

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

Remerciements	1
Résumé	5
Abstract	7
Chapitre 1 : Introduction	9
1.1 La spermatogénèse	9
1.1.1 Rappels sur le testicule	10
1.1.2 La phase de multiplication	11
1.1.3 La méiose	12
1.1.4 La spermogénèse	14
1.2 Structure et fonction du spermatozoïde	16
1.2.1 Anatomie du spermatozoïde	16
La tête	17
Le flagelle	19
1.2.2 Fonction du spermatozoïde	21
1.3 L'infertilité masculine	21
1.3.1 Les différents phénotypes d'infertilité masculine	22
Liée à la quantité	22
Liée à la forme	23
Liée à la mobilité	23
1.3.2 La génétique de l'infertilité	24
Les causes fréquentes	24
Les nouveaux gènes	24
1.4 Les techniques d'analyses génétiques	24
1.4.1 Les puces	24
Les puces à SNP, le génotypage... (titre à revoir)	24
Du tissu au transcriptome, le différentiel d'expression	24
1.4.2 Le séquençage NGS	25
La capture des parties à séquencer, avantage et inconvénients	26
L'amplification	27
La réaction de séquence	29
1.5 L'analyse bioinformatique des données de NGS	32
1.5.1 Les données fournies par le NGS	33

Un <i>read</i> c'est quoi ?	33
Le format FASTQ	33
1.5.2 L'alignement	34
1.5.3 L'appel des variants	35
1.5.4 L'annotation des variants, filtrage et priorisation	36
1.5.5 Conclusion NGS	39
Chapitre 2 : Investigation génétique et physiologique de la globo-zoospermie	41
Chapitre 3 : MutaScript	43
3.1 Introduction	43
3.2 Matériel & Méthodes	44
3.2.1 Récupération des données ExAC, filtrage et pré analyse	44
3.3 Résultats	44
3.3.1 Définition de la formule de score	44
3.3.2 Analyse de la corrélation	44
3.3.3 Analyse HPO	44
3.4 Conclusion	44
Conclusion	45
Annexe A : The First Appendix	47
Annexe B : The Second Appendix, for Fun	49
References	51

Liste des tableaux

1.1 Durée de vie moyenne des cellules germinales humaines	9
---	---

Table des figures

1.1	Schéma anatomique du testicule humain :	10
1.2	Les différentes phases de la spermatogénèse (À CHANGER!!!!!)	12
1.3	Les différentes étapes de la méiose gamétique masculine	13
1.4	Les différentes phases de la division cellulaire	14
1.5	Schéma simplifié d'un enjambement chromosomique	14
1.6	Principales étapes et modifications structurales lors de la spermogénèse	16
1.7	Anatomie du spermatozoïde	17
1.8	Schéma de la compaction de l'ADN dans les cellules somatiques et dans les spermatozoïdes	19
1.9	Structure simplifiée de l'axonème d'après Inaba (2003)	20
1.10	Structure du flagelle d'un spermatozoïde d'après Borg et al. (2010) . .	21
1.11	Présentation de la taille des reads et du nombre de reads par run en fonction de la technologie de séquençage utilisée	26
1.12	Présentation des différentes stratégies d'amplification de l'ADN dans le cadre du NGS	29
1.13	Exemple de séquençage CRT tel qu'il est effectué par Illumina	30
1.14	Exemple de séquençage SNA tel qu'il est effectué par Ion Torrent . .	31
1.15	Exemple de séquençage SBL tel qu'il est effectué par SOLiD	32
1.16	présentation d'un fichier FASTQ (FIGURE A CHANGER)	34
1.17	Représentation des erreurs d'appel de type IR et AR en fonction de la plateforme de séquençage et du logiciel d'appel	36
1.18	Diagramme de Venn des prédictions de pathogénicités de six logiciels	38

Remerciements

Je remercie ...

- Les rapporteurs
- Les membres du jury
- Pierre
- Nicolas
- L'équipe BCM
- Kevin Keurcien Thomas Florient
- L'équipe GETI
- La BGM
- Mes amis
- Ma famille
- Dadette et Marco
- Simon
- Aurélien
- Mes parents
- Ma soeur
- Estelle
- Noham

Cette thèse est dédiée à Fabien le québécois

Résumé

Résumé de ma thèse

Second paragraph of abstract starts here.

Abstract

Même chose en anglais

Chapitre 1

Introduction

1.1 La spermatogénèse

La spermatogenèse des mammifères est un processus long et complexe contrôlé par plusieurs mécanismes étroitement liés ((Gnessi, Fabbri, & Spera, 1997, KIERSZEN-BAUM (1994)), **Sharpe1994 à trouver !!!**). C'est au cours de celle-ci qu'à partir de cellules germinales, seront produits les spermatozoïdes matures. Ce processus est divisé en trois phases principales : La phase de multiplication, la phase de division (appelée la méiose) et la phase de maturation. Chez les hommes, ces étapes se déroulent en continue dans la paroi des tubes séminifères du testicule depuis la puberté jusqu'à la mort et implique trois types de cellules germinales : les spermatogonies, les spermatocytes et les spermatides. Le temps nécessaire pour obtenir un spermatozoïde mature à partir de cellules germinales est de 74 jours et la production quotidienne de spermatozoïde est d'environ 45 million par testicules (JOHNSON, PETTY, & NEAVES, 1980). Le cycle spermatogénétique est défini comme la succession chronologique des différents stades de différenciation d'une génération de cellules germinales (depuis la spermatogonie jusqu'au spermatozoïde). Chacune des étapes 35du cycle spermatogénétique a une durée fixe et constante selon les espèces.

Table 1.1 – Durée de vie moyenne des cellules germinales humaines

Cellules germinales	Durée de vie moyenne (jours)
Spermatogonies Ap	16-18
Spermatogonie B	7.5-9
Spermatocytes primaires	23
Spermatocytes secondaires	1
Spermatides	1

1.1.1 Rappels sur le testicule

Les testicules sont les organes sexuels masculins. Ils possèdent deux fonctions principales (plus ou moins exprimées selon les périodes de la vie de l'individu) : une fonction endocrine caractérisée par la synthèse des hormones stéroïdes sexuelles masculines (la stéroïdogenèse) et une fonction exocrine au cours de laquelle seront produits les gamètes masculins. Chez un individu adulte en bonne santé, le testicule présente une forme ovoïde ayant un volume moyen de 18 cm³. Chez l'homme, comme chez la plupart des mammifères terrestres, ils sont localisés sous le pénis dans une poche de peau appelée scrotum et reliés à l'abdomen par le cordon spermatique (**Figure : 1.1**). Cette externalisation des testicules permet leur maintien à une température plus basse que celle du reste du corps nécessaire à la spermatogenèse.

L'intérieur du testicule contient des tubes séminifères enroulés ainsi que du tissu entre les tubules appelé espace interstitiel. Les tubes séminifères sont de longs tubes compactés sous forme de boucles et dont les deux extrémités débouchent sur le *rete testis* (**Figure : 1.1**). C'est le long des parois du tube séminifère que se déroulera l'ensemble des étapes de la spermatogenèse.

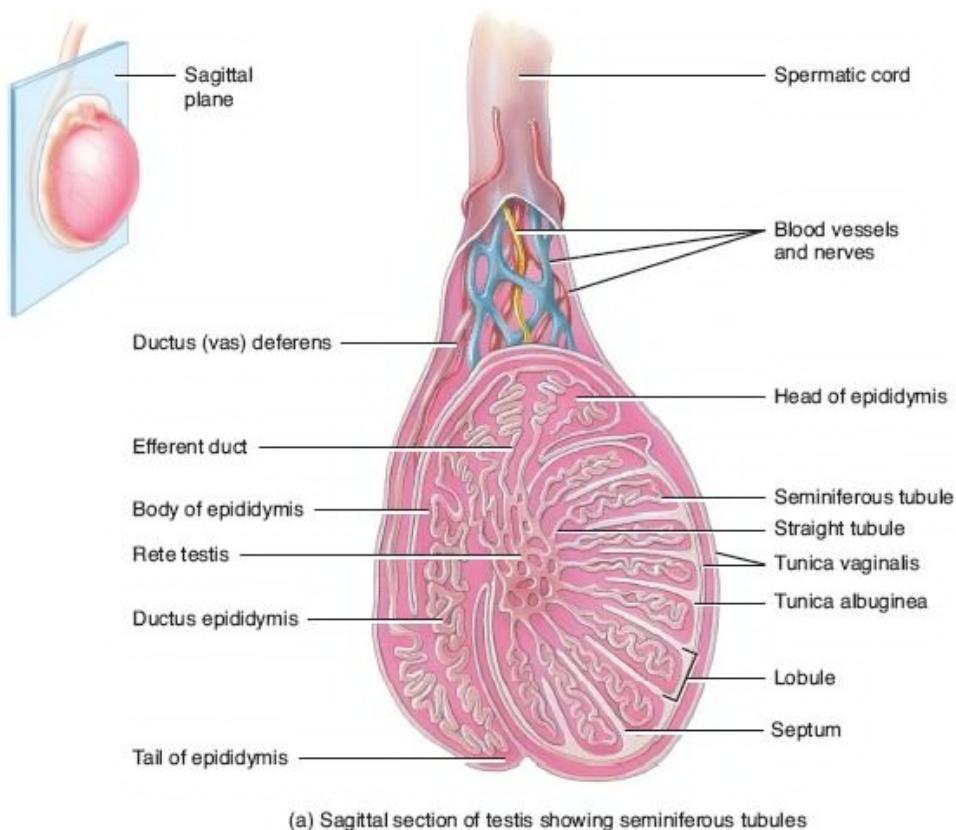


Figure 1.1 – Schéma anatomique du testicule humain :

1.1.2 La phase de multiplication

La phase de multiplication est la phase au cours de laquelle les spermatogonies se divisent par mitoses pour aboutir au stade de spermatocytes primaires. Les spermatogonies sont des cellules diploïdes à l'origine de l'ensemble des autres cellules germinales humaines. Pour cela, elles vont s'auto-renouveler par mitose successive afin de maintenir une production continue de spermatozoïdes tout au long de la vie de l'individu. Ces cellules sont localisées dans le compartiment basal des tubes séminifères. Les analyses histologiques ont permis de distinguer trois types de spermatogonies en fonction de leur contenu en hétérochromatine ((Clermont, 1963, Clermont (1966), Goossens & Tournaye (2013))) :

1. Les spermatogonies de type A dark (ou Ad)
2. Les spermatogonies de type A pale (ou Ap)
3. Les spermatogonies de type B

Chez l'Homme, les spermatogonies Ad ont une activité mitotique au cours de la spermatogénèse et servent de réserve. Elles vont au cours d'une première mitose former une spermatogonie Ad et un spermatogonie Ap (**Figure : 1.2**). Cette propriété permet à la fois de se différencier en spermatocytes tout en constituant un compartiment de réserve de spermatogonies Ad pour la régénération de la population de cellules germinales au sein de l'épithélium séminifère. L'entrée en division des spermatogonies Ap se fait par groupes cellulaire tous les 16 jours. Les cellules d'une même génération maintiennent entre elles des ponts cytoplasmiques jusqu'à la spermiogénèse ce qui permet la synchronisation parfaite du développement gamétique de toutes les cellules filles issues d'un groupe de spermatogonies Ap. Ce phénomène est appelé onde spermatogénétique. Chaque spermatogonie Ap va, lorsqu'elle se divise par mitose, former deux spermatogonies B qui elles-mêmes se diviseront en deux spermatocytes primaires diploïdes (**Figure : 1.2**).

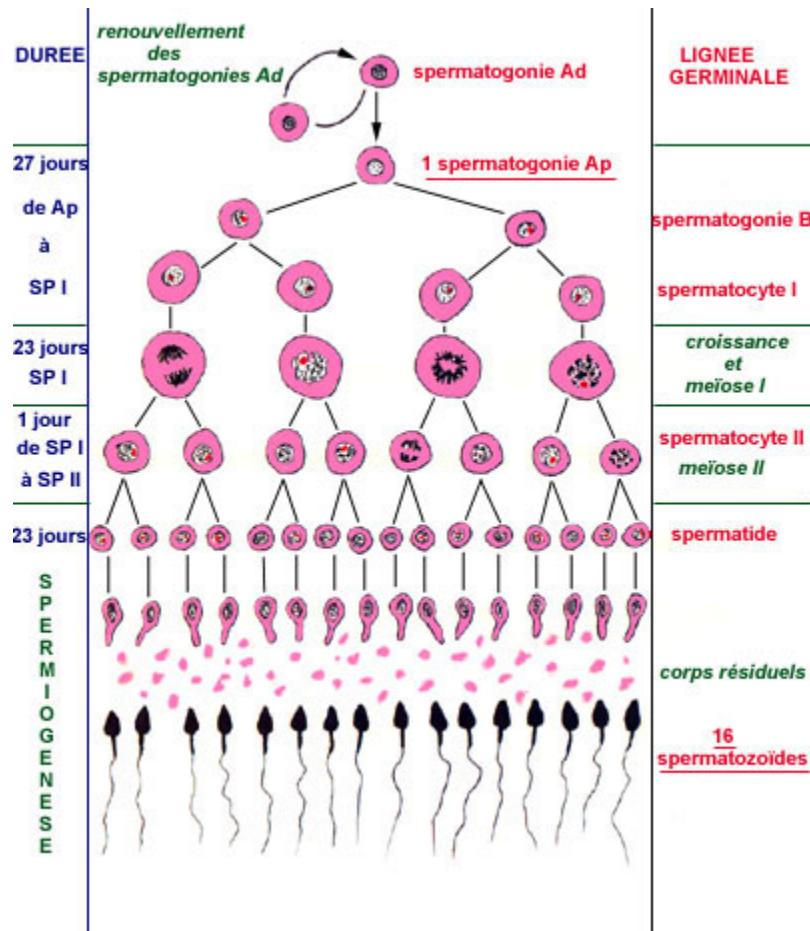


Figure 1.2 – Les différentes phases de la spermatogénèse (À CHANGER!!!!)

1.1.3 La méiose

La méiose, ou phase de maturation, est l'étape au cours de laquelle, à partir de cellules diploïdes (les spermatogones B) vont se former des cellules haploïdes, les spermatocytes secondaire (spermatocytes II). Ce résultat est le fruit de deux divisions successives (**Figure :** @ref(fig :méiose)) appelée respectivement méiose réductionnelle ou méiose I (MI) et méiose équationnelle ou méiose II (MII). La MI va séparer les chromosomes homologues, produisant deux cellules et réduisant la ploïdie de diploïde à haploïde (d'où son nom *réductionnelle*). En plus de son rôle de division vu précédemment, la méiose joue un rôle clef dans le brassage génétique (mélange des gènes) et ce, grâce à deux mécanismes de brassage : le brassage inter-chromosomique, lorsque les chromosomes sont séparés et le brassage intra-chromosomique impliquant notamment des enjambements chromosomiques (crossing-over) (**Figure :** 1.5).

