Construction de la donnée génétique à des fins médicales

REVUE MÉDECINE ET PHILOSOPHIE

Yannis Duffourd[*]
[*] FHU-TRANSLAD, CHU Dijon Bourgogne, UMR1231 GAD, Inserm, Dijon, France

RÉSUMÉ
Les données de génétique sont un ensemble vaste et hétérogène de données personnelles et médicales. Leur constitution peut requérir des chemins complexes, impliquant aussi bien des techniques de biologie moléculaire, des méthodes bioinformatiques, statistiques, des stratégies d'interprétation, incluant à chacune de ces étapes une profonde connaissance de la biologie médicale, de la génétique humaine et des populations.
La construction de ces données implique différentes expertises et spécialistes. Partant de l'échantillon biologique et de l'extraction de la molécule d'ADN, les techniques d'études du génome, et particulièrement le séquençage à haut débit, permettent de transformer une information moléculaire en une information bioinformatique sous la forme d'une séquence dite « brute ». Différents traitements permettent d'en extraire toujours plus d'informations, de les enrichir, jusqu'à l'obtention d'une liste réduite de variations d'intérêt du génome étudié. La valeur de la donnée génétique individuelle est parfois mise en évidence par son agrégation avec les données génétiques d'autres individus. Les variations identifiées en regard du phénotype du patient sont étudiées, celles pouvant expliquer la pathologie du patient sont sélectionnées. D'autres données, sans lien avec la pathologie et pas nécessairement souhaitées par le patient, sont également disponibles. Enfin les variations génétiques d'intérêt seront restituées au patient par l'intermédiaire d'un, mais le cycle de vie de la donnée génétique ne s'arrête pas ici, d'autres analyses pouvant être réalisées afin d'approfondir l'étude du génome du patient.

MOTS-CLÉS : Bioinformatique, Données, Séquençage, Génétique, Génomique, Analyse.
DOI : 10.51328/220505

Introduction

Le génome humain recèle de nombreux secrets encore non révélés. Cependant notre connaissance de celui-ci s'agrandit de jours en jours. Grâce à cette évolution et grâce aux avancées technologiques concernant l'étude et l'analyse du génome, la génétique médicale a subi plusieurs révolutions d'ordre scientifique, mais aussi technologique. Il existe environ 8000 maladies génétiques rares impactant à l'échelle de la population globale des millions d'individus à travers le monde. Outre un enjeu scientifique important, l'étude du génome humain est donc également un réel enjeu médical, économique et sociétal.

Étudier le génome humain nécessite l'emploi et la mise en œuvre de beaucoup de connaissances scientifiques. Ces techniques d'étude du génome ont beaucoup évolué, partant du simple caryotype vers les puces à ADN, et le séquençage du génome lui-même. Ce séquençage a

été mis en œuvre par plusieurs concepts technologiques d'envergure. D'abord par le séquençage par la méthode de Sanger (Sanger et al. 1977) considérée comme la technologie de séquençage de première génération, et ses successives améliorations jusqu'à l'avènement du séquençage de deuxième génération (Margulies et al. 2005, Bentley et al. 2008) permettant l'étude du génome à l'échelle individuelle pour un coût relativement accessible. Bien qu'il existe désormais des technologies de troisième génération, dans cet article, nous allons donc nous concentrer à l'étude des données génétiques principalement reliées à la technologie du séquençage à haut-débit.

La donnée génétique reste une notion très vague, désignant successivement plusieurs types d'informations que nous allons décrire à travers cet article. La construction de la donnée génétique reste un point essentiel pour estimer sa valeur, cette construction passe par beaucoup de phases différentes dont une qui sera centrale ici : l'analyse bioinformatique.

La bioinformatique est la science qui permet de transformer une donnée biologique en une information. Décrite pour la première fois en 1970 par Paulien Hogeweg et Ben hesper, elle est multidisciplinaire par essence, et implique la création et l'utilisation d'algorithmes pour l'analyse de données biologiques. Elle est donc une approche incluant l'informatique, les mathématiques, des méthodes statistiques, ainsi qu'une profonde compréhension des problématiques biologiques. Elle est constituée de nombreux domaines d'application différents, et seule une fraction de la bioinformatique se consacre à la génétique humaine.

Construire la donnée génétique

De la données biologique à la donnée bioinformatique

Le séquencage à haut-débit est la technique privilégiée d'étude du génome à l'heure actuelle. Alors que plusieurs acteurs se partagent le marché international des séquenceurs, la société Illumina reste le leader incontesté du marché actuel. C'est donc cette technologie de séquençage qui servira de guide à travers la description du processus de génération de la donnée.

L'ADN reste une molécule complexe présente au sein des noyaux de la plupart des cellules humaines. Le passage de cette information biologique à une information informatique constituée de la séquence elle-même, a été le fruit d'un long développement technologique au cours des dernières décennies. Ce développement est toujours d'actualité et en constante évolution afin de permettre d'aller toujours plus vite, toujours plus précisément, et à un coût de plus en plus faible.

Les molécules d'ADN sont extraites du noyau de la cellule par des techniques de biologie moléculaire aujourd'hui très standardisées. Cet ADN doit être préparé sous la forme d'une librairie, pour qu'il puisse être utilisé par un séquenceur haut-débit.

Dans cette optique de préparation de la librairie, les techniques de seconde génération de séquençage nécessitent de fragmenter les molécules d'ADN en morceaux d'une taille plus petite, d'une longueur de l'ordre de quelques centaines de bases.

En effet, les technologies actuelles ne sont guère capables de séquencer des molécules ayant une grande taille. Les raisons étant la plupart du temps dues à des limites technologiques induites par l'utilisation de nucléotides marqués par des fluorophores qui sont des molécules fragiles et dont la conservation est délicate à long terme.

Cependant, les technologies de séquençage dites de « 3^{e} génération » mettent en œuvre des possibilités de séquençage de fragments d'ADN beaucoup plus longs incluant des tailles jusqu'à plusieurs millions de bases. Elles ne nécessitent donc pas cette étape de fragmentation.

La méthode de fragmentation peut varier selon les besoins et l'application choisie, mais généralement elle se divise en plusieurs choix technologiques : fragmentation par sonication[1], fragmentation enzymatique, méthode d'amplicons[2], etc. Ce choix repose sur des critères dépendant de l'application désirée de l'expérience de séquençage. Ainsi pour des approches de séquençage pangénomique, on optera plutôt pour de la sonication ou une fragmentation enzymatique. La fiabilité de la sonication n'est plus à démontrer, mais sa mise en œuvre est parfois complexe. Ainsi, l'approche enzymatique autrefois délaissée pour les différents biais qu'elle engendrait est de plus en plus répandue, car plus simple à l'utilisation et maintenant largement compétitive avec la technologie de sonication.

