Périodique de formation

CORRESPONDANCES EN Onco-Hématologie

RACONTÉ À JULIETTE

ADN, WES & co

Pr Marie-Christine Béné *(Nantes)*

DOSSIER

Exploration du génome dans les hémopathies malignes

Coordonné par le Dr Thierry Leblanc (Paris)

… tout le sommaire ➔

EDIMARK PRESSE ÉDITION MÉDIA

DPC

Société éditrice : EDIMARK SAS
CPPAP : 0124 T 88680 - ISSN : 1954-4820
Bimestriel
Prix du numéro : 46 €

Vol. XVII - n° 4 - juillet-août 2022

Revue indexée dans la base internationale ICMJE

sommaire

À découvrir
Application mobile Edimark !

Correspondances en Onco-Hématologie

Éditeur : Edimark SAS
Siège social : 44, rue de Prony, CS 10107, 75017 Paris
Société détenue à 100 % par la SAS PHILI@MEDICAL ÉDITIONS
Représentant légal et Directeur des publications : Claudie Damour-Terrasson

Président et Directeur de la publication
Claudie Damour-Terrasson

Rédacteur en chef
Noël Milpied

Rédacteurs en chef adjoints
Marie-Christine Béné et Sylvain Choquet

Comité de rédaction
Pierre Feugier, Luc-Matthieu Fornecker, Romain Guièze, Thierry Leblanc, Laurence Legros, Xavier Leleu, Aurore Perrot, Emmanuel Raffoux, Cédric Rossi

Comité scientifique
L. Adès, A. Baruchel, D. Bordessoule, P. Colombat, P. Cornillet-Lefebvre, F. Cymbalista, E. Deconinck, A. Delmer, H. Dombret, J. Feuillard, J. Gabert, C. Haioun, R. Herbrecht, O. Hermine, M. Hunault-Berger, N. Ifrah, C. Lacombe, M. Lafage-Pochitaloff, F.-X. Mahon, J.-P. Marolleau, A. Martin, M. Michallet, J.-F. Rossi, G. Socié, E. Solary, L. Sutton, X. Troussard, A. Turhan

Comité de lecture
H. Avet-Loiseau, J.-N. Bastié, J.-O. Bay, K. Bouabdallah, D. Bouscary, G. Cartron, G. Damaj, V. Delwail, C. Faucher, T. Fest, M. Fontenay, L. Fouillard, R. Itzykson, A. Jaccard, F. Jardin, T. Lamy, V. Lévy, B. Lioure, K. Maloum, M. Maynadié, M. Mohty, P. Moreau, F. Morschhauser, N. Mounier, S. Le Gouill, C. Pautas, A. Pigneux, S. Raynaud, C. Récher, M. Renaud, F. Rigal-Huguet, P. Soubeyran, O. Tournilhac, V. Ugo, N. Vey, M.-C. Woronoff-Lemsi

Fondateur : Claudie Damour-Terrasson

Fondateur scientifique : Noël Milpied

Rédaction - Infographie - Média
Directeur des rédactions : Magali Pelleau
Secrétaire général de rédaction : Laurence Ménardais
Premiers secrétaires de rédaction : Anne-Claire Blanchet, Fleur-Élodie Buffet, Virginie Condamine, Carole Hurviez
Rédacteurs-réviseurs : Sylvie Duverger, Isabelle Mora
Premier rédacteur graphiste : Dino Perrone
Chef de service Infographie : Hélène Burczynski
Rédacteurs graphistes : Stéphanie Dairain, Thibault Menguy, Romain Meynier
Infographiste : Claire Thibouméry
Dessinateur d'exécution : David Véas
Infographiste multimédia : Christelle Ochin
Webmaster : Mouna Issaadi-Allem

Commercial
Directeur des opérations : Jennifer Lévy-Benkemoun
Directeur d'unité : Rim Boubaker
Directeur d'éditions scientifiques unité oncologie : Charlotte Tovar

Régie publicitaire et annonces professionnelles
Valérie Glatin – Tél. : 01 46 67 62 77

Abonnements
Responsable/responsable adjoint : Badia Mansouri/ Florence Lebreton
Tél. : 01 46 67 62 74/87 – Fax : 01 46 67 63 09

EDIMARK PRESSE ÉDITION MÉDIA
DPC
Tél. : 01 46 67 63 00
E-mail : contacts@edimark.fr
Site Internet : www.edimark.fr

© décembre 2006 - Edimark SAS
Imprimé en France - POINT 44
94500 Champigny-sur-Marne
Dépôt légal : à parution

PEFC 10-31-3162
IMPRIM'VERT
REVUE BAC DE TH

Adhérent au SPEPS
Revue indexée dans la base ICMJE

Un génome ou des génomes ?

