

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 globo- zoospermie	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 Utilisation de la pipeline dans des cas familiaux :	13
Description des familles	13
Resultats des exomes	14
Etude d’une large cohorte de patients MMAF	20
Chapitre 5 : MutaScript	25
Conclusion	27
Chapitre 6 : The First Appendix	29
References	31

Liste des tableaux

4.1 Tableau recapitulatif des familles séquencées 14

Table des figures

4.1	Nombre de transcrits filtrés car ils sont annotés NMD	15
4.2	Nombre d'individus la cohorte contrôle constituée pour chaque famille de l'analyse	16
4.3	Comparaison du pouvoir discriminant de chaque filtre employé. . . .	17
4.4	Comptage des SNVs et indels retrouvés par patients avec leur génotypes associés	18
4.5	Nombre de gènes passant l'ensemble des filtres par famille	19

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 publié 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 description] 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'intégralité 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 ayant 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 [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 précédemment 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 possible, 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 résultat du séquençage exomique de treize individus provenant de quatre familles différentes (Table : 4.1) :

Table 4.1 – Tableau recapitulatif des familles séquencées

Familly	Individuals	Phenotype	Year	Platform	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

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

1. **L'union des variants** : Dans ces différentes études, nos patients ont à chaque fois au moins un frère présentant le même phénotype. Ainsi nous avons pu formuler l'hypothèse d'une cause génétique commune entre les différents frères d'une même famille et donc filtrer l'ensemble des variants qui ne sont pas partagés par l'ensemble des membre de la fraterie.
2. **Genotype des variants** :
3. **Impact du variant** : on filtre les variants intonique et ceux tombant dans les sequences UTRs. de même les variants synonymes ne sont pas conservés car n'impactent pas la séquences protéique. pour les variant faux sens (changement d'un seul aa de la séquence protéique) il est plus difficile de se décider [TODO insert citation] nous avons donc utiliser les logiciels SIFT et Polyphen et filtrer l'ensemble des fauxsens prédit comme *tolerated* par SIFT et *benign* par Polyphen.
4. **Transcrits NMD** : Le mécanisme NMD (*nonsense-mediated decay*) a pour but de controler la qualité des ARNm cellulaires chez les eucaryotes (Y.-F. Chang, Imam, & Wilkinson, 2007) en éliminant les ARNm qui comportent un codon stop prématuré (Baker & Parker, 2004), pouvant être le résultat d'une erreur de transcription, d'une mutation ou encore d'une erreur d'épissage. Il est donc eu probable que les variants présents sur transcrits annotés NMD soient responsables du phénotype. Nous avons donc filtré l'ensemble des variants chevauchant **uniquement** des transcrits annotés NMD. Cette première étape de filtre permet à elle seule de filtrer systematiquement les variants de 2587 à 3212 transcrits (**Figure : 4.1**) en fonction des individus soit, entre 7261 et 10872 variants différent par individus (**Figure : 4.3**)