Il est ensuite nécessaire de fixer de manière covalente chacune des séquences à étudier à un support de séquençage. Celui-ci peut être de différentes natures en fonction de la technologie de séquençage choisie (Goodwin et al. 2016). Il est souvent nécessaire de fixer des adapteurs aux extrémités des molécules d'intérêt. Ceux-ci sont des séquences artificielles permettant la fixation des séquences au support de séquençage via des techniques de biologie moléculaire, elles permettent également l'amplification de la librairie.

Puis une étape importante de la préparation de la librairie consiste en l'amplification de celle-ci. Il est actuellement impossible pour les technologies de séquençage de deuxième génération de ne séquencer qu'une seule molécule isolée. Celle-ci doit donc être clonée afin d'en multiplier les copies, ce clone constitué de plusieurs milliers de copies est également appelé un cluster.

En effet, la technologie repose sur un principe de séquençage par ajout de bases nucléotidiques complémentaires à la séquence d'intérêt. Ces bases sont marquées à l'aide de fluorophores. Lors de leur incorporation à la séquence, le fluorophore excité par un faisceau laser d'une longueur d'onde déterminée va émettre une intensité lumineuse qui pourra être mesurée et qui témoignera de l'incorporation de cette base. Cependant, l'intensité lumineuse émise par un fluorophore porté par une seule molécule n'est pas détectable par les technologies actuelles. Il est donc nécessaire de multiplier la séquence d'intérêt afin de rendre détectable l'incorporation simultanée de plusieurs milliers de fluorophores. Le processus d'incorporation de ces molécules marquées permet de cliver le fluorophore lorsque la mesure est réalisée et de recommencer à réincorporer la base suivante, exciter le fluorophore et d'en mesurer

[1] La méthode de fragmentation par sonication consiste en un bombardement des molécules d'ADN par des ultra-sons dans des conditions appropriées permettant de casser ces molécules. En appliquant une fréquence, une durée et une température spécifique, il est possible d'obtenir de manière reproductible des fragments d'ADN d'une longueur choisie.

[2] La sélection des portions d'ADN par amplicons est une méthode basée sur l'utilisation de la PCR (Polymerase Chain Reaction) permettant de reproduire une portion choisie de séquences de la molécule d'ADN à l'aide de primers spécifiques. Ainsi sont obtenus des millions de copies de la séquence d'intérêt qui seront ensuite sélectionnées pour en permettre le séquençage.

l'intensité lumineuse résultante. Cette série d'étapes constituée de l'ajout de nucléotides marqués, l'élimination des nucléotides non incorporés par lavage, l'excitation du fluorophore, la mesure de l'intensité lumineuse résultante, et le clivage du fluorophore et du bloqueur constitue un cycle de séquençage de la technologie de seconde génération de séquençage, aussi appelée « wash & scan ».

Bien sûr les détails techniques peuvent changer selon le constructeur et donc la technologie de séquençage choisie, mais le principe du « wash scan » reste celui qui définit ces technologies de deuxième génération de séquençage.

Le séquenceur n'est donc finalement qu'un gros automate de biologie moléculaire, capable de réaliser des millions de réactions en parallèle et qui va prendre en photo ces réactions pour identifier les bases séquencées. Celui-ci va produire une quantité très importante d'images : plusieurs dizaines de Téraoctets pour un génome humain. A partir de cet instant, la molécule d'ADN a en quelques sortes été photographiée et transformée en une donnée informatique, constituée de simples pixels de couleurs différentes. La bioinformatique entre donc en jeu à cet instant dans le processus de génération de la donnée génétique.

Étant donné le volume informatique considérable de ces images, celles-ci ne vont pas être conservées sous ce format. Elles vont immédiatement être analysées par un algorithme permettant d'extraire la position de chacun des clusters, et l'intensité lumineuse émise par chacun des fluorophores différents. Cette opération sera réalisée à chaque cycle de séquençage. Le support étant fixe, un cluster de séquençage aura toujours les mêmes coordonnées dans l'image, et ainsi en analysant les clusters successivement sur ces images, il est simple de reconstituer la séquence d'un cluster de molécules d'intérêt. Cette étape est appelée le « *base calling* », et permet donc de transformer la donnée sous la forme d'une image en une donnée textuelle constituée de la séquence nucléotidique elle-même. Cette étape est réalisée en partie par le séquenceur lui-même, la fin de celle-ci étant généralement réalisée par une machine de calcul indépendante du séquenceur.

Ces fichiers contenant les séquences constituent la donnée dite « brute » de la bioinformatique du séquençage à haut-débit. Elle est généralement proposée sous la forme d'un fichier standardisé au format fastq. (Andrews et al. 2010) Ces fichiers peuvent contenir des centaines de millions de séquences dans le cas d'une expérience de séquençage du génome humain.

A partir de cette étape, le séquenceur n'entre plus en œuvre dans le traitement de la donnée, celui-ci peut donc être utilisé pour réaliser une nouvelle expérience de séquençage indépendante. Le traitement de la donnée va maintenant impliquer du matériel informatique dédié, ainsi que des algorithmes créés, choisis et paramétrés grâce à l'expertise du bioinformaticien en fonction de l'application de séquençage choisie.

L'analyse bioinformatique

Données brutes. L'informatique est un monde de possibilités infinies. Ainsi, représenter des données génétiques d'une manière informatique peut donner lieu à beaucoup d'interprétations différentes. Il reste cependant nécessaire de pouvoir comparer les données issues du séquençage de patients différents entre elles. Et pour réaliser cette comparaison de manière efficace, il est nécessaire de représenter les données dans des formats standardisés de fichiers informatiques.

C'est donc pour cette raison qu'un certain nombre de formats spécifiques de fichiers existent et évoluent. Tous les acteurs du séquençage à haut-débit sont fortement encouragés à utiliser ces formats de fichiers, mais leur évolution n'est pas non plus interdite. Ces formats seront cités et décrits lors des différentes étapes de l'analyse bioinformatique les impliquant.

Le premier format de fichier spécifique du séquençage à haut-débit est le format fastq évoqué plus tôt. Celui-ci permet de représenter la donnée dite « brute » de séquençage, soit la séquence en elle-même mais aussi la qualité de celle-ci. En effet chacune des bases séquencées par le séquenceur haut-débit se voit attribuer un score de qualité calculé en fonction de l'intensité du cluster, de sa netteté, des interférences détectées dans l'image, du bruit de fond, etc. Ce score est donc inclus dans ce fichier. Chaque séquence sera identifiée par un identifiant unique permettant de tracer son analyse.

