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

Chapitre 1 : Delete line 6 if you only have one advisor **1**

Remerciements **3**

Résumé **5**

Chapitre 2 : Introduction **7**

Chapitre 3 : Investigation génétique et physiologique de la globozoospermie **9**

Chapitre 4 : Mise en place d'une stratégie pour l'analyse des données exomiques – application en recherche clinique **11**

4.1 Intro 11

4.2 Résultats 12

4.2.1 Description de la pipeline 12

4.2.2 utilisations de la pipeline pour l'identifications de variants pathogènes et l'identification de nouveaux gènes impliqués dans l'infertilité 13

Etude familiale MMAF -> DNAH1 14

Etude familiale echec de fécondation -> PLCzeta 15

Etude familiale azoospermie : SPINK2 17

Etude d'une large cohorte de patients MMAF 22

Chapitre 5 : MutaScript **27**

Conclusion **29**

Chapitre 6 : The First Appendix **31**

References **33**

Liste des tableaux

4.1 Comptage des variants communs à B1 et B2 15
4.2 Comptage des variants communs à B1 et B2 18
4.3 Liste des variants ayant passé l'ensemble des filtres 19

Table des figures

4.1 Comptage des variants retrouvés sur les frères B1 et B2 avec leur génotypes et prédiciotn VEP associés 16
4.2 Arbre généalogique des deux frères azoospermes B1 et B2 17
4.3 Comptage des variants retrouvés sur les frères B1 et B2 avec leur génotypes associés 19
4.4 Expression du gène *SPINK2* dans plusieurs tissus 20
4.5 Représentation du gène *SPINK2* 20

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 l'impact que celui-ci aura sur la séquence codante de l'ensemble des transcrits qu'il chevauche. 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 (**? ? ?**) et 1000Genomes (**? ? ?**) 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.
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 utilisations de la pipeline pour l'identifications de variants pathogènes et l'identification de nouveaux gènes impliqués dans l'infertilité

Etude familiale MMAF -> DNAH1

Table 4.1 – Comptage des variants communs à B1 et B2

Variant type	Genotype	Count
SNV	Heterozygous	33211
SNV	Homozygous	33195
Indel	Heterozygous	1069
Indel	Homozygous	1196

Etude familiale echec de fécondation -> PLCzeta

[TODO : Redéfinir brievement l'echec d'activation ovocitaire] le rôle de

Cette étude se concentre donc sur le phénotype d'infertilité de deux frères Tunisiens (PLCZ_1 et PLCZ_2) issus d'un union consanguin et présentant un phénotype d'echec total de fécondation même lors de tentative d'ICSI. Après avoir effectué un séquençage exomique pour ces deux frères, nous avons appliqué notre pipeline d'analyse sur les résultats du séquençage. Compte tenu de l'hitorique de consanguinité familiale, nous nous sommes concentré uniquement sur les variants homozygotes. De même, ces deux patients présentant le même phénotype et étant frères, nous avons filtré l'intégralité des variants n'étant observés dans les deux exomes. nous ensuite aussi confronté les variants restant à notre base de données de variants identifié sur 132 individus sains ou résentant un autre phénotype d'infertilité masculine filtrant ainsi les variants retrouvés homozygotes chez les individus contrôles. Après avoir effectué l'intégralité des filtres, seul 1 variants impactant 1