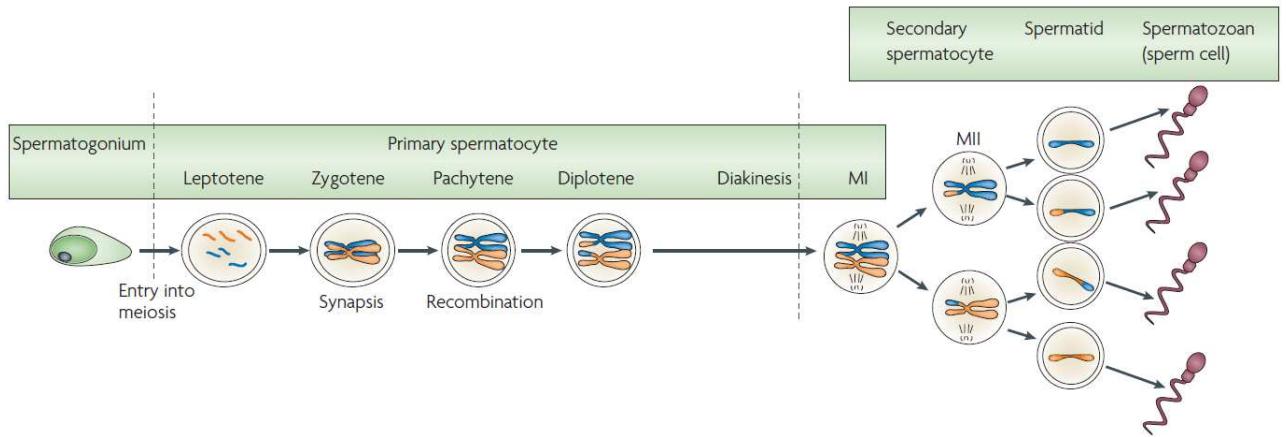


Figure 1.3 – Les différentes étapes de la méiose gamétique masculine :
D'après Sasaki et Matsui, 2008

La méiose est initiée dès la fin de la phase de multiplication à partir des spermatocytes primaires issus de la division des spermatogonies de type B. Ces cellules nouvellement formées se situent dans le compartiment basal du tube séminifère. C'est là qu'ils vont tout d'abord subir une interphase (stade préleptotène) durant entre 2 à 4 jours. Au cours de cette phase a lieu la réplication de l'ADN. Cette réplication se fait lorsque l'ADN est à l'état de chromatine, pendant la phase S (pour synthèse) de l'interphase. À l'issue de cette phase, chaque chromosome sera composé de deux chromatides reliés entre elles par le centromère, le matériel génétique de chaque cellule ayant donc été multiplié par 2. Par la suite, ces cellules vont subir deux divisions méiotiques, chacune composées de 4 étapes distinctes (**Figure : 1.4**) :

1. La prophase, caractérisée par la condensation de la chromatine formant ainsi les chromosomes.
2. La métaphase, phase au cours de laquelle les chromosomes vont s'aligner à l'équateur de la cellule pour former la plaque équatoriale.
3. L'anaphase, les chromatides soeurs (ou les chromosomes homologues en fonction de la phase méiotique) vont se séparer et migrer aux pôles opposés de la cellule.
4. La télophase, qui est l'étape finale, les chromosomes se décondensent et l'enveloppe nucléaire se reforme autour des chromosomes. La cellule mère se sépare alors en deux cellules filles.

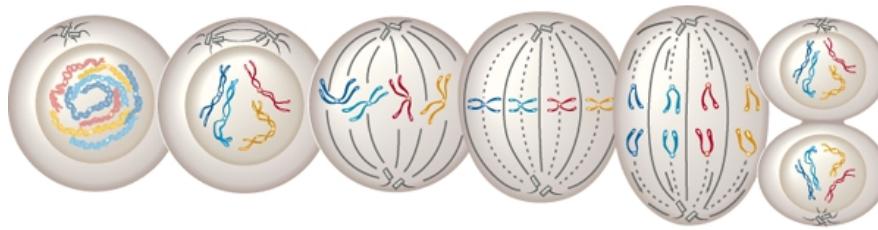


Figure 1.4 – Les différentes phases de la division cellulaire : De la prophase (à gauche) à la télophase (à droite)

La première division méiotique aboutit à la formation des spermatocytes secondaires (spermatoctyes II). À ce stade, les cellules sont haploïdes et chaque chromosome est composé de deux chromatides sœurs. Après, cette brève étape (environ 1 jour) ainsi qu'une très courte interphase sans réPLICATION de l'ADN, les spermatoctyes II vont entrer en deuxième division méiotique. Cette deuxième division est très semblable à une division mitotique. La prophase II, à la différence de la prophase I, est très courte. Lors de cette étape, les chromosomes constitués de chromatides sœurs se dirigent vers la plaque équatoriale. En métaphase II, les chromosomes s'alignent au niveau de leurs centromères. En anaphase II, les chromatides sœurs se séparent l'une de l'autre et migrent vers les pôles opposés des spermatoctyes II. Lors de la télophase II, on observe la formation de cellules filles haploïdes appelées spermatides, contenant chacune n chromosomes.

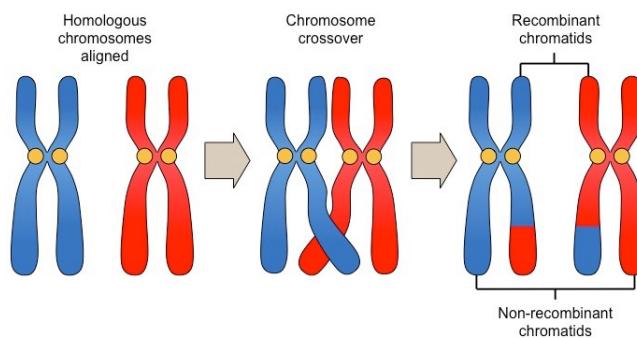


Figure 1.5 – Schéma simplifié d'un enjambement chromosomique

1.1.4 La spermiogénèse

La spermiogénèse est la phase finale de la spermatogénèse. Elle dure environ 23 jours chez l'humain et peut être subdivisée en sept étapes (**Figure : 1.6**). La spermiogénèse définit la cytodifférenciation des spermatides en spermatozoïdes. C'est au cours de cette phase que les caractéristiques morphologique et fonctionnelles du spermatozoïde seront déterminées (Clermont & Oko 1993 à trouver!!!). Elle est caractérisée par 3

événements majeurs : la formation de l'acrosome, la compaction de l'ADN nucléaire et la formation du flagelle. Le développement de l'acrosome et la formation du flagelle commence au niveau des spermatides rondes (Escalier et al., 1991). Pendant l'élongation de la spermatide, le noyau se condense et devient hautement polarisé (Hamilton, D. W., Waites, 1990).

Les spermatides sont situées dans le compartiment adluminal, à proximité de la lumière du tube séminifère. Ce sont de petites cellules (8 à 10 µm) que l'on peut schématiquement diviser en trois classes :

1. Les spermatides rondes (**Figure : 1.6 1-2**) : L'identification de ces cellules représente une difficulté technique. Elles ont cependant pu être décrites en détail par différentes techniques de coloration sous microscope optique (Clermont, 1963, Papic, Katona, & Skrabalo (1988), Schenck & Schill (n.d.), Adelman & Cahill (1989), World Health Organization (1992)). Plusieurs études animales ont pu démontré le potentiel des spermatides rondes à donner la vie à des individus sains et fertiles, (a Ogura, Matsuda, & Yanagimachi, 1994), A. Ogura, Matsuda, Asano, Suzuki, & Yanagimachi (1996), Sasagawa & Yanagimachi (1997)], la même chose ayant été également observée plus récemment chez l'homme (A. Tanaka et al., 2015) bien que le taux de fécondation et d'implantation soit extrêmement faible (Asimakopoulos, 2003). Ils possèdent un noyau rond avec une chromatine pâle et homogène. C'est à partir de ces étapes que démarre la biogénèse de l'acrosome avec la production par l'appareil de Golgi des vésicules pro-acrosomales (phase de Golgi). Les deux centrioles contenus dans le cytoplasme vont se déplacer au futur pôle caudal. Le centriole proximal est inactif alors que le centriole distal donne naissance à un ensemble de microtubules à l'origine de l'axonème du futur flagelle.
2. Les spermatides en élongation (**Figure : 1.6 3-4**) : peuvent aussi donner naissance avec un meilleur taux que les spermatides rondes et engendrerai théoriquement moins de risques d'anomalies génétiques ((Asimakopoulos, 2003)). **A compléter**
3. Les spermatides en condensation (**Figure : 1.6 5-7**) : C'est le stade final de la différentiation de la spermatide en spermatozoïde. À ce stade le noyau est très allongé, avec une partie caudale globulaire et une partie antérieure saillante. La chromatine est sombre et condensée. L'axonème va continuer à s'allonger pour former le flagelle mature. Les différentes organelles inutiles pour la physiologie spermatique et l'excès de cytoplasme vont former la gouttelette cytoplasmique qui va se détacher et donner le corps résiduel qui va ensuite être phagocyté par les cellules de Sertoli (Hermo, Pelletier, Cyr, & Smith, 2010).

Une fois ces étapes de différentiation finies, les spermatides sont relachées en tant que spermatozoïdes dans la lumière du tube séminifère. Ce procédé est appelé spermiation.

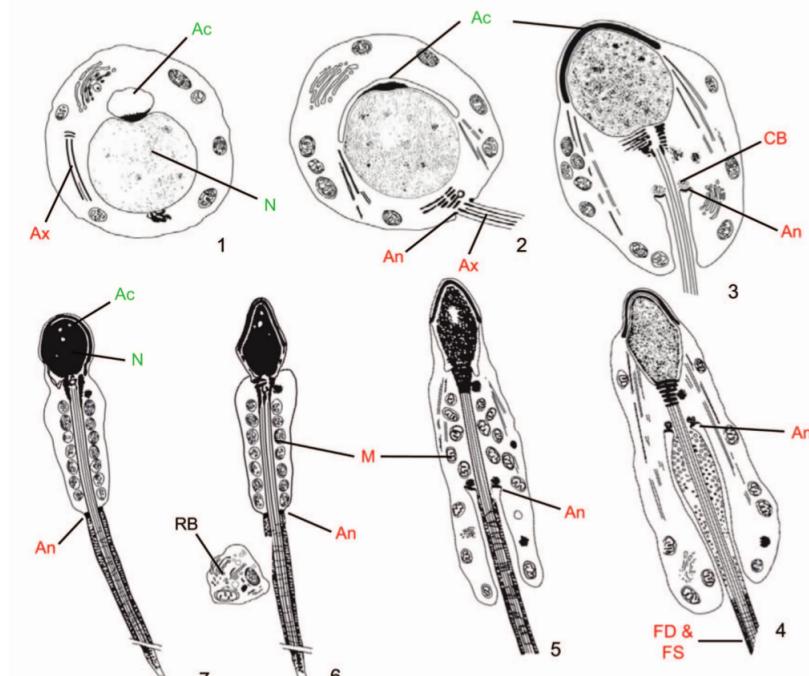


Figure 1.6 – Principales étapes et modifications structurales lors de la spermiogénèse : 1. La spermatide immature avec un gros noyau arrondi. La vésicule acrosomale est attachée au noyau, l'ébauche du flagelle n'atteint pas le noyau. 2. La vésicule acrosomale a augmenté de taille et apparaît aplatie au niveau du noyau. Le flagelle entre en contact avec le noyau. 3-7. Formation de l'acrosome, condensation du noyau et développement des structures flagellaires. Ac, acrosome ; Ax, axonème ; CC, corps chromatoïdes ; CR, corps résiduel ; FD, fibres denses ; GF, gaine fibreuse ; M, mitochondrie ; Ma, manchette. D'après Touré et al., 2011

1.2 Structure et fonction du spermatozoïde

1.2.1 Anatomie du spermatozoïde

Une fois Il est composé de deux parties principales : La tête et le flagelle (**Figure : 1.7**).

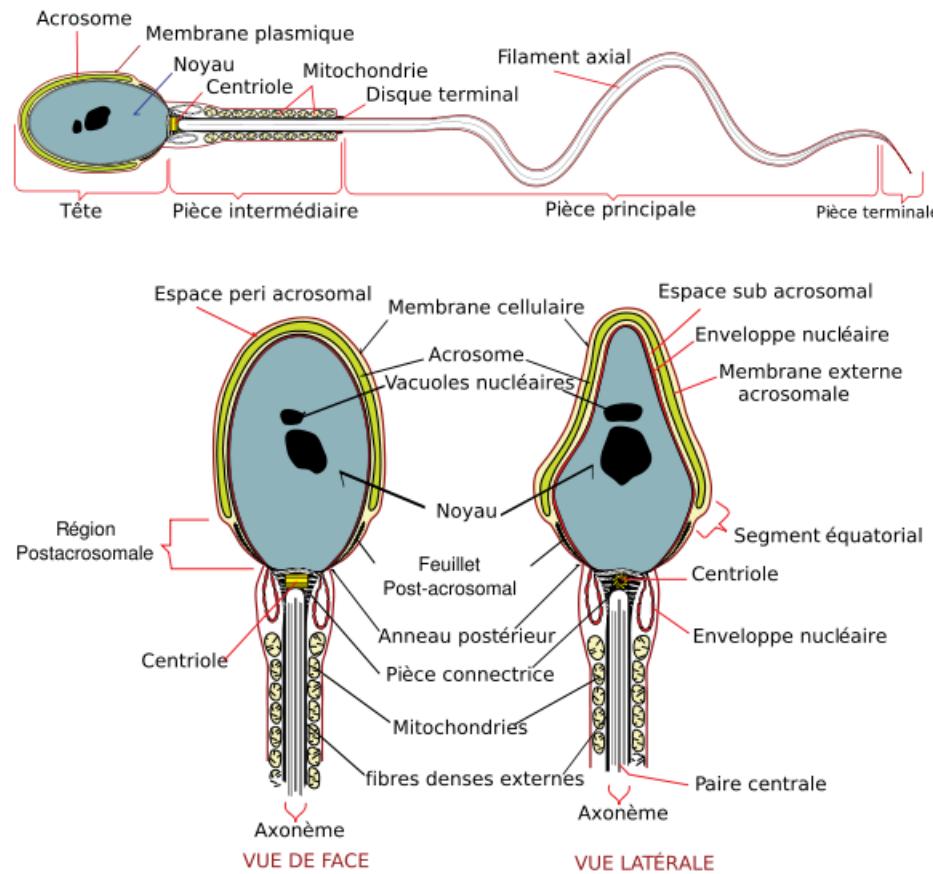


Figure 1.7 – Anatomie du spermatozoïde

La tête

1. **L'acrosome** : C'est une vésicule de sécrétion géante située dans la moitié supérieure de la tête du spermatozoïde. Elle se développe à partir de l'appareil de Golgi lors de la spermiogénèse. Au cours de sa formation, l'acrosome forme tout d'abord un granule sphérique qui se colle sur la partie apicale du noyau. En s'aplatissant contre celui-ci, l'acrosome va prendre une forme hémisphérique recouvrant la membrane nucléaire formant la coiffe céphalique... Le rôle de l'acrosome est fondamental dans le processus de fécondation puisqu'il permet d'excréter notamment l'acrosine, une enzyme de digestion permettant au spermatozoïde de pénétrer la zone pellucide qui entoure les ovocytes. Ce processus de relargage est appelé réaction acrosomale.
2. **L'acroplaxome** : TODO !!!
3. **Le noyau** : C'est une structure cellulaire présente dans la majorité des cellules eu-