Évidemment, les séquences générées ne sont pas parfaites. Cette imperfection du séquençage est due aux différentes limites technologiques actuelles. L'un des principaux biais est la présence d'erreurs de séquençage. En effet, les technologies de séquençage utilisant du matériel biologique pour réaliser les réactions de séquençage, celles-ci peuvent faire des erreurs, par exemple en incorporant une mauvaise base lors d'un cycle de séquençage. Ce type d'erreur est très majoritairement aléatoire, et mesurable simplement. Il varie selon la technologie, les données constructeurs Illumina spécifient par exemple 0,1 % d'erreur par base séquencée. (Stoler et al. 2021) D'autres types de biais sont présents dans ces données brutes : présence d'adapteurs, bases de mauvaise qualité, séquences trop courtes, etc.

Ces biais peuvent être en partie corrigés lors de l'analyse bioinformatique, et c'est généralement la première étape d'une analyse bioinformatique réussie. Il est donc nécessaire de les détecter et d'en estimer l'impact sur la future analyse, puis de les corriger lorsque ceci est possible.

Il existe quelques logiciels capables de représenter la qualité des données : FASTQC (Andrews et al. 2010), MultiQC (Ewels et al. 2016).

Ces logiciels calculent certaines métriques permettant de représenter la qualité du jeu de données à analyser. Parmi elles, on peut observer la qualité moyenne des bases, la présence de séquences répétées, la longueur moyenne des séquences ou encore leur taux de duplication.

Il existe également plusieurs méthodes pour corriger ces différents biais, celles-ci sont implémentées à travers plusieurs logiciels : Trimmomatic (Bolger et al. 2014), fastp (Chen et al. 2018), CutAdapt (Martin et al. 2011), etc.

La correction de ces données reste une étape déterminante dans le processus d'analyse des données. En effet, les données comportant moins de biais induiront moins de faux-positifs et de faux-négatifs dans les résultats de l'analyse. Ceci implique évidemment de se séparer d'une partie des données, la partie la moins qualitative, et peut donc avoir un impact sur le résultat. Mais le gain en ter-

mes de qualité reste de très loin un avantage conséquent dans le bon déroulement de l'analyse et la pertinence des résultats de celle-ci.

Alignement. Il y a évidemment plusieurs façons d'étudier le génome humain, mais la plus répandue reste la comparaison du génome d'un individu à un génome de référence. Cette séquence de référence est stockée dans un autre format de donnée classiquement utilisé : le fasta. Celui-ci n'est pas spécifique du séquençage à haut-débit et permet donc de stocker des séquences, principalement des séquences de référence de génome, notamment celle du génome humain.

Ce génome de référence constitue une séquence consensuelle du génome humain, il ne représente pas un individu en particulier, mais plutôt la séquence moyenne du génome humain. En effet, les différences significatives entre 2 individus sont de l'ordre de plusieurs millions de bases, constituées principalement par le polymorphisme humain. Ainsi grâce à un assemblage de plusieurs individus témoins, un génome de référence a été constitué. Ce génome de référence a été établi lors d'un des plus grands projets que l'humanité ait porté : le Human Genome Project. Réalisé en une douzaine d'années pour un coût de plusieurs milliards de dollars, celui-ci a consisté en une collaboration internationale inédite de centaines de laboratoires ayant pour objectif d'identifier la séquence du génome humain à l'aide de la technologie disponible à cette époque : le séquençage de Sanger. La séquence a été publiée en 2001 (International Human Genome Sequencing Consortium et al. 2001), mais un grand nombre de « trous » ont été comblés dans les années suivantes pour aboutir à une première version stable et utilisable de la séquence de référence en 2003. Cette collaboration internationale a aussi été l'objet d'une forme de compétition par certaines compagnies privées souhaitant être les premières à découvrir la séquence du génome humain et de nouveaux gènes afin de breveter ceux-ci. En 1995, les responsables scientifiques du projet ont fait le choix d'une décision politique forte : le génome humain est déclaré public et patrimoine de l'humanité. Chaque nouvelle séquence doit être publiée très rapidement, c'est le principe des Bermudes, annihilant les possibilités de brevets et d'exploitation commercial de la séquence du génome humain.

Ce génome humain évolue car notre connaissance du génome croît à travers le temps. Ainsi la séquence de référence est détenue par un consortium : le GRC (Genome Reference Consortium) composé de l'EBI (*European Bioinformatics Institute*), du NCBI (*National center for Biotechnology Information*), du *Sanger Institute*, et de l'université de Washington à Saint-Louis. Ce consortium réalise un travail important pour toute la communauté scientifique de maintien des génomes de référence.

Actuellement, nous sommes dans la vingtième version de cette séquence, celle-ci est appelée GRCh38. La séquence de ce génome est librement accessible depuis un ensemble de portails web, n'importe quel individu peut le télécharger librement et gratuitement.

Ainsi cette comparaison va consister en un alignement des séquences sur ce génome de référence. Le but étant de déterminer, pour chacune des séquences spécifiques de l'individu étudié, à quelle localisation génomique cette séquence est reliée.

L'application principale du séquençage de génomes humains dans le cadre de la médecine consiste en l'identification de différences entre la séquence de l'individu étudié et la séquence de référence, puis en l'identification de potentielles différences qui pourraient expliquer un phénotype ou une pathologie génétique. Ainsi, par nature des différences existent entre ces 2 séquences, principalement reliées au polymorphisme humain. Il est donc nécessaire de permettre des différences ponctuelles lors de cette phase d'alignement, afin d'identifier ces différences appelées variations.

Il existe beaucoup de logiciels permettant de réaliser cette étape, parmi les plus connus figurent BWA (Li et al. 2009), Bowtie2 (Langmead et al. 2012)

Le format BAM (*Binary Alignement Map*) est probablement le format le plus connu dans le cadre de l'utilisation du séquençage à haut-débit. En effet, celui-ci va décrire l'alignement de la séquence (Li et al. 2009) et sera utilisé par tous les acteurs de la chaine d'analyse des données. Ce fichier est généralement utilisé pour la suite de l'analyse bioinformatique, mais aussi à des fins de visualisation.