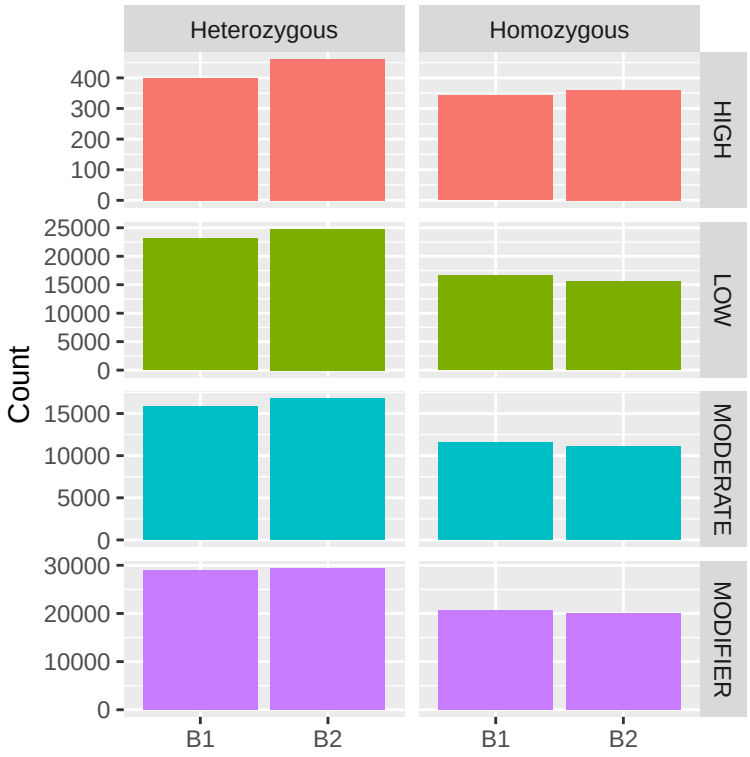


Figure 4.1 – Comptage des variants retrouvés sur les frères B1 et B2 avec leur génotypes et prédiciotn VEP associés

Etude familiale azoospermie : SPINK2

Description de la cohorte familiale Comme nous avons pu le voir, l'azoospermie est un phénotype d'infertilité masculine caractérisé par l'absence de spermatozoïde dans l'éjaculat. En 2017, notre équipe a publiée des résultats liant une mutation du gène *SPINK2* à ce phénotype. Pour cette étude, nous avons effectué un séquençage exomique de deux frères azoospermes (B1 et B2) issus d'un union consanguin puisque leurs parents sont cousin au deuxième degré (**Figure :** 4.2). Étant donné l'historique consanguin de cette famille nous avons emis l'hypothèse d'un phénotype à transmission autosomique recessif.

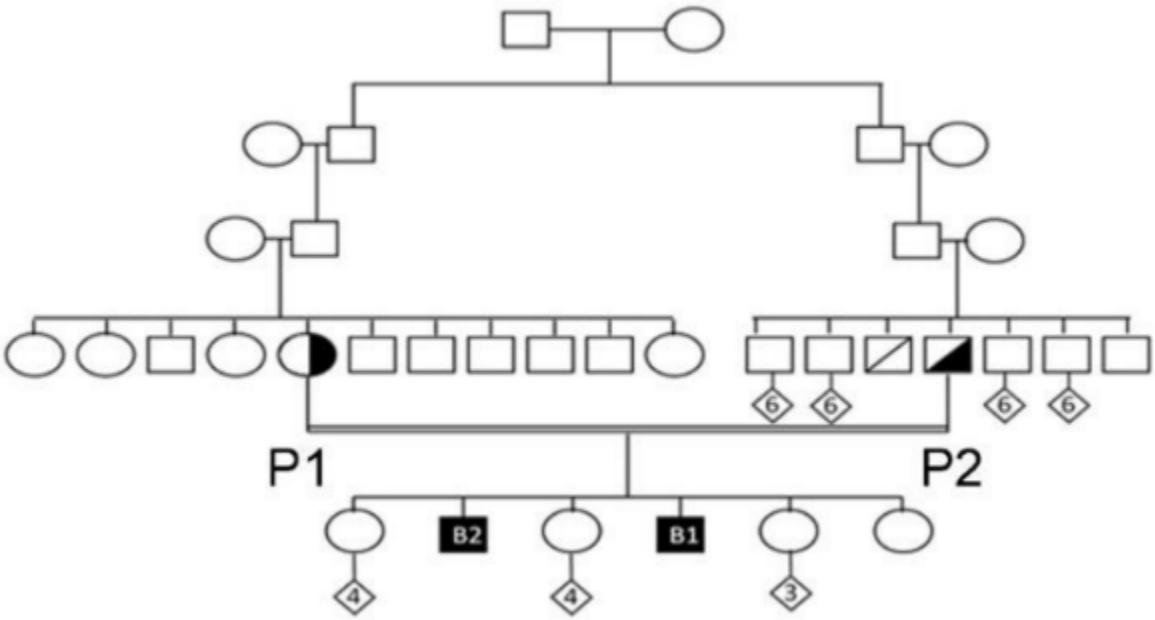


Figure 4.2 – Arbre généalogique des deux frères azoospermes B1 et B2 : Sur cette arbre nous pouvons observés la parentés des parents (P1 et P2), cousins au deuxième degré

Séquençage WES et analyse bioinformatique Le séquençage exomique des deux frères B1 et B2 fût réalisé au Mount Sinaï Institutes en [TODO : mettre l'année du séquençage]. Suite à celà nous avons, comme pour les études précédentes nous avons appliqué notre pipeline d'analyse nous permettant de recansser un total de 88546 variants différents pour B1 et 66521 pour B2 (**Figure :** 4.3). Ayant emmit l'hypothèse d'une cause génétique commune expliquant le phénotype des deux frères, nous nous sommes concentrés uniquement sur les variants homozygotes communs aux deux frères réduisant ainsi notre liste de variant à un total de 26825 variants dont 25963 SNVs et 862 indels (**Table :** 4.2). À partir de cette liste, nous avons soustrait l'ensemble des variants prédit par VEP comme n'impactant pas un transcrit codant pour une protéine. De même, nous avons filtrer l'ensemble des variant présent