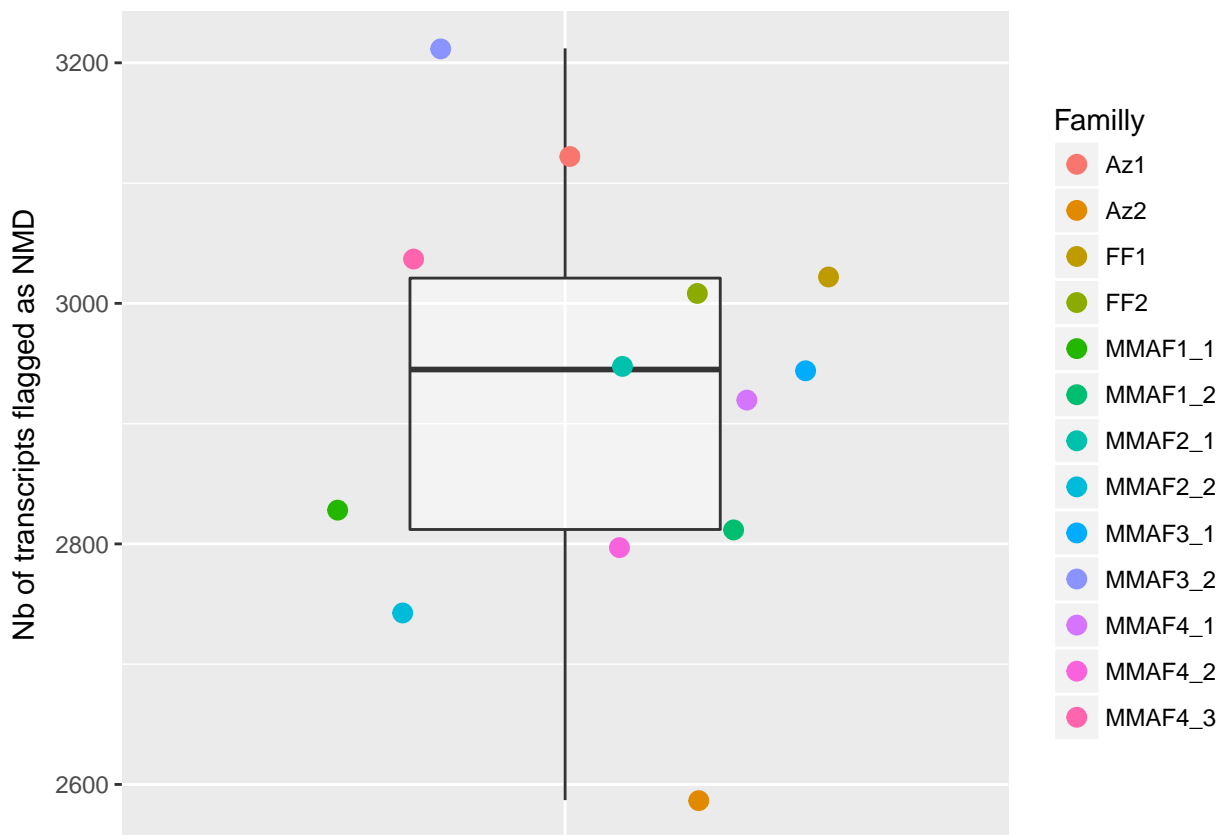


Figure 4.1 – Nombre de transcrits filtrés car ils sont annotés NMD : Chaque point représente un individu séquençé, la couleur et la forme du point dépend de la famille d'origine de l'individu

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 amené à séquencer ... [TODO mettre chiffre exacte]. L'ensemble de ces individus peuvent être soit sains soit présenter l'un des ... phénotypes suivant : ... [TODO insert pheno]. Ces phénotypes étant très différents, il n'est pas aberrant d'émettre l'hypothèse qu'ils que leurs causes génétiques soient différentes. De même, les variants recherchés étant rares, il est peu probable qu'un individu porte les variants de deux phénotypes différents. Ainsi, pour chacune de nos ... [todo insert nb of fam] 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** : 4.2). 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.

