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 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 Utilisation de la pipeline dans des cas familiaux :

Description des familles

Dans cette partie, je me concentre sur l'analyse bioinformztique des résultat du séquençage exomique

1. *Fam azoo* : Pour cette étude, nous avons effectué un séquençage exomique de

Table 4.1

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

deux frères azoospermes (Az1 et Az2) issus d'un union consanguin puisque leurs parents sont cousin au deuxième degré.
2. *Fam fert fail* : PLCZ
3. *Fam MMAF 1-4* : DNAH1

Resultats des exomes

Dans ce chapitre nous nous concentrerons sur l'analyses des résultats du séquençage exomique de [TODO : STOP HARDCODING FOLLOWING VALUES] 13 individus provenant de 6 familles distinctes. La première de ces famille (Famille Az) est composée de deux frères (Az1 et Az2) tout deux porteur d'un phénotype d'azoospermie non-obstructive. La deuxième famille (famille FF) est elle composée de deux frères (FF1 et FF2) infertiles dont les spermatozoïdes ne parviennent pas à féconder les ovocytes. Les quatre dernières familles (MMAF1, MMAF2, MMAF3 et MMAF4) sont toutes composées de deux frères infertiles (sauf MMAF4 qui comprend trois frères) tous atteins d'un syndrome MMAF caractérisé par des anomalies morphologiques multiples au niveau du flagelle spermatique. Un séquençage exomique fût réalisé pour l'ensemble des individus de ces quatres familles.

Pour l'ensemble des individus de ces quatre familles nous avons appliqué notre pipeline d'analyse de sorte à obtenir pour chaque patient une liste de SNV et d'indel avec leur génotype associé (**Figure :** 4.1). Parmi ces variants, nous avons filtrés l'ensemble de ceux chevauchant uniquement des transcrits annotés *NMD* (*non sense mediated decay*) par VEP qui représentent entre XXX et YYY (TODO)% des transcrits (**Figure :** 4.2) afin de nous concentrer sur les variants ayant la plus forte probabilité d'avoir un effet sur une protéine. Chacun de ces variants a été confronté à une liste de variants homozygotes observés sur les individus sains ou présentant un phénotype différent des patients étudiés de notre bases de données interne (**Figure : ??**). Cette