Table 4.2 – Comptage des variants communs à B1 et B2

Variant type	Genotype	Count
SNV	Heterozygous	26634
SNV	Homozygous	25963
Indel	Heterozygous	852
Indel	Homozygous	862

dans les séquences UTRs ou encore ceux causant une substitution synonymes ou une substitution faux-sens prédite "benign" par PolyPhen2 et "tolerated" par SIFT. Suite à cela, nous avons filtrer les variants fréquents dans la population générale en filtrant l'ensemble des variant ayant une MAF ≥ 0.01, de même, nous avons confrontés les variants restant à une base de donnée interne ressencant les variant de 83 individus non atteint d'azoospermie afin de filtrer les variants homozygotes retrouvés dans cette basse de données de contrôle.

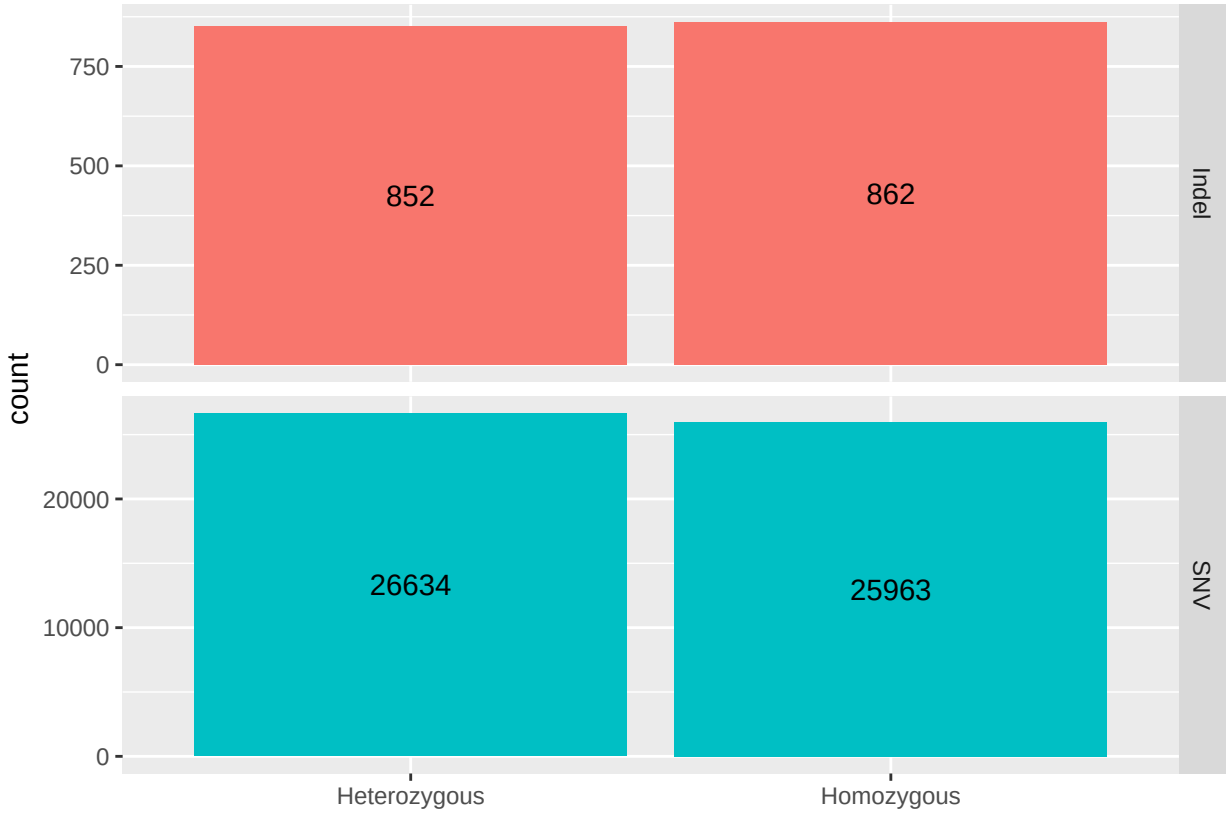


Table 4.3 – Liste des variants ayant passé l'ensemble des filtres

Chromosome	Position	Reference allele	Alterated allele	Gene
4	57686748	G	C	SPINK2
4	44683156	G	T	GUF1