caryotes. Il contient l'essentiel du matériel génétique. Le noyau du spermatozoïde est caractérisé par une compaction extrêmement importante de l'ADN. Dans les cellules somatiques l'ADN est enroulé par unité de 146 paires de bases autour d'un octamère d'histones dit de cœur (H2A, H2B, H3 et H4) afin d'organiser les 3 milliards de paires de bases du génome humain dans un noyau de quelques microns (**Figure : 1.8**). L'ADN des spermatides va subir une réorganisation chromatinnienne plus importante au cours de la spermatogénèse afin d'augmenter sa compaction. Ainsi, les octamères d'histones présents dans les cellules somatiques sont remplacées par deux protéines riches en arginine et en cystéine PRM1 et PRM2. Ces protéines sont appelées des protamines (**Figure : 1.8**). L'intégrité des deux protéines composant ce dimère est nécessaire pour la procréation (Cho et al., 2001). Cette compaction extrême permet de réduire la taille du noyau, mais aussi de protéger l'ADN d'agents de dégradation comme l'oxydation des bases. Parallèlement à cette condensation chromatinnienne se produit un arrêt des processus de transcription cellulaire (Kierszenbaum & Tres, 1978). Le noyau du spermatozoïde est donc un noyau au repos, transcriptionnellement inactif (Ward, 1994)

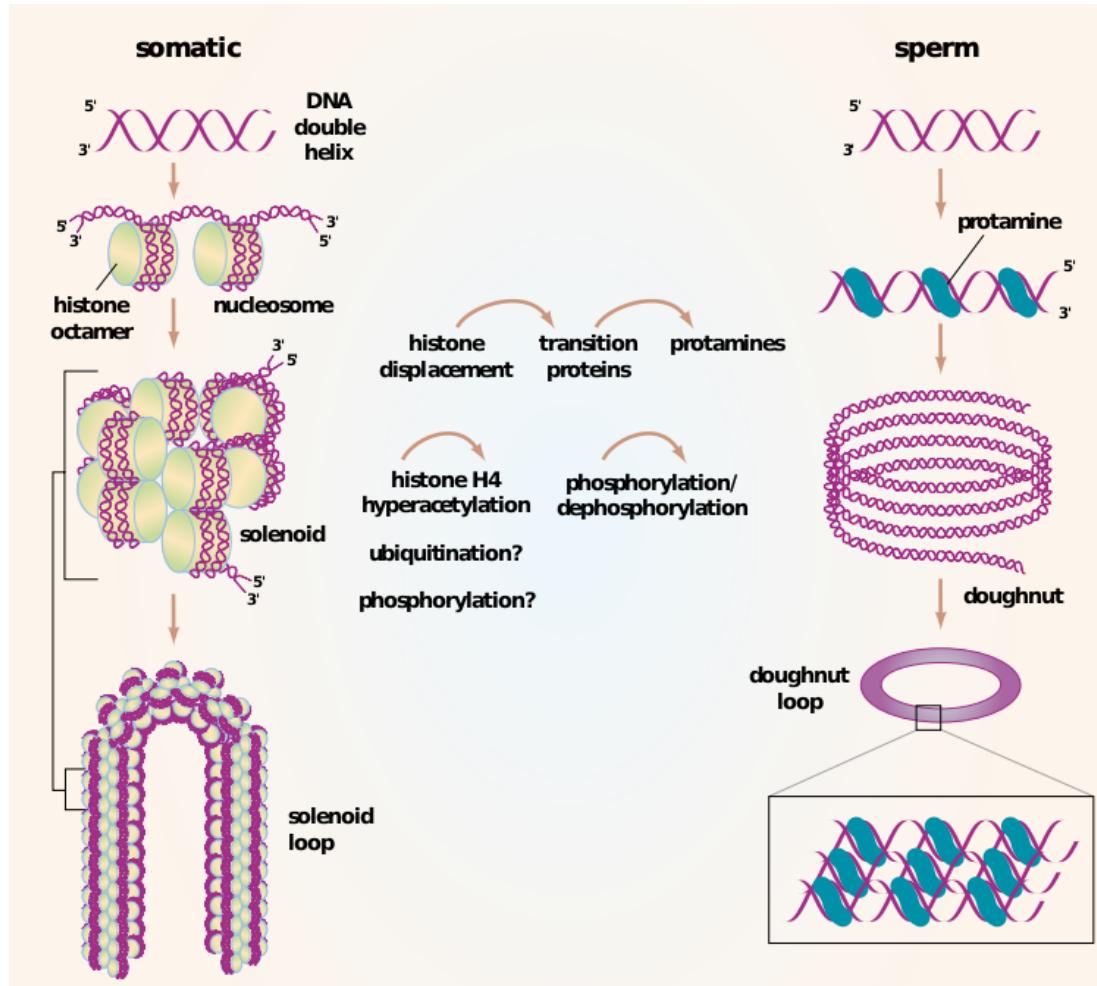


Figure 1.8 – Schéma de la compaction de l'ADN dans les cellules somatiques et dans les spermatozoïdes : D'après Braun (2001)

Le flagelle

Le flagelle représente la queue du spermatozoïde. Celui-ci permet, par mouvement d'oscillation à haute vitesse, le déplacement du spermatozoïde. Cette mobilité est générée par un cytosquelette interne extrêmement conservé durant l'évolution appelée l'axonème. Celui-ci est composé de neuf doublets de microtubules périphériques et de deux doublets internes (Inaba, 2003) (Figure : 1.9), on parle alors de structure “9 + 2”. Les doublets externes sont reliés entre eux par des ponts de nexine et au doublet central par des ponts radiaires.

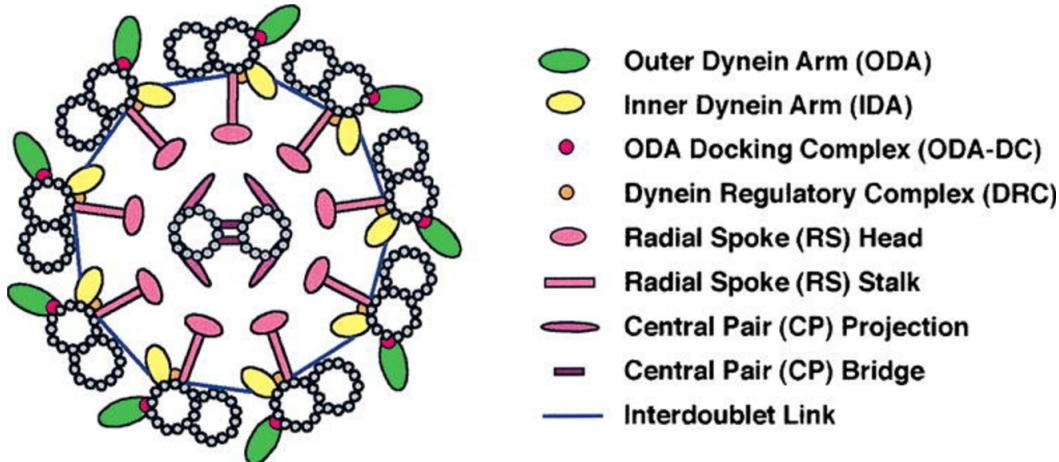


Figure 1.9 – L’axonème est constitué de neuf doublets de microtubules périphériques reliés entre eux par des liens de nexine d’un doublet central relié aux doublets périphériques par des ponts radiaires

Le flagelle su spermatozoïde peut être divisé en trois partie distinctes (**Figure : 1.10**) :

1. **La pièce intermédiaire** : Elle fait jonction avec la tête du spermatozoïde et est composée de la gaine de mitochondrie qui fournira une partie de l’énergie nécessaire au battement flagellaire (grâce à la phosphorylation oxydative qui produit de l’ATP), l’axonème qui se prolonge dans la pièce principale et un ensemble de neuf faisceaux de fibres denses.
2. **La pièce principale** : Ici, la gaine de mitochondrie a disparue ainsi que deux des faisceaux de fibres denses présents dans la pièce intermédiaire. On note cependant la présence d’une structure supplémentaire, la gaine fibreuse. Cette gaine entoure l’axonème et comporte deux épaissements diamétralement opposés, appelées colonnes longitudinales sur lesquelles s’insère les fibres denses 3 et 8. C’est le long de la gaine fibreuse qu’est produit la majorité de l’énergie nécessaire au glissement des microtubules (Eddy, 2007).
3. **La pièce terminale** : Elle est située au niveau de l’extrémité distale du flagelle et ne contient que l’axonème (Inaba, 2003).

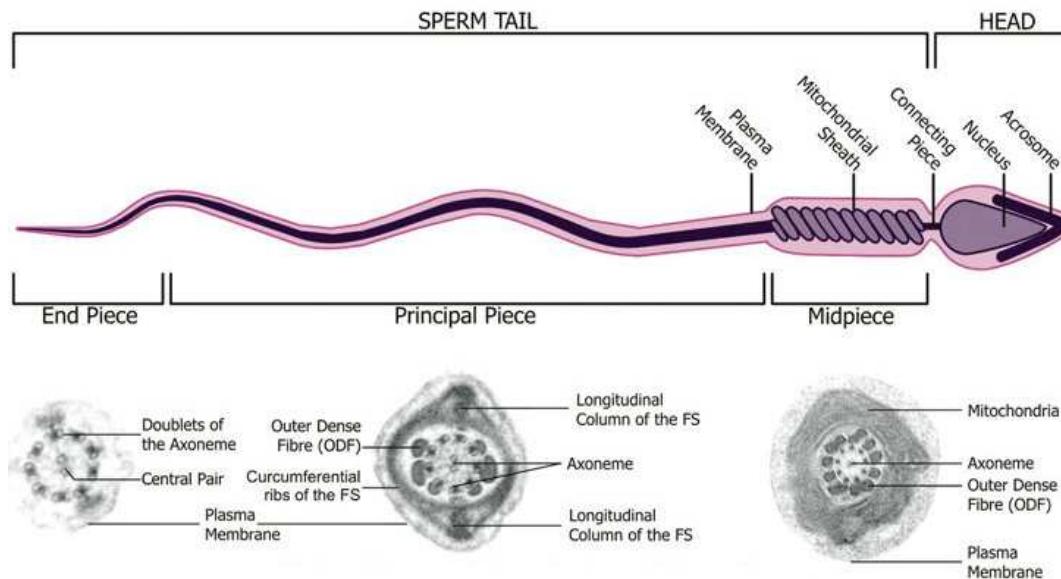


Figure 1.10 – Structure du flagelle d'un spermatozoïde d'après Borg et al. (2010) : Coupes transversales en microscopie électronique. Le flagelle se compose de trois parties : la pièce intermédiaire, contenant les mitochondries, la pièce principale et la pièce terminale. L'axonème, en position centrale, parcourt tout le flagelle. Des structures périaxonémiales sont observables : les fibres denses dans la pièce intermédiaire et principale, et la gaine fibreuse dans la pièce principale seulement.

1.2.2 Fonction du spermatozoïde

En plus d'être unique dans sa morphologie, le spermatozoïde l'est aussi dans sa fonction puisque c'est la seule cellule produite de manière endogène et dont l'action est exercée de manière exogène.

1.3 L'infertilité masculine

L'organisation mondiale de la santé définit l'infertilité comme étant : “une pathologie du système reproductif définie par l'échec d'une grossesse clinique après 12 mois ou plus de rapports sexuels réguliers non protégés” (Who.int. 2013-03-19. Retrieved 2013-06-17). Environ 10-15% des couples humains sont considérés infertiles. On estime que dans la moitié des cas, la cause sous-jacente est masculine. Les facteurs causaux sous-jacents de l'infertilité masculine peuvent être attribués à des toxines environnementales, des troubles systémiques tels que la maladie hypothalamo-hypophysaire, les cancers testiculaires et l'aplasie des cellules germinales. Les facteurs génétiques, y compris les aneuploïdies et les mutations de gènes uniques, contribuent également

à l'infertilité masculine. Cependant, aucune cause n'est identifiée dans 10-20% des cas.

The entire process (of spermatogenesis) is tightly synchronized and integrated, so that pathological conditions which produce even very small deviations are likely to lead to infertility (Barratt, 1995)

Barratt, C.L.R. (1995) Spermatogenesis. In Grudzinsky, J.G. and Yovich, J.L. (eds) Gametes : the spermatozoon. Cambridge University Press, Cambridge

1.3.1 Les différents phénotypes d'infertilité masculine

Liée à la quantité

Immature germ cells are present in ejaculates of subjects with a normal sperm count (Michael and Joel, 1937; Tomlinson et al., 1992), oligozoospermia (Mac Leod, 1970; Tomlinson et al., 1993), or azoospermia (Kurilo et al., 1993) and the presence of immature germ cells increases as the sperm count decreases (Sperling and Kaden, 1971) Michael, M. and Joel, K. (1937) Zellformen in normalen und pathologischen Ejakulaten und ihre klinische Bedeutung. Schweiz. Med. Wsch., 33, 757.

Tomlinson, M.J., Barratt, C.L.R., Bolton, A.E. et al. (1992) Round cells and sperm fertilizing capacity : the presence of immature germ cells but not seminal leukocytes are associated with reduced success of in vitro fertilization. Fertil. Steril., 58, 1257–1259. MacLeod, J. (1970) The significance of deviations in human sperm morphology. In : Rosemberg, E. and Paulsen, C.A. (eds) The human testis. Plenum, New York, pp. 481–494.

Tomlinson, M.J., Barratt, C.L.R. and Cook, I.D. (1993) Prospective study of leukocytes and leukocyte populations in semen suggests they are not a cause of male infertility. Fertil. Steril., 60, 1069–1075

Kurilo, L.F., Liubashevskaya, I.A., Dubinskaia, V.P. and Gaeva, T.N. (1993) Karyological analysis of the count of immature germ cells in the ejaculate. Urol. Nefrol. (Mosk.), 2, 45–47.