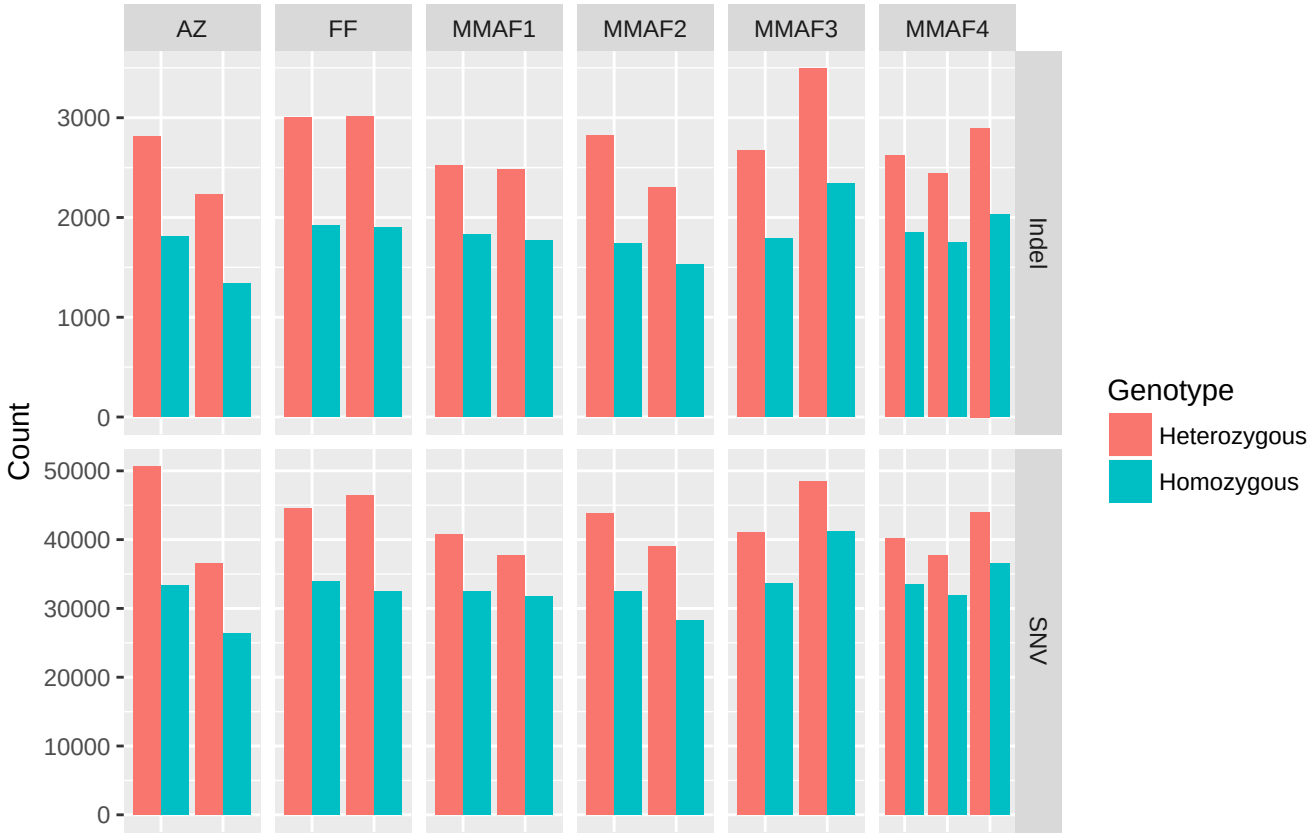
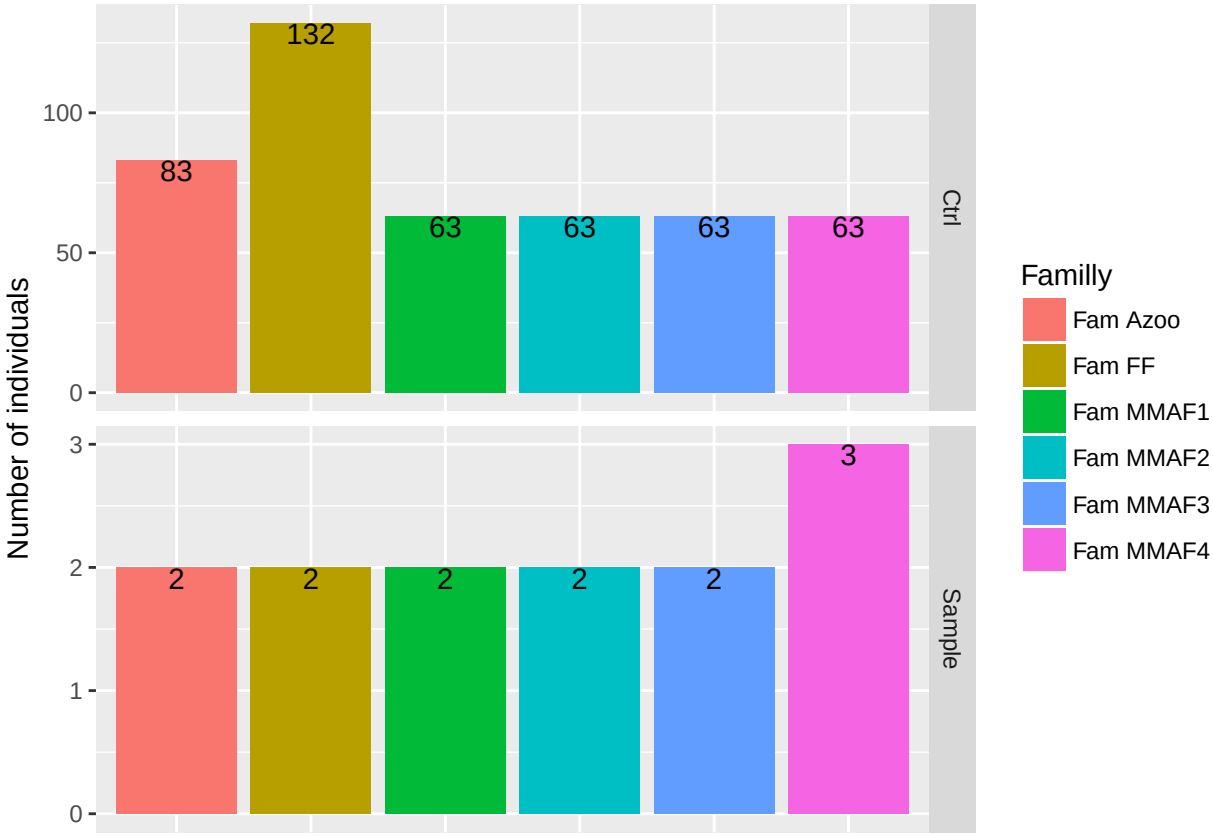