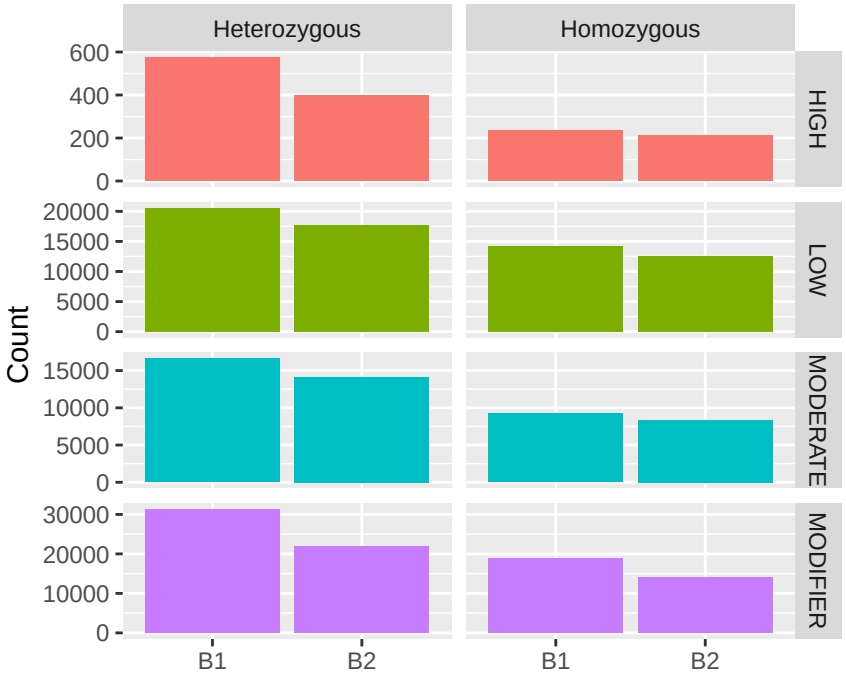


Figure 4.3 – Comptage des variants retrouvés sur les frères B1 et B2 avec leur génotypes associés

Après avoir appliqué l'ensemble de ces filtres, nous sommes arrivés à une liste de 2 variants impactant 2 gènes différents : *SPINK2* et *GUF1* (**Table :** 4.3). Parmi ces deux gènes, seul *SPINK2* était décrit comme fortement exprimé dans le testicule [INS2RER FIGURE ACEVIEW]. Nous avons d'ailleurs pu confirmer cette forte expression par RT-PCR quantitative en temps réel forte à la fois chez l'Homme (**Figure :** 4.4 - **A**) et chez la souris (**Figure :** 4.4 - **B**). Ces données ont donc fait de *SPINK2* le seul candidat évident pouvant expliquer ce phénotype. Le variant partagé par les deux frères : Chr4 :57686748G>C n'a été recencé dans aucune des bases de données que sont ExAC, 1000Genomes et ESP6500. Le gène *SPINK2* est localisé sur le chromosome 4 et contient 4 exons (**Figure :** 4.4). Sa localisation intronique à 3 pb du $2^{ième}$ exon indique que ce variant pourrait avoir un effet sur l'épissage de l'ARNm