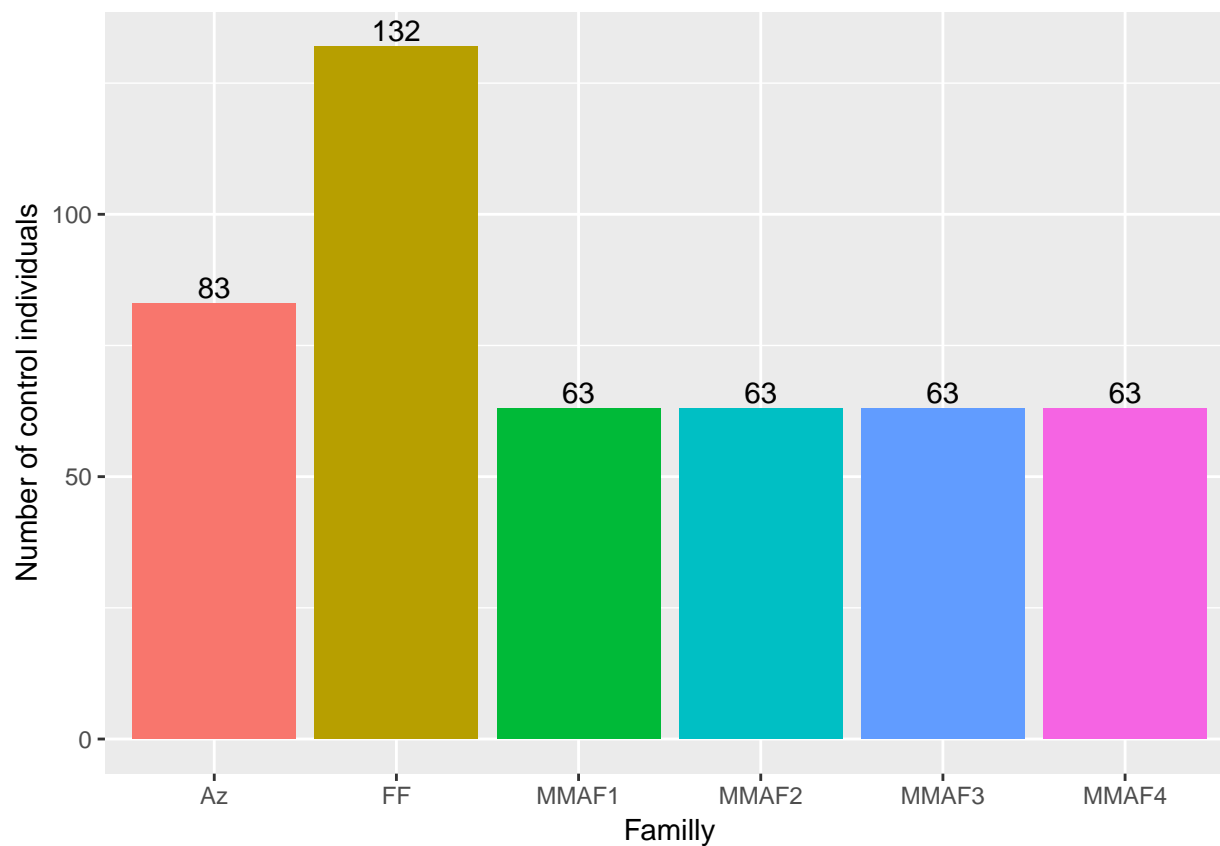


Figure 4.2 – Nombre d'individus la cohorte contrôle constituée pour chaque famille de l'analyse

Afin de comparer le pouvoir discriminant de chacun de ces filtres, nous avons compté le nombre de variant filtrés par chacun d'entre eux (**Figure : 4.3**)

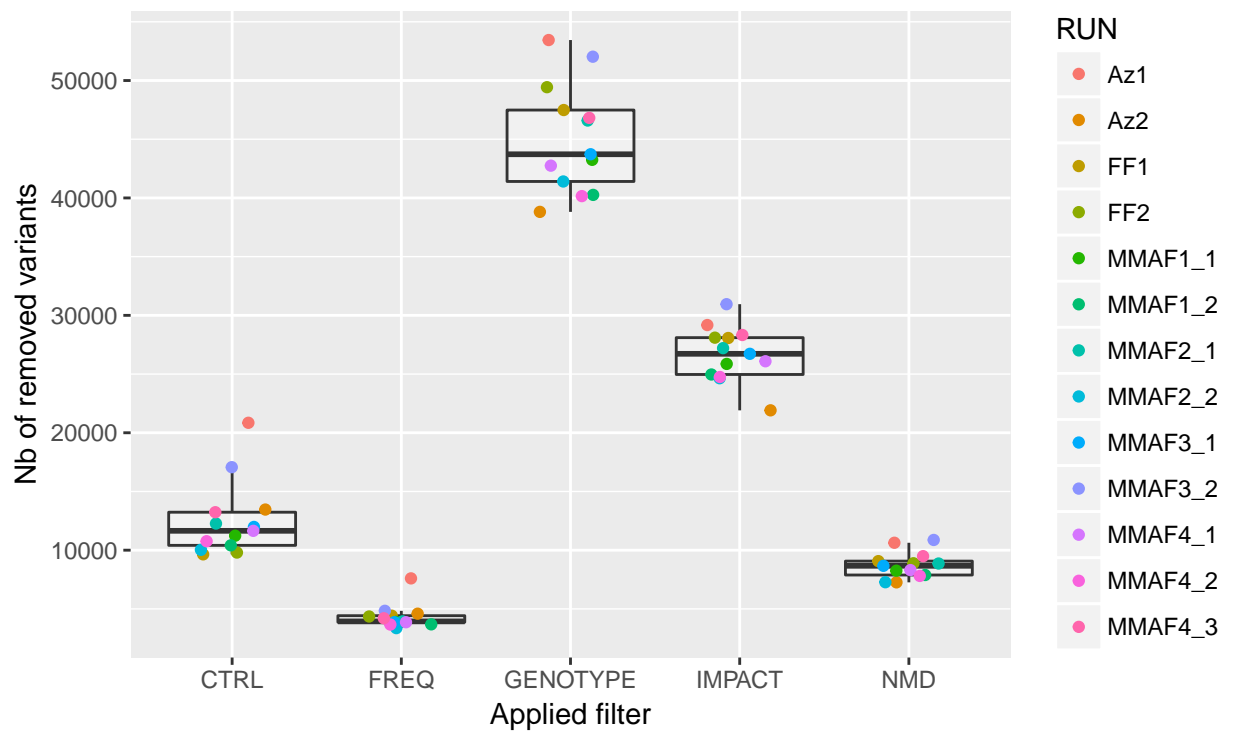


Figure 4.3 – Comparaison du pouvoir discriminant de chaque filtre employé.

