DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : 1er juillet 2019

Nom de famille et prénom de l'auteur : DELHOMME Tiffany

Titre de la thèse : « *Utiliser la nature systématique des erreurs dans les données NGS pour détecter efficacement les mutations : méthodes de calcul et application à la détection précoce du cancer* ».

Résumé

Le cancer est une des causes majeures de décès à travers le monde, avec plus de 10 millions de morts au cours de l'année 2018, ce qui correspond à environ un décès sur 6 étant dû au cancer. Le cancer peut être défini comme une maladie génétique complexe, les cellules cancéreuses pouvant être qualifiées d'"épaves" génétiques, portant dans leur génome les abberations génétiques étant traces de l'évolution d'une cellule normale vers une cellule anormale, se divisant plus rapidement, plus efficacement, sans relâche, et ayant perdu sa capacité biologique, signe d'une tumeur. Le génome, quelle que soit la nature de la cellule, correspond à l'ensemble de son matériel génétique. Le génome d'une cellule est codé dans son ADN (Acide DésoxyriboNucléique), sous forme d'un enchaînement de nucléotides, aussi appelés bases. Cet ADN peut être séquencé, afin d'en déterminer la séquence, ou l'enchaînement des bases, et être comparé à un génome de référence, afin d'en identifier les variations. La première séquence complète du génome humain a été proposée en 2001 par un gigantesque effort international, le Human Genome Project, et actuellement, c'est le Genome Reference Consortium qui définit le génome humain de référence, qui est basé sur l'ADN d'un individu particulier, faisant naître l'idée de la construction d'un génome de référence plus représentatif de l'ensemble des individus. Les variations de l'ADN peuvent être séparées en deux groupes, selon leur source. Les mutations constitutionnelles, caractérisent les mutations issues de la lignée germinale des cellules, c'est à dire les mutations présentes dans les cellules germinales ayant données naissance au zygote, et donc présentes dans toutes les cellules de l'individu. D'un autre côté, les mutations acquises au cours de la vie d'un individu dans une des cellules constituantes du soma sont appelées mutations somatiques, et ne sont présentes dans toutes les cellules d'un individu. La caractérisation exhaustive et précise des variations de l'ADN, ou mutations, peut aider à progresser dans de nombreux champs liés à la génomique du cancer. Le séquençage nouvelle génération, (NGS en anglais pour Next Generation Sequencing) est actuellement la technique la plus efficace pour déterminer une séquence ADN, dû à la réduction des coûts et durées des expériences comparées à la méthode de séquençage traditionnelle de Sanger. Malgré de tels avantages, le plus grand inconvénient des NGS est leur capacité à être enclin aux erreurs de séquençage, qui rendent les données potentiellement difficilement exploitables. La détection de mutations à partir de données NGS reste donc encore un problème difficile, en particulier pour les mutations somatiques présentes en très faible abondance, comme lorsque l'on essaye d'identifier des mutations sous-clonales d'une tumeur, des mutations somatiques dans des tissus normaux, des mutations dérivées de la tumeur dans l'ADN circulant libre, dues à des proportions de mutations potentiellement au même niveau que les erreurs de séquençage, les rendant difficilement distinguables. Par exemple, pour le cas de l'ADN circulant libre, la présence de mutations dérivées de la tumeur est due à la dégradation des cellules et à la libération de leur ADN, et donc la proportion attendue de ces mutations est très faible étant donnée la différence de proportion entre cellule tumorales et cellules normales.

Dans cette thèse, nous nous sommes intéressés à la nature systématique des erreurs dans les données NGS afin d'identifier de manière efficace les mutations. Nous avons développé des méthodes de calcul, et nous présentons diverses applications pour la détection précoce du cancer.

Dans une première partie, nous avons développé un nouvel algorithme d'appel de variant, dans le but d'être capable d'identifier les mutations présentes en faible abondance. Pour cela, nous modélisons,