Figure 4.1 – Comptage des SNVs et indels retrouvés par patients avec leur génotypes associés


Number of individuals
100
50
0
3
2
1
0
83
132
63
63
63
63
2
2
2
2
2
3
Ctrl
Sample
Familly
Fam Azoo
Fam FF
Fam MMAF1
Fam MMAF2
Fam MMAF3
Fam MMAF4

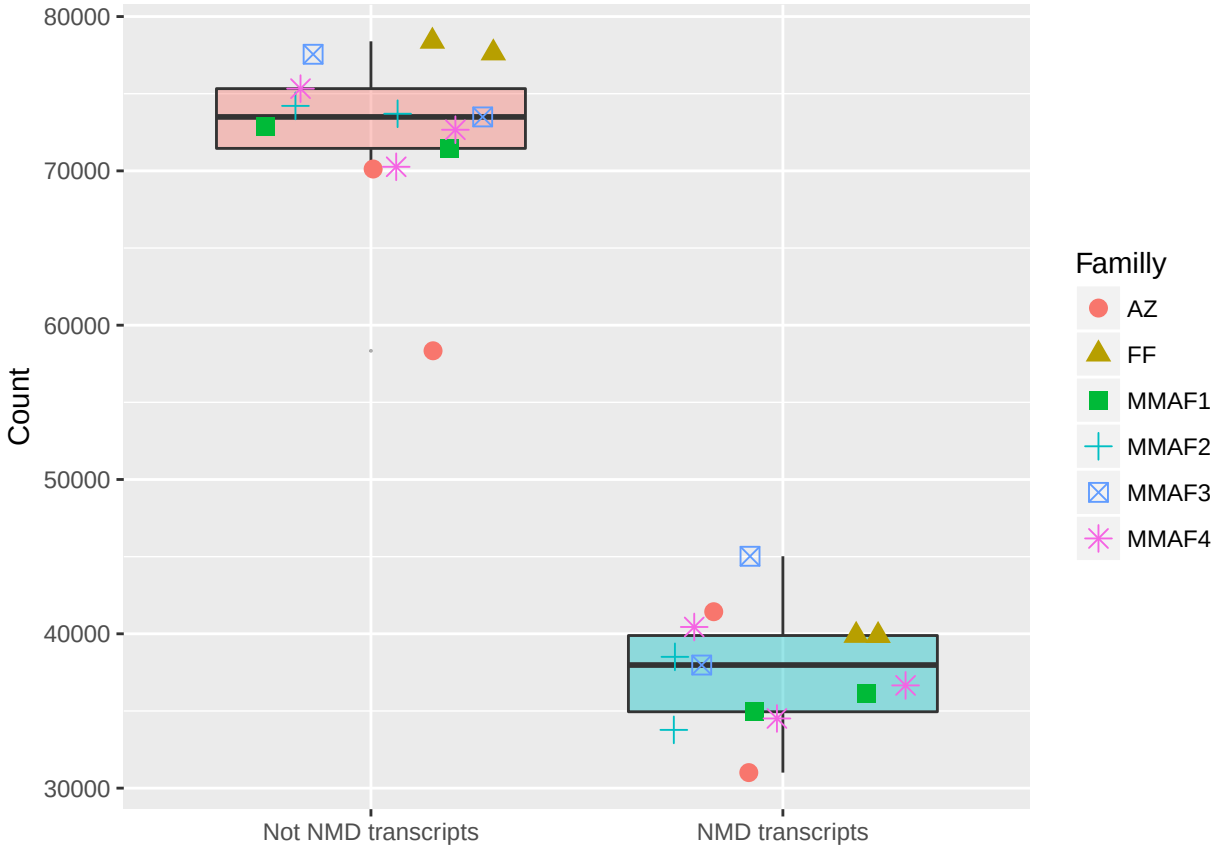


Figure 4.2 – Nombre de transcrits NMD et non NMD impacté par exomes : Chaque point représente un individu séquencé, la couleur et la forme du point dépend de la famille d'origine de l'individu