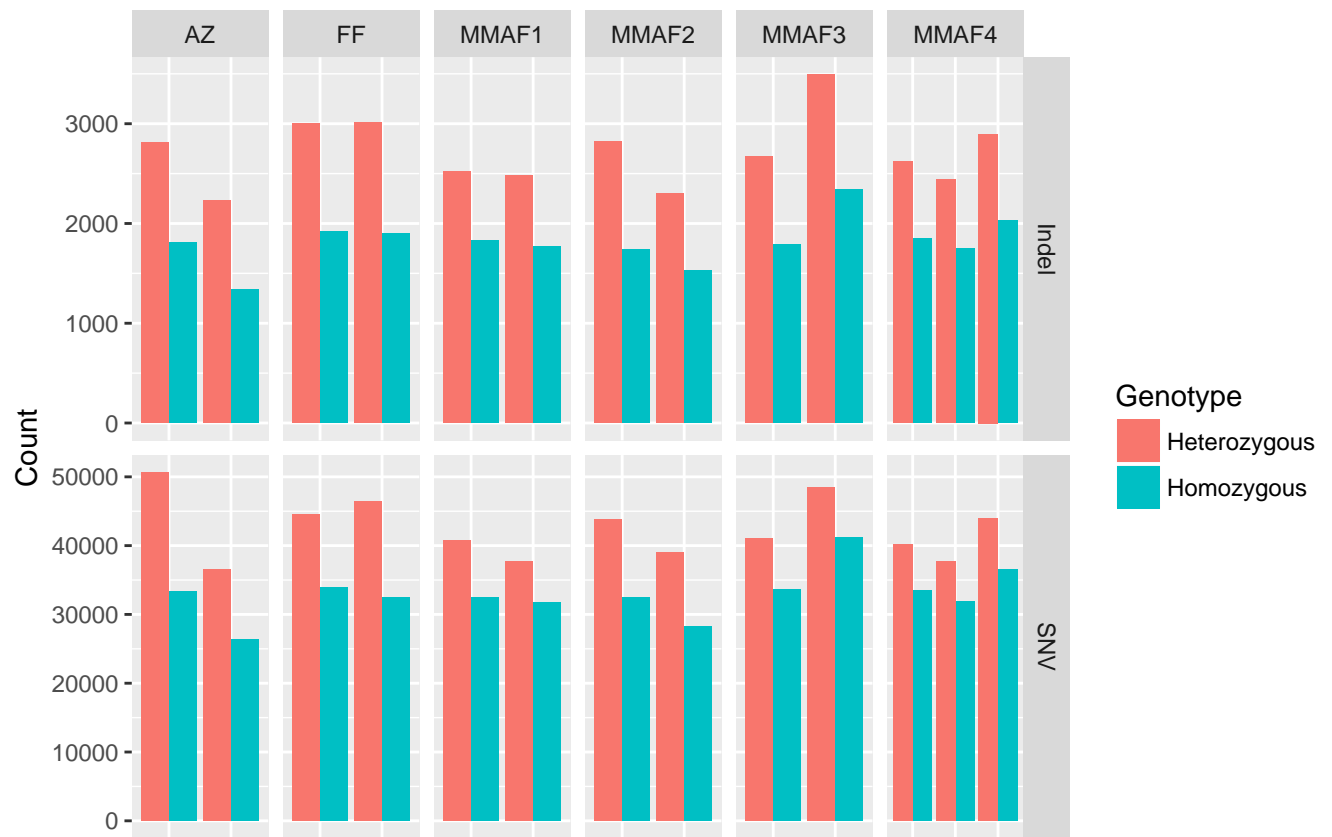


Figure 4.4 – Comptage des SNVs et indels retrouvés par patients avec leur génotypes associés