Sperling, K. and Kaden, R. (1971) Meiotic studies of the ejaculated seminal fluids of humans with normal sperm count and oligospermia. Nature, 232, 481

In humans, spermatogenic arrest was considered a hopeless condition for couples desiring to conceive. However, the documented success of intracytoplasmic sperm injection (ICSI; Palermo et al., 1992) has pointed to using this technique to inject spermatids into oocytes (Edwards et al. 1994; Ogura et al., 1994) Palermo, G., Joris, H., Devroey, P. and Van Steirteghem, A.C. (1992) Pregnancies after intracytoplasmic sperm injection of a single spermatozoon into an oocyte. Lancet, 340, 17–18.

Edwards, R.G., Tarin, J.J., Dean, N. et al. (1994) Are spermatids injections into human oocytes now mandatory ? Hum. Reprod., 9, 2217–2219.

Ogura, A., Matsuda, J. and Yanagimachi, R. (1994) Birth of normal young after electrofusion of mouse oocytes with round spermatids. Proc. Natl. Acad. Sci. USA, 91, 7460–7462

Spermatogenic arrest, the inability of spermatogenetic cells to develop into male gametes within the gonads, has been reported in 4–30% of testicular biopsies of patients with severe oligospermia or azoospermia (Wong et al., 1973; Levin, 1979; Colgan et al., 1980; Soderstrom and Suominen, 1980; Nomen et al., 1984) Wong, T.W., Strauss, F.H. and Worne, N.E. (1973) Testicular biopsy in male infertility : I. Testicular causes of infertility. Arch. Pathol. Lab. Med., 95, 151–159.

Levin, H.S. (1979) Testicular biopsy in the study of male infertility. Hum. Pathol., 10, 569–579

Colgan, T.J., Bedar, Y.C., Strawbridge, H.T.G. et al. (1980) Reappraisal of the value of the testicular biopsy in the investigation of infertility. Fertil. Steril., 33, 56–60.

Soderström, K.O. and Suominen, J. (1980) Histopathology and ultrastructure of meiotic arrest in human spermatogenesis. Arch. Pathol. Lab. Med., 104, 476–482.

Soderström, K.O. and Suominen, J. (1980) Histopathology and ultrastructure of meiotic arrest in human spermatogenesis. Arch. Pathol. Lab. Med., 104, 476–482.

Spermatogenic arrest can occur at any stage of germ cell formation ; primary spermatocyte arrest is most prominent, followed by spermatid arrest, and least commonly, spermatogonial arrest. Arrest at primary spermatocyte stage can be incomplete, so that a few secondary spermatocytes or spermatids are observed (Girgis et al., 1969) Girgis, S.M., Etriby, A., Ibrahim, A.A. and Kahil, A. (1969) Testicular biopsy in azoospermia. A review of the last ten years' experience of over 800 cases. Fertil. Steril., 20, 467–477.

Liée à la forme

Teratozoospermie

La globozoospermie La globozoospermie est une anomalie des spermatozoïdes caractérisé par une tête ronde dépourvue d'acrosome et d'une pièce intermédiaire désorganisée (Singh, n.d., Pedersen & Rebbe (1974))

Liée à la mobilité

Sperm motility is necessary for the transport of male DNA to eggs in species with both external and internal fertilization.

1.3.2 La génétique de l'infertilité

Les causes fréquentes

Les microdélétions du chromosome Y

Anomalies chromosomiques

Mutations CFTR

Les nouveaux gènes

1.4 Les techniques d'analyses génétiques

L'acide désoxyribonucléique (ADN) a été identifié comme étant le porteur de l'information génétique par Oswald Theodore Avery en 1944. Sa structure en double hélice composée par quatre bases, la thymine, l'adénine, la guanine et la cytosine fut caractérisée en 1953 par James D. Watson et Francis Crick

1.4.1 Les puces

1. Bref historique de la technologie
2. A quoi ça sert
3. Comment ça marche

Les puces à SNP, le génotypage... (titre à revoir)

Du tissu au transcriptome, le différentiel d'expression

1.4.2 Le séquençage NGS

Le terme séquençage de l'ADN fait référence à l'ensemble des techniques permettant de déterminer l'ordre des nucléotides adénine (A), thymine (T), cytosine (C) et guanine (G) de l'intégralité ou d'une partie d'une molécule d'ADN. Avant de parler des nouvelles technologies de séquençage (NGS) faisons un bref historique du séquençage de l'ADN. En 1977 Frederick Sanger développe une technologie de séquençage d'ADN basée sur la méthode *chain-termination*. Ce procédé est désormais connu sous le nom de séquençage Sanger. D'autre méthode furent développées à la même période, notamment celle de Walter Gilbert basée sur la modification chimique de l'ADN, cependant sa grande efficience et sa faible utilisation de la radioactivité permirent au séquençage Sanger de s'imposer comme référence dans la "première génération" de séquenceur à application de commerciale et de recherche (Wikipédia). Apparu en 1998, les instruments de séquençage automatique ainsi que les logiciels associés utilisant le séquençage par capillarité et la technologie Sanger furent les outils principaux qui permirent la compléction du *human genome project* en 2001 (Collins, Morgan, & Patrinos, 2003).

Contrairement à la méthode Sanger, le NGS *lit* des fragments d'ADN, provenant d'un génome entier, de manière aléatoire. On parle alors de séquençage de génomes entiers ou *whole genome sequencing* (WGS). Pour cela, la molécule d'ADN est "coupée" en plusieurs fragments d'une taille donnée. Ce sont ensuite ces fragments qui seront, après une étape d'amplification spécifique aux différentes plateformes, séquencés simultanément. C'est pourquoi on parle souvent de séquençage parallèle massif pour décrire le NGS. Le produit de ce séquençage est appelé *read*. Cette technologie est avantageuse de par la masse de *reads* qu'elle produit et par son faible cout par bases séquencées (Metzker, 2010). Ces caractéristiques ont permis au séquençage Haut-débit d'être couramment utilisé dans le domaine de la recherche clinique.

La taille des *reads* obtenus par séquençage NGS est nettement inférieure à celle atteinte par le séquençage Sanger. À l'heure actuelle, les *reads* obtenus par séquençage NGS ont une taille comprise entre 50 et 500 pb pour la plupart des plateformes contre ... obtenus par Sanger (**Figure : 1.11**), c'est pour cela que les résultats du séquençage NGS sont appelés des *reads* courts ou *short reads*.

Étant donné que le NGS produit à l'heure actuelle des *reads* courts la notion de couverture est importante et représente l'un des critères majeurs à considérer dans l'analyse des données (D. Sims, Sudbery, Ilott, Heger, & Ponting, 2014). La couverture est définie comme le nombre de *reads* qui, après l'étape d'alignement ou d'alignement, se chevauchent les uns les autres au sein d'une région génomique spécifique. Par exemple, une couverture de 30x pour le gène XXXX signifie que chaque nucléotide de ce gène est chevauché par au moins *reads* distincts.

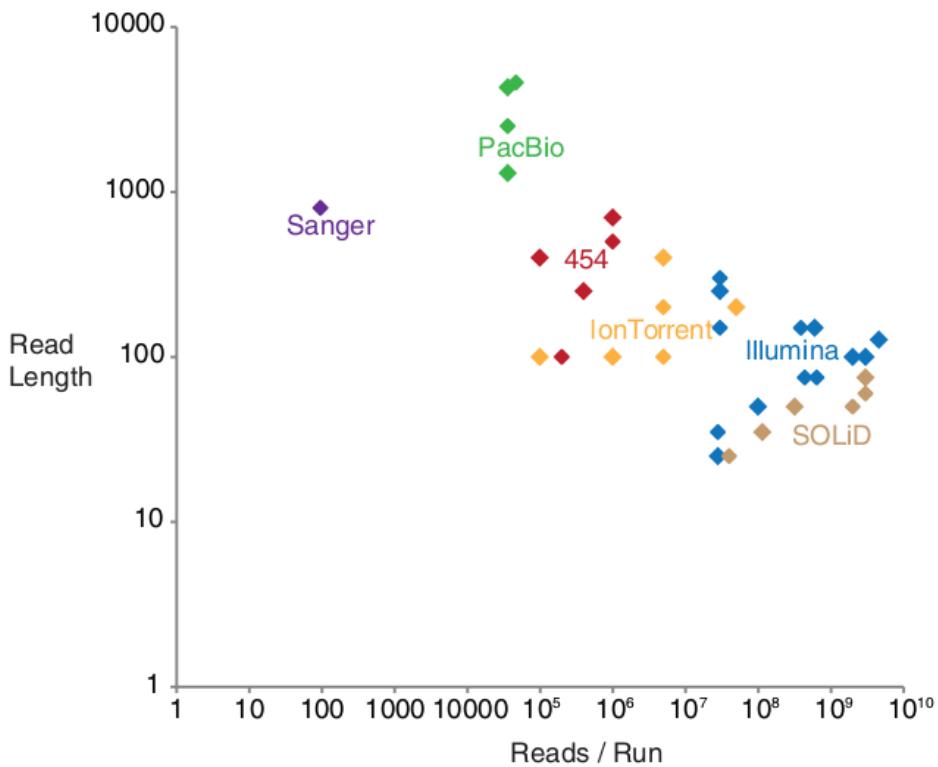


Figure 1.11 – Présentation de la taille des reads et du nombre de reads par run en fonction de la technologie de séquençage utilisée d'après Brendan et. al, (2014) : Sequencing space based on read length (in bases) and number of reads per run. Points represent official platform/chemistry combination releases and are color-coded based on the platform family. To see this illustration in color, the reader is referred to the web version of this article at www.liebertpub.com/wound

La capture des parties à séquencer, avantage et inconvenants

Pour de nombreuse application, il peut être intéressant de ne séquencer qu'une partie du génome et non pas son intégralité. Dans cette sous partie de génome ciblé on peut trouver par exemple : une région génomique spécifique à laquelle une pathologie a déjà été associé, l'ensembles des exons de certains gènes candidats, ou encore l'intégralité des exons de l'ensemble des gènes codant pour une protéine. Dans ce dernier cas on parle alors de séquençage exomique ou *whole exome sequencing* (WES). Les principaux avantages du WES par rapport au WGS sont son cout réduit ainsi qu'une masse de données moins importantes à stocker et à analyser. En effet, l'ensemble de l'exome ne représente qu'environ 1% du génome entier. Pour ces raisons, le WES considéré comme le standard dans le cadre de recherche sur des pathologies génétiques et se révèle être

un outil puissant pour l'identification de variants associés à des pathologies (S. B. Ng et al., 2010). Le procédé de séquençage est identique au WGS, il est simplement précédé d'une étape d'enrichissement au cours de laquelle les exons sont capturés par hybridation à des sondes. De fait les exons capturés sont donc dépendant du kit de capture utilisé, cette technique permet donc de séquencer uniquement les exons connus et ciblés par les sondes. Il faut également noter que depuis quelques années, plusieurs études ont remis en cause l'intérêt du WES au profit du WGS, notamment car le WGS fournit une meilleure couverture sur l'exome que le WES (Lelieveld, Spielmann, Mundlos, Veltman, & Gilissen, 2015, Meienberg, Bruggmann, Oexle, & Matyas (2016)), de plus le WES montre une plus grande sensibilité au pourcentage de GC contenu dans la région à séquencer et à la sélection des kits de capture utilisés (Meienberg et al., 2016). Ainsi, bien que le WES soit encore à l'heure actuelle le choix privilégié dans la majorité des études (citation...), la réduction des couts de séquençage et de stockage des données, il est possible que le WGS remplace totalement le WES ainsi que l'ensemble des techniques impliquant la capture de séquences ciblées (Meienberg et al., 2016).

L'amplification

Dans la plupart des technologies, la phase de séquençage est précédée par une étape d'amplification de l'ADN. Cette amplification se fait dans la grande majorité des cas sur une surface solide excepté pour la PCR en émulsion qui s'effectue en phase aqueuse. Elle permet d'obtenir dans une région définie plusieurs milliers de copie du même fragment d'ADN, appelés des clones. Cette étape assure que le signal émis lors du séquençage pourra être distingué du bruit. Chacun de ces *spots* d'amplification appelés aussi centre de réaction, se retrouve donc être le représentant d'un unique fragment d'ADN et sera ensuite séquencé parallèlement aux autres *spots*. Une plateforme de séquençage pouvant gérer plusieurs millions de ces centres de réactions simultanément, séquençant ainsi plusieurs millions de molécules d'ADN en parallèle, donnant ainsi le nom à ces techniques qualifiées de séquençage massif en parallèle. Cette étape d'amplification est généralement précédée d'une phase de fragmentation de l'ADN. Cette fragmentation peut être physique, enzymatique ou bien chimique. Ce sont les résidus d'ADN résultant de cette fragmentation qui seront ensuite amplifié. Il existe quatre stratégies utilisées pour le clonage de l'ADN dans le cadre du NGS :

1. **La PCR en émulsion ou emPCR (Figure : 1.12 - a)** : Le patron d'ADN fragmenté simple brin est lié à une séquence adaptatrice complémentaire et est capturé par une gouttelette aqueuse appelée micelle contenant une bille recouverte d'adaptateur complémentaire à celui fixé sur le fragment d'ADN ainsi que tous les composant nécessaire à la réaction de PCR. En respectant un ratio nombre de molécule d'ADN / nombre de billes, on va fixer un seul fragment d'ADN sur chaque bille. Chacune de ces billes seront donc, en fin de réaction, recouverte par plusieurs milliers de copies de la même séquence d'ADN.