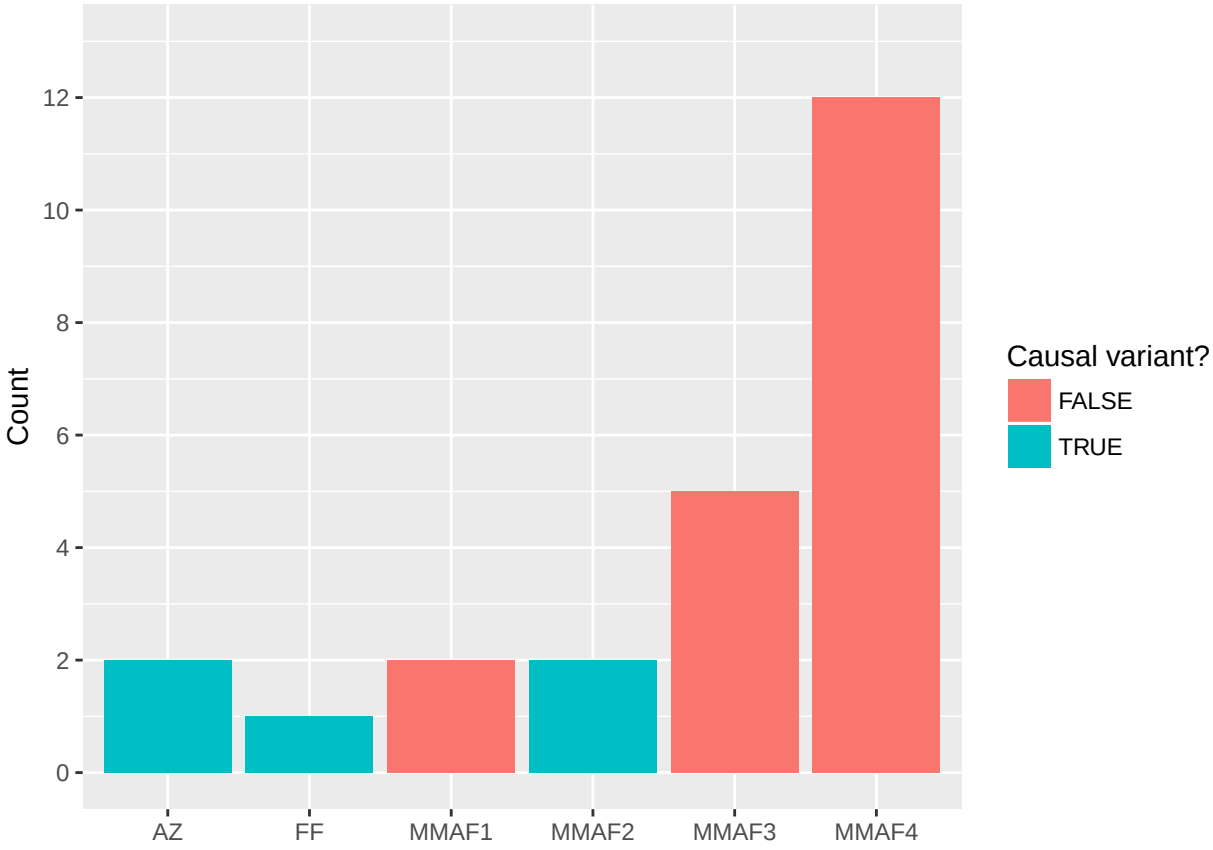
