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

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 zones 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 nombre de *reads forward* (+) et *reverse* (-) sur lesquels est observé l'allèle de **référence** (R) à une position donnée.

   b. **V+ et V-** : À l'inverse de R+ et R-, ces comptages correspondent au nombre de *reads forward* et *reverse* sur lesquels est observé un allèle 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 vus précédemment. Tout d'abord, les positions ayant une couverture < 10 sur l'un des deux *strands* sera considérée comme de faible qualité, celles ayant une couverture < 10 sur les deux *strands* seront exclus. Ensuite pour chaque variant, des appels indépendants 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 bonne 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 ce pipeline est qu'elle

conserve l'ensemble des variants identifiés dans les études effectuées précédemment permettant d'ajouter aux annotations la fréquence 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 ultérieur.

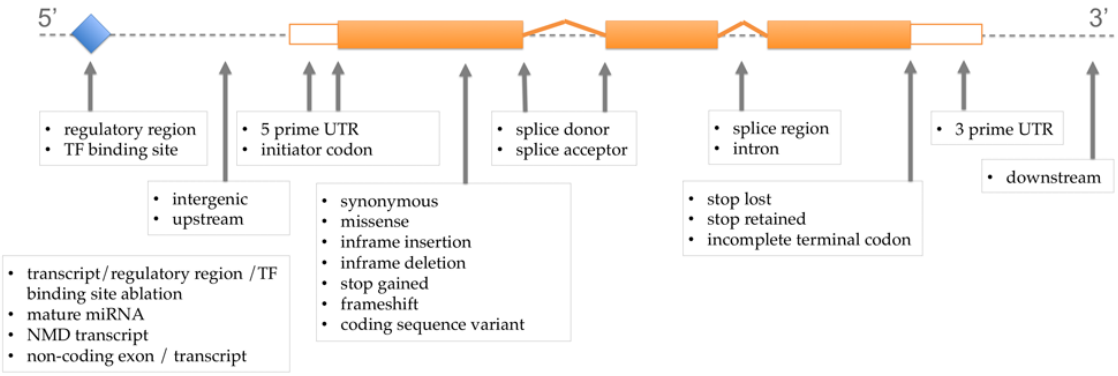

Figure 4.1 – Listes des différentes conséquences prédites par VEP et leur positionnement sur le transcrit d'après [VEP site](http ://www.ensembl.org/info/genome/variation/consequences.jpg)

4. **Le filtrage des variants** : L'étape de filtrage est extrêmement 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 besoins de l'étude. Afin de rendre son utilisation le plus efficace possible, nous avons souhaité définir des paramètres par défauts pertinent dans la plupart des études 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.

Table 4.1 – Liste simplifiée des conséquences prédites par VEP avec leur description et impact associée

VEP consequence	VEP impact	Description
Splice acceptor / donor	HIGH	A splice variant that changes the 2 base region at the 3' / 5' end of an intron
Stop gained	HIGH	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript
Frameshift	HIGH	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three
Stop lost	HIGH	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript
Start lost	HIGH	A codon variant that changes at least one base of the canonical start codo
Inframe insertion / deletion	MODERATE	An inframe non synonymous variant that inserts / deletes bases into in the coding sequenc
Missense	MODERATE	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved
Splice region	LOW	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron
Stop retained	LOW	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains
Synonymous	LOW	A sequence variant where there is no resulting change to the encoded amino acid
5 / 3 prime UTR	MODIFIER	A UTR variant of the 5' / 3' UTR
Intron	MODIFIER	A transcript variant occurring within an intron
NMD transcript	MODIFIER	A variant in a transcript that is the target of NMD
Non coding transcript	MODIFIER	A transcript variant of a non coding RNA gene

4.2.2 Utilisation du 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'éjaculât

2. **Échec 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 récapitulatif des familles séquencées et de leur phénotype

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

Resultats des exomes

Résultat de l'alignement 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* composant les exomes de chaque individu peut varier en fonction de plusieurs paramètres et n'est donc pas égale pour chaque patient bien que l'ordre de grandeur reste le même exceptés, c'est à dire une médiane de 91438630 *reads*. Seul les deux frères AZ1 et AZ2 se distinguent près de 3 fois plus de *reads* que pour les autres patients. Cette différence peut être expliqué car ces deux patients sont les deux seuls à voir été séquencé au Mount Sinaï Institut or leur protocole d'amplification précédent le séquençage contient un nombre de cycles de PCR supérieur à ceux appliqué au Génopole d'Évry où ont été séquencé les autres patients (**Table :** 4.2, **Figure :** 4.2 - **A**).

L'ensemble de nos exomes ayant été réalisés en *paired-end*, les deux extrémités de chaque fragment sont séquencés chaque *end* d'un même *read* peut donc être considéré comme un *read* à part entière qui sont alignées **indépendamment** le long du génome de référence. L'information fournit par le *paired-end* n'étant utilisé qu'à *posteriori* en tant que critère qualité. La première étape du contrôle qualité des *reads* consiste filtrer les *reads* ne s'étant pas aligné sur le génome. Ces *reads* sont extrêmement minoritaires puisqu'ils représentent entre 1.2 et 5.5 % des *reads* de nos individus (**Figure :** 4.2 - **B**).

Une fois cela fait, nous vérifions la "compatibilité" des deux *ends* composant chacun des *reads* s'étant correctement alignés. Un *reads* 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. Les *reads* dont les deux *ends* se sont alignées mais ne remplissant pas ces conditions seront dit "Non compatible", ceux dont une seule des deux *ends* s'est alignés seront appelés "orphelins". Dans nos analyses, seuls les *reads* compatibles sont conservés, c'est à dire environs ... % (médiane) des *reads* s'étant correctement alignés. (**Figure :** 4.2 - **C**).

La dernière étape de ce contrôle-qualité consiste à analyser le nombre de site auxquels se sont alignés les *reads*. En effet, certaine zone du génome étant dupliqué, l'une des problématiques des *short-reads* est qu'il est possible que ceux-ci s'alignent à plusieurs régions différentes du génome. Afin d'éviter toute ambiguïté, seul ceux s'étant aligné sur un site unique sont conservés pour la suite des analyses. Ces *reads* représente entre 92.3 et 96.9 % des *reads* ayant passé les précédents filtres (**Figure :** 4.2 - **C**).

Les *reads* ayant passé l'ensemble des critères qualité mentionnés précédemment seront ensuite utilisés pour effectuer l'appel des variants.