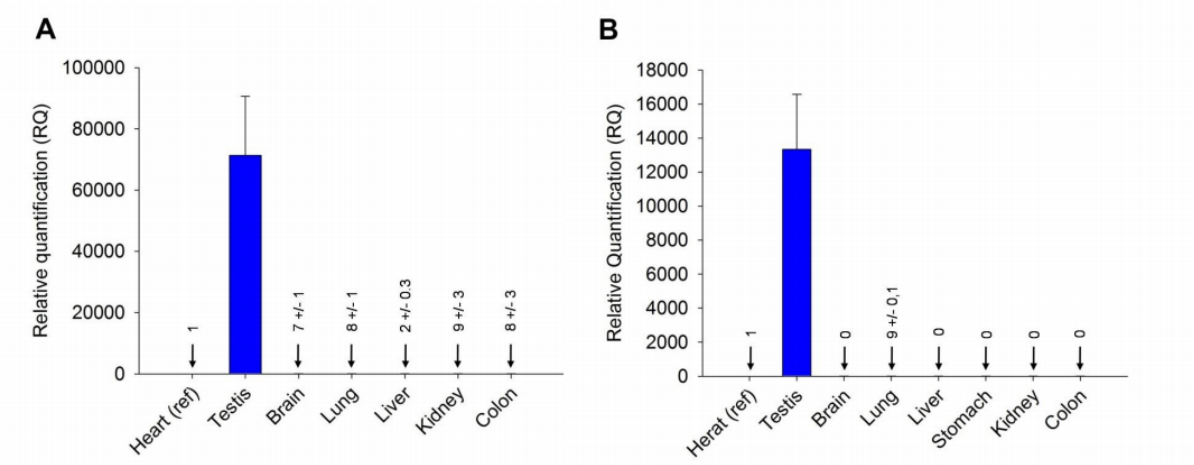


Figure 4.4 – Expression du gène *SPINK2* dans plusieurs tissus : On peut constater que chez l'humain (**A**) comme chez la souris (**B**), le gène *SPINK2* a non seulement une forte expression exclusive au testicule

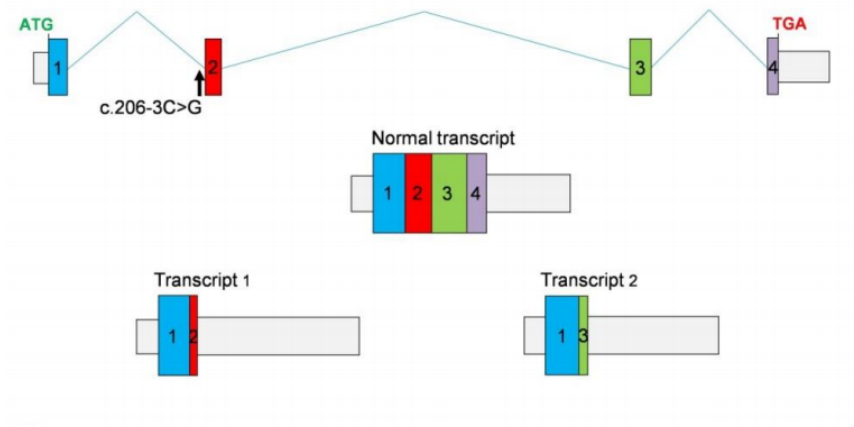


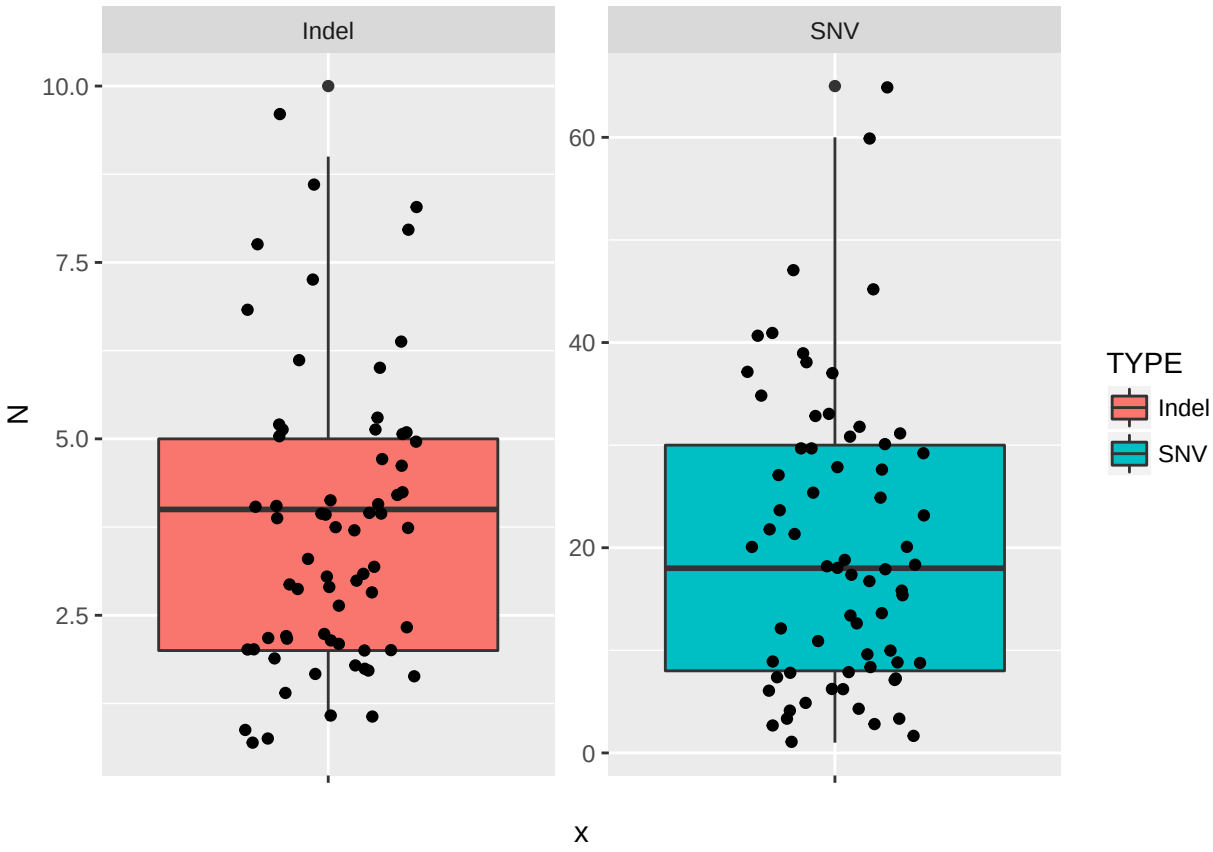
Figure 4.5 – Représentation du gène *SPINK2* : L'épissage du gène *SPINK2* crée un transcrit de 4 exons (Transcrit normal). Cependant, le variant c.206-3C>G observé chez les frères B1 et B2 crée un nouveau site accepteur d'épissage ajoutant 2 nucléotides à l'exon 2 induisant un décalage du cadre de lecture menant à un codon stop 3 nucléotides plus loin (Transcrit 1) et / ou causant le saut de l'exon 2 menant à un codon stop prématuré (Transcrit 2)