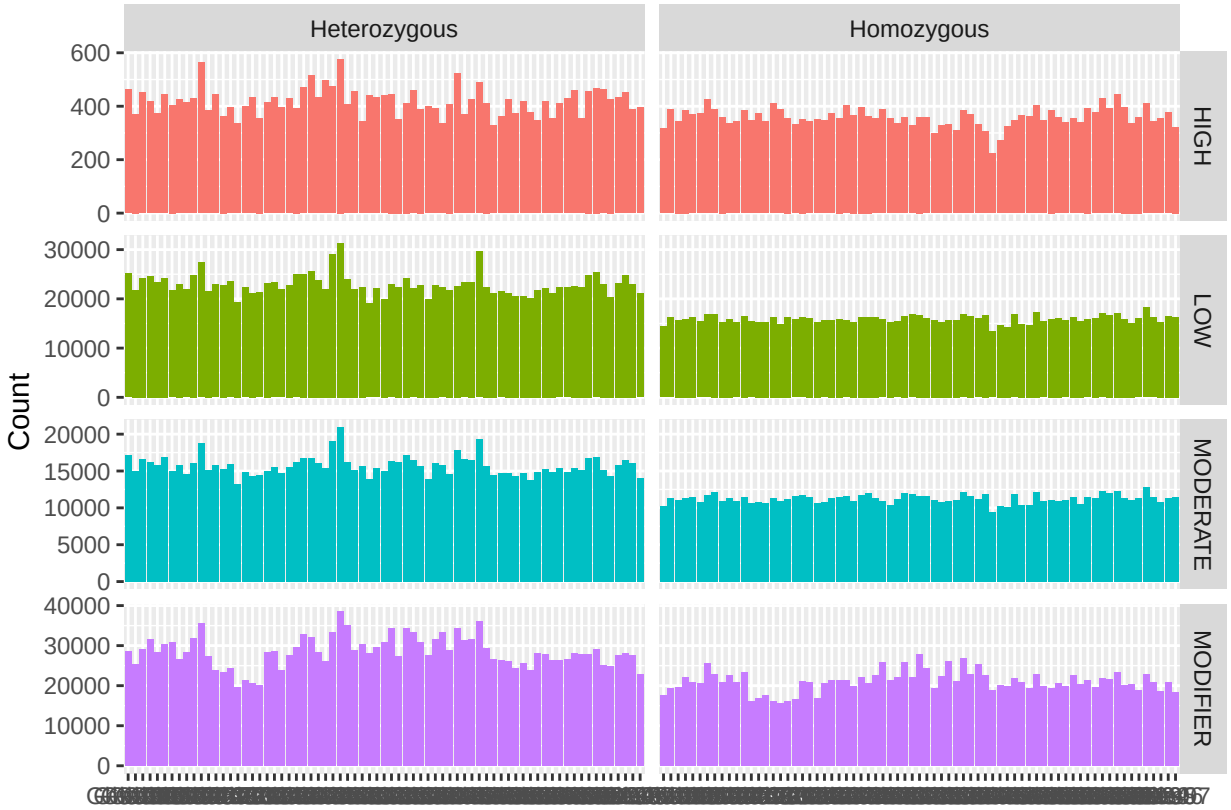
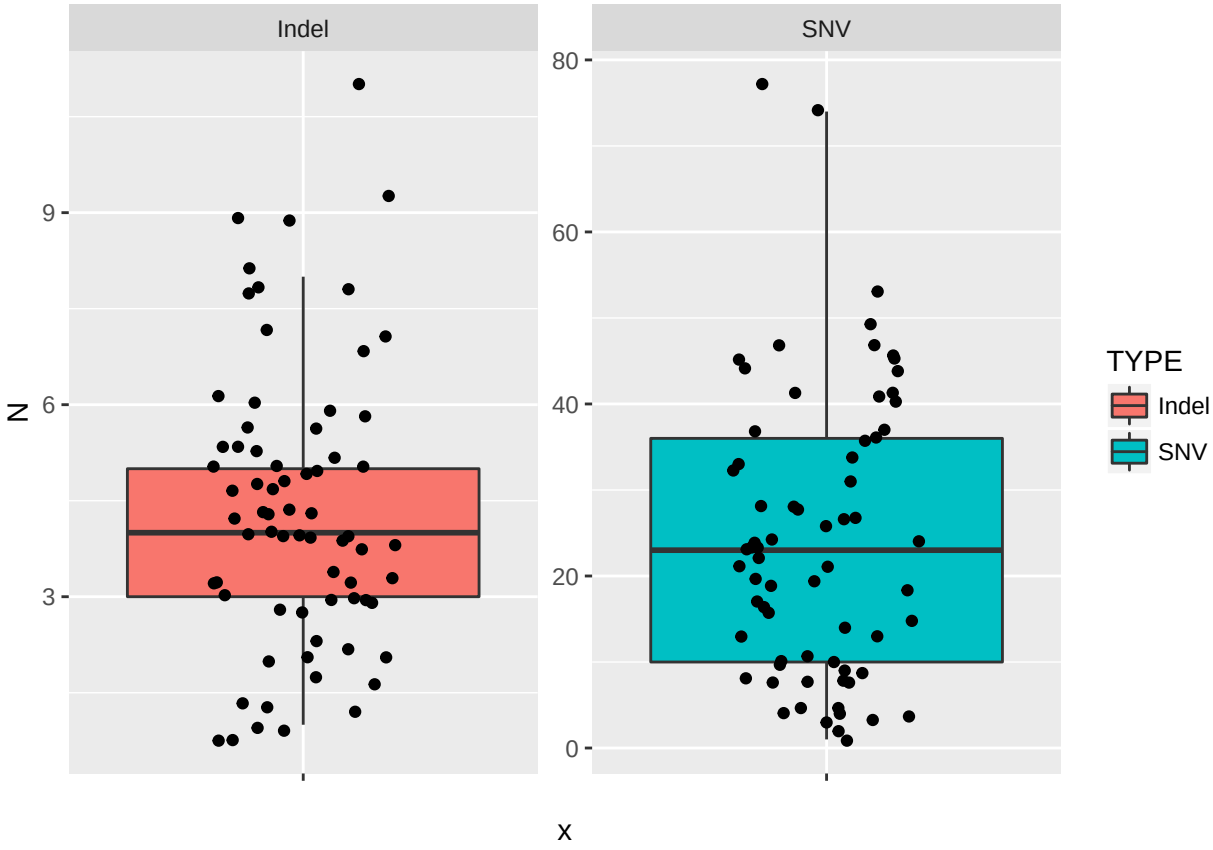