La problématique de l'alignement est assez centrale dans l'analyse bioinformatique, elle est imparfaite car nécessite d'effectuer des calculs très complexes en des temps très courts. Ainsi des méthodes bioinformatiques d'alignement existaient bien avant l'avènement du séquençage haut-débit, utilisant des algorithmes particulièrement performants dans la précision de l'alignement. Mais ceux-ci restent trop lents pour aligner plusieurs centaines de millions de séquences et sont donc inadaptés face à la problématique du séquençage haut-débit. Ainsi de nouveaux algorithmes heuristiques ont été créés et sont utilisés, impliquant des biais dans les résultats d'alignement des séquences.

Ces biais sont connus et peuvent être en partie corrigés par des méthodes bioinformatiques complexes. L'un des principaux biais réside dans la présence de séquences dupliquées anormalement lors des différentes étapes de préparation de la librairie de séquençage. Par conséquent, certaines séquences peuvent être comptées plusieurs fois et donc influencer de façon néfaste la détection des variations. Il est donc nécessaire d'identifier ces séquences et de les marquer afin de considérer cette duplication artificielle dans la comptabilisation des séquences lors de la recherche de variations. (Li et al. 2009)

D'autres biais sont malheureusement présents dans les données après alignement, certaines séquences peuvent ne pas être alignées mais avoir un score d'alignement non nul, le score de qualité des bases peut se retrouver biaisé en fonction de la technologie utilisée (Ni et al. 2016), les sites d'insertions / délétions de bases sont souvent médiocrement alignés nécessitant un potentiel réalignement. (Tian et al. 2016) Ces biais sont corrigés par l'application d'algorithmes classiques consistant principalement à modéliser les biais statistiques observés, établir une méthode de correction basée sur ces modèles et à modifier les données en conséquence. Bien que ces corrections soient largement adoptées par la communauté scientifique, il subsiste certaines discussions quant à leur efficacité. Ainsi le choix de la méthode bioinformatique peut-être conditionné par une étude approfondie des arguments de la littérature, et en les confrontant à l'application recherchée lors de l'expérience de séquençage. Les méthodes de corrections à utiliser ne seront pas les mêmes si

l'utilisateur cherche à réaliser un assemblage de novo du génome étudié ou s'il s'attarde à séquencer un seul gène du génome humain.

Recherche de variations. La dernière partie de l'analyse bioinformatique consiste en la recherche effective des variations génétiques, de leur annotation et de l'identification de variations potentiellement impliquées en pathologie humaine. Cette partie de l'analyse est communément appelée « *variant calling* ».

Il existe de nombreux types de variations de la séquence du génome humain pouvant être impliquées dans une pathologie génétique. Parmi elles, on peut compter :

- les variations ponctuelles ou de petites tailles, impliquant généralement une base, ou plus rarement quelques bases.
- les variations de structure, impliquant des modifications de taille variable pouvant aller jusqu'à plusieurs milliers à des millions de bases.
- les évènements plus complexes et plus rares, tels que les répétitions de motifs, les insertions d'éléments mobiles, les variations somatiques ou en mosaïque, etc.

Chacun de ces types de variation nécessite une méthode de détection dédiée, et il existe une multitude de logiciels pouvant rechercher chacun de ces types :

- pour les variations ponctuelles : GATK Haplotype Caller (Poplin et al. 2017), FreeBayes (Garrison et al. 2012), DeepVariant (Poplin et al. 2017), Strelka (Saunders et al. 2012), Mutect (Cibilskis et al. 2013), etc.
- pour les variations de structure : Lumpy (Layer et al. 2014), Manta (Chen et al. 2016), ControlFreeC (Boeva et al. 2012), etc.
- pour les évènements plus complexes et rares : ExpansionHunter (Dolzhenko et al. 2017), HLAScan (Ka et al. 2017), Mobster (Thung et al. 2014), etc.

Aucun de ces logiciels n'est parfait, et leurs résultats dépendent très fortement des méthodes choisies pour traiter les données avant d'arriver à cette étape. Dans la grande majorité des cas, ces outils se veulent très permissifs afin de limiter au mieux le nombre de faux négatifs, et donc de variations existantes au sein du génome étudié, mais non détectées. Cela implique donc souvent un très grand nombre d'évènements détectés, incluant souvent de nombreux faux positifs. Ces faux-positifs sont largement éliminés en réalisant les différentes étapes d'analyse des données décrites précédemment. Il reste cependant des évènements de bonne qualité qui sont des faux positifs à l'issue de l'analyse. Ces variations pourront être filtrées via des méthodes plus manuelles à la fin de l'interprétation des données.

Les pathologies génétiques humaines sont généralement provoquées, dans le cas de maladies rares, par un très faible nombre d'évènements génétiques, en général une à deux variations, rarement plus. Cependant le séquençage d'un génome humain va rendre compte d'un très grand nombre de différences par rapport au génome de référence. Quelques millions de variations vont effectivement être détectées. Il faut donc parmi cette « botte de foin » de variations, rechercher « l'aiguille » qui va potentiellement expliquer le phénotype du patient dont le génome a été séquencé. Considérant en plus la présence non-négligeable de faux positifs, il est donc nécessaire d'établir des stratégies de sélection ou de filtre des variations détectées.

On peut ainsi considérer deux grandes catégories de filtres :

- les filtres sur critères techniques
- les filtres sur critères biologiques.

Les filtres sur critères techniques sont principalement dédiés à l'élimination des faux-positifs. Il consiste en la détermination de paramètres basés sur des métriques principalement bioinformatiques. Par exemple, il est possible d'éliminer les variations qui ne sont pas supportées par un nombre suffisant de séquences. Ces filtres, qui sont souvent spécifiques de la technologie et des méthodes utilisées, sont principalement choisis par le bioinformaticien assisté par un biologiste, ou un spécialiste de la technologie.

Les filtres sur critères biologiques sont eux dédiés à la sélection de variations d'intérêt en regard de l'application de séquençage choisie. L'analyse bioinformatique des données nécessite donc d'être complétée par des informations biologiques supplémentaires afin de réaliser ces filtres. Classiquement, ces annotations sont formées par la localisation de la variation (gène, transcrit, protéine, domaine, codon, etc.) mais aussi par les conséquences biologiques de la variation (création d'un codon stop prématuré, perte d'un codon initiateur de la transcription, etc.) ou encore par sa fréquence dans des bases de données de population générale, des scores de prédictions, etc. Cette annotation peut également être considérée comme une partie de la donnée génétique même si, ici, elle n'est ni personnelle, ni identifiante car elle concerne des informations générales sur la structure et l'organisation du génome humain (i.e. les coordonnées des gènes sur le génome humain, la fonction de ceux-ci, les différents transcrits d'un gène)

Sur la base de ces annotations, des filtres sont définis par le biologiste. Ils sont grandement dépendants du but initial du séquençage réalisé. Par exemple, dans le cas de maladies rares, les filtres cherchent à identifier des variations peu fréquentes en population générale, et potentiellement situées dans un gène d'intérêt médical.