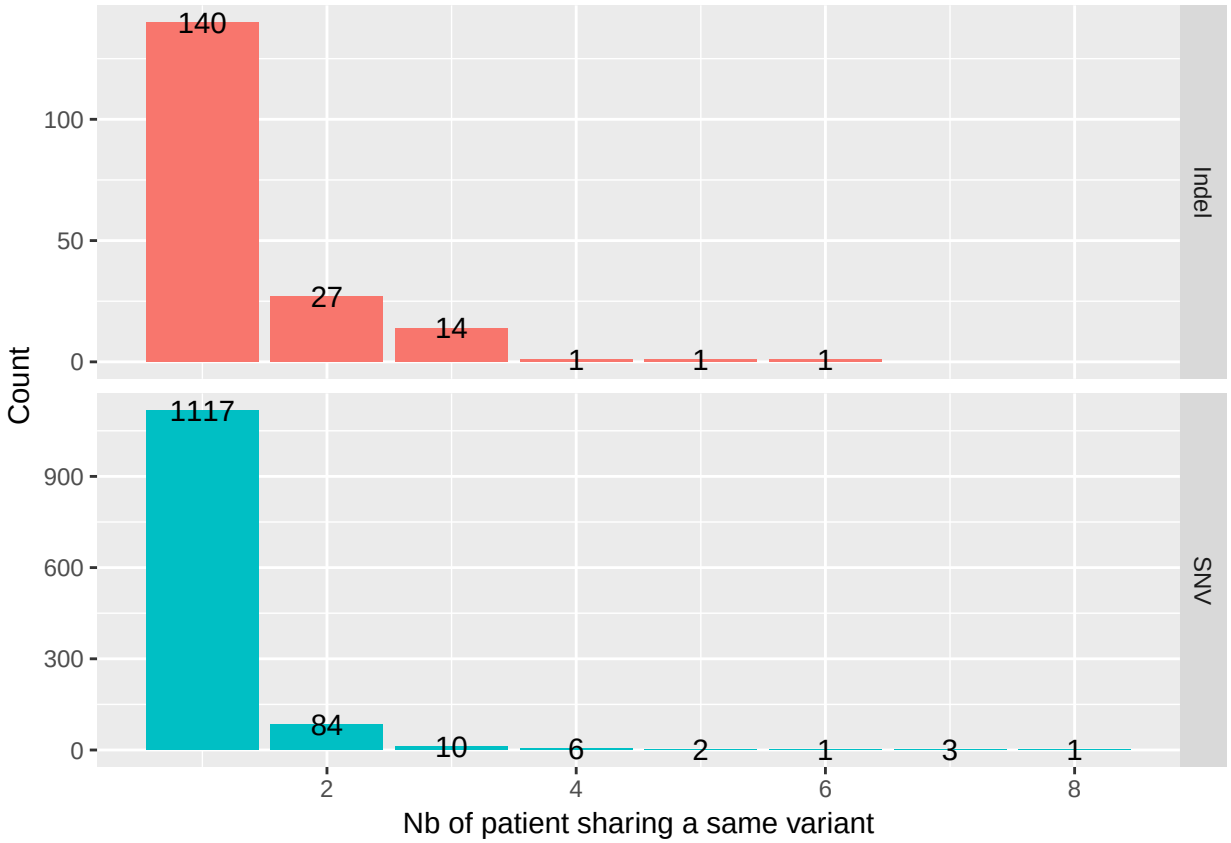
Estimation de l'incidence du gène *SPINK2* chez des patients azoospermes
Afin de déterminer l'incidence des variants du gène *SPINK2* dans la population azoospermiques nous avons séquencé l'intégralité des séquences codantes de ce gène chez 611 individus comprenant 210 patients azoospermiques, 393 oligozoospermiques et présentant un phénotype non spécifié. Parmi cet ensemble de patient, seul le