pour chaque position génomique et chaque potentiel changement de base, (variations nucléotidiques, insertions et délétions) les erreurs systématiques au sein d'échantillons multiples en utilisant une régression négative binomiale robuste, que nous avons adaptée à nos données. Puis, nous utilisons les paramètres estimés du modèle d'erreurs pour calculer, pour chaque échantillon, la probabilité de faire partie du modèle d'erreur (p-value), que nous transformons en utilisant la correction de Benjamini-Hochberg pour les tests multiples (nous obtenons une q-value). Nous rapportons finalement une q-value en échelle Phred, la statistique QVAL=-10*log10(q-value). Cette statistique QVAL est ensuite confrontée à un seuil d'acceptation s, et seuls les échantillons vérifiant QVAL>s sont considérés comme mutés pour l'altération. Nous avons implémenté efficacement notre algorithme en utilisant le langage dédié nextflow, qui permet à la fois le déployement de pipelines sur plusieurs types d'environnements (local, sur un cluster ou sur le cloud) de manière transparente pour l'utilisateur, la reproductibilité des analyses en utlisant le système de conteneurs (tel que Docker) et une interaction avec GitHub, une plateforme en ligne permettant de stocker du code informatique sous forme de versions. Nous avons aussi validé notre méthode d'appel de variants sur des données à la fois simulées et issues de séquençage. Premièrement, nous avons validé notre méthode sur des données simulées, en utilisant les données de séquençage du gène *TP53* à partir d'ADN circulant libre provenant de 125 échantillons de plasma d'individus sains, en introduisant des mutations aléatoirement à l'aide de BAMsurgeon. Cet outil permet de précisément connaître les mutations qu'on cherche à détecter tout en gardant le bruit de fond des données réelles. Nous avons montré que la sensibilité de notre méthode dépend du taux d'erreur, et que plus la différence entre la fraction allélique d'une mutation et le taux d'erreur pour cette mutation est grande, plus la sensibilité est haute. Nous avons aussi rapporté une excellente sensibilité globale de needlestack, elle est par exemple supérieure à 99% pour les mutations ayant une fraction allélique d'au moins 1%. Nous avons aussi comparé notre méthode à la méthode shearwaterML, développé au Sanger Institute, et nous avons montré que le taux de faux positifs de needlestack ne dépendais pas du taux d'erreur, contrairement à shearwaterML. Nous avons aussi validé notre méthode sur des données réelles. Nous avons pour cela détecter les mutations du gène *TP53* à partir du séquençage de l'ADN circulant libre (cfDNA) issu des échantillons de plasma de 35 individus atteints d'un cancer du poumon, puis nous avons essayé de valider ces mutations dans la tumeur. En effet, il est attendu que les mutations du cfDNA d'un individu, notamment les mutations délétères, sont des mutations dérivées de la tumeur de ce même individu. Nous avons montré que les 12 mutations délétères que nous avons détectées avec needlestack dans le cfDNA de ces individus ont toutes été validées dans la tumeur. Finalement, nous avons aussi estimé la performance de needlestack pour des mutations constitutionnelles à partir de 62 échantillons de sang, et nous avons comparé ces résultats à la méthode GATK, une des plus utilisées actuellement sur ce type de données. Nous avons utilisé une puce à ADN, disponible pour 33 individus, en tant que «référence». Nous avons montré une excellente concordance entre needlestack, GATK, et la puce ADN (>95%) et aussi une excellente concordance entre needlestack et GATK sans tenir compte de la puce, qui peut ne pas être représentative de l'ensemble des mutations (97% de concordance pour les mutations d'un nucléotide et 70% pour les insertions et délétions.).

Dans une deuxième partie, nous avons développé des méthodes de filtrage de variants, dans le but réduire les potentielles fausses mutations détectées par l'appel de variant. Nous avons développé deux méthodologies, la première basée sur des résumés statistiques et la seconde sur de l'apprentissage automatique. La première méthode a été développée pour des données présentant des fortes couvertures de séquençage, particulièrement pour des mutations en faible abondance, et utilise cinq résumés statistiques pour identifier les fausses découvertes : (i) le nombre de mutation spar librairie, qui, s'il est anormalement grand, peut indiquer un problème de séquençage, (ii) la validation dans une librairie indépendante du même échantillon, afin d'identifier les erreurs de PCR qui ne sont pas répliquées, (iii) le biais de brin, (iv) la distance génomique d'un vrai variant, (iv) la probabilité que l'altération soit de faible confiance en terme de séquençage. Cette dernière a demandé un développement particulier, avec l'introduction d'une nouvelle statistique, que nous avons nommée LCAP pour *Low Confidence Alteration Probability*. Cette statistique est décrit comme suit :

$$LCAP = \sum_{p=p_{obs}}^{p_{max}} \left[\prod_{i=0}^{p-1} \binom{k-2i}{2} \left[\prod_{j=0}^{k-p-1} (2N-2j) \right] \frac{1}{p!\,k! \binom{2N}{k}} \right]$$