Ces filtres sont souvent appliqués automatiquement par le bioinformaticien et permettent l'obtention d'une liste restreinte de variations, de l'ordre de quelques centaines, permettant ainsi de filtrer 99,9 % des variations.

Il reste parmi ces patients à déterminer les variations (1 à 2) qui expliquent précisément le phénotype du patient dont le génome a été séquencé. Ce travail repose très souvent sur une curation manuelle de la liste restreinte des variations par un généticien spécialiste qui va mettre en parallèle le phénotype précis du patient et la variation étudiée.

Si ces variations sont jugées pertinentes en regard du phénotype du patient, alors des vérifications vont être effectuées pour contrôler la réelle présence de la variation. Ces vérifications consistent en une visualisation des variations elles-mêmes à l'aide d'un outil bioinformatique spécialisé (IGV : Integrated Genomic Viewer). Celui-ci va permettre d'observer la variation et de déterminer si celle-ci semble vraisemblable. Enfin une variation sera dans la plupart des cas vérifiée in vitro par une technologie alter-

native de référence, principalement via un séquençage de Sanger.

Enfin, dans un cadre diagnostique les laboratoires de biologie médicale doivent répondre à la norme ISO15189. Celle-ci spécifie les exigences de qualité et de compétences qui sont applicables aux laboratoires de biologie médicale. Dans le cadre du séquençage à haut-débit pangénomique, il est nécessaire de démontrer la capacité du laboratoire à détecter des variations préalablement connues sur des échantillons contrôles via une étape de validation de méthode préalable à la mise en place de l'examen. Ceci permet de confirmer la compétence du laboratoire, de son personnel et de ses équipements, incluant la bioinformatique et l'interprétation des données.

Le choix de l'analyse, des logiciels et de leurs paramètres

L'analyse bioinformatique des données repose sur un ensemble de choix déterminants dans l'obtention des résultats et de leur pertinence. Ces choix sont cruciaux pour le succès de l'analyse et par conséquent de l'examen médical. Par exemple, changer un seul paramètre du logiciel d'alignement, en modifiant le nombre de « mismatch » autorisés lors de l'alignement pourrait conduire à la détection d'un nombre arbitrairement plus faible de variations, et donc potentiellement à la perte de beaucoup de vrais positifs parmi celles-ci.

Il est donc nécessaire de se reposer sur une expertise avérée conduite par une étude minutieuse de la littérature scientifique et de l'étude des problématiques à élucider. L'analyse de données de séquençage à haut-débit est réalisée depuis maintenant plusieurs années, et l'expérience accumulée lors des analyses passées est un atout incommensurable. Ainsi, plusieurs instituts mettent librement à disposition des recommandations basées sur leurs expériences préalables. C'est le cas, par exemple, du Broad Institute à Cambridge aux Etats-Unis. Cet institut est l'un des précurseurs de l'utilisation des technologies de séquençage à haut-débit dans le cadre de la génétique médicale. Ils mettent donc à disposition un guide des bonnes pratiques à mettre en œuvre pour la bonne conduite de ce genre d'analyse. (DePristo et al. 2011)

De nos jours, la plupart des pipelines d'analyse bioinformatique de données de séquençage à haut-débit sont très fortement inspirés de ces travaux, et de ce guide en particulier. Ces méthodes sont évidemment supportées par un très grand nombre de publications scientifiques attestant de leur pertinence et de leur efficacité.

Cependant ce guide ne reste qu'un ensemble de recommandations, et les laboratoires sont relativement libres de choisir leurs méthodes d'analyse bioinformatique. Dans le cadre de laboratoires d'analyse médicale, ceux-ci doivent néanmoins justifier d'une accréditation selon la norme NF ISO 15189, demandant entre autres de justifier les méthodes et les résultats obtenus selon des exigences réglementaires.

Ce genre de guide est aussi assez spécifique d'une problématique biologique à résoudre (i.e. trouver la cause d'une maladie génétique rare) et doit donc être adapté pour répondre très précisément à la problématique biologique spécifique de l'examen si celle-ci est différente. Cette adaptation repose généralement sur l'expertise d'un ou plusieurs bioinformaticiens de l'équipe. Ceux-ci sont chargés d'analyser les problématiques à résoudre, d'étudier la littérature scientifique reliée et de proposer une solution technique et scientifique. Cette solution peut aller du simple choix d'un logiciel d'analyse dédié, jusqu'à la création d'une nouvelle méthode spécialisée et de son implémentation à travers un nouveau logiciel.

Le choix d'un logiciel repose sur plusieurs concepts fondamentaux. Généralement, le logiciel choisi doit avoir démontré que ses résultats sont pertinents, souvent sous la forme d'une publication scientifique dans un journal à comité de lecture. Dans l'essentiel des cas, ces logiciels sont dits « open source », impliquant la mise en ligne du code source du logiciel et des algorithmes dédiés. Outre la transparence que cela apporte à l'utilisateur, l'open source permet de consulter le code, de vérifier si des erreurs ne sont pas présentes, de détecter tout code malicieux, et éventuellement de proposer des améliorations. Tous ces facteurs conduisent à une meilleure confiance en la méthode et le logiciel, une certaine dynamique d'évolution et une certaine indépendance vis à vis du créateur. Tous les logiciels utilisés dans le cadre de l'analyse de données génétiques sont évalués par les bioinformaticiens, souvent à l'aide de jeux de données de test dont les résultats sont publiquement connus et vérifiés. Il est très important pour le bioinformaticien de comprendre le fonctionnement des logiciels utilisés, ainsi que l'effet de chaque paramètre afin de déterminer l'ensemble des réglages optimaux pour obtenir les meilleurs résultats dans l'analyse bioinformatique.

Un autre point essentiel des méthodes d'analyse bioinformatique est la reproductibilité des données. Il est important dans le cas d'une analyse de génétique de pouvoir reproduire un résultat. Ceci implique donc que les algorithmes utilisés soient fiables et maintenus pour pouvoir répondre à des problématiques technologiques qui évoluent fortement. En effet, en environ 15 ans d'existence de la technologie de séquençage à haut-débit de deuxième génération, pas moins d'une trentaine de séquenceurs différents ont pu être mis sur le marché, retirés ou mis à jour. L'évolution technique est très importante et très rapide, et le débit est toujours plus important. Ces vérifications de reproductibilité sont également vérifiées lors de la validation de méthode nécessaire à la mise en place de l'accréditation ISO15189 du laboratoire, incluant des répliquats ainsi que des échantillons dont les variations sont connues.