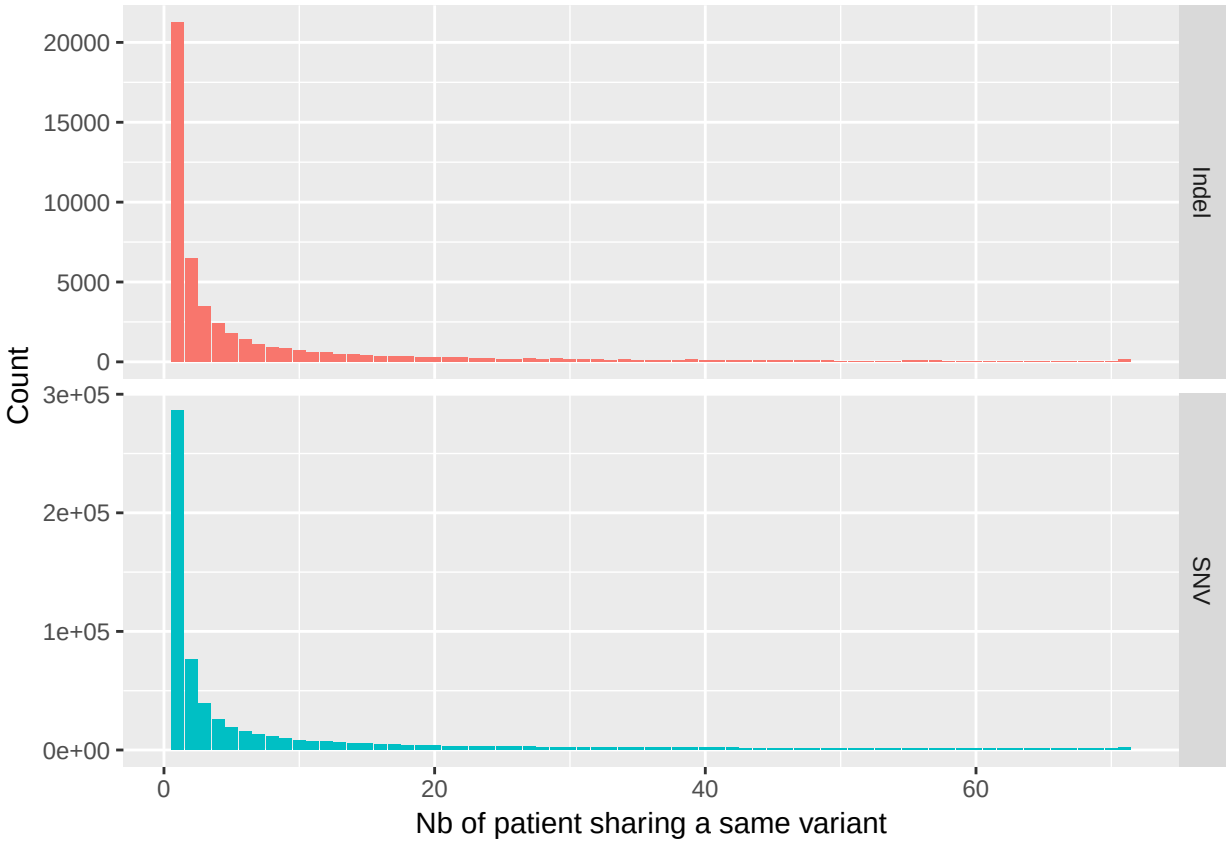
patient 105 (P105) présentant un phénotype d'oligozoospermie s'est révélé porteur d'un variant non répertorié dans les bases de données et présentant un impact prédit comme délétère. Ce variant, c.1A>T, présent à l'état hétérozygote chez P105 affecte le codon start décalant ainsi le démarage de la traduction produisant ainsi une protéine tranquée.