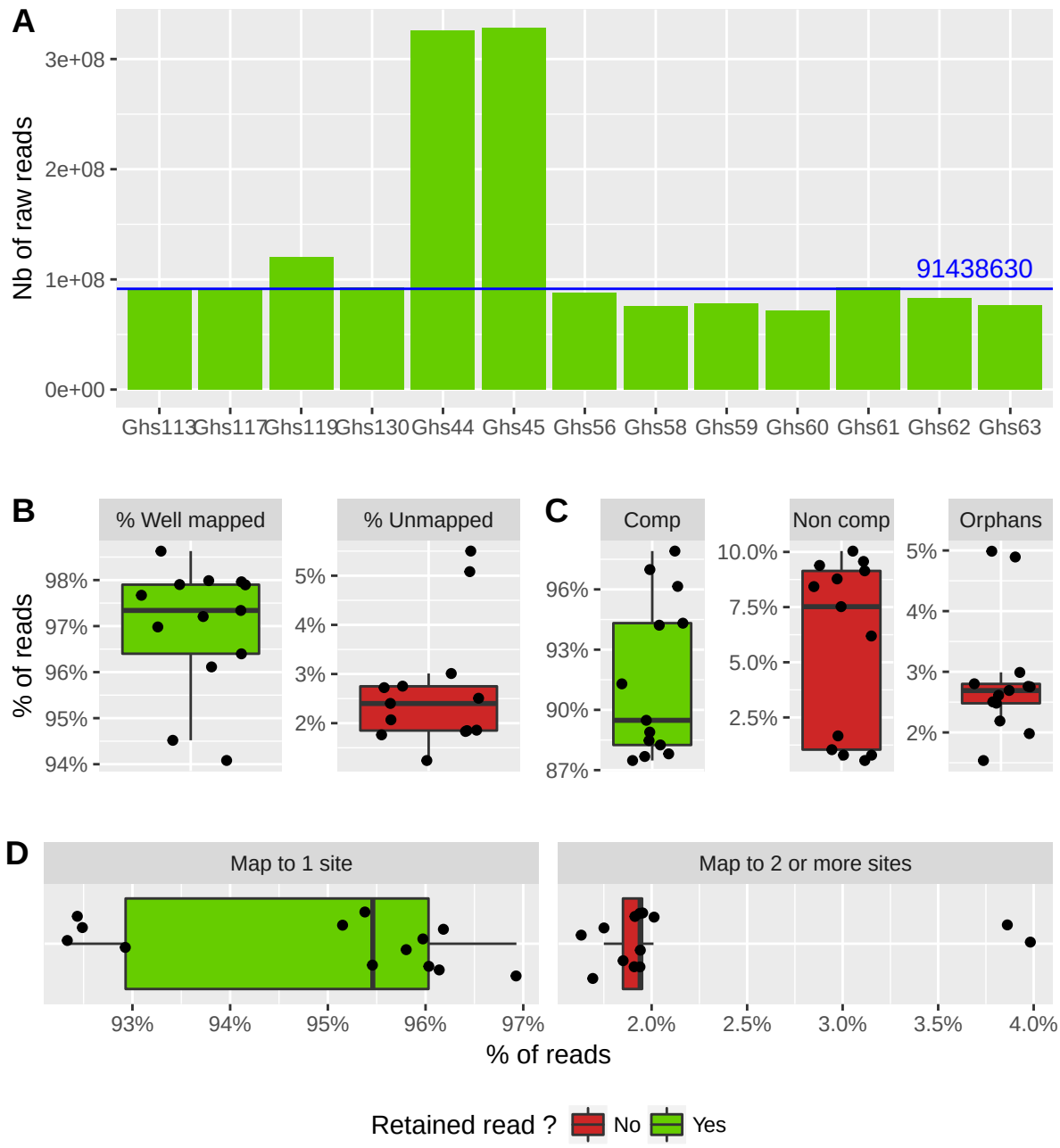

Figure 4.2 – Processus simplifié du contrôle qualité des *reads* : Pour chacun des graphiques, les *reads* représentés en vert sont conservés tandis que ceux en rouge sont filtrés. **A** : Quantité de *reads* bruts générés pour chaque patient au cours de l'étape de séquençage. La médiane des *reads* est représentée en bleue. **B** : Pourcentage pour chaque individu de *reads* s'étant aligné correctement et ne s'étant pas alignés sur le génome de référence. **C** : Distribution pour chaque patient des *reads* compatibles (Comp), non compatibles (Non comp) et orphelins (Orphans). **D** : Présentation pour chaque *reads* du nombre de site auxquels ils s'alignent

Résultat de l'appel des variants Comme dit précédemment, l'appel des variants fait suite à l'alignement et consiste à comparer la séquence d'un individu 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 indépendants Le premier sera effectué en utilisant uniquement les *reads forward* et le second le *reads reverse*. Encore une fois, plusieurs filtres sont appliqués de sorte à conserver uniquement les variants les plus qualitatifs.

Tout d'abord, nos appels sont classés en trois catégories :

1. **Les appels *double strand* (DS) :** Qualifie les positions ayant une couverture ≥ 10 sur **les deux** strands. Ces appels sont ceux sont ceux ayant la meilleure qualité
2. **Les appels *single strand* (SS) :** Ces appels définissent les positions pour lesquels **un des deux** *strands* présentent une couverture ≤ 10. Dans ce cas, ce *strand* est ignoré et l'appel est effectué uniquement en utilisant le second *strand*.

3. **Les appels *non strand* (NS) :** Les positions NS sont celles pour lesquelles la couverture est ≤ 10 sur **les deux** strands. Aucun appel n'est effectué à ces positions.

Dans nos données, les appels SS sont majoritaires et représentent environ 48.1 % de nos appels (contre 35.6 % d'appels DS). Au vus de l'importance de ces appels, nous avons fait le choix de les conserver afin de ne pas filtrer une quantité trop importante de données. Ces appels seront cependant considérés comme étant de faible qualité, de fait, leurs analyses et interprétation seront plus précautionneuses En revanche, au vus de la trop grande incertitude de l'appel des variants NS, ceux-ci sont systématiquement filtrés éliminant ainsi entre 10.3 et 18.7 % des positions appelées (**Figure :** 4.3 - **A**).

Un second filtre est appliqué aux variants ayant été précédemment appelés DS. Celui-ci consiste à comparer les appels effectués indépendamment sur chacune des deux *ends* et à vérifier leur concordance, c'est à dire que les deux appels soit identique. Les appels discordant et ambigus sont filtrer, ainsi environs 86.3 % des variants DS passent ce filtre. Il est intéressant de noter que bien que les variants *single strand* (SS) soient conservés, on peut s'attendre à ce qu'environ 13.7 % de ceux-ci soient aberrants, ceux-ci n'ayant pu subir le même contrôle que les SS (**Figure :** 4.3 - **B**).

Pour l'ensemble des variants ayant passé les filtres énoncés ci-dessus, 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 établir une liste de SNVs et d'indels avec leur génotype associé. Pour chacun de nos 13 patients les ordres de grandeur du nombre de variants

appelés sont identique. Ainsi pour chaque patient nous avons appelés environ 43670 variants hétérozygotes (41044 SNVs et 2626 indels) et 65040 variants homozygotes (32520 SNVs et 1809 indels) (**Figure :** 4.3 - **C**).