L'analyse biologique des leucémies aiguës à la recherche de mutations acquises des cellules leucémiques a commencé avec l'étude du caryotype et la découverte du chromosome Philadelphie associé à la leucémie myéloïde chronique *[1, 2]*. Cette approche s'est révélée par la suite extrêmement fructueuse, même si aucun Ph2 n'a pu concurrencer le Ph1…, et a été suivie de la découverte de nombreuses autres anomalies chromosomiques acquises, récurrentes et souvent spécifiques d'un sous-type de leucémie. La détection de celles-ci a ensuite été facilitée par l'hybridation in situ avec révélation par fluorescence (FISH) et la mise en évidence de transcrits de fusion par RT-PCR. Les étapes ultérieures ont été moléculaires avec l'analyse des ARN et de l'ADN de la cellule leucémique, et les progrès de la génétique ont permis des approches de plus en plus exhaustives, allant de la recherche ciblée d'anomalies à l'étude de l'ensemble de l'exome puis du génome des cellules leucémiques. Ces différentes approches mises aujourd'hui à la disposition des cliniciens, localement ou via des plateformes dédiées, comme celles de Plan France médecine génomique 2025, permettent désormais de décrire le "paysage génomique" d'un type précis de leucémie *[3]*. Les résultats délivrés par ces approches génétiques se sont accompagnés d'une confrontation, parfois brutale, entre les hématologues cliniciens et la génétique fondamentale. L'interprétation de la valeur clinique d'un variant peut de fait impliquer, avec l'identification d'anomalies potentiellement constitutionnelles, non seulement le patient, mais aussi l'ensemble de ses apparentés.

L'objet de ce dossier est de faire le tour des différents aspects liés à ces analyses génétiques. Les 2 premiers articles reviennent sur l'ADN et la notion de génome. Comme d'habitude, Marie-Christine Béné nous métamorphose tous en "Juliettes ravies" en nous racontant l'histoire passionnante de la découverte de l'ADN. Pierre Sujobert revient sur la variabilité du génome dans les cellules d'un individu et la complexité des évolutions clonales qui concernent aussi bien les tissus sains que les cancers, nous ôtant, si besoin était, nos dernières illusions, sur l'unicité du génome chez un individu et sa stabilité dans les cellules normales. L'expérience vécue de nos collègues nantais illustre la complexité pour les hématologues cliniciens de la prise en compte de ces variants germinaux qui, outre les aspects humains et éthiques, nécessite une bonne connaissance des approches génétiques. Celles-ci sont détaillées par Pierre Hirsch qui nous rappelle les avantages et limites des différentes techniques, alors que Yoann Vial nous apporte le point de vue du biologiste sur ces analyses en illustrant, par des cas concrets, l'analyse des variants identifiés. Les approches génomiques non ciblées (WES et WGS) génèrent par ailleurs une masse de données telle que seule une analyse bio-informatique est capable de la traiter et d'en tirer des résultats utiles au clinicien ou à la recherche. Christophe Antoniewski et son équipe ont bien voulu se charger de nous expliquer les différentes étapes d'une analyse bio-informatique et nous apporter les données minimales nécessaires aux cliniciens pour pouvoir, au moins, discuter avec les bio-informaticiens. Le lexique, fourni gracieusement, facilite grandement la lecture de l'article et pourra être conservé précieusement…

Espérons que ces approches non ciblées nous permettront d'avancer dans la connaissance des événements génétiques associés à la progression leucémique et d'élucider enfin les cas non encore compris, soit en mettant en évidence de nouveaux types d'atteintes de l'expression des gènes, soit, en particulier pour les prédispositions génétiques, en démontrant l'implication de nouveaux gènes.

Enfin, il n'aurait pas été possible pour un numéro consacré à la génétique au sens large du terme de ne pas aborder les problèmes légaux et éthiques liés à ces approches. Merci à Alice Hallopeau et à Sandrine de Montgolfier de nous rappeler la loi et de discuter l'ensemble des enjeux éthiques.

Bonne lecture !