Les contraintes de temps sont elles aussi de plus en plus exigeantes, le délai de rendu d'un examen devant être le plus court possible. Ainsi les méthodes utilisées doivent être toujours plus efficaces tout en conservant une précision élevée. Il semble donc évident que l'évolution des méthodes, algorithmes et logiciels bioinformatiques doive suivre ces contraintes impliquant de nouveaux matériels informatiques spécialisés (GPU, FPGA, etc.) et de nouvelles méthodes innovantes (Machine learning, Deep learning, etc). Bien que les temps d'analyse dépendent de beaucoup de facteurs (Type et quantité de processeurs, nombre de cœurs[3], fréquence des processeurs, quantité de mémoire vive, fréquence de la mémoire vive, espace disque disponible, performance des disques durs, etc.), l'emploi par exemple d'une méthode spécialisée par GPU permet de diviser le temps d'analyse

[3] Un cœur (core) de processeur est une sous-unité fonctionnelle du processeur capable d'exécuter des instructions de manière autonome. Un processeur est constitué de un à plusieurs cœurs indépendants, permettant d'exécuter parallèlement des taches informatiques, impliquant une potentielle diminution du temps de calcul et d'exécution.

par un facteur allant de 5 à 20 fois. Dans notre expérience, nous avons eu la chance de tester des GPU Tesla V100 (©Nvidia), nous avons pu constater la diminution du temps d'analyse d'un génome de patient séquencé à une profondeur moyenne de 30x de 4 jours de calcul sur notre matériel de calcul habituel, à moins de 2 heures. Cependant, ces matériels restent coûteux, difficiles à mettre en œuvre et nécessitent un développement important pour être utilisés dans de bonnes conditions.

Le phénotype du patient

La donnée génétique n'est pas seulement constituée de la séquence du génome du patient étudié. Elle est évidemment multiple et comprend une donnée essentielle pour le diagnostic génétique : le phénotype détaillé du patient.

En effet, le phénotype du patient est un aspect essentiel pour permettre l'interprétation finale des données. Il est quasiment impossible de déterminer la cause génétique d'une pathologie humaine si le phénotype du patient n'est pas décrit de manière précise et pertinente. C'est en cette description que réside en partie l'expertise des médecins prescripteurs qui demeure essentielle dans l'examen génétique.

Plus le phénotype est précis et détaillé, plus la sélection et l'interprétation des variations sera pertinente à l'issue de l'analyse bioinformatique.

Cependant, une difficulté importante demeure dans l'agrégation et la description de ce type de données. La richesse de la langue française permet un large éventail de termes différents permettant de décrire une même idée, cela ne facilite donc pas la comparaison de phénotypes et la normalisation de leurs descriptions.

Ainsi des initiatives intéressantes existent, créées par la communauté scientifique dans le but de normaliser la description d'un phénotype humain. Parmi elles, on peut citer HPO (Human Phenotype Ontology) qui est une ontologie standardisée des anomalies phénotypiques chez l'humain. Chaque terme de la base décrit une anomalie phénotypique, avec un niveau de granularité croissant. HPO décrit environ 13 000 termes et 156 000 annotations relatives aux pathologies génétiques. HPO fait partie intégrante d'une initiative internationale : GA4GH (Global Alliance for Genomics and Health) visant à définir et à mettre en œuvre les standards techniques et organisationnels de la génomique médicale dans le monde.

L'interdisciplinarité que nécessite la conduite d'une analyse complète d'une expérience de séquençage à haut-débit de génome humain est une condition importante de la bonne réussite de cette analyse. En effet, les trois corps de métier (médecins prescripteurs, Biologistes et Bioinformaticiens) nécessaires à celle-ci se doivent de travailler à l'unisson pour réussir cette analyse. Par exemple, les paramètres de filtre du pipeline bioinformatique sont établis par les bioinformaticiens, mais selon les paramètres demandés par les biologistes et les généticiens. Ceux-ci définissent ces paramètres et les bioinformaticiens les implémentent et les testent, puis les adaptent à nouveaux en fonction du retour des biologistes et des généticiens. Ce cycle de mise au point est indispensable à chaque nouvelle implémentation d'une analyse.

De la même façon, les biologistes ont besoin de la description clinique fournie par les généticiens pour interpréter de manière efficace les données. Celles-ci seront également présentées dans une réunion de concertation pluri-disciplinaire permettant de solliciter l'expertise clinique des généticiens à propos de la pertinence des variations sélectionnées par les biologistes en regard du phénotype du patient.

Cette interdisciplinarité reste essentielle à la conduite de l'analyse. Des logiciels commencent à voir le jour cherchant à simuler ces expertises combinées et d'automatiser les interprétations. Mais à l'heure actuelle les résultats proposés par ces solutions informatiques sont loin d'être aussi pertinents que les l'interprétation réalisée par un ou plusieurs humains.

Vie de la donnée

La donnée génétique est donc multiple et complexe, constituée à la fois de la séquence du génome humain et des informations qui en découlent, d'informations personnelles et de la description du phénotype du patient. Cette donnée est « vivante », elle évolue, d'abord collectée par le généticien, puis sous la forme de matériel biologique, elle est transformée grâce aux avancées technologiques humaines en une donnée informatique qui subira diverses étapes avant d'être réduite à la cause possible de la pathologie du patient, donnée finale qui lui sera finalement remise à la fin de l'examen.

Au cours de ce cycle, nous avons vu que la donnée prend plusieurs formes et qu'elle évolue dépendamment de son état d'analyse et des besoins des acteurs de l'examen médical. Un aspect important réside dans le stockage des données de séquences issues des séquenceurs haut-débit. Ce stockage nécessite d'être volumineux, performant et sécurisé. En effet le volume représenté par les données brutes issues du séquençage du génome d'un patient (à 30x) représente environ 200 Go. La totalité des données incluant les données brutes, les données alignées, les variations peut représenter jusqu'à 500 Go par patient. Cette volumétrie couplée à une production à l'échelle nationale peut rapidement entraîner des besoins importants de matériel informatique. (i.e. pour 20000 génomes humain par an : 10 Po de données) Ces données sont également précieuses, car elles nécessitent afin d'être produites une mise à disposition importante de matériel, technologies et personnels qualifiés. Il est donc très important d'assurer la fiabilité du stockage utilisé pour accueillir et analyser les données. Celui-ci doit aussi être performant afin d'assurer des temps d'analyse raisonnables et compatibles avec un examen médical. L'évolution des matériels en termes de performance et de fiabilité est donc un facteur très important de la bonne conduite des analyses, et du maintien en vie de la donnée. Les innovations et la recherche concernant le matériel informatique sont importantes, cependant il faut garder en tête que la fiabilité reste prioritaire par rapport à la rapidité et l'innovation, ainsi les nouvelles technologies prometteuses (ordinateurs basés sur l'information quantique, l'intelligence artificielle, les blockchains, etc.) nécessitent d'être mises au point et stabilisées avant de pouvoir assurer une partie de l'analyse de ces données. Enfin, la sécurité de ces données se doit également d'être assurée car d'une part, il s'agit de données personnelles et donc soumises à réglementation (RGPD), et d'autre part il s'agit également de données médicales qui sont soumises à une réglementation stricte en ce qui concerne leur accès et leur conservation. La démocratisation des tests génétiques peut conduire à un risque important de