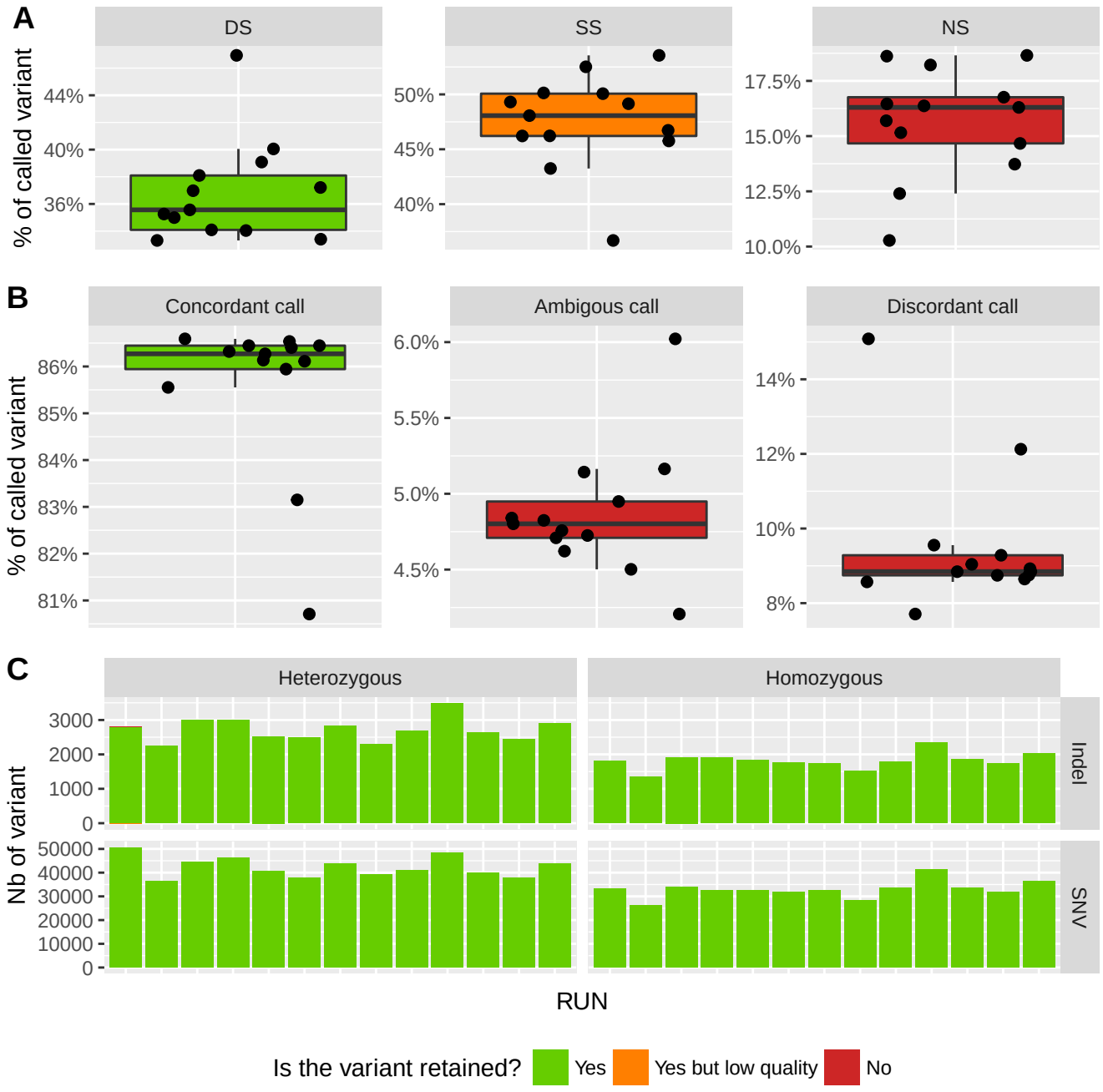

Figure 4.3 – Contrôle qualité des variants appelés : Pour chacun des graphiques, les variants représentés en vert et en orange sont conservés tandis que ceux en rouge sont filtrés. **A** : Distribution du *stranding* des appels pour chaque patient. **B** : Comparaison des appels entre les deux *ends* des variants appelés DS. **C** : Distribution des SNVs et indels en fonction de leur génotype pour chaque patients (représentés par une barre

Résultats de l'annotation L'annotation des variants appelés consiste à ajouter un maximum d'informations sur les variants. Ces informations seront ensuite utilisées afin de filtrer et / ou prioriser les variants. Dans ces analyses nous avons utilisé le logiciel *Variant Effect Predictor* (VEP) (W. McLaren et al., 2016) qui va à la fois prédire l'effet qu'auront ces variants sur l'ensemble des transcrits (et gènes) qu'ils chevauchent, ajouter, lorsqu'elle est disponible, la fréquence de chacun de ces variants dans les bases de données ExAC, 1000Genomes (1KG) et ESP6500. Pour finir VEP nous permettra de connaitre les prédictions de pathogénicités fournies par SIFT et PolyPhen pour les variants faux-sens.

Après avoir annoter nos variants par VEP, nous avons pu constater que pour chaque patient 24975 gènes sont en moyenne affecté par au moins un variant pour en moyenne 122735 transcrits (soit environs 5 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 (**Table :**4.1). 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 autres, les variants créant un codon stop ou encore ceux causant un décalage du cadre de lecture (**Table :**4.1), se retrouvent en quantité extrêmement faible puisqu'ils ne représentent en moyenne que 0.15 % des variants, soit une moyenne de 466 hétérozygotes et 370 homozygotes par patient) (**Figure :** 4.4 - **B**).

Parmi ces variants, certains étaient déjà recensés dans une des trois base donnée (ExAC, ESP et 1KG). Ainsi, on peut observer qu'entre 38.6 et 55.5 % de nos variant étaient listés dans ExAC et entre 33.1 et 43.8 % dans ESP. En revanche environ 87.1 % d'entre eux sont recensés dans 1KG (**Figure :** 4.4 - **C**) (À discuter !!!!!).

(À discuter !!!!!) (**Figure :** 4.4 - **D**)

LES FIGURES SUR LA FRÉQUENCE SONT À DISCUTER CAR LEUR INTERPRÉTATION ME LAISSE PERPLEX (SURTOUT LA PROPORTION DE NOS VARIANTS PRÉSENTS DANS 1KG)

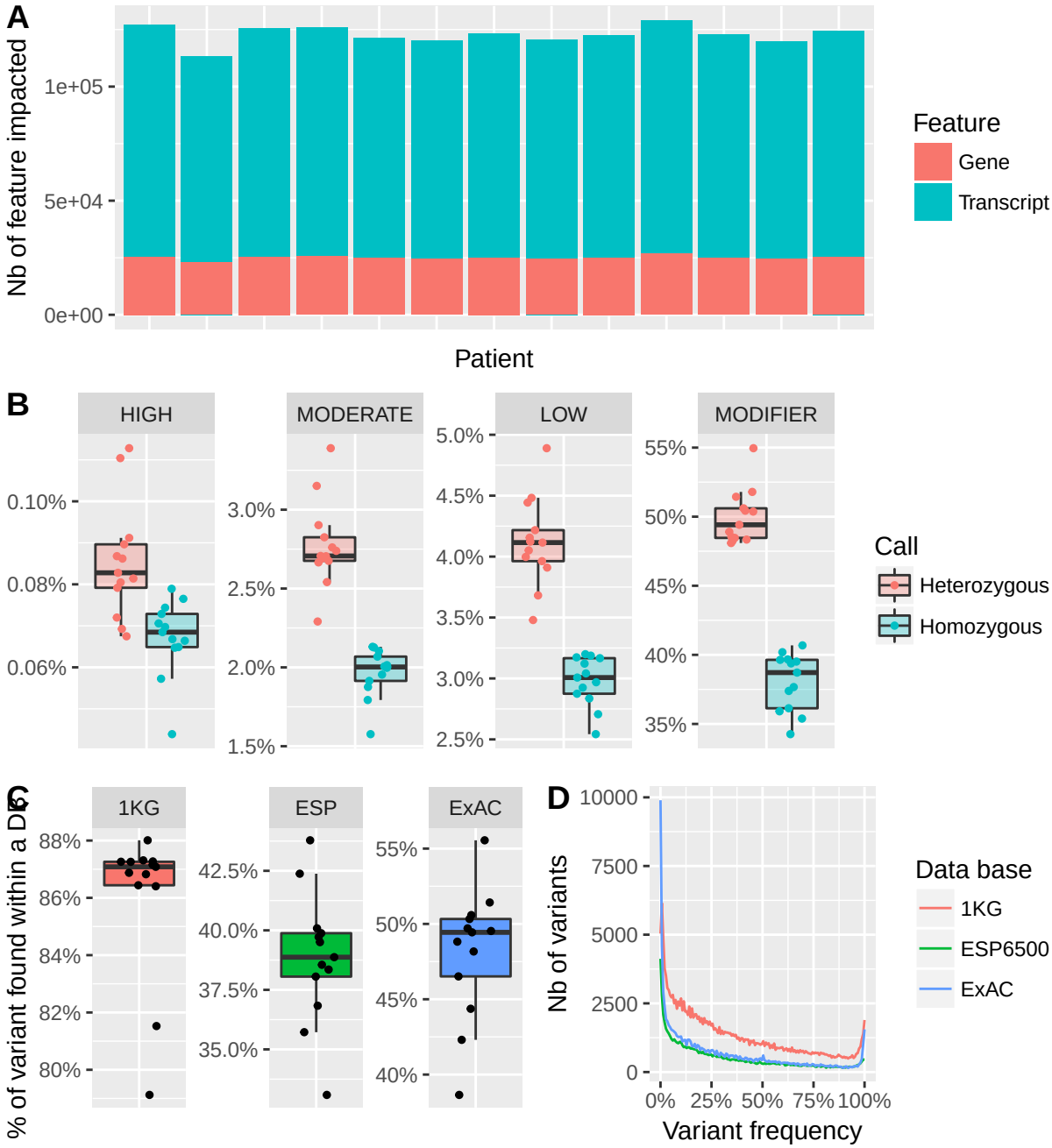

Figure 4.4 – Annotation des variants par VEP : **A** : Quantification du nombre de gènes (en bleu) / transcrits (en rose) impactés par au moins un variant pour chaque patient chacun représentés par une barre. **B** : Distribution des impact HIGH MODERATE LOW et MODIFIER en fonction des patients et du génotype du variant. **C** : Pourcentage de nos variants retrouvés au sein des trois bases de données : ExAC, ESP et 1KG. **D** : Distribution des fréquences de nos variants au sein des trois bases de données : ExAC, ESP et 1KG