2. **L'amplification par pont sur face solide (Figure : 1.12 - b)** : Les fragments d'ADN sont liés à des séquences adaptatrices et liée par une de leurs extrémités à une amorce fixée sur un support solide. Du fait de la dilution, les molécules d'ADN se trouvent éloignées les unes des autres. L'extrémité libre du fragment interagit avec les amorces situées à proximité formant une structure en pont, d'où le nom de PCR en pont ou *bridge-PCR*. La PCR va alors synthétiser un deuxième brin complémentaire aux fragments immobilisés sur le support. En procédant à des cycles de température comme pour une réaction PCR classique, on obtient à l'emplacement de chaque molécule initiale un massif de molécules fixées sur la plaque, toutes identiques à la molécule initiale.
3. **Amplification par modèle mobile ou *walking-template* (Figure : 1.12 - c)** : L'ADN fragmenté est lié à un adaptateur et lié à une amorce complémentaire fixée sur un support solide. Le brin complémentaire du fragment sera synthétisé par PCR à partir de l'amorce fixée. La molécule double brin nouvellement formée sera ensuite partiellement dénaturée permettant à l'extrémité libre de se fixer à une séquence amorce voisine. Des amorces *reverse* sont ensuite utilisées pour resynthétiser un fragment d'ADN libre à partir des fragments fixés sur le support.
4. **(Figure : 1.12 - d) : PAS DU TOUT COMPRIS LE MECHANISME !!!**

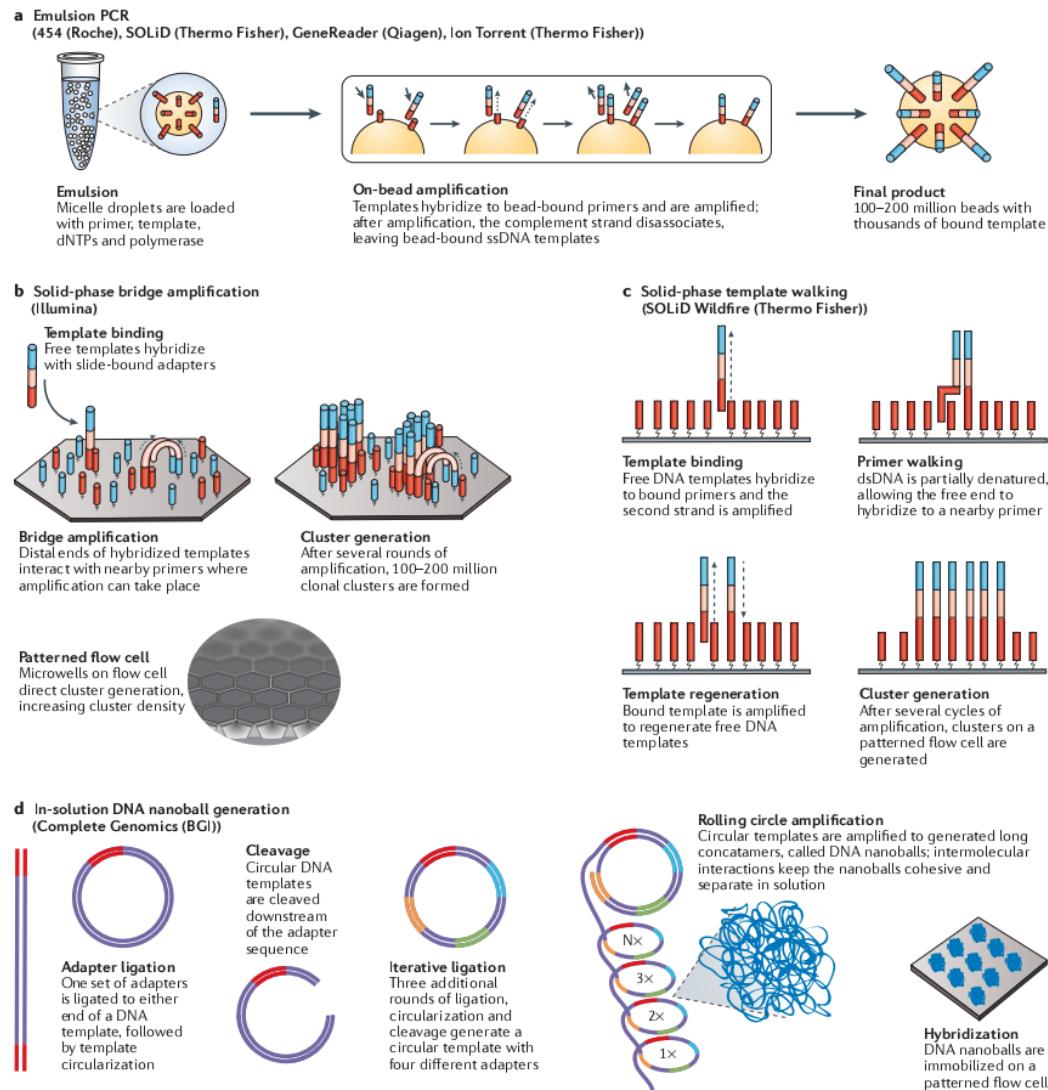


Figure 1.12 – Présentation des différentes stratégies d'amplification de l'ADN dans le cadre du NGS d'après [@Goodwin2016] : **a** : PCR en émulsion. **b** : amplification par pont. **c** : Amplification par modèle mobile. **d** :

La réaction de séquence

La réaction de séquence est l'étape suivant l'amplification et consiste à déterminer l'ordre dans lequel se succèdent les nucléotides de l'ensemble des clones générés dans la phase d'amplification. Il existe deux technologies principales permettant le séquençage de *reads* courts :

1. Séquençage par synthèse (SBS) : Ce type de séquençage regroupe l'ensemble

des méthodes utilisant l'ADN polymérase pour synthétiser de l'ADN. En 2016, Sahra Goodwin et ses collègues ont différenciées deux catégories de séquençage par synthèse (Goodwin, McPherson, & McCombie, 2016) :

- Terminaison par cycle réversible, cyclic reversible termination (CRT)** (**Figure : 1.13**) : Cette méthode est caractérisée par son utilisation de molécules d'acides terminatrices auxquelles le groupement 3' – OH est modifié de sorte à éviter l'élongation (J. Guo et al., 2008), on parlera de groupement 3' – bloqué. Le processus est initialisé une amorce est liée au fragment d'ADN et permettra l'initialisation de la polymerisation. À chaque cycle, un mélange comprenant l'ensemble des quatre désoxyribonucléotides (dNTPs), préalablement étiquetés par un fluorophore et 3' – bloqué sont mis en contact du fragment. Après l'incorporation d'un seul dNTP au fragment, les dNTP non liés sont éliminés et la nature du dNTP ajouté est identifiée grâce à son fluorophore. Le fluorophore et le groupement 3' – bloqué sont retirés permettant ainsi à un nouveau cycle de commencer.

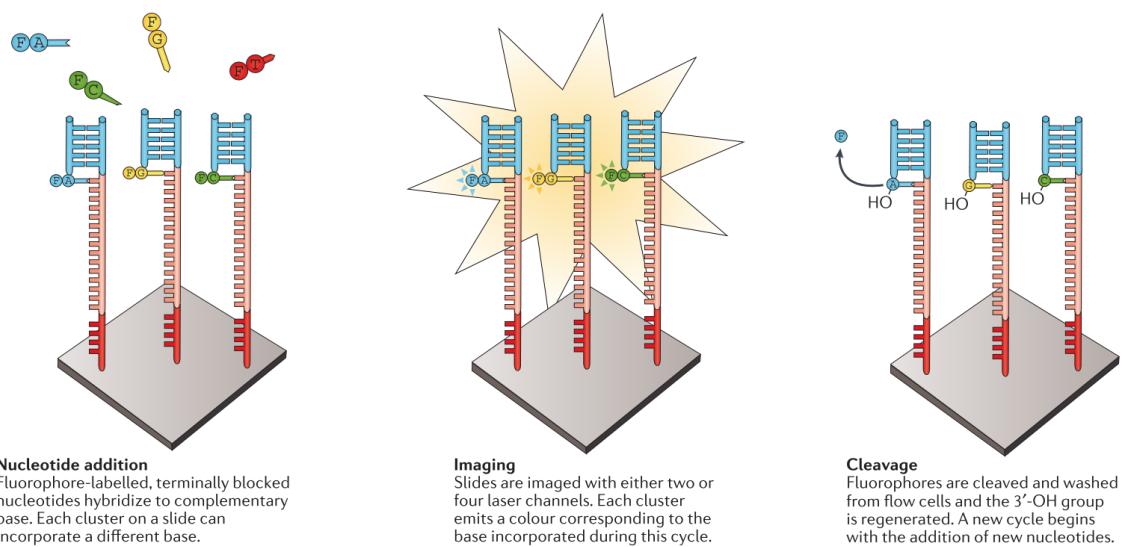


Figure 1.13 – Exemple de séquençage CRT tel qu'il est effectué par Illumina d'après [Goodwin2016] : a : ajout d'un dNTP labellisé par un fluorophore et 3'-bloqué. b : identification du dNTP ajouté grâce au fluorophore. c : le fluorophore est clivé du dNTP et le groupement 3'-OH est reformé à partir du groupement 3'-bloqué permettant ainsi l'élongation

- Addition de nucléotide unique, single nucleotide addition (SNA)** (**Figure : 1.14**) : L'initialisation de la méthode SNA est identique à celle de la méthode CRT. La différence se fait donc au moment de la phase d'élongation. Contrairement à la méthode CRT, le mélange contenant les dNTPs ne contient qu'un seul type de dNTP. Quatre mélanges différents sont donc présentés successivement au fragment d'ADN à séquencer, ceux-ci se fixeront uniquement s'ils sont complémentaires à

la séquence. Ces dNTPs n'ont donc pas besoin d'être 3' – bloqué puisqu'un seul dNTP est ajouté à chaque itération. Après avoir présenté un mixe, vérifie si un dNTP s'est lié au fragment. Lors des séquences homopolymériques (plusieurs nucléotides identiques successifs dans la séquence), plusieurs dNTP sont donc lié simultanément, cela sera détecté car le signal émis sera proportionnel au nombre de nucléotides ajoutés.

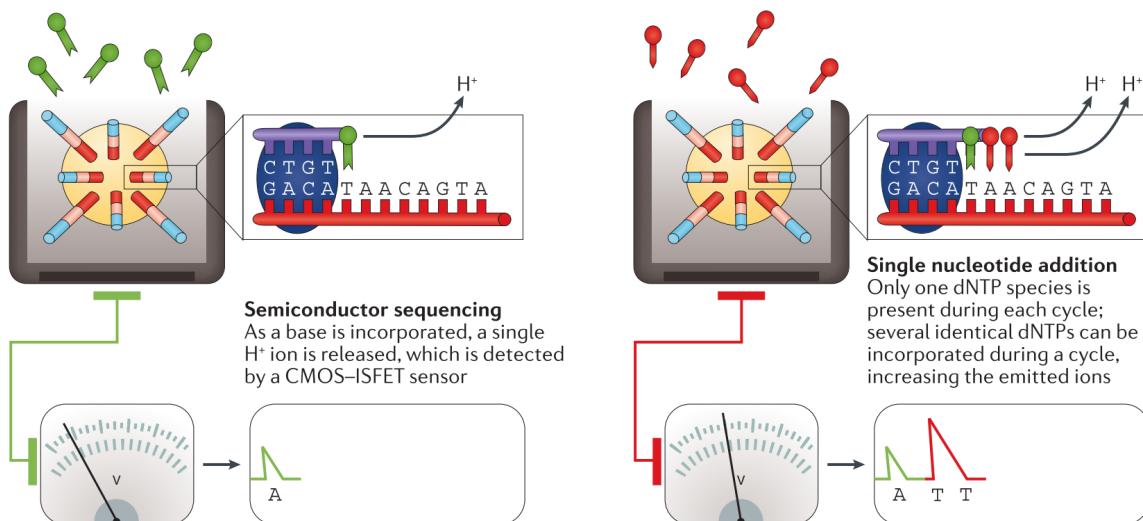


Figure 1.14 – Exemple de séquençage SNA tel qu'il est effectué par Ion Torrent d'après [@Goodwin2016] : **a** : Mise en présence du patron d'ADN à séquencer avec un mix contenant un seul type de dNTP, si le dNTP est complémentaire au patron, il se fixe et libère un proton permettant d'identifier la liaison. **b** : Dans d'homopolymère, plusieurs nucléotides identiques successifs, autant de proton sont relâché que de constituant de bases constituant l'homopolymère, le signal émit est donc plus fort permettant d'identifier le nombre des dNTPs liés

2. **Séquençage par ligation (SBL)** : Par définition, cette méthode est basée sur l'hybridation et la ligation de l'ADN (Tomkinson, Vijayakumar, Pascal, & Ellenberger, 2006) d'une sonde liée à un fluorophore. Ce processus utilise les caractéristiques de la ligase, une enzyme qui a pour fonction de catalyser la liaison de deux brins d'ADN par des liaison phosphodiester. La sonde est constituée d'une ou deux bases connues, on parle alors de *one-base-encoded probes* ou de *two-bases-encoded probes* suivis d'une succession de bases "dégénérées" ou universelle, c'est à dire, des bases capables de s'apparier avec n'importe laquelle des quatre bases de l'ADN.

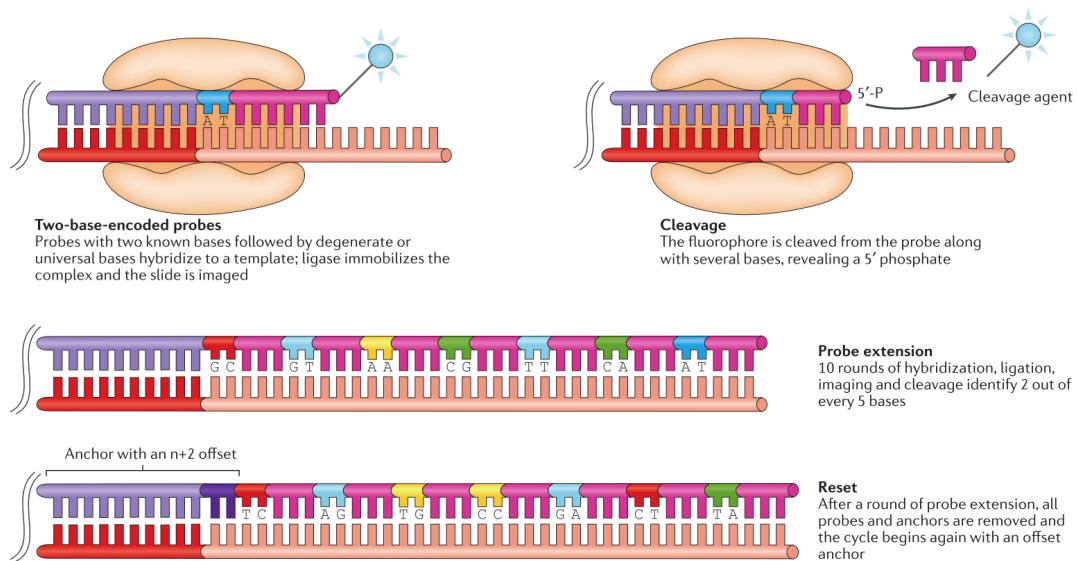


Figure 1.15 – Exemple de séquençage SBL tel qu'il est effectué par SOLiD d'après [Goodwin2016] :

1.5 L'analyse bioinformatique des données de NGS

La stratégie consistant à séquencer en parallèle plusieurs milliers de *reads* court a engendré plusieurs nouveaux défis bioinformatique dans l'analyse et l'interprétation des données de séquençage et la recherche de variants dans le génome humain (Wold & Myers, 2007, M. Q. Yang et al. (2009)). Ces techniques ont été appliquées dans différents contextes, notamment la métagénomique (J. Qin et al., 2010), la détection de SNPs (Van Tassell et al., 2008) et de variants structuraux (Alkan et al., 2010, Medvedev, Stanciu, & Brudno (2009)) mais également dans des études portant sur la méthylation de l'ADN (K. H. Taylor et al., 2007), l'analyse de l'expression des ARNs messagers (Sultan et al., 2008), dans la génétique du cancer (Guffanti et al., 2009) et la médecine personnalisée (Auffray, Chen, & Hood, 2009). Cependant, pour l'ensemble de ces applications, la grande quantité de données générées par chaque analyse pose plusieurs défis informatiques (Horner et al., 2009). En effet, les progrès techniques des dernières décennies ont rendu possible le séquençage de plusieurs millions des *reads* d'ADN en un temps relativement court et à couts raisonnable. Ainsi, l'émergence du séquençage haut débit et notamment du WGS et du WES a permis de réunir une quantité jusqu'à présent inégalé d'information sur les variations génétiques, et d'une manière plus générale, sur les gènes et leurs fonctions (Mardis, 2008, Bentley (2006)). Cependant, de par leur nature et leur quantité, l'acquisition de ces nouvelles données a engendrée de nouvelles problématiques qui freinent les biologistes dans leurs recherches.