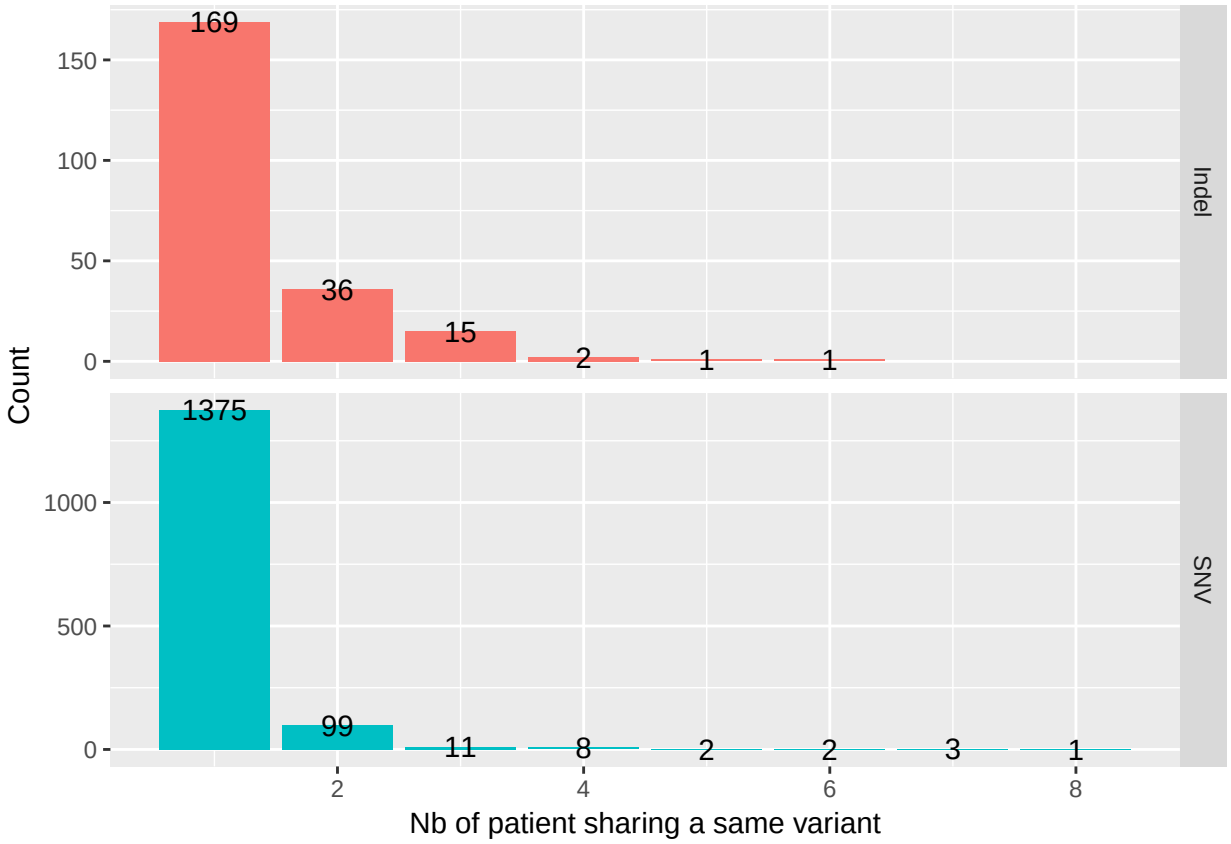