T. Leblanc
Service d'hématologie pédiatrique, hôpital Robert-Debré, Paris.

Références

1. *Nowell P, Hungerford D. A minute chromosome in human chronic granulocytic leukemia. Science 1960;132:1497.*

2. *Rowley JD. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature 1973;243:290-3.*

3. *Kogure Y al. Whole-genome landscape of adult T-cell leukemia/lymphoma. Blood 2022;139(7):967-82.*

T. Leblanc n'a pas précisé ses éventuels liens d'intérêts en relation .

Analyse bio-informatique des variants somatiques tumoraux à partir du séquençage d'exomes ou de génomes complets

Bioinformatics analysis of somatics tumor variants from whole exome or genome sequencing

***CONECT-AML*[1], *L. Bellenger*[2,3], *N. Naouar*[2,3], *C. Antoniewski*[2,3]**

Résumé

Médecins et biologistes ont un rôle moteur à jouer pour améliorer la pertinence des analyses bio-informatiques des séquençages profonds (NGS) et permettre leur standardisation, indispensable en contexte clinique. Mais cela requiert une bonne compréhension des chaînes de traitement des données. Nous présentons ici de manière pratique un *workflow* d'identification de mutations somatiques à partir du séquençage NGS d'échantillons de tissus normal et tumoral. Les notions abordées sont détaillées dans un glossaire associé et un script de *workflow* permet de reproduire si besoin l'analyse dans un serveur Galaxy.

Mots-clés : Variants somatiques – NGS – Bio-informatique – Galaxy.

Summary

Physicians and biologists have a driving role to play in improving the relevance of bioinformatics analyzes of next generation sequencing (NGS) datasets and enabling their standardization, which is essential in a clinical context. To do so, a clear understanding of data processing is required. Here we present a practical workflow to identify somatic mutations from NGS of normal and tumoral tissue samples. The concepts covered are detailed in an associated glossary and a workflow script allows to reproduce the process on a Galaxy server if necessary.

***Keywords:* *Somatic variants – NGS* – Bioinformatics – Galaxy.**

Le séquençage à grande profondeur (*next-generation sequencing*, NGS) *[1]* permet d'identifier des mutations somatiques ou germinales dans le génome de cellules cancéreuses. Il est également à la base d'une médecine dite de précision, permettant de proposer à chaque patient un médicament, une thérapie ou un essai clinique optimaux en fonction des variants génétiques détectés dans sa tumeur *[2, 3]*. L'analyse bioinformatique des génomes requiert l'accès à des ressources informatiques adaptées, mais c'est loin d'être le seul défi à relever. En effet, des expertises très variées sont nécessaires pour choisir le protocole de séquençage, actionner les outils informatiques et les chaînes de traitements appropriés aux questions posées, et traiter les nombreux facteurs confondants observés dans les données de NGS.

L'intégration de ces expertises se fait idéalement au sein de plateformes multidisciplinaires dont les cliniciens peuvent se sentir simples utilisateurs. Pourtant, leur participation active à l'analyse est nécessaire. Par exemple, les méthodes d'analyse sont loin d'être standardisées et leur mise œuvre nécessite des connaissances cliniques pour éviter les contresens méthodologiques. Par ailleurs, face à la masse d'informations récoltées, la sélection des résultats pertinents demande une compréhension avancée des modèles cliniques et biologiques.

Cet article ne dresse pas un panorama des méthodes de recherche de variants somatiques dans les cancers, mais s'attache à illustrer de façon pratique et synthétique le déroulé de ces étapes. Les lecteurs sans expertise avancée trouveront également en annexe (glossaire, page 217) une explication des notions couramment utilisées en analyse bio-informatique de NGS (termes soulignés). Notre objectif est de fournir des clés aux biologistes et aux cliniciens pour approfondir leur compréhension de l'analyse bio-informatique de variants somatiques dans les cancers et les inciter à y contribuer activement.