1.5.1 Les données fournies par le NGS

Un *read* c'est quoi ?

Après la phase d'amplification, chaque clone est analysé puis, la séquence composant chacun de ces clones est déterminée. La taille de cette séquence varie en fonction de la plateforme de séquençage (**Figure : 1.11**) mais est généralement comprise entre 40 et 150 pb pour le NGS. Deux types de *reads* :

1. ***Read single-end*** :
2. ***Read paired-end*** : Avec le séquençage de *paired-end reads* les deux extrémités (les *ends*) du fragment d'ADN sont désormais séquencées. La distance séparant les deux extrémités du *read* étant connue, cela permet aux aligneurs d'utiliser cette information afin d'améliorer leur précision, notamment dans les zones répétées (H. Li et al., 2008). En plus de SNP, ce format permet de mettre en évidence des variants structuraux (Korbel et al., 2009).

Le format FASTQ

Le format FASTQ (**Figure : 1.16**) est le format de donnée le plus couramment retourné par les séquenceurs haut-débit à l'heure actuelle. Sa création est cependant antérieure à l'émergence du NGS puisqu'il fut inventé à la fin du XX^{ème} siècle par Jim Mullikin au Wellcome Trust Sanger Institute alors que le séquençage commençait à prendre de l'ampleur grâce à des projets tels que le Projet Génome Humain. La quantité de données générées par ces programmes nécessitait une analyse automatisée, c'est ainsi que chaque base séquencée s'est vue associé un score de qualité appelé *Phred-score*. Chaque séquence générera ainsi deux fichiers, un fichier FASTA contenant les séquences et un fichier QUAL contenant les scores *Phred* associés à chaque base du fichier FASTA. Plus tard, afin de n'avoir à manipuler qu'un seul fichier, les fichiers FASTA et QUAL furent fusionnés en ce que l'on appelle désormais le fichier FASTQ. Ce format est aujourd'hui le plus utilisé par les différents séquenceurs. On peut noter certaines différences dans les formats FASTQ provenant des différentes plateformes puisqu'à l'époque, aucune spécification officielle n'avait été donnée (Cock, Fields, Goto, Heuer, & Rice, 2009).

```

@HC9D00P01AN1VB rank=0000246 x=156.0 y=3301.0 length=309
ACACATACGCACTGGCGTAAAGGGCGCGCAGGGCGTCAGAGCGCTGGTGCTAAAGTCCACCGCTTAACGGTGGAGGCCTG
+HC9D00P01AN1VB
FFFFFFFFFFFFGD554A6911144442AAABDFFIIIIIIIIIIIIIIIIHHHFFFFFFFA@CFFDFDFC??CCFFFFFFI
@HC9D00P01AWYAE rank=0000402 x=258.0 y=772.0 length=373
ACACATACGCACTGGCATAAAGGGCACGTAGCGGATTGTAAGTCAGGGGTGAATCCGGGCGTCAACCTCGGAAGTCCT
+HC9D00P01AWYAE
IIIIIIIIIIHHHII;666HHHIIIIIIIIICCIIEEEFDC2//.<-//93.....---9?CCCCFEEECCCIIIIDI
@HC9D00P01A3C8R rank=0000675 x=331.0 y=1081.0 length=373
ACACATACGCACTGGGTTAAAGGGTGCCTAGGCCTTAAGTCAGGGGTGAATCTGGAGCTCAACTCCAGAACTGCCTT
+HC9D00P01A3C8R
IIIIIIIIII3//...---4AIIIECCE466GH974EEIAC@.0004.000>9@CEEEIIIIIIIIIIIIIIIIHHI
@HC9D00P01AW8TJ rank=0000926 x=261.5 y=2133.0 length=373
ACACATACGCACTGGGTGTAAGCGCACGTAGCGGATTGTAAGTCAGGGGTGAATCTGGAGCTCAACTCCAGAACTGCCT
+HC9D00P01AW8TJ
IIIIHHHHHHHHIII;;IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@HC9D00P01AUI8Y rank=0000952 x=230.0 y=2656.0 length=372
ACACATACGCACTGGCATAAAGAGCGCGTAGGCCTTGTAGTCGAGTGTGAAAGCCCTGGCTTAACCCGGGAAGCGCGC
+HC9D00P01AUI8Y
IIIIIIIIIIHHHIII;;IIIIIIIIIIIIIIIIIIIIIIIIIIII?666DHMHFEEIIIIC;555994?FIGI:
@HC9D00P01AUAIW rank=0000977 x=228.0 y=226.0 length=372

```

Figure 1.16 – présentation d'un fichier FASTQ (FIGURE A CHANGER) : **a** : identifiant du *read*. **b** : séquence du *read*. **c** : score de qualité associé

1.5.2 L'alignement

L'alignement constitue la première étape de l'analyse des données de NGS lorsqu'un génome de référence est disponible. L'objectif de l'alignement est de déterminer la position correcte de chacun des *reads* séquencés le long du génome de référence. Cette référence est souvent construite à partir des données de séquençage de plusieurs donneurs et ne représente donc pas la séquence d'un individu en particulier mais est censé représenter la séquence consensus d'une espèce donnée. Par exemple, la séquence de référence humaine GRCh37 (*Genome Reference Consortium human build 37*) a été créés à partir de 13 volontaires anonymes New-Yorkais. Dès lors, cette référence servira de patron aux aligneurs afin qu'ils replacent correctement les différents *reads* des individus séquencés. Cette étape peut être comparée à la reconstruction d'un puzzle dans laquelle les *reads* seraient les pièces et le génome de référence le modèle. Elle constitue probablement l'étape la plus importante de l'analyse des données issues du séquençage haut débit (Flicek & Birney, 2009) car elle est la base sur laquelle reposent l'ensemble des étapes effectuées en aval, notamment l'appel des variants (R. Nielsen, Paul, Albrechtsen, & Song, 2011). Cependant, l'étape d'alignement peut être sujette à de nombreuses erreurs dont certaines proviennent directement des erreurs devenues lors de l'étape de séquençage, d'autres, sont dues aux caractéristiques des régions séquencées comme par exemple les séquences répétées (Ben Langmead & Salzberg, 2012) qui pourront entraîner l'alignement d'un même *read* à plusieurs régions du génome (Treangen & Salzberg, 2013). De nombreux aligneurs ont émergé afin de répondre au mieux à cette problématique tel que Bowtie (B. Langmead, Trapnell, Pop, & Salzberg, 2009), Bowtie2 (Ben Langmead & Salzberg, 2012), BWA, NovoAlign, MAGIC (Su et al., 2014). De nombreuses études ont cependant montré de grandes différences entre ces aligneurs, au niveau du temps de calcul, de leur coût en mémoire et

de leur taux d'erreur (Ruffalo, Laframboise, & Koyutürk, 2011, Thankaswamy-Kosalai, Sen, & Nookae (2017), S. Bao et al. (2011)).

1.5.3 L'appel des variants

L'appel des variants, ou *variant calling*, fait référence à l'ensemble des méthodes permettant d'identifier des SNVs ou des indels à partir des résultats de l'alignement. Cette étape est souvent différenciée de l'alignement, cependant, les résultats de l'appel étant extrêmement dépendant de l'alignement, il est conseillé d'effectuer son appel en tenant compte de l'aligneur choisi (R. Nielsen et al., 2011, M. A. DePristo et al. (2011), Lunter & Goodson (2011)). On appellera variants toutes différences de séquence observées entre un individu et la séquence de référence utilisée. Pour reprendre la comparaison avec la construction d'un puzzle, cette étape consiste à détecter quels sont les pièces qui présentent des différences avec le modèle. De nombreux logiciels d'appel des variants, ou *caller*, basés sur des algorithmes différents ont émergés ces dernières années pour répondre à cette problématique. Parmi les plus connus on note SAMtools (H. Li et al., 2009), Genome Analysis Tool Kit - HaplotypeCaller (GATK-HC) (McKenna et al., 2010), Freebayes, SOAPindel et TVC . Les quatre premiers cités, peuvent être utilisés pour analyser des données provenant de tout type de plateforme de séquençage contrairement à TVC qui a été développé spécifiquement pour les données provenant de Ion Proton. Les données issues de NGS peuvent présenter un taux d'erreur important. Ce taux d'erreur est multifactoriel et inclus notamment les erreurs de l'alignement. L'un des éléments clef à prendre en compte pour pouvoir effectuer un appel de qualité est la couverture de la position appelée (D. Sims et al., 2014). Cependant, malgré la prise en compte de cet élément, l'appel de variants reste un processus difficile souvent lié à plusieurs erreurs. Plusieurs de ces erreurs sont même directement liées à la plateforme de séquençage utilisée en amont, et les différents logiciels ne présentent pas les mêmes performances en fonction de ces différentes plateforme (Hwang, Kim, Lee, & Marcotte, 2015), c'est pourquoi il convient d'adapter le logiciel d'appel en fonction de la plateforme de séquençage utilisée préalablement. Les erreurs d'appel sont généralement classées en deux catégories principales et certains aligneurs auront tendance à être plus sujets à l'un de ces types d'erreur qu'à l'autre (**Figure : 1.17**) :

1. Oubli de l'allèle de référence (**IR**, *ignore the reference allele*) : représente un variant appelé homozygote correspondant en réalité à un variant hétérozygote composé de l'allèle de référence et d'un allèle variant.
2. Ajout de l'allèle de référence (**AR**, *adding the reference allele*) : représente un variant appelé hétérozygote composé de l'allèle de référence et d'un allèle variant correspondant en réalité à un variant homozygote composé de deux allèles variants.

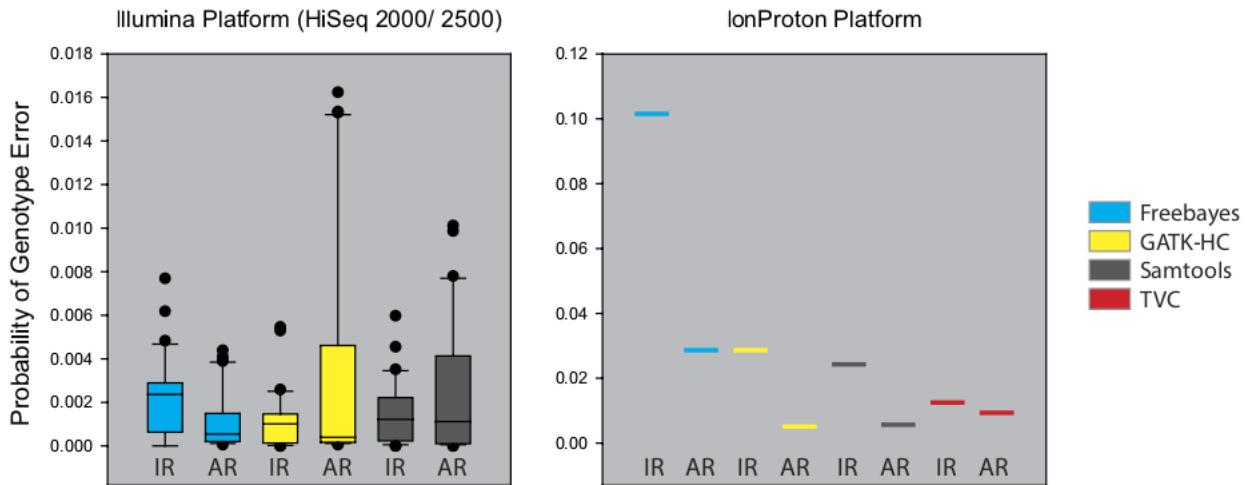


Figure 1.17 – Représentation des erreurs d'appel de type IR et AR en fonction de la plateforme de séquençage et du logiciel d'appel d'après [Hwang2015] : Pour la plateforme Illumina, on peut voir que Freebayes préfère les appels variant-homozygote tandis que GATK-HC et Samtools préfèrent les appels hétérozygotes. Pour la plateforme Ion Proton, les 4 logiciels ont une préférence pour les erreurs de type IR

De même que pour l'aligneur, le choix du logiciel d'appel est crucial car il existe de nombreuses différences dans les variants appelés par différents logiciels se basant sur les mêmes données brutes (Baes et al., 2014, O’Rawe et al. (2013), Rosenfeld, Mason, Smith, Wallin, & Diekhans (2012)). En effet, en 2013, une étude comparant les résultats de 5 caller montrait que seulement 57,4% des variants étaient appelés par les 5 caller et que 80,7% des variants étaient appelés par au moins 3 d'entre eux. Ce taux chutait drastiquement pour les indels puisque la concordance était cette fois seulement de 26,8% pour les indels non retrouvés par les 3 *caller* (O’Rawe et al., 2013). Ces résultats sont cependant à pondérés avec une étude de 2015 comparant 4 *caller* et montrant que 91,7% des SNVs séquencés sur une plateforme Illumina étaient appelés par 3 *caller*, cependant, pour les variants séquencés sur Ion Proton, seulement 27,3% des variants étaient appelés par au moins 3 *caller* et 57,4% des variants n'étaient appelés que par un seul des *caller* (Hwang et al., 2015).