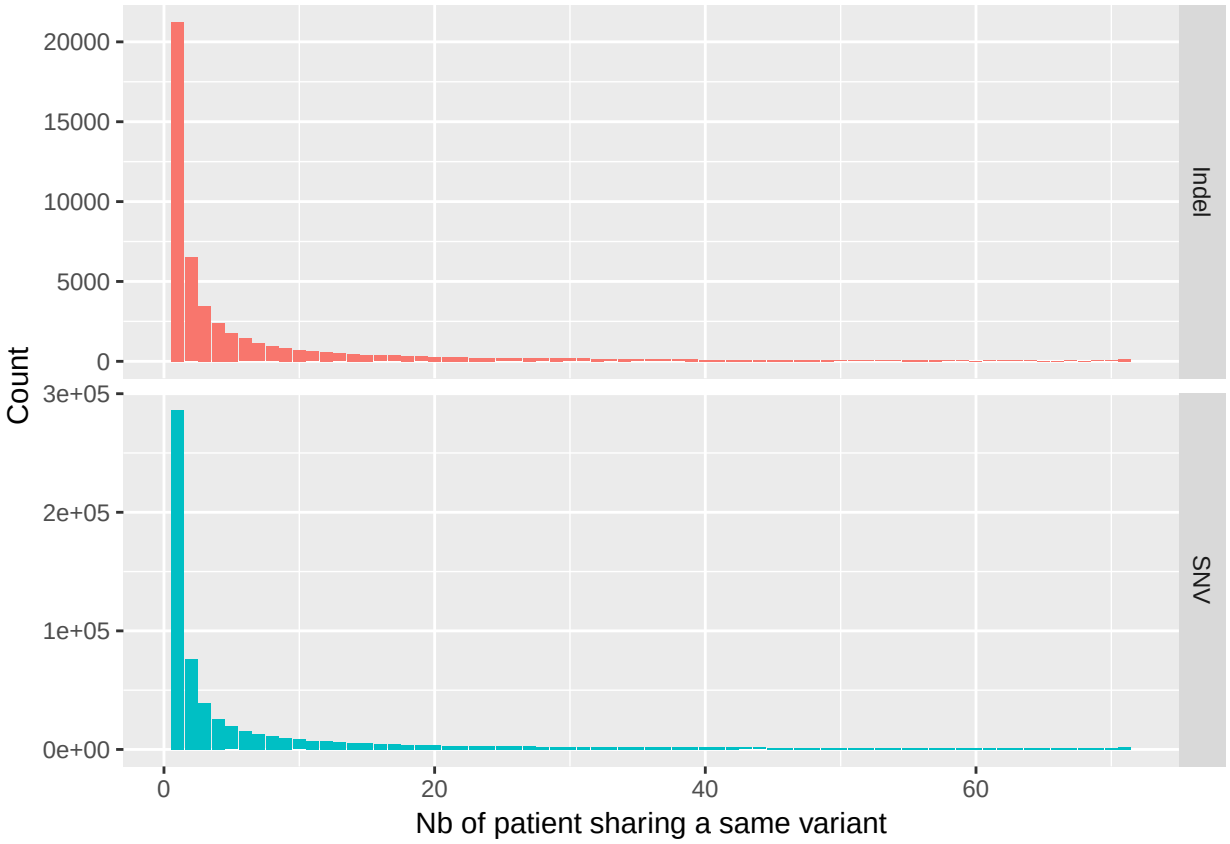
Figure 4.3 – Nombre de gènes passant l'ensemble des filtres par famille