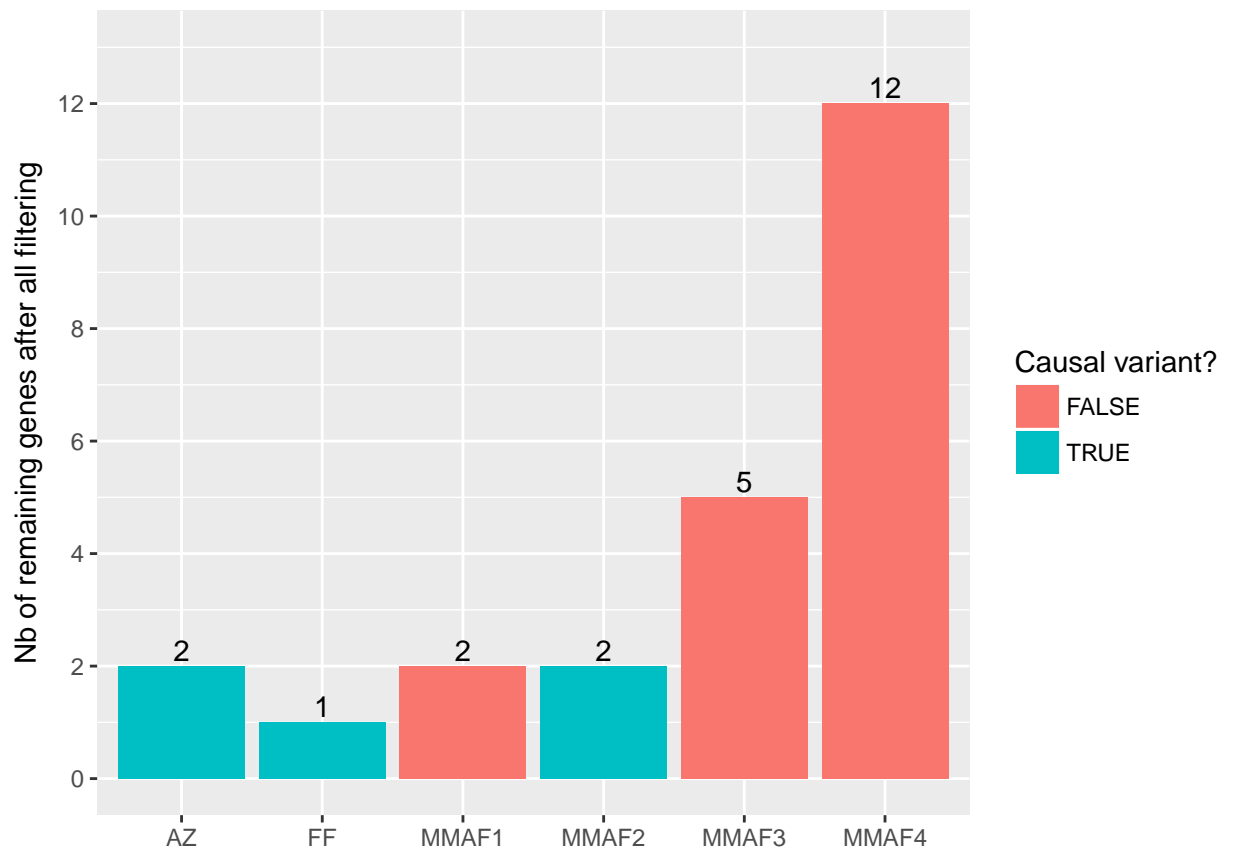
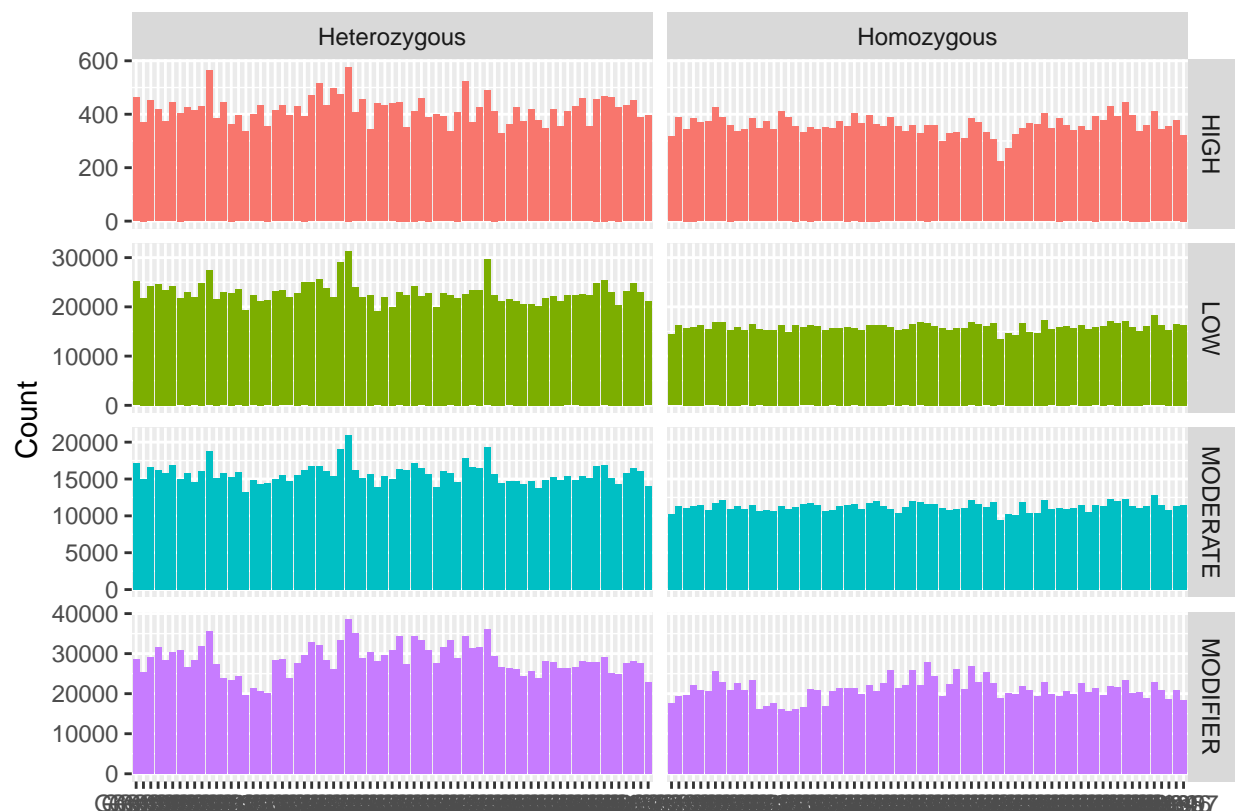