[1] *COllaborative NEtwork on research for Children and Teenagers with Acute Myeloid Leukemia (www.conect-aml.fr).*
[2] *ARTbio, Sorbonne université, CNRS FR 3631, Inserm US 037, Paris.*
[3] *Institut français de bioinformatique (IFB).*

Étapes de l'analyse du génome ou de l'exome entier

La recherche de variants somatiques à partir de données de séquençage d'ADN tumoral implique séquentiellement la filtration des lectures du séquençage brut, leur alignement avec le génome humain et la sélection des alignements valides, l'identification de variants candidats et leur annotation et priorisation en fonction des questions posées. Chaque étape de cette chaîne de traitement – ou *workflow* – peut être réalisée avec un ou plusieurs outils différents, eux-mêmes réglés avec des valeurs de paramètres spécifiques. Pour la pratique, nous avons choisi de ne présenter en détail que le *workflow* utilisé par la plateforme d'analyse ARTbio ***(figure, p. 212)*** au sein du consortium de recherche CONECT-AML pour la recherche de variants somatiques dans des séquençages génomiques complets (*whole genome sequencing*, WGS) des leucémies aiguës myéloblastiques, à partir de 2 échantillons de moelle osseuse prélevés respectivement au diagnostic et à la rémission complète après traitement. Il est important de noter que le même *workflow* peut être appliqué à des séquençage d'exomes (*whole exome sequencing*, WES).

Traitement et filtration des données brutes

À l'heure actuelle, les WGS et WES en oncologie sont principalement obtenus avec des appareils Illumina® et la technique de *paired-end sequencing*. Dans cette configuration, un échantillon génère une paire de fichiers compressés au format fastq de quelques dizaines (WES) à centaines (WGS) de gigaoctets chacun. Le traitement des fichiers fastq débute par le calcul de métriques permettant d'évaluer la qualité du séquençage. FastQC, le programme le plus utilisé pour ces opérations, calcule les qualités moyennes par position dans les lectures, les qualités moyennes des lectures entières, le taux de nucléotides indéterminés, la distribution du contenu en bases GC des lectures, le taux de duplication des lectures, ainsi que la présence de séquences surreprésentées ou provenant des adaptateurs utilisés pour construire la librairie de fragments génomiques.
Ces indicateurs sont utilisés par les programmes de filtration des fichiers fastq afin d'en éliminer les lectures de mauvaise qualité pouvant induire l'identification de faux variants. Dans la plateforme ARTbio, les fichier fastq sont traités afin d'éliminer les lectures contenant des séquences d'adaptateurs, comportant plus de 10 % de nucléotides inconnus, ou plus de 50 % de nucléotides de qualité inférieure à 5 (sur une échelle de 40). Dans la plupart des cas, ces critères éliminent tout au plus 5 à 10 % des lectures. C'est suffisant à ce stade car, d'une part, une sélection trop forte sur la qualité des lectures induit un biais (les lectures Illumina® de forte qualité sont enrichies en A et T) et, d'autre part, une filtration plus efficace est effectuée sur les lectures une fois qu'elles sont alignées sur le génome.

Alignement des séquences sur un génome de référence

L'étape suivante consiste à aligner les lectures sur un génome de référence, de préférence GRCh38/hg38, voire GRCh39/hg39 disponible depuis juin 2020. Les principaux programmes d'alignement de lectures Illumina® sont BWA *[4, 5]* et Bowtie2 *[6]*. Nous utilisons BWA-mem avec les paramètres précisés dans le glossaire. Les aligneurs génèrent un fichier compressé BAM où les alignements sont représentés suivant le format SAM et triés selon leurs coordonnées génomiques. Sauf indication contraire, les fichiers BAM contiennent également les lectures de séquences non-alignées et peuvent ainsi se substituer aux fichiers fastq pour la conservation des données de séquençage. Le format SAM tient avec le format VCF *(*lire, *Détection des variants SNV et indels)* une place centrale dans l'analyse bioinformatique des variants. Il est donc recommandé de le connaître dans le détail (voir le glossaire).

Post-traitement des alignements

Si un variant génomique se traduit dans un fichier BAM par des alignements anormaux sur le génome de référence, il en est de même pour les lectures de séquences erronées, ou comportant des motifs nucléotidiques impossibles à aligner de manière unique. Le post-traitement des alignements a pour but d'éliminer ces artefacts techniques, sources de faux variants. Une 1re tâche est de réduire à un seul les alignements de lectures issues d'une même molécule d'ADN par amplification PCR, dont le comptage biaiserait l'estimation des fréquences alléliques des variants. Sans dispositif expérimental pour incorporer un identifiant moléculaire unique (*unique molecular identifiers*, UMI) dans les fragments génomiques avant amplification, on suppose que 2 fragments séquencés dérivent de la même molécule amplifiée et que les alignements de leurs couples de lectures respectifs sont identiques. Les outils de déduplication Picard MarkDuplicates et samtools-markdup sont basés sur cette assertion. Notons que cette étape n'est pas effectuée pour les approches de séquençage pour lesquelles les couvertures sont si élevées que la présence de fragments génomiques de même séquence, mais issus d'amplicons différents, devient vraisemblable. C'est le cas par exemple des séquençages NGS de panels de gènes.

