous avons ajouté mais aussi sur nos connaissance du mode de transmission du phénotype :

1. **Filtre 1 : L'union des variants :** Dans ces différentes études, nous avons à chaque fois séquencé des duos ou des trios d'individus provenant de même frateries et étant caractérisés par le même phénotype. Ainsi nous avons pu formuler l'hypothèse d'une cause génétique commune entre les différents patients d'une même famille et donc filtrer l'ensemble des variants qui ne sont pas partagés par l'ensemble des membre de la fraterie [TODO : discussion de l'efficacité du filtre].

2. **Filtre 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 ... et ... variants par individus (**Figure : ??**, **??**).

3. **Filtre 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. **Filtre 4 : Les transcrits non pertinents :** Au cours de nos analyses nous nous sommes concentré uniquement sur les transcrits codant pour une protéine. Ainsi, l'ensemble des transcrits annotés comme étant non codant furent filtrés. De même 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. Dès lors, ces transcripts furent eux aussi filtrés. Ainsi, nous avons pu retirer de nos listes de variants l'ensemble des mutations impactant **uniquement** des transcrits non codant et / ou annoté NMD. Cette étape de filtre permet à elle seule de systematiquement filtrer entre 36576 et 44581 transcrits différents par patients, soit une moyenne de NaN variants par individus (**Figure :** 4.6).

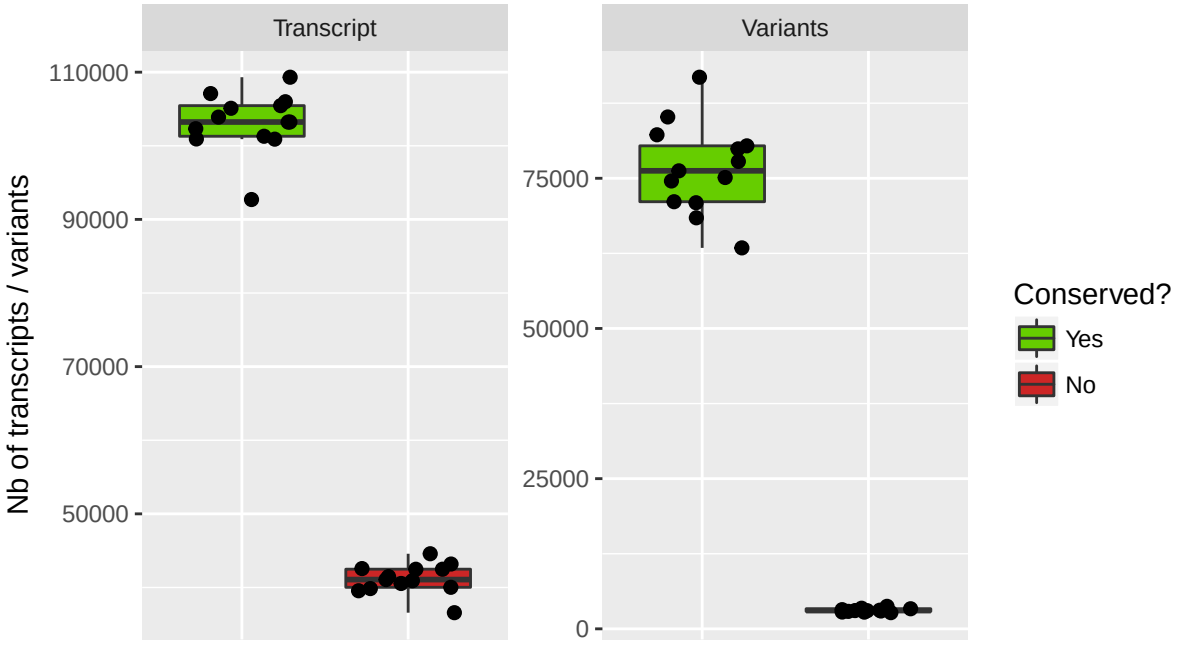

Figure 4.6 – Filtrage des transcrits jugés "non pertinents" et des variants les chevauchant : Pour chaque patients nous avons filtrer les transcrits jugés "non pertinents" pour l'analyse, c'est à dire ceux ne codant pas pour une protéine et ceux annoté NMD. Dès lors, l'intégralité des variants chevauchant uniquement des transcrits non pertinents ont put systématiquement être filtrés (boites rouges). les autres furent conservés (boites vertes)

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 :** 4.7). 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.