Résultats du filtrage 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 autres d'avoir connaissance de leurs l'impacts sur les différents transcrits qu'ils chevauchent ou encore leur fréquence dans la population générale. Désormais, afin de ne conserver que les variants ayant la plus forte probabilité d'être responsable du phénotype de ces patients, nous avons appliqué successivement six filtres basés à la fois sur les différentes annotations que nous avons ajoutées mais aussi sur nos connaissances 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 fratries 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 membres de la fratrie.

2. **Filtre 2 : Génotype des variants :** Dans ces études, nous avons émis l'hypothèse d'une transmission récessive du phénotype. Ainsi, seul les variants homozygotes ont été conservés. (**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 intronique et ceux tombant dans les séquences UTRs. De même les variants synonymes ne sont pas conservés (exceptés 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 (Kumar, Henikoff, & Ng, 2009) et Polyphen (Adzhubei et al., 2010) et filtré l'ensemble des faux-sens 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 contrôler 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 transcrits 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 systématiquement filtrer entre 36576 et 44581 transcrits différents par patients, soit une moyenne de NaN variants par

individus (**Figure :** 4.5).

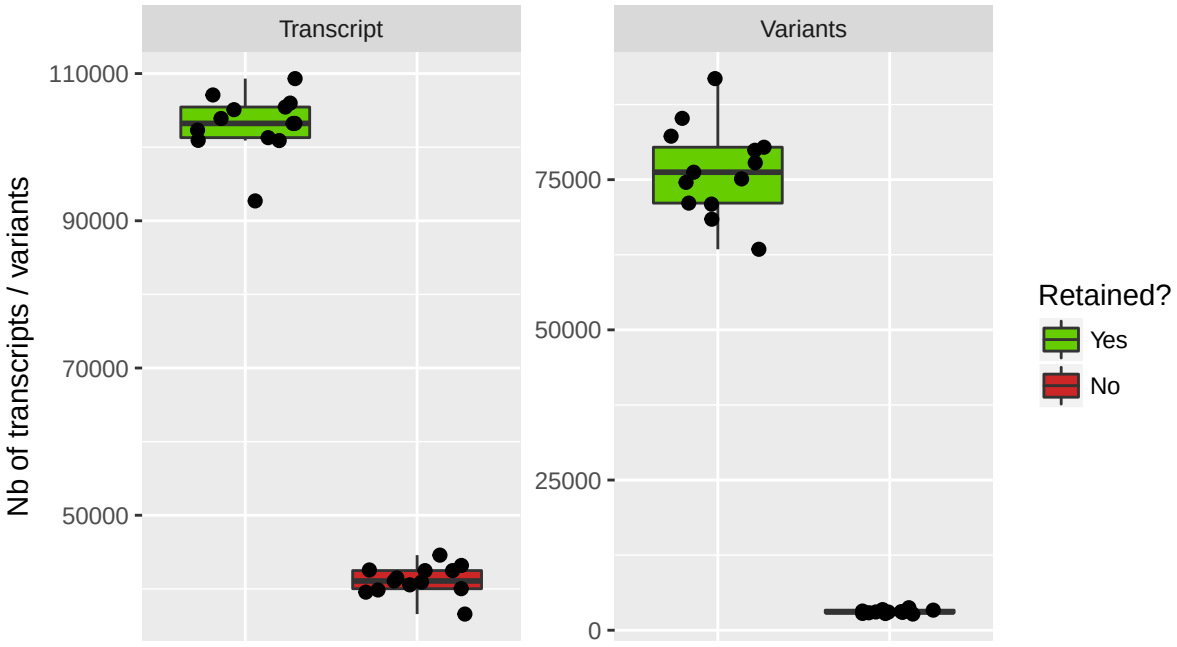

Figure 4.5 – Filtrage des transcrits jugés "non pertinents" et des variants les chevauchant : Pour chaque patients nous avons filtrer les transcrits jugés "non pertinents" pour l'analyse, c'est à dire ceux ne codant pas pour une protéine et ceux annoté NMD. Dès lors, l'intégralité des variants chevauchant uniquement des transcrits non pertinents ont put systématiquement être filtrés (boites rouges). les autres furent conservés (boites vertes)

5. **Fréquence 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 effet, il est peu probable qu'un retrouvé fréquemment 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 amenés 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 aberrant d'émettre l'hypothèse qu'ils que leurs causes génétiques le soient également. De même, les variants recherchés 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 :** 4.6). Dès lors, nous avons pu filtrer l'ensemble des variants retrouvés à la fois chez nos patients et observés à l'état homozygote dans la cohorte contrôle.