1.5.4 L'annotation des variants, filtrage et prioritisation

Traditionnellement, les scientifiques développaient leur expertise dans un nombre de pathologie et de gènes associés limité. L'emergence du NGS a totalement remis en cause cette pratique, dès lors qu'il est désormais courrant de retrouver plus de 30.000 variants différents par exome. Afin de pouvoir lier un variant à une pathologie, il

est désormais indispensable d'annoter cet ensemble de variant, c'est à dire d'associer à ces variants l'ensemble des informations qui les caractérisent afin de pouvoir les replacer dans leur contexte biologique. Ces informations serviront ensuite d'indicateur afin de filtrer ou prioriser un variant. Cette dernière étape de l'analyse est elle aussi cruciale puisqu'elle permet de réduire le nombre de variant à considérer On peut généralement distinguer deux niveaux d'annotation d'un variant :

1. **Au niveau du variant** : Ce niveau d'annotation regroupe l'ensemble des informations **spécifiques** à un variant
 - a. **Information issues des résultats du séquençage** : la couverture du variant ainsi que la qualité qui lui est associée peuvent permettre de considérer un variant comme étant. Le génotype associé à ce variant est également une information importante.
 - b. **La fréquence du variant dans la population générale** : l'emergence du séquençage haut-débit a permis de gros consortium tel que ESP6500 [CITATION], 1KG [CITATION]. Ces consortium ont été mis à disposition du public de données de séquençage exomique de 6503 individus pour ESP et de 2504 pour la phase 3 du 1000Genomes. On peut également noter l'*Exome Aggregate Consortium* (ExAC) (Lek et al., 2016) qui n'a effectué aucun séquençage mais qui a récupéré les données de plusieurs gros jeux (notamment 1000Genome et ESP) afin de les appliquer à la même analyse bioinformatique harmonisant ainsi les données provenant de 60.706 individus non apparentés. Cette masse d'information permet de faire une idée de la fréquence d'un variant dans la population générale et même au sein de sous-populations humaines.
 - c. **Son impact sur le transcript** : Dans la plupart des analyses phénotype-génotype, les chercheurs se limitent au variant chevauchant des transcrits codants pour une protéine. Il est donc important de savoir l'impact d'un variant sur ce transcript, c'est à dire si le variant va causer une mutation synonyme, un faux-sens... Des logiciels tels que *Variant Effect Predictor* (VEP) (W. McLaren et al., 2016), SnpEff (Cingolani et al., 2012) ou encore ANNOVAR [@] vont prédire l'impact qu'aura un variant sur les différents transcrits qu'il chevauche. D'autre logiciel tels que SIFT (P. Kumar, Henikoff, & Ng, 2009), PROVEAN (Y. Choi, Sims, Murphy, Miller, & Chan, 2012), Polyphen2, ou encore CADD vont chercher à prédire la pathogénicité de ce variant, c'est à dire la probabilité que ce variant soit délétère pour l'individu qui le porte. Bien que cette information soit importante, elle est à pondérer étant donné le peu de concordance qu'il existe entre les prédictions de ces différents logiciels (**Figure : 1.18**).

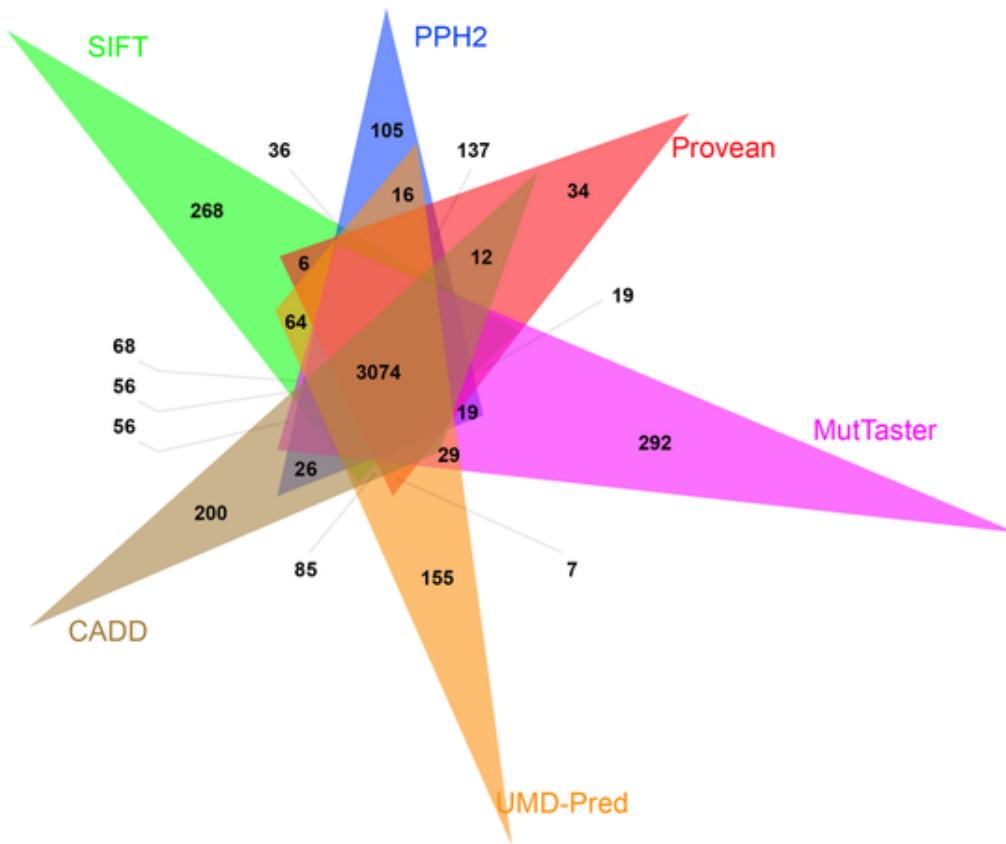


Figure 1.18 – Diagramme de Venn des prédictions de pathogénicités de six logiciels d'après [Salgado2016] :

2. **Au niveau de l'unité génétique** : DÉCRIRE UNITÉ GÉNÉTIQUE (gène, transcript). L'annotation au niveau de l'unité génétique consiste à récupérer l'ensemble des informations disponibles non plus sur le variant uniquement mais sur la ou les unités génétiques qu'il impacte. Ce “dézoom” permet d'ajouter des informations complémentaires particulièrement utiles notamment lorsque plusieurs informations sont disponibles sur le variant lui-même. En pratique, la plupart des variants connus pour impliquer une pathologie sont des variants privés, c'est à dire spécifiques à une famille ou un individu limitant ainsi la quantité d'information disponible sur ce variant. Élargir l'annotation au niveau des unités génétiques impactées par des variants permet d'augmenter considérablement la quantité d'information disponible et permet donc d'améliorer la capacité des algorithmes à filtrer et / ou prioriser les variants rendant donc les analyses plus efficaces. On peut relever certains logiciels tels que le *Protein ANalysis THrough Evolutionary Relationships* (PANTHER) (Mi et al., 2017) qui permet par exemple de classer une liste de gènes en fonction de leurs fonctions moléculaires, des processus biologiques et des voies de signalisation dans lesquels ils sont impliqués. On peut également noter *the Human Phenotype Ontology project* (HPO) (Köhler et al., 2014) qui fournit une classification (à compléter). Plus récemment, on a pu voir émerger des “scores mutationnels” tels que RVIS (Petrovski et al., 2013)

ou encore le pLI (Lek et al., 2016). En se basant sur les bases de données telle que ESP ou encore ExAC, ces scores permettent de classer les gènes en fonction de leur tolérance (ou intolérance) aux variations avec l'idée sous-jacente que "les gènes impliqués dans des pathologies à transmission Mendéliennes" devraient être moins tolérants aux variations que les autres.

Comme nous l'avons vu, l'accumulation de ces informations est extrêmement importante puisqu'elle permet aux biologistes de faire face à la masse de données générées par le NGS l'a aidant ainsi dans ses prises de décisions. Il est à noter que la plupart de ces informations sont extrêmement dépendantes du jeu de gènes utilisés, les prédictions seront donc différentes si l'on se base sur les gènes RefSeq, Ensembl ou UCSC (D. J. McCarthy et al., 2014, S. Zhao & Zhang (2015)) bien que les gènes du *Consensus Coding Sequence project* (CCDS) soient bien représentés par ces trois listes (K. D. Pruitt et al., 2009). De même, pour une même liste de gène, de nombreuses différences seront observées en fonction du ou des logiciels de prédition utilisés (D. J. McCarthy et al., 2014, Salgado 2016).

1.5.5 Conclusion NGS

A key challenge in genomics is to understand the phenotypic consequence of genomic variation. With the advent of next-generation sequencing technologies, the challenge is no longer to generate DNA sequence data, but to interpret them. Currently, the molecular basis of roughly 3700 Mendelian diseases has been elucidated, and a similar number of named Mendelian or suspected Mendelian diseases awaits elucidation [1] [1] A new face and new challenges for Online Mendelian Inheritance in Man

En moins de 10 ans, les technologies NGS sont passées du séquençage de panel de gènes (environ 100 Mb pour le Roche GS FLX system) au séquençage de génome entiers (environ 1500 GB pour l'Illumina Hiseq 4000) et d'une utilisation exclusive à la recherche à la routine clinique. Cependant, et ce malgré son succès dans le domaine de la génomique et de la post-génomique, plusieurs problématiques découlent de cette technologie. Il reste au NGS plusieurs problématiques à résoudre

Malgré les dizaines de milliers d'exomes et de génomes ayant été jusqu'à présent étudiés, notre compréhension des mécanismes moléculaires qui sous-tendent la variété génétique humaine reste particulièrement limitée, et ce particulièrement dans le contexte de l'analyse de pathologies génétiques.

Cependant, cette quantité de données produites crées de nouvelles problématiques pour les généticiens qui se retrouvent désormais face au "déluge de données génétiques" (Schatz & Langmead, 2013) ce qui se retrouve être un frein dans la compréhension et l'interprétation des réseaux de gènes et leurs implications dans des pathologies, la limitation de cette technologie n'étant plus le séquençage d'un, de plusieurs, ou de l'ensemble des gènes, mais plutôt l'analyse et l'interprétation de la masse de données générées.

De même, nous avons pu voir que le séquençage par NGS se déroulait en plusieurs étapes mêlant à la fois des techniques de biologie moléculaire pour l'amplification par exemple, et des techniques d'informatiques et mathématiques, comme pour l'alignement.

De nombreux efforts sont fait pour palier la contrainte instaurée par les *reads courts* dans le cadre d'analyse génomique, cependant les solutions informatique et Bioinformatique proposée jusqu'à présent sont bien en dessous des besoins créés pour l'analyse des données NGS (J. D. McPherson, 2009).

Le séquençage nouvelle génération (NGS) a apporté avec lui des opportunités sans précédent dans le domaine de la recherche en génomique. Il a pu être appliqué à une grande variété de contexte avec notamment le séquençage de génome entier, ou *Whole Genome Sequencing* (WGS)) ou encore le séquençage exomique, le *Whole Exome Sequencing* (WES). Cependant, certaines de ses caractéristiques techniques tel que la production de plusieurs milliards de ***reads courts***, bien qu'elles soient en partie responsable de son succès, sont aussi à l'origine de nouvelle problématique, notamment dans l'analyse et l'interprétation des données.

cf Evaluation of next-generation sequencing software in mapping and assembly partie CHALLENGES AND PROSPECTS

Chapitre 2

Investigation génétique et physiologique de la globozoospermie

Chapitre 3

MutaScript

3.1 Introduction

Up to a few years, linkage analysis using genetic markers or Sanger sequencing were massively used to identify genetic disorder in clinical research. These technic were extremely time and money expensive and in 2011 the genetic cause of over 3,500 Mendelian disorders remained unknown [1]. The advent of New Generation Sequencing (NGS) has immediately initiated the new era of clinical research in genomic by bringing the sequencing of entire genomes (WGS) or exome (WES) economically feasible for many small laboratories. This technological leap has permitted a great advance in the field of clinical research linking more than ... genetic cause to a mendelian disorder [need citation here]. However, despite all of that, NGS bring with him new problematics. Indeed, with over 30,000 variants per individual exome and surrounding 10,000 of which are predicted to lead to a nonsynonymous substitution, a modification of a splice sites, or to a small insertion / deletion (indel) [need citation], finding the disease-causing variant became to be the daily labor-intensive task of geneticists. Even more that this new task regrouping among other things informatics and statistic is far from traditional geneticist skills. Usually NGS data analysis is divided in three principal steps. The first one is the raw data pre-processing which mainly consists to mapping raw data to a reference genome (may be cite some soft which do that ?). Secondly, the analysis of mismatches between sample data and the reference sequence draw up a list of Single Nucleotide Variation (SNV) and small insertion / deletion (Indels). As said previously this list can enumerate more than 30,000 variation per individual (may be cite some soft which do that ?). Finally, the analyze of these variants, including the annotation and filtering, is often weakness of genotype-phenotype analysis. Indeed, in most of case it is not enough to obtain a small human interpretable list and the causal variant is drowned in a multitude of non-causal variant. To improve the quality of variant filtering and so the efficiency of phenotype-genotype analysis we develop MutaScript. This score rank transcript considering their probability of being link to a mendelian disorder. For that we assumed that most mutated transcripts in general

population were probably not involved in any severe Mendelian disease. MutaScript score of a transcript combines the allelic frequency of each variant displayed on ExAC [2] database overlapping this transcript as well as their impact predicted by Variant Effect Predictor (VEP) [36]. So, higher is the variant impact, higher will be its MutaScript score contribution. Ranking transcript (or gene) considering their variation load was already done by the Residual Variation Intolerance Score (RVIS) [4] and the Probability of loss-of-function (LoF) Incoherency (pLI) [2] scores. However, MutaScript differs from them on several points. Firstly, like pLI, MutaScript give a score to each protein-coding transcript where RVIS only scored genes which leads to a loss of information from RVIS. Moreover, RVIS and pLI only consider LoF variants, mainly splice donor / acceptor variants, nonsense or frameshift variants which represent only ... % of ExAC variants. MutaScript consider all variants whatever its consequence, and it weight its score contribution considering the predicted deleteriousness of its impact. Also, we noticed a strong correlation between the RVIS score of a gene and the pLI score of a transcript and the CDS size of this gene (or transcript). Because this correlation is mostly due to a bias of their score formula and not because a biological fact, we build MutaScript to avoid this correlation which can lead to interpretation errors. To validate our scores we used the Human Phenotype Ontology (HPO) [5] and demonstrate that MutaScript is more...