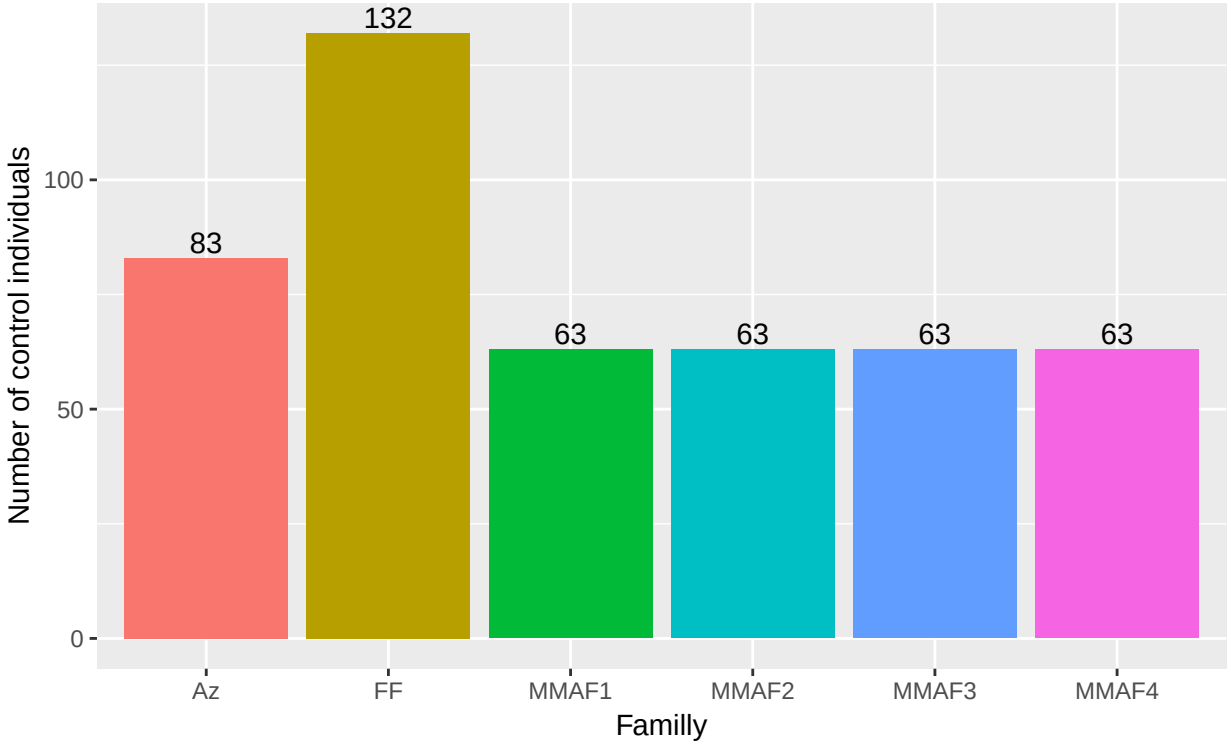

Figure 4.7 – Nombre d'individus composant la cohorte contrôle de chaque famille : Ici, chaque barre représente une famille et sa hauteur est déterminée par le nombre d'individus composant la cohorte contrôle à laquelle elle a été confronté. Chaque individus de la cohorte contrôle a été séquencés en WES par notre équipe. Afin d'être considéré comme "contrôle" et intégrer cette cohorte, 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

[TODO : COMPARAISON DE L'EFFICACITE DES FILTRES]

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 : ??**). Ainsi, la cause génétique expliquant le phénotype d'une famille a pu être mis en évidence dans ... familles sur ... [TODO] (**Figure :** 4.8). 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.