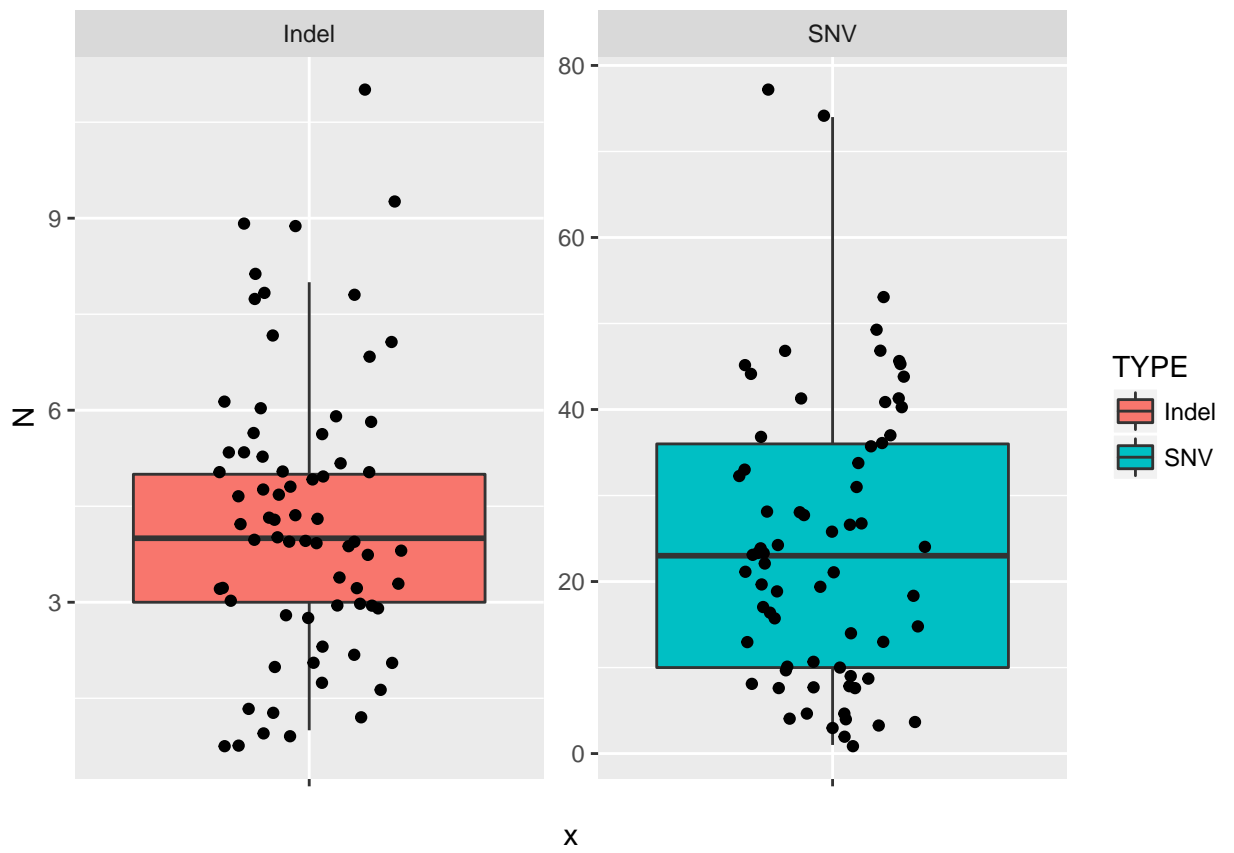
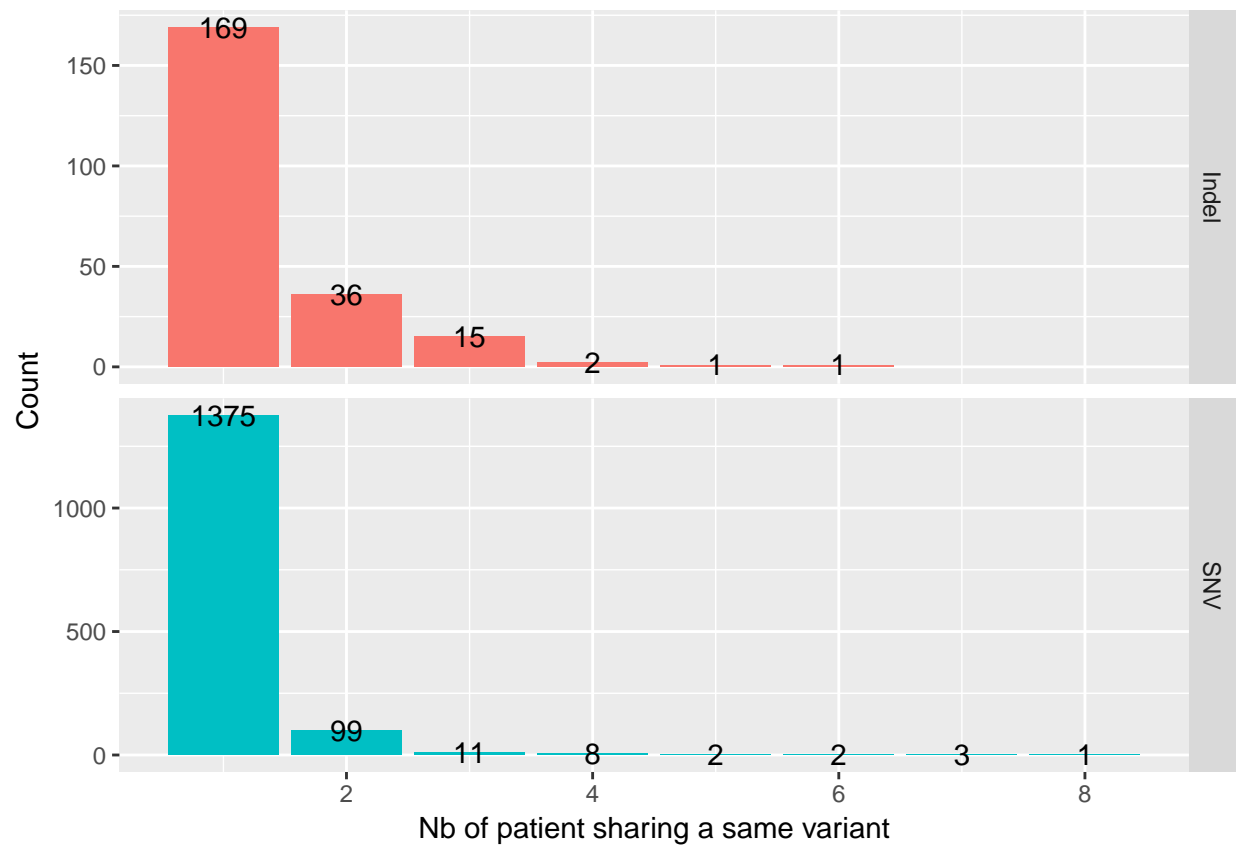