vol ou d'utilisation frauduleuse des données, incluant des données présentes dans des bases de données publiques. Des démonstrations de désanonymisation de données publiques prétendues anonymisées ont déjà été réalisées (Gymrek et al. 2013) et mettent en évidence les faiblesses de la sécurité informatique et la capacité identifiante des données de génétique.

La compétence du bioinformaticien reste une condition essentielle de la réussite de l'analyse. En France, les bioinformaticiens restent un corps de métier assez rare. Même si beaucoup d'universités proposent des formations diplômantes (Licence, Master, DU), le profil type d'un bioinformaticien reste rare parmi les étudiants formés soit à la biologie et qui vont vers l'informatique, soit formés à l'informatique et qui vont vers la biologie. Ainsi de nombreux laboratoires, plateformes et autres structures réalisant des analyses bioinformatiques ont du mal à combler les postes vacants. Même si la concentration et l'automatisation des taches constitueront probablement l'avenir de la bioinformatique, ceci reste un danger pour la bonne conduite des analyses bioinformatiques des laboratoires de biologie médicale.

Enfin, malgré tous les efforts mis en balance dans les différents compartiments de la réalisation d'un examen génétique, il reste cependant une proportion non négligeable des cas où la cause de la pathologie peut ne pas être déterminée. Dans ce cas, et en fonction de l'analyse réalisée, il est nécessaire de réfléchir à l'objectif de l'expérience de séquençage réalisée. Changer la stratégie d'analyse en passant par exemple à une structure familiale plus complète (i.e. trio) peut faciliter l'identification de nouveaux candidats en permettant une vérification directe de la ségrégation familiale de la variation, ou en identifiant des variations de novo qui ne seraient donc pas héritées des parents. Agréger des cas similaires non-reliés peut également permettre l'identification de nouveaux gènes ou variations d'intérêt encore non-associés à une pathologie humaine. (Bamshad et al. 2011) Dans ce cas, des systèmes de partage de données efficaces existent et sont efficaces afin de constituer des cohortes de patients ayant des phénotypes similaires. (Bruel et al. 2019)

La remise en cause de la méthode d'analyse doit également être effectuée. Ainsi les méthodes utilisées, la qualité de l'expérience de séquençage, la qualité du phénotypage, les annotations apportées à l'analyse et les filtres doivent être en adéquation avec la problématique adressée.

De nouvelles techniques innovantes peuvent également être utilisées pour apporter plus d'informations à l'expérience globale d'étude du génome d'un patient, c'est le cas par exemple de l'approche RNASeq. Cette technologie permet de d'identifier et de quantifier les différents transcrits dans la cellule, conduisant à la production de nouvelles informations, telles que l'expression différentielle de certains gènes, l'identification de transcrits alternatifs, de nouveaux gènes de fusion ou encore de l'expression d'allèles spécifiques. (Cappuccio et al. 2020)

L'utilisation d'autres technologies peut également contribuer à l'amélioration globale des connaissances du génome humain et à la résolution des cas les plus complexes. Ces technologies sont variées : séquençage de troisième génération, HiC, analyse du profil de méthylation, cartographie optique, etc. L'objectif lointain du déploiement de toutes ces analyses reste l'intégration de ces données. L'intégration de données hétérogènes permet en combinant plusieurs approches d'extraire plus d'informations que la somme des informations produites individuellement par ces approches.

Malgré tout ceci, il faut également considérer que notre connaissance du génome humain reste imparfaite et que celle-ci évolue rapidement avec le temps. L'avantage d'un séquençage pan-génomique réside également dans la possibilité d'une réanalyse ultérieure en regard de l'avancée des connaissances (Nambot et al. 2017). La conservation et l'agrégation de celles-ci sera donc un futur enjeu majeur de l'avancée des connaissances et de la maitrise des données génétiques.

La France a choisi de s'engager vers un programme de médecine génomique en mettant en place un grand plan : le Plan France Médecine Génomique 2025 : PFMG 2025.

Celui-ci vise la mise en place de plateformes de séquençage à haut-débit de génomes humains dans la cadre des maladies génétiques rares et du cancer. Ainsi deux plateformes ont été créées (AURAGEN, SeqOIA) et ont pour objectifs de développer ces approches dans un cadre diagnostique d'ici 2025. Les acteurs de ce plan devront certainement prendre en compte tous ces aspects afin de mener à bien les objectifs de cet ambitieux projet.

Références

Andrews S. (2010) A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011 Sep 27;12(11):745-55. doi: 10.1038/nrg3031. PMID: 21946919.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill

MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008 Nov 6;456(7218):53-9. doi: 10.1038/nature07517. PMID: 18987734; PMCID: PMC2581791.

Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012 Feb 1;28(3):423-5. doi: 10.1093/bioinformatics/btr670. Epub 2011 Dec 6. PMID: 22155870; PMCID: PMC3268243.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. Epub 2014 Apr 1. PMID: 24695404; PMCID: PMC4103590.

Bruel AL, Vitobello A, Mau-Them FT, Nambot S, Duffourd Y, Quéré V, Kuentz P, Garret P, Thevenon J, Moutton S, Lehalle D, Jean-Marçais N; Orphanomix Physicians' Group, Garde A, Delanne J, Lefebvre M, Lecoquierre F, Trost D, Cho M, Begtrup A, Telegrafi A, Vabres P, Mosca-Boidron AL, Callier P, Philippe C, Faivre L, Thauvin-Robinet C. 2.5 years' experience of GeneMatcher data-sharing: a powerful tool for identifying new genes responsible for rare diseases. Genet Med. 2019 Jul;21(7):1657-1661. doi: 10.1038/s41436-018-0383-z. Epub 2018 Dec 19. PMID: 30563986.