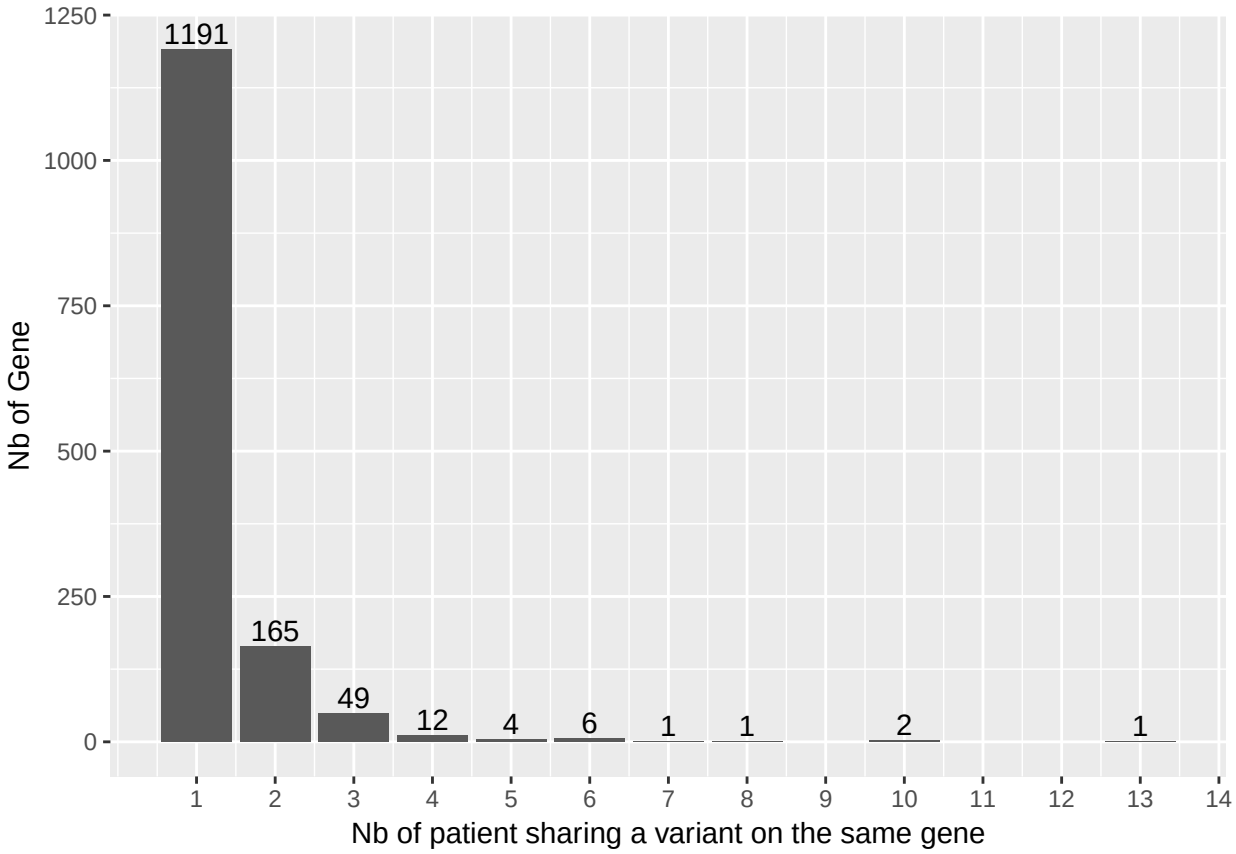
Etude d'une large cohorte de patients MMAF


Heterozygous
Homozygous
HIGH
LOW
MODERATE
MODIFIER
Count
600
400
200
0
30000
20000
10000
0
20000
15000
10000
5000
0
40000
30000
20000
10000
0


Indel
SNV
N
9
6
3
80
60
40
20
0
X
TYPE
Indel
SNV


169
36
15
2
1
1
1375
99
11
8
2
2
3
1
Count
Indel
SNV
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


1250
1000
750
500
250
0
Nb of Gene
1191
165
49
12
4
6
1
1
2
1
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