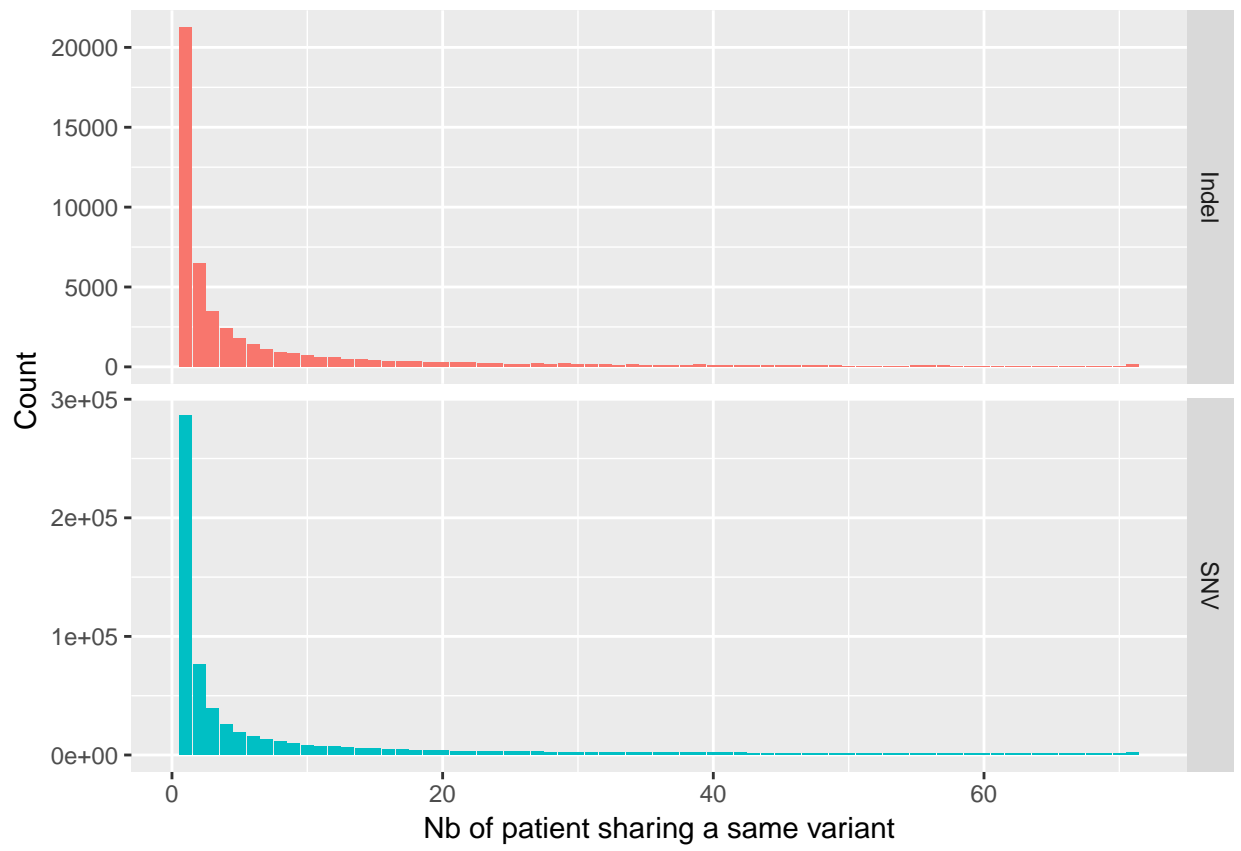
Figure 4.5 – Nombre de gènes passant l'ensemble des filtres par famille