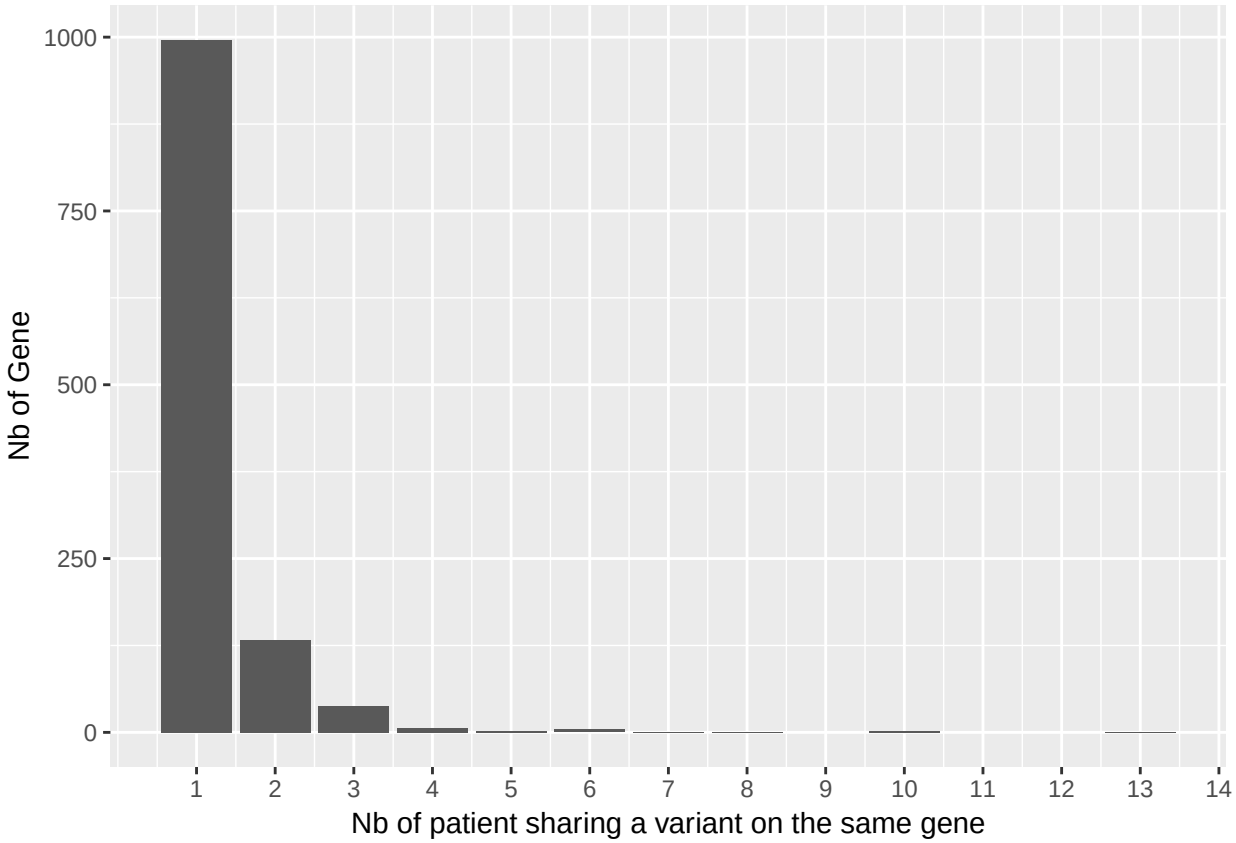
Autres résultats Dans cette même étude nous avons égallement pu étudié les caractéristiques reproductives de souris KO *Spink2*$^{-/-}$ et de souris hétérozygotes *Spink2*$^{+/-}$. Ainsi, nous avons pu confirmer que les souris mâles KO étaient infertiles et présentaient un phénotype d'azoospermie ainsi qu'une diminution de la taille de leurs testicules tandis que les hétérozygotes étaient parfaitement fertiles bien leur nombre de spermatozoïdes par mL de sperme soit plus faibles que les souris sauvages. Les souris femelles, elles, présentaient des caractéristiques reproductives tout à fait normal. [TODO : finir cette partie].

Etude d'une large cohorte de patients MMAF




140
27
14
1
1
1
1117
84
10
6
2
1
3
1
100
50
0
900
600
300
0
Count
Indel
SNV
2
4
6
8
Nb of patient sharing a same variant


20000
15000
10000
5000
0
Indel
3e+05
2e+05
1e+05
0e+00
SNV
Count
0
20
40
60
Nb of patient sharing a same variant


1000
750
500
250
0
Nb of Gene
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Nb of patient sharing a variant on the same gene

Chapitre 5

MutaScript

Conclusion

Chapitre 6

The First Appendix

References

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