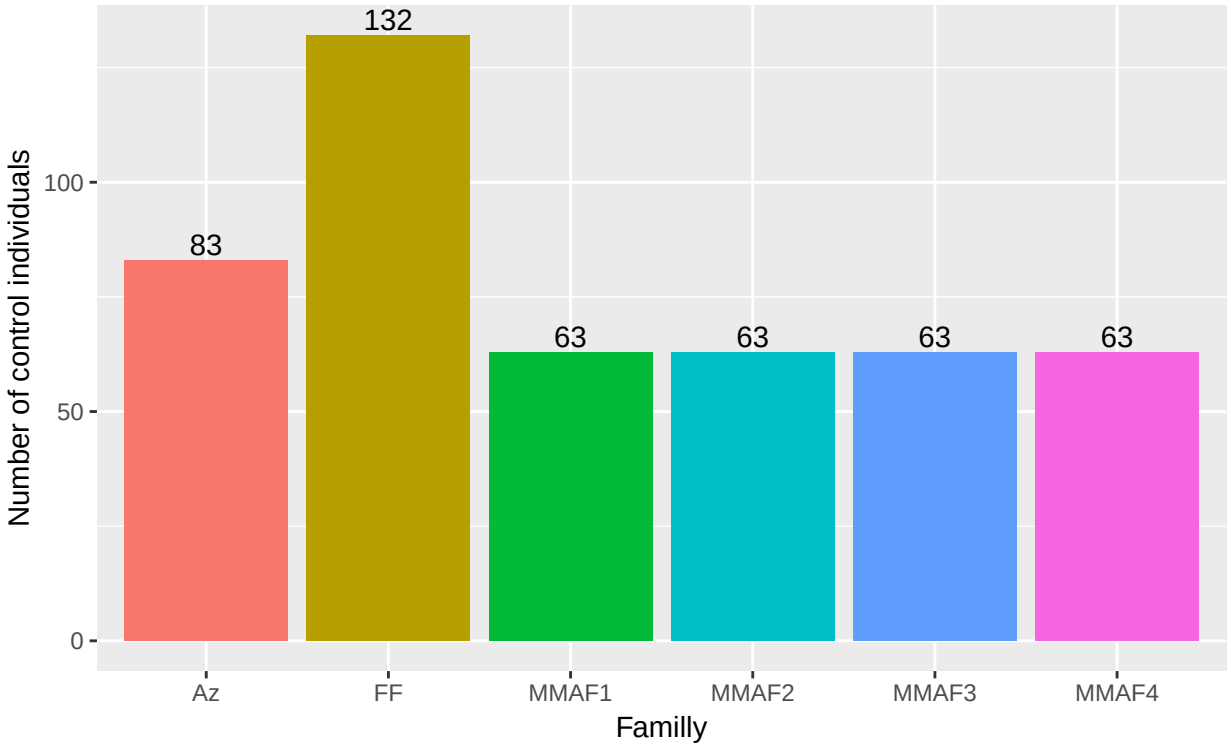

Figure 4.6 – Nombre d'individus composant la cohorte contrôle de chaque famille : Ici, chaque barre représente une famille et sa hauteur est déterminée par le nombre d'individus composant la cohorte contrôle à laquelle elle a été confronté. Chaque individu de la cohorte contrôle a été séquencés en WES par notre équipe. Afin d'être considéré comme "contrôle" et intégrer cette cohorte, un individu 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

Comme on pouvait s'y attendre, ces six filtres ont un pouvoir discriminant extrêmement différent (**Figure :** 4.7). En effet, tandis que le filtre "Transcript relevance" (filtre n°4) éliminer en moyenne 3.9 % des variants de chaque individu tandis que le filtre "Variant impact" (filtre n° 3) élimine jusqu'à 90.1 % de ces mêmes variants (**Figure :** 4.7 - **A**). Cette différence n'est pas surprenante. En effet, comme nous l'avions vu plus tôt, les variants de la catégorie VEP MODIFIER qui regroupe entre autres les variants chevauchant les séquences UTRs et introniques (**Table :**) représentent en moyenne ... % des variants de nos patients (**Figure :** 4.4 - **A**). Ceux-ci étant tous filtrés, on s'attendait donc à une valeur aussi élevée. On peut également constater l'importance de la cohorte contrôle qui, je le rappelle, permet de filtrer l'ensemble des variants homozygotes observés en son sein, puisque ce filtre permet retirer entre 76.5 et 88.4% des variants de chaque individus (**Figure :** 4.4 - **A**).

Cependant, regarder uniquement le pourcentage de variants filtrés par chaque filtre révèle une information partielle. En effet, dans ce cas de figure, on observe la quantité de variant éliminé par chaque filtre indépendamment les uns des autres. Ainsi, un même variant peut donc être filtrer par plusieurs filtres. Dès lors, il faut également analyser la quantité de variants filtrés **spécifiquement** par chaque filtre. Ainsi, on peut constater que le classement des filtres en fonctions de leur stringeance reste quasi identique (**Figure :** 4.7 - **B**) il est tout de même intéressant de noter que désormais le filtre "Variant impact" apparait moins efficace que les filtres "Ctrl" et "Genotype" en filtrant spécifiquement une moyenne de 253 variants par individu contre 423 pour le filtre génotype et 882 pour le filtre "Ctrl". Ainsi, ce dernier devient celui filtrant spécifiquement le plus de variants avec entre 364 et 1060 variants spécifiquement filtrés par patients confirmant ainsi l'importance de ce filtre dans nos analyses. Aussi, les filtres "Transcript relevance", "Union" et "Frequency" apparaissent désormais comme étant anecdotiques en comparaison aux trois autres filtres puisqu'ils filtrent au maximum 43 variants spécifiques (**Figure :** 4.7 - **B**).

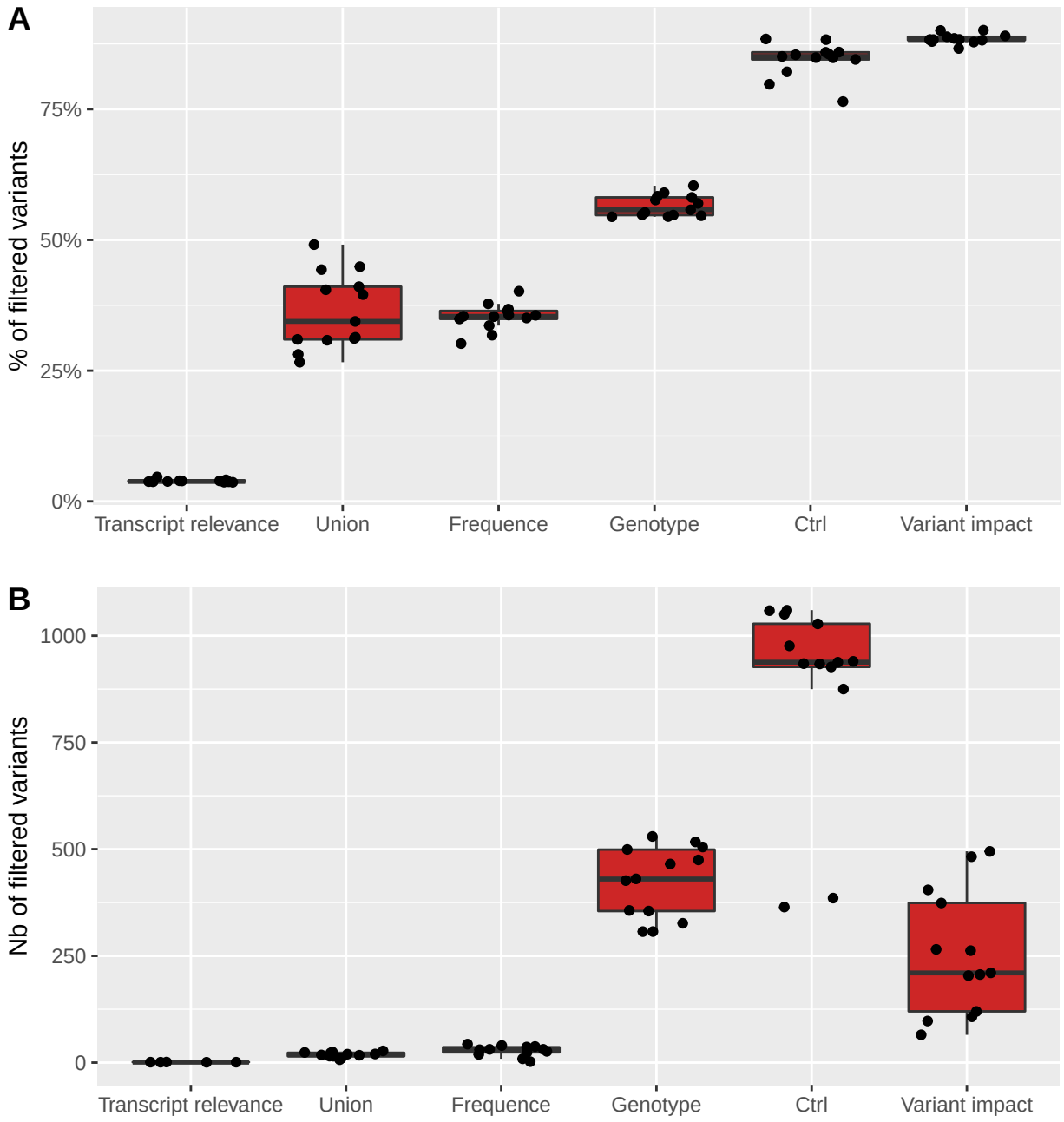

Figure 4.7 – Comparaison de l'efficacité de chacun des six filtres utilisés : **A** : Comparaison du pourcentage de variants filtrés par chacun des six filtres indépendamment les uns des autres pour chaque patient (représenté par les points. Dès lors, un même variants peut être filtré par plusieurs filtres. **B** : Comparaison du nombre de variant filtrés spécifiquement par chacun des filtres. Ici, un variant ne peut-être filtré que par un seul filtre

Après avoir appliqué l'ensemble de ces filtres, seuls quelques variants subsistent nous permettant d'obtenir une liste de gènes restreinte pour chaque famille (**Table :** 4.3) et ainsi de tirer des conclusions quant au variant responsable du phénotype.