Cappuccio G, Sayou C, Tanno PL, Tisserant E, Bruel AL, Kennani SE, Sá J, Low KJ, Dias C, Havlovicová M, Hančárová M, Eichler EE, Devillard F, Moutton S, Van-Gils J, Dubourg C, Odent S, Gerard B, Piton A, Yamamoto T, Okamoto N, Firth H, Metcalfe K, Moh A, Chapman KA, Aref-Eshghi E, Kerkhof J, Torella A, Nigro V, Perrin L, Piard J, Le Guyader G, Jouan T, Thauvin-Robinet C, Duffourd Y, George-Abraham JK, Buchanan CA, Williams D, Kini U, Wilson K; Telethon Undiagnosed Diseases Program, Sousa SB, Hennekam RCM, Sadikovic B, Thevenon J, Govin J, Vitobello A, Brunetti-Pierri N. De novo SMARCA2 variants clustered outside the helicase domain cause a new recognizable syndrome with intellectual disability and blepharophimosis distinct from Nicolaides-Baraitser syndrome. Genet Med. 2020 Nov;22(11):1838-1850. doi: 10.1038/s41436-020-0898-y. Epub 2020 Jul 22. PMID: 32694869.

Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i890. doi:10.1093/bioinformatics/bty560

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016 Apr 15;32(8):1220-2. doi: 10.1093/bioinformatics/btv710. Epub 2015 Dec 8. PMID: 26647377.

Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Mar;31(3):213-9. doi: 10.1038/nbt.2514. Epub 2013 Feb 10. PMID: 23396013; PMCID: PMC3833702.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491-8. doi: 10.1038/ng.806. Epub 2011 Apr 10. PMID: 21478889; PMCID: PMC3083463.

Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, van Blitterswijk M, Narzisi G, Ajay SS, Rajan V, Lajoie BR, Johnson NH, Kingsbury Z, Humphray SJ, Schellevis RD, Brands WJ, Baker M, Rademakers R, Kooyman M, Tazelaar GHP, van Es MA, McLaughlin R, Sproviero W, Shatunov A, Jones A, Al Khleifat A, Pittman A, Morgan S, Hardiman O, Al-Chalabi A, Shaw C, Smith B, Neo EJ, Morrison K, Shaw PJ, Reeves C, Winterkorn L, Wexler NS; US–Venezuela Collaborative Research Group, Housman DE, Ng CW, Li AL, Taft RJ, van den Berg LH, Bentley DR, Veldink JH, Eberle MA. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. 2017 Nov;27(11):1895-1903. doi: 10.1101/gr.225672.117. Epub 2017 Sep 8. PMID: 28887402; PMCID: PMC5668946.

Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1;32(19):3047-8. doi: 10.1093/bioinformatics/btw354. Epub 2016 Jun 16. PMID: 27312411; PMCID: PMC5039924.

Garrison, E., Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907.

Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016 May 17;17(6):333-51. doi: 10.1038/nrg.2016.49. PMID: 27184599.

Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013 Jan 18;339(6117):321-4. doi: 10.1126/science.1229566. PMID: 23329047.

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001). https://doi.org/10.1038/35057062

Ka S, Lee S, Hong J, Cho Y, Sung J, Kim HN, Kim HL, Jung J. HLAscan: genotyping of the HLA region using next-generation sequencing data. BMC Bioinformatics. 2017 May 12;18(1):258. doi: 10.1186/s12859-017-1671-3. PMID: 28499414; PMCID: PMC5427585.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357-9. doi: 10.1038/nmeth.1923. PMID: 22388286; PMCID: PMC3322381.

Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014 Jun 26;15(6):R84. doi: 10.1186/gb-

2014-15-6-r84. PMID: 24970577; PMCID: PMC4197822.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60. doi: 10.1093/bioinformatics/btp324. Epub 2009 May 18. PMID: 19451168; PMCID: PMC2705234.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8. PMID: 19505943; PMCID: PMC2723002.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005 Sep 15;437(7057):376-80. doi: 10.1038/nature03959. Epub 2005 Jul 31. Erratum in: Nature. 2006 May 4;441(7089):120. Ho, Chun He [corrected to Ho, Chun Heen]. PMID: 16056220; PMCID: PMC1464427.s

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1), pp. 10-12. doi:https://doi.org/10.14806/ej.17.1.200

Nambot S, Thevenon J, Kuentz P, Duffourd Y, Tisserant E, Bruel AL, Mosca-Boidron AL, Masurel-Paulet A, Lehalle D, Jean-Marçais N, Lefebvre M, Vabres P, El Chehadeh-Djebbar S, Philippe C, Tran Mau-Them F, St-Onge J, Jouan T, Chevarin M, Poé C, Carmignac V, Vitobello A, Callier P, Rivière JB, Faivre L, Thauvin-Robinet C; Orphanomix Physicians' Group. Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis. Genet Med. 2018 Jun;20(6):645-654. doi: 10.1038/gim.2017.162. Epub 2017 Nov 2. PMID: 29095811.

Ni S, Stoneking M. Improvement in detection of minor alleles in next generation sequencing by base quality recalibration. BMC Genomics. 2016 Feb 27;17:139. doi: 10.1186/s12864-016-2463-2. PMID: 26920804; PMCID: PMC4769523.

Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, Gross SS, Dorfman L, McLean CY, DePristo MA. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018 Nov;36(10):983-987. doi: 10.1038/nbt.4235. Epub 2018 Sep 24. PMID: 30247488.

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., ... Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. BioRxiv, 201178.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463-7. doi: 10.1073/pnas.74.12.5463. PMID: 271968; PMCID: PMC431765.

Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012 Jul 15;28(14):1811-7. doi: 10.1093/bioinformatics/bts271. Epub 2012 May 10. PMID: 22581179.

Stoler N, Nekrutenko A, Sequencing error profiles of Illumina sequencing instruments, NAR Genomics and Bioinformatics, Volume 3, Issue 1, March 2021, lqab019.

Thung DT, de Ligt J, Vissers LE, Steehouwer M, Kroon M, de Vries P, Slagboom EP, Ye K, Veltman JA, Hehir-Kwa JY. Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biol. 2014;15(10):488. doi: 10.1186/s13059-014-0488-x. PMID: 25348035; PMCID: PMC4228151.

Tian S, Yan H, Kalmbach M, Slager SL. Impact of post-alignment processing in variant discovery from whole exome data. BMC Bioinformatics. 2016 Oct 3;17(1):403. doi: 10.1186/s12859-016-1279-z. PMID: 27716037; PMCID: PMC5048557.