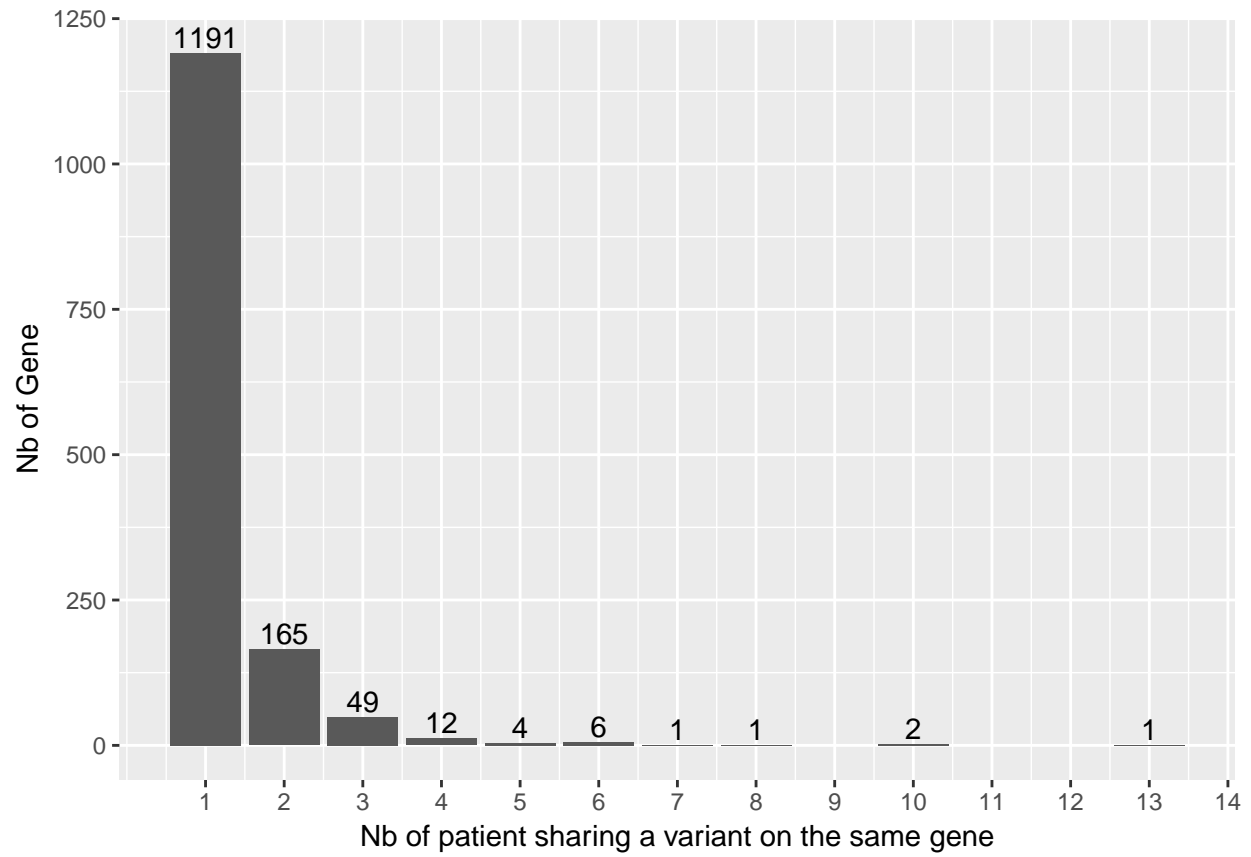
Etude d'une large cohorte de patients MMAF











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>