1. **Famille AZ** : Parmi les 2 gènes restant pour cette famille, *SPINK2* est apparu comme étant un candidat évident. Notamment son expression étant spécifique au testicule tandis que celle de *GUF1* est ubiquitaire (TODO fig). De plus, des mutations du gène *Spink2* chez la souris avait déjà été identifiée comme induisant des défauts de la spermatogenèse (Lee et al., 2011).
2. **Famille FF** : Pour cette famille, le gène *PLC*ζ1 a passé l'ensemble des filtres. Nos connaissances sur la fonction de se gène et notamment son rôle dans l'activation ovocytaire (TODO : REF) on fait de ce gène le candidat idéal pour expliquer le phénotype de ces deux frères.

3. **Famille MMAF1** : L'analyse bibliographique des 2 gènes ayant passé l'ensemble des filtres n'a ici pu nous permettre de d'affirmer que l'un de ces gènes étaient responsable du phénotype MMAF de ces 2 frères.

4. **Famille MMAF2** : À l'issue des filtres, 2 gènes ressortaient chez ces deux frères : *MYH11* et *DNAH1*. Or, notre équipe ayant déjà, il y a quelques années établit le lien entre des mutations du gène *DNAH1* et le syndrome MMAF (Ben Khelifa et al., 2014) ce gène s'est révélé être un candidat idéal pour expliquer le phénotype de ces 2 frères. De plus, l'implication de *MYH11* dans le phénotype de dissection aortique (Imai et al., 2015) l'ont écarté des candidats pour le phénotype MMAF.

5. **Famille MMAF3** : Comme pour les gènes de la famille MMAF2, l'analyse bibliographique des 5 gènes ayant ici passé les filtres de même que l'étude de leurs expressions ne nous a pas permis de conclure que l'un d'entre eux étaient responsable du phénotype MMAF de ces 2 frères.

6. **Famille MMAF4** : Seul le gène *TGIF2* a passé l'ensemble des filtres pour la famille MMAF4. L'expression ubiquitaire de ce gène n'en font pas un candidat idéal. Cependant une étude de 2011 effectuée sur le wallaby décrit que la protéine TGIF2 localise spécifiquement dans le cytoplasme du spermatide, ainsi que dans le corps résiduel et la pièce intermédiaire du flagelle du spermatozoïde mature (Hu, Yu, Shaw, Renfree, & Pask, 2011). Ces données pourraient corréler avec le phénotype MMAF de ces 3 frères.

Table 4.3 – Liste des gènes ayant passé l'ensemble des filtres pour chaque famille

AZ	FF	MMAF1	MMAF2	MMAF3	MMAF4
SPINK2	PLCZ1	PLA2G4B	MYH11	PCSK5	TGIF2
GUF1		JMJD7-PLA2G4B	DNAH1	WEE2	
				GBP2	
				FCGR3A	
				ZFYVE28	

Discussion

L'analyse de ces 6 familles nous a permis de mettre en évidence l'efficacité de notre pipeline d'analyse puisque pour 3 d'entre elles (soit 50%) le variant causal a pu être identifié avec certitude (**Figure :** 4.8) et les résultats publiés dans trois revus dont je suis co-auteur :

1. **Famille AZ** : **SPINK2 deficiency causes infertility by inducing sperm defects in heterozygotes and azoospermia in homozygotes** : Dans cet article j'ai effectué 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 partie des 611 individus analyser ainsi que contribué à l'extraction de l'ARN testiculaire des souris pour l'analyse fonctionnelle du gène *Spink2* sur le modèle murin.

2. **Famille FF** : **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, effectué l'intégralité des analyses bioinformatiques des données d'exomes effectués sur deux frères infertiles présentant des échecs de fécondation.

3. **Famille MMAF2** : **Whole-exome sequencing of familial cases of multiple morphological abnormalities of the sperm flagella (MMAF) reveals new DNAH1 mutations** : Dans cet article j'ai, comme précédemment, effectué l'ensemble des analyses bioinformatiques des données d'exomes effectués sur deux frères infertiles présentant des échecs de fécondation.

Pour une d'entre elle, un candidat potentiel a pu être mis en évidence avec le gène *TGIF2* et notre équipe travaille actuellement sur la caractérisation de ce gène afin de savoir s'il peut effectivement expliquer le phénotype MMAF de cette famille (**Figure :** 4.8).

Pour les 2 familles restantes, aucun variant n'a pu pour l'instant expliquer leur phénotype. L'explication la plus vraisemblable est que le variant ait été filtré par

l'un de nos six filtres, probablement celui consistant à filtrer l'ensemble des variants hétérozygotes. En effet, l'hypothèse d'un variant causal homozygote était extrêmement crédible pour les familles AZ, FF et MMAF2 étant donné l'historique consanguin de ces 3 familles dont les parents sont à chaque fois apparentés. En revanche rien ne laisse supposé une telle chose pour les familles restantes. Cependant, le filtre des variants hétérozygotes pour l'ensemble des patients de ces 3 familles a été maintenu en première intention afin de faciliter les analyses en réduisant au maximum le nombre de variant. Au vus des résultats il apparait clair que les variants responsablent de leur phénotype aient été filtrés pour au moins 2 de ces familles. Dès lors, l'ensemble des analyses effectuées lors de l'étape de filtrage doivent être refaites en changeant les paramètres de filtrage. Cette fois-ci, les variants hétérozygotes seront conservés et les gènes sur lesquels au moins deux variants hétérozygotes seront recensés seront analysés en priorité. En effet, bien que les analyses exomiques nous fournissent en l'état pas d'informations suffisante pour savoir si ces deux variants sont présent sur le même allèle ou bien sur deux allèles différents, cela pourrait-être la signature de variants hétérozygotes composites. C'est donc sur ces analyses que se concentre actuellement notre équipe.