3.2 Matériel & Méthodes

3.2.1 Réécuation des données ExAC, filtrage et pré analyse

3.3 Résultats

3.3.1 Définition de la formule de score

3.3.2 Analyse de la corrélation

3.3.3 Analyse HPO

3.4 Conclusion

Conclusion

Annexe A

The First Appendix

In the main Rmd file

In Chapter ?? :

Annexe B

The Second Appendix, for Fun

References

- Adelman, M. M., & Cahill, E. M. (1989). *Atlas of sperm morphology* (p. 123). ASCP Press.
- Alkan, C., Kidd, J. M., Marques-bonet, T., Aksay, G., Hormozdiari, F., Kitzman, J. O., ... Eichler, E. E. (2010). Personalized Copy-Number and Segmental Duplication Maps using Next-Generation Sequencing. *Nature Genetics*, 41(10), 1061–1067. <http://doi.org/10.1038/ng.437>. Personalized
- Asimakopoulos, B. (2003). Is There a Place for Round and Elongated Spermatids Injection in, 1(1), 1–6.
- Auffray, C., Chen, Z., & Hood, L. (2009). Systems medicine : the future of medical genomics and healthcare. *Genome Medicine*, 1(1), 2. <http://doi.org/10.1186/gm2>
- Baes, C. F., Dolezal, M. A., Koltes, J. E., Bapst, B., Fritz-Waters, E., Jansen, S., ... Gredler, B. (2014). Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. *BMC Genomics*, 15(1), 948. <http://doi.org/10.1186/1471-2164-15-948>
- Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X., & Song, Y.-Q. (2011). Evaluation of next-generation sequencing software in mapping and assembly. *Journal of Human Genetics*, 56(May), 406–414. <http://doi.org/10.1038/jhg.2011.62>
- Bentley, D. R. (2006). Whole-genome re-sequencing. *Current Opinion in Genetics and Development*, 16(6), 545–552. <http://doi.org/10.1016/j.gde.2006.10.009>
- Cho, C., Willis, W. D., Goulding, E. H., Jung-Ha, H., Choi, Y. C., Hecht, N. B., & Eddy, E. M. (2001). Haplloinsufficiency of protamine-1 or -2 causes infertility in mice. *Nature Genetics*, 28(1), 82–6. <http://doi.org/10.1038/88313>
- Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE*, 7(10). <http://doi.org/10.1371/journal.pone.0046688>
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide poly-

- morphisms, SnpEff. *Fly*, 6(2), 80–92. <http://doi.org/10.4161/fly.19695>
- Clermont, Y. (1963). The cycle of the seminiferous epithelium in man. *American Journal of Anatomy*, 112(1), 35–51. <http://doi.org/10.1002/aja.1001120103>
- Clermont, Y. (1966). Renewal of spermatogonia in man. *American Journal of Anatomy*, 118(2), 509–524. <http://doi.org/10.1002/aja.1001180211>
- Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6), 1767–1771. <http://doi.org/10.1093/nar/gkp1137>
- Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project : Lessons from Large-Scale Biology. *Science*, 300(5617), 286–290. <http://doi.org/10.1126/science.1084564>
- 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>
- Eddy, E. M. (2007). The scaffold role of the fibrous sheath. *Society of Reproduction and Fertility Supplement*, 65, 45–62. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/17644954>
- Escalier, D., Gallo, J. M., Albert, M., Meduri, G., Bermudez, D., David, G., & Schrevel, J. (1991). Human acrosome biogenesis : immunodetection of proacrosin in primary spermatocytes and of its partitioning pattern during meiosis. *Development (Cambridge, England)*, 113(3), 779–788. Retrieved from <http://dev.biologists.org/content/develop/113/3/779.full.pdf>
- Flicek, P., & Birney, E. (2009). Sense from sequence reads : methods for alignment and assembly. *Nature Methods*, 6(11 Suppl), S6–S12. <http://doi.org/10.1038/nmeth0610-479b>
- Gnessi, L., Fabbri, A., & Spera, G. (1997). Gonadal peptides as mediators of development and functional control of the testis : An integrated system with hormones and local environment. *Endocrine Reviews*, 18(4), 541–609. <http://doi.org/10.1210/er.18.4.541>
- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age : ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17(6), 333–351. <http://doi.org/10.1038/nrg.2016.49>
- Goossens, E., & Tournaye, H. (2013). Adult stem cells in the human testis. *Seminars in Reproductive Medicine*, 31(1), 39–48. <http://doi.org/10.1055/s-0032-1331796>
- Guffanti, A., Iacono, M., Pelucchi, P., Kim, N., Soldà, G., Croft, L. J., ... De Bellis,

- G. (2009). A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. *BMC Genomics*, 10(1), 163. <http://doi.org/10.1186/1471-2164-10-163>
- Guo, J., Xu, N., Li, Z., Zhang, S., Wu, J., Kim, D. H., ... Ju, J. (2008). Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, 105(27), 9145–9150. <http://doi.org/10.1073/pnas.0804023105>
- Hamilton, D. W., Waites, G. M. H. (1990). *Cellular and Molecular Events in Spermiogenesis* (p. 334). Cambridge University Press. Retrieved from <http://www.cambridge.org/us/academic/subjects/medicine/obstetrics-and-gynecology-reproductive-medicine/cellular-and-molecular-events-spermiogenesis>
- Herimo, L., Pelletier, R. M., Cyr, D. G., & Smith, C. E. (2010). Surfing the wave, cycle, life history, and genes/proteins expressed by testicular germ cells. Part 3 : Developmental changes in spermatid flagellum and cytoplasmic droplet and interaction of sperm with the zona pellucida and egg plasma membrane. *Microscopy Research and Technique*, 73(4), 320–363. <http://doi.org/10.1002/jemt.20784>
- Horner, D. S., Pavesi, G., Castrignano', T., Meo, P. D. O. de, Liuni, S., Sammeth, M., ... Pesole, G. (2009). Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in Bioinformatics*, 11(2), 181–197. <http://doi.org/10.1093/bib/bbp046>
- Hwang, S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, 5(December), 17875. <http://doi.org/10.1038/srep17875>
- Inaba, K. (2003). Molecular Architecture of the Sperm Flagella : Molecules for Motility and Signaling. *Zoological Science*, 20(9), 1043–1056. <http://doi.org/10.2108/zsj.20.1043>
- JOHNSON, L., PETTY, C. S., & NEAVES, W. B. (1980). A Comparative Study of Daily Sperm Production and Testicular Composition in Humans and Rats. *Biol Reprod*, 22(5), 1233–1243. Retrieved from <http://www.biolreprod.org/content/22/5/1233.short>
- KIERSZENBAUM, A. L. (1994). Mammalian Spermatogenesis *< i>in Vivo</i>* and *< i>in Vitro</i>* : A Partnership of Spermatogenic and Somatic Cell Lineages*. *Endocrine Reviews*, 15(1), 116–134. <http://doi.org/10.1210/edrv-15-1-116>
- Kierszenbaum, A. L., & Tres, L. L. (1978). RNA transcription and chromatin structure during meiotic and postmeiotic stages of spermatogenesis. *Federation Proceedings*, 37(11), 2512–6. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/>

357185

- Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., ... Snyder, M. (2009). Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *October*, 318(5849), 420–426. <http://doi.org/10.1126/science.1149504.Paired-End>
- Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., ... Robinson, P. N. (2014). The Human Phenotype Ontology project : linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(Database issue), D966–74. <http://doi.org/10.1093/nar/gkt1026>
- 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>
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359. <http://doi.org/10.1038/nmeth.1923>
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3), R25. <http://doi.org/10.1186/gb-2009-10-3-r25>
- 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>
- Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. a, & Gilissen, C. (2015). Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Human Mutation*, 36(8), 815–22. <http://doi.org/10.1002/humu.22813>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <http://doi.org/10.1093/bioinformatics/btp352>
- Li, H., Ruan, J., Durbin, R., Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores Mapping short DNA sequencing reads and calling variants using mapping quality scores, 1851–1858. <http://doi.org/10.1101/gr.078212.108>
- 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>
- Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3), 133–141. <http://doi.org/10.1016/j.tig.2007.12>.

007

- McCarthy, D. J., Humburg, P., Kanapin, A., Rivas, M. a, Gaulton, K., Cazier, J.-B., & Donnelly, P. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, 6(3), 26. <http://doi.org/10.1186/gm543>
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... DePristo, M. A. (2010). The Genome Analysis Toolkit : a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–303. <http://doi.org/10.1101/gr.107524.110>
- 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>
- McPherson, J. D. (2009). Next-generation gap. *Nature Methods*, 6(11s), S2–S5. <http://doi.org/10.1038/nmeth.f.268>
- Medvedev, P., Stanciu, M., & Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, 6(11s), S13–S20. <http://doi.org/10.1038/nmeth.1374>
- Meienberg, J., Bruggmann, R., Oexle, K., & Matyas, G. (2016). Clinical sequencing : is WGS the better WES ? *Human Genetics*, 135(3), 359–362. <http://doi.org/10.1007/s00439-015-1631-9>
- Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, 11(1), 31–46. <http://doi.org/10.1038/nrg2626>
- Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., & Thomas, P. D. (2017). PANTHER version 11 : expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, 45(D1), D183–D189. <http://doi.org/10.1093/nar/gkw1138>
- Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Abigail, W., Lee, C., ... Shendure, J. (2010). Targeted Capture and Massively Parallel Sequencing of twelve human exomes. *Nature*, 461(7261), 272–276. <http://doi.org/10.1038/nature08250.Targeted>
- 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>
- Ogura, a, Matsuda, J., & Yanagimachi, R. (1994). Birth of normal young after electrofusion of mouse oocytes with round spermatids. *Proceedings of the National Academy of Sciences of the United States of America*, 91(16), 7460–7462. <http://doi.org/10.1073/pnas.91.16.7460>
- Ogura, A., Matsuda, J., Asano, T., Suzuki, O., & Yanagimachi, R. (1996). Mouse oocytes injected with cryopreserved round spermatids can develop into normal

- offspring. *Journal of Assisted Reproduction and Genetics*, 13(5), 431–434. <http://doi.org/10.1007/BF02066177>
- O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., ... Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines : practical implications for exome and genome sequencing. *Genome Medicine*, 5(3), 28. <http://doi.org/10.1186/gm432>
- Papic, Z., Katona, G., & Skrabalo, Z. (1988). The cytologic identification and quantification of testicular cell subtypes. Reproducibility and relation to histologic findings in the diagnosis of male infertility. *Acta Cytologica*, 32(5), 697–706. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/3421018>
- Pedersen, H., & Rebbe, H. (1974). Fine structure of round-headed human spermatozoa. *Journal of Reproduction and Fertility*, 37(1), 51–4. <http://doi.org/10.1530/JRF.0.0370051>
- Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., Goldstein, D. B., Davydov, E., ... Lisacek, F. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genetics*, 9(8), e1003709. <http://doi.org/10.1371/journal.pgen.1003709>
- Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., ... Lipman, D. (2009). The consensus coding sequence (CCDS) project : Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, 19(7), 1316–1323. <http://doi.org/10.1101/gr.080531.108>
- Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, S., Manichanh, C., ... Yang, H. (2010). A human gut microbial gene catalog established by metagenomic sequencing. *Nature*, 464(7285), 59–65. <http://doi.org/10.1038/nature08821.A>
- Rosenfeld, J. A., Mason, C. E., Smith, T. M., Wallin, C., & Diekhans, M. (2012). Limitations of the Human Reference Genome for Personalized Genomics. *PLoS ONE*, 7(7), e40294. <http://doi.org/10.1371/journal.pone.0040294>
- Ruffalo, M., Laframboise, T., & Koyutürk, M. (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, 27(20), 2790–2796. <http://doi.org/10.1093/bioinformatics/btr477>
- Sasagawa, I., & Yanagimachi, R. (1997). Spermatids from mice after cryptorchid and reversal operations can initiate normal embryo development. *Journal of Andrology*, 18(2), 203–209. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9154515>
- Schatz, M. C., & Langmead, B. (2013). The DNA Data Deluge : Fast, efficient genome sequencing machines are spewing out more data than geneticists can analyze. *IEEE Spectrum*, 50(7), 26–33. <http://doi.org/10.1109/MSPEC.2013.6545119>
- Schenck, U., & Schill, W. B. (n.d.). Cytology of the human seminiferous epithelium.

- Acta Cytologica*, 32(5), 689–96. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/3421017>
- Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage : key considerations in genomic analyses. *Nature Reviews. Genetics*, 15(2), 121–32. <http://doi.org/10.1038/nrg3642>
- Singh, G. (n.d.). Ultrastructural features of round-headed human spermatozoa. *International Journal of Fertility*, 37(2), 99–102. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/1349598>
- 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>
- Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., ... Yaspo, M.-L. (2008). A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science*, 321(5891), 956–960. <http://doi.org/10.1126/science.1160342>
- Tanaka, A., Nagayoshi, M., Takemoto, Y., Tanaka, I., Kusunoki, H., Watanabe, S., ... Yanagimachi, R. (2015). Fourteen babies born after round spermatid injection into human oocytes. *Proceedings of the National Academy of Sciences*, 112(March 2014), 201517466. <http://doi.org/10.1073/pnas.1517466112>
- Taylor, K. H., Kramer, R. S., Davis, J. W., Guo, J., Duff, D. J., Xu, D., ... Shi, H. (2007). Ultradeep Bisulfite Sequencing Analysis of DNA Methylation Patterns in Multiple Gene Promoters by 454 Sequencing. *Cancer Research*, 67(18), 8511–8518. <http://doi.org/10.1158/0008-5472.CAN-07-1016>
- Thankaswamy-Kosalai, S., Sen, P., & Nookaew, I. (2017). Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics*. <http://doi.org/10.1016/j.ygeno.2017.03.001>
- Tomkinson, A. E., Vijayakumar, S., Pascal, J. M., & Ellenberger, T. (2006). DNA Ligases : Structure, Reaction Mechanism, and Function. *Chemical Reviews*, 106(2), 687–699. <http://doi.org/10.1021/cr040498d>
- Treangen, T. J., & Salzberg, S. L. (2013). Repetitive DNA and next-generation sequencing : computational challenges and solutions. *Nat Rev Genet.*, 13(1), 36–46. <http://doi.org/10.1038/nrg3117.Repetitive>
- Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., ... Sonstegard, T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*,

- 5(3), 247–252. <http://doi.org/10.1038/nmeth.1185>
- Ward, W. S. (1994). The structure of the sleeping genome : implications of sperm DNA organization for somatic cells. *Journal of Cellular Biochemistry*, 55(1), 77–82. <http://doi.org/10.1002/jcb.240550109>
- Wold, B., & Myers, R. M. (2007). Sequence census methods for functional genomics. *Nature Methods*, 5(1), 19–21. <http://doi.org/10.1038/nmeth1157>
- World Health Organization. (1992). *WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction*. (3th ed, p. 128). Cambridge University Press.
- Yang, M. Q., Athey, B. D., Arabnia, H. R., Sung, A. H., Liu, Q., Yang, J. Y., ... Deng, Y. (2009). High-throughput next-generation sequencing technologies foster new cutting-edge computing techniques in bioinformatics. *BMC Genomics*, 10 Suppl 1, I1. <http://doi.org/10.1186/1471-2164-10-S1-I1>
- Zhao, S., & Zhang, B. (2015). A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*, 16(1), 97. <http://doi.org/10.1186/s12864-015-1308-8>