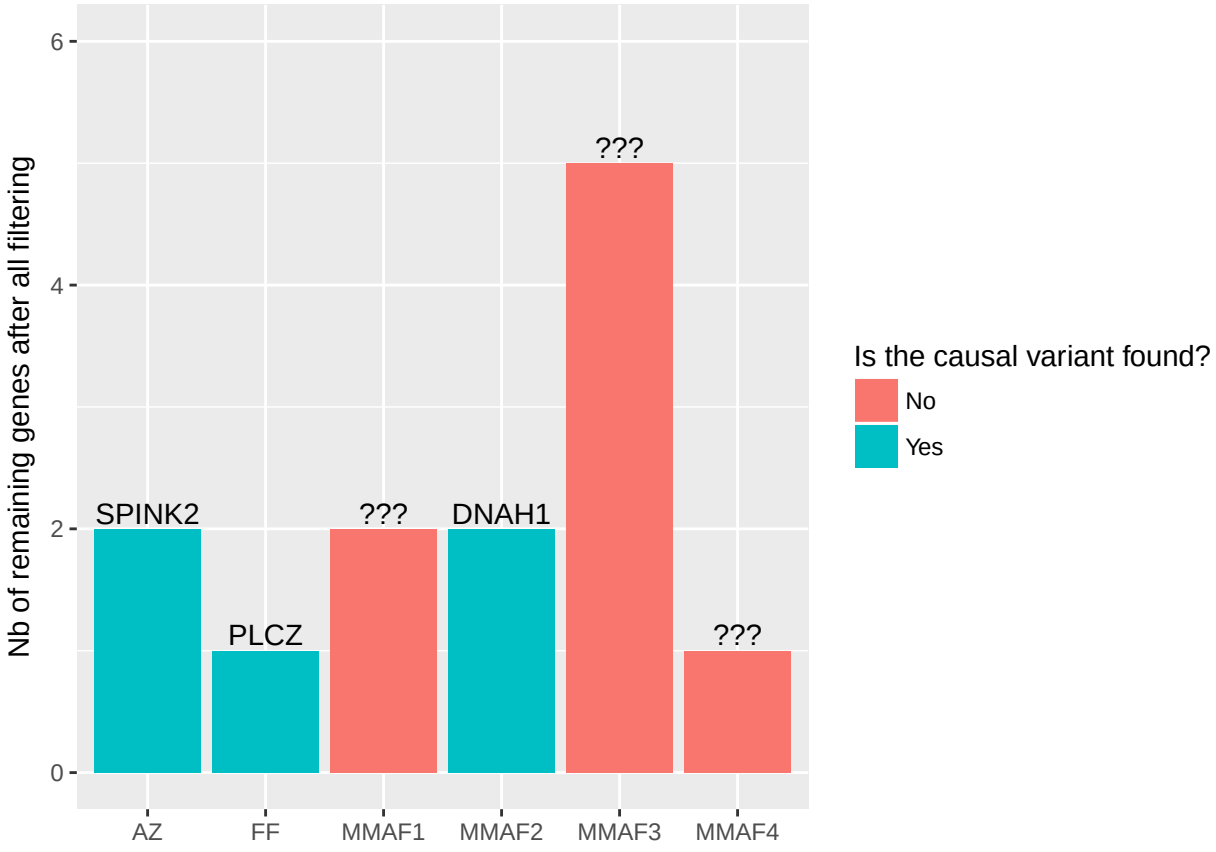

Figure 4.8 – Nombre de gènes passant l'ensemble des filtres par famille : Chaque barre représente une des familles analysées. La hauteure de cette barre correspond au nombre de gènes ayant passé l'ensemble des filtres pour chaque famille. Les barres bleues caractérisent les familles pour lesquelles le 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)

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