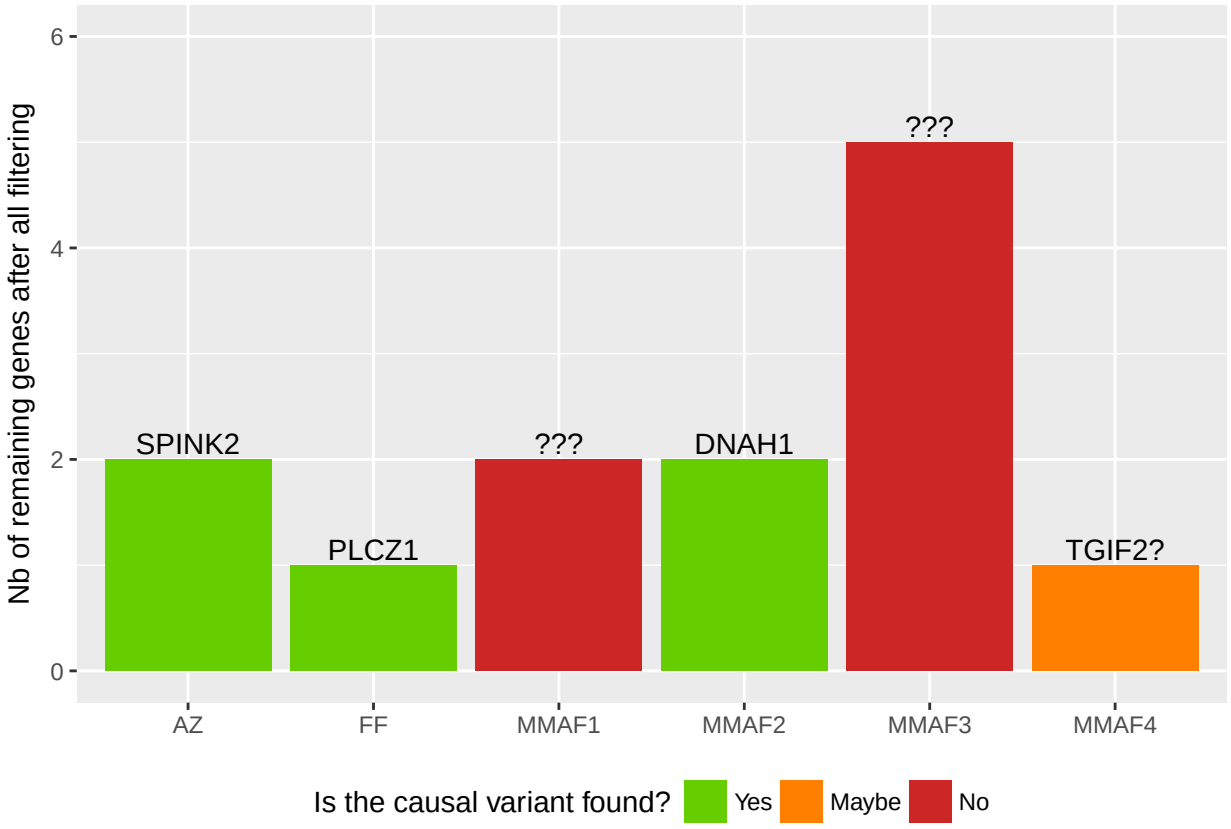

Figure 4.8 – Nombre de gènes passant l'ensemble des filtres par famille : Chaque barre représente une des familles analysées. La hauteur de cette barre correspond au nombre de gènes ayant passé l'ensemble des filtres pour chaque famille. Les barres vertes caractérisent les familles pour lesquelles le 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). La barre orange caractérise la famille pour laquelle un candidat potentiel a été identifié (le symbole du gène est écrit au-dessus suivit d'un " ?"). Les barres rouges indiquent 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

Description de la cohorte

Historique : après avoir mis en évidence DNAH1 -> MMAF notre équipe s'est en partie spécialisé dans ce syndrome.

ainsi, entre (année) et année, notre équipe a effectué le séquençage de ... individus présentant ce phénotype afin d'en établir la cause génétique. parmi ces patients, la majorité provenait d'Afrique du Nord, cependant ... vfenaient de et de ... ces séquençage ont été effectué dans ... centres diférents que sont (listes des centre de séquençage) et sur ... plateforme : liste des plateformes

Table 4.4 – Liste des différents projets de séquençages effectués

Place	Year	Nb of sequenced individuals
MountSinai	2012	2
Strasbourg	2012	13
Genoscope	2013	13
Genoscope	2014	28
Genoscope	2015	6

Application de la pipeline - Résultats

Après avoir appelé les variants de nos ... patients, nous avons obtenu une total de 677228 variants différents comprenant 628793 SNVs et 48435. Ces variants étant répartit entre chaque patients qui portaient environs chacun 81618 SNV et 5148 indels dont 0 % étaient homozygote. Comme on peut le voir, la proportion de chaque appel est relativement homogène lorsque l'on compare les patients ayant été séquencés dans le même centre la même année. Cependant, il est possible de noter de grandes disparités lorsque l'on compare les données provenant de différents centres ou bien du même centre avec plusieurs années de différences. Ces écarts peuvent-être causés par plusieurs facteur, tel que les différents kits de capture d'exons qui on put être utilisés puisque ... (todo lister les différents kit de capture dans une table) en revanche nous pouvons écarté un effet dus à la platforme de séquençage ou encore le modèle de séquenceur puisque tout ces projets ont été réalisés sur des Illumina HiSeq2000 (**Table :** 4.4) (**Figure :** 4.9 - **A**).

Le même constat peut être effectué lorsque l'on compare la qualité des appels puisque plus les projets de séquençage s'avèrent être récent, plus la proportion d'appel *Single Strand* s'avère être faible tandis que la proportion d'appel *Double Strand* (DS) est élevée. Ceci est une bonne chose, car, bien que ces deux appels soient conservés dans les analyses ulterieures, les appels DS sont de meilleur qualité que les appels SS. Cette augmentation des appels DS au cours du temps pourrait s'expliquer par une amélioration des protocole de séquençage qinsi que des kit de capture. En revanche cela est à pondérer avec le taux croissant d'appels *No-strand* (NS) au fur et à mesure des années pour atteindre environs ... % en (... Année) avec un projet réalisé au Génoscope. Ces derniers appels étant systématiquement filtrés, ils n'altèreront en rien les résultats obtenus en aval ormis le fait qu'ils réduisent la quantité des données utilisées (**Figure :** 4.9 - **B** et **C**).