Elle correspond à la probabilité de former au moins *p* paires en tirant aléatoirement *k* entités dans une urne contenant 2N entités. En effet, nous avons séquencés chaque échantillon deux fois indépendamment, donc nous nous attendons à retrouver les vraies mutations dans les deux librairies. Cette statistique nous permet d'identifier les altérations pour lesquelles la probabilité d'observer la mutation dans les deux librairies peut

être dû à un évènement aléatoire (LCAP<0.05). Nous montrons une application de cette méthode à des données de cfDNA afin de développer un biomarqueur du cancer du poumon à petites cellules, un des cancers présentant le plus faible taux de survie après diagnostique (application qui est aussi discutée dans une troisième partie de la thèse). La deuxième méthode de filtrage des variants a été développée pour des mutations constitutionnelles, et est basée sur de l'apprentissage automatique. L'apprentissage automatique permet d'«apprendre» les caractéristiques de données connues (dans notre cas les caractéristiques de vraies et fausses mutations) afin de prédire le statut de données inconnues. Notre méthode se base sur l'algorithme du «random forest», dont l'idée est de construire des arbres de décisions d'après les données connues et d'utiliser les prédictions de ces arbres sur des données inconnues pour en prédire le statut. En tant que données connues, référentes, nous avons utilisé environ 11000 mutations détectées à partir du séquençage de 86 gènes chez 55 individus, correspondant à environ 8000 vraies mutations, 2000 fausses mutations et 1000 mutations non classées (nous avons utliser une validation avec une technologie indépendante pour identifier leur statut). Nous avons entraîné notre algorithme sur ces données et ensuite nous l'avons appliqué aux mutations détectées chez environ 1500 individus additionnels. En utilisant une validation croisée, nous avons pu estimer la performance de notre méthode d'apprentissage automatique et la comparer à celle de filtres qui utilisent des seuils de statistiques et qui sont traditionnellement utilisés. Nous avons rapporté une augmentation de sensibilité de 4% avec un taux de fausses découvertes fixé à 2%, et une baisse du taux de fausse découverte de 1.3% avec une sensibilité fixée à 95%.

Dans une troisième et dernière partie, nous présentons quatre applications des méthodes présentées dans les deux premières parties, dans le but de proposer une détection efficace des mutations dans le cfDNA issu d'échantillons de patients atteints d'un cancer, afin d'estimer la possibilité d'utiliser le cfDNA comme un biomarqueur du cancer, et potentiellement comme un biomarqueur du cancer en stade précoce. La première étude propose d'utiliser les mutations du gène *KRAS* à partir d'échantillons de sang pour identifier des cas de cancer du pancréas. La seconde étude décrit UroMuTERT, un test basé sur la détection des mutations au niveau du promoteur du gène *TERT*, afin d'identifier des patients atteints d'un cancer urothélial (cancer de la vessie), à partir d'échantillon de sang et d'urine. La troisième étude s'intéresse à la détection des mutations du gène *TP53* dans le cfDNA de patients atteints d'un cancer du poumon à petites cellules, afin de développer un biomarqueur de détection précoce de ce type de cancer. Bien que le sensibilité de ce biomarqueur était acceptable (environ 50%), la spécificité étant supérieure à celle attendue pour un biomarqueur du cancer (généralement faible due à une faible prévalence dans la population génerale) : environ 11% des individus non atteints de ce cancer présentaient une mutation dans le gène *TP53*. Nous nous sommes donc intéressés dans une quatrième étude à l'amélioration de ce biomarqueur (étude présentée aussi dans la deuxième partie de la thèse), en combinant les mutations du gène *TP53* et du gène *RB1*, les deux gènes les plus souvent mutés dans ce type de cancer, et en utilisant des méthodes de filtrage de variants efficaces ainsi qu'en développement un score génétique (publication en préparation).

Finalement, nous discutons des potentielles limites de nos approches, comme l'alignement et le génome de référence, dont les variations n'ont pas été testées ici. Nous discutons aussi du problème de la validation des mutations, qui reste une opération nécessaire mais difficile à obtenir de manière exhaustive. Nous proposons aussi diverses applications non présentées dans la thèse, comme la détection des mutations somatiques dans les tissus normaux ou la détection précise des mutations pour étudier l'hétérogéneité tumorale.