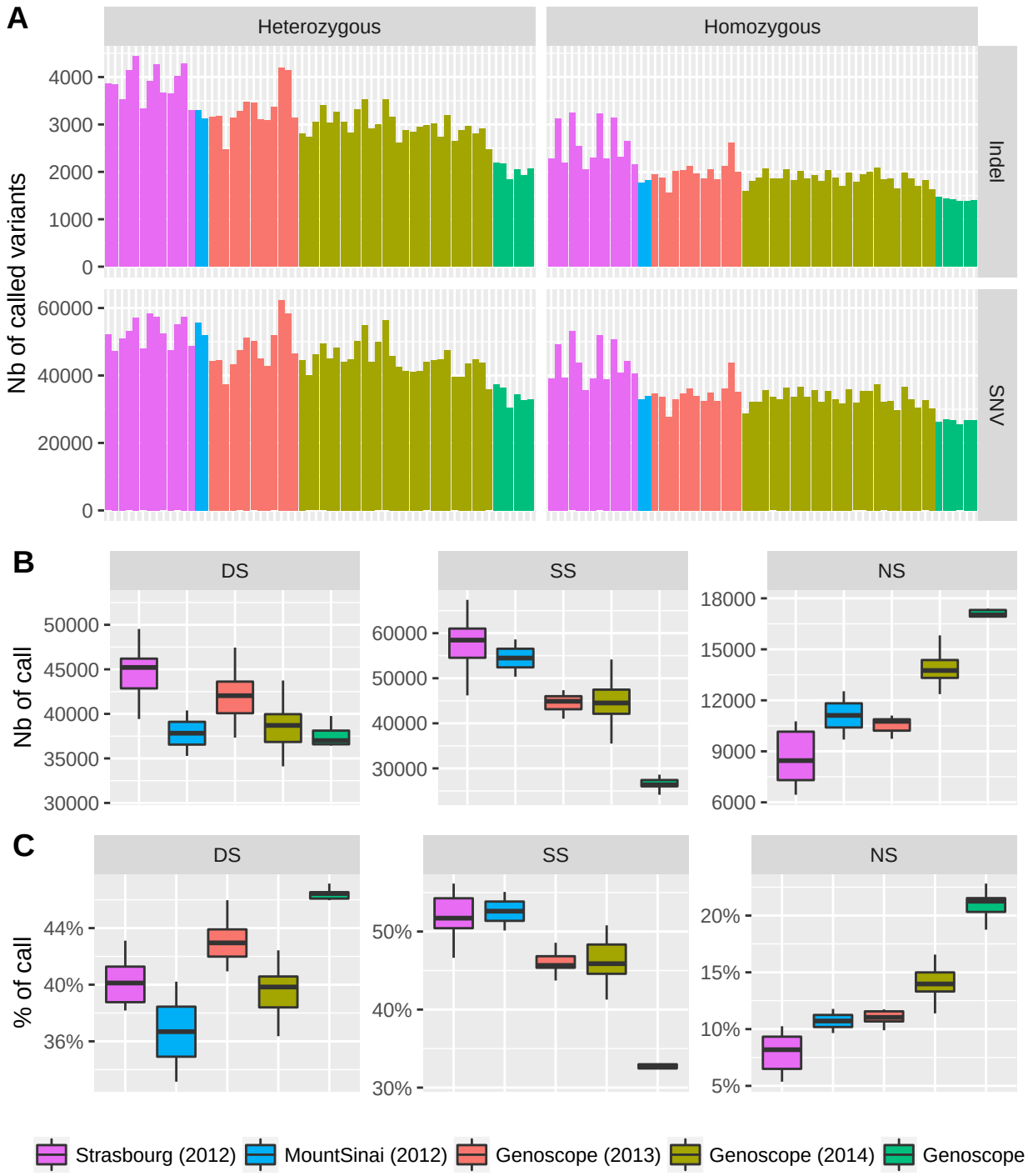

Figure 4.9 – Résultats de l'appel des variants par individus et par projet de séquençage : Chaque couleur définit un projet de séquençage caractérisé par un centre de séquençage et une année. **A** : Quantification pour chaque individus (représentés par les barres) du nombre de variants (SNVs et Indels) appelés homozygotes et hétérozygotes. **B** : Quantification des appels *Double Strand* (DS), *Single Strand* (SS) et *No strand* (NS) pour chaque projet de séquençage. **C** : Même chose en pourcentage

Après avoir appliqué les mêmes filtres que ceux décrit précédement à l'exception du filtre n°... Union puisqu'ici nous avons uniquement des individus non apparentés, nous avons put obtenir une liste de 1568 variants différents composés de 1359 SNVs et 209 indels et impactant un total de 1306 gènes distincts. Ces variants étant répartis sur l'ensemble de nos ... patients ceux-ci portaient en moyenne 25 SNVS et 4 indels, de sorte que chacun d'entre eux avaient entre 1 et 73 gènes impactés par au moins un variants.

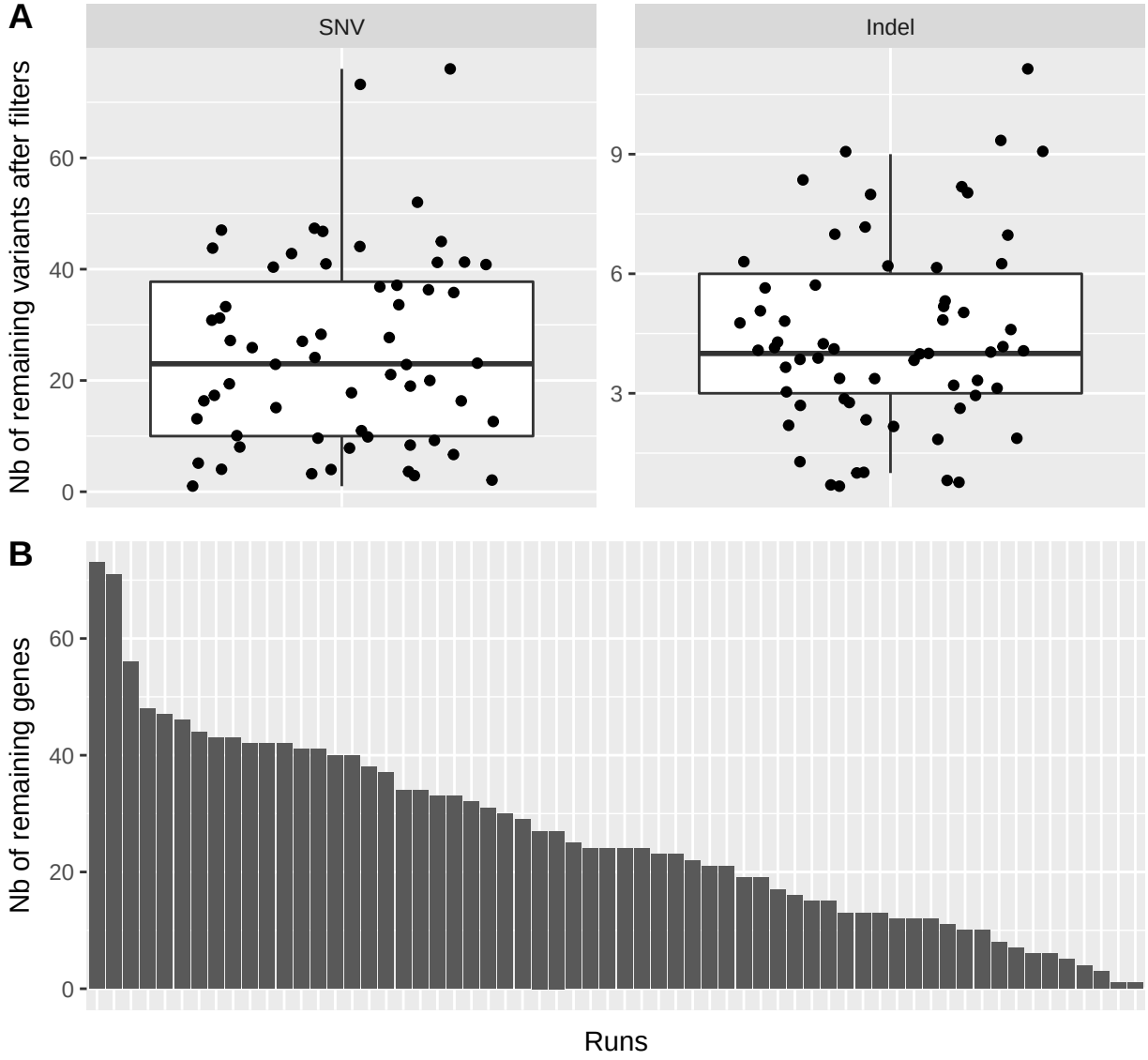

Figure 4.10 – TODOOOOOOOOOOOOOOOOOO : **A** : Quantification du nombre de SNVs et indels ayant passé l'ensemble des filtres pour chaque patients. **B** : Nombre de gènes impactés par au moins un variant ayant passé les filtres pour chaque individus représentés par les barres

Analyse des listes de gènes

Chapitre 5

MutaScript

Conclusion

Chapitre 6

The First Appendix

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., ... Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–9. http://doi.org/10.1038/nmeth0410-248

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

Ben Khelifa, M., Coutton, C., Zouari, R., Karaouzène, T., Rendu, J., Bidart, M., ... Ray, P. F. (2014). Mutations in DNAH1, which encodes an inner arm heavy chain dynein, lead to male infertility from multiple morphological abnormalities of the sperm flagella. *American Journal of Human Genetics*, *94*(1), 95–104. http://doi.org/10.1016/j.ajhg.2013.11.017

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

Hu, Y., Yu, H., Shaw, G., Renfree, M. B., & Pask, A. J. (2011). Differential roles of TGIF family genes in mammalian reproduction. *BMC Developmental Biology*, *11*, 58. http://doi.org/10.1186/1471-213X-11-58

Imai, Y., Morita, H., Takeda, N., Miya, F., Hyodo, H., Fujita, D., ... Komuro, I. (2015). A deletion mutation in myosin heavy chain 11 causing familial thoracic aortic dissection in two Japanese pedigrees. *International Journal of Cardiology*, *195*, 290–292. http://doi.org/10.1016/j.ijcard.2015.05.178

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(7), 1073–1081. http://doi.org/10.1038/nprot.2009.86

Lee, B., Park, I., Jin, S., Choi, H., Kwon, J. T., Kim, J., ... Cho, C. (2011). Impaired spermatogenesis and fertility in mice carrying a mutation in the Spink2 gene expressed

predominantly in testes. *The Journal of Biological Chemistry*, *286*(33), 29108–17. http://doi.org/10.1074/jbc.M111.244905

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

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