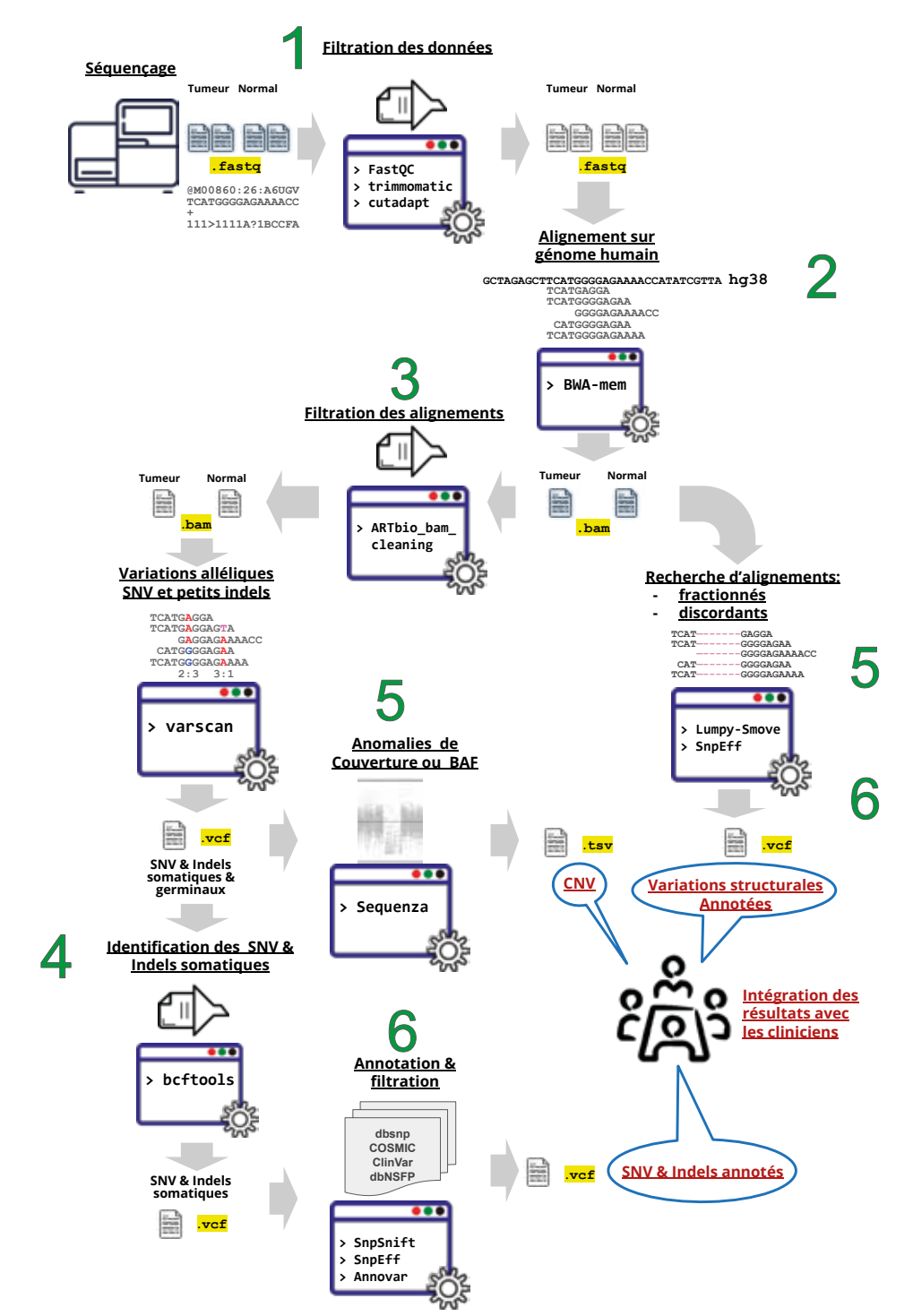

Figure. Schéma d'une analyse de séquences NGS pour la recherche de mutations somatiques dans un échantillon tumoral. Les titres soulignés identifient les différentes étapes de l'analyse. Les numéros verts identifient la section dans laquelle l'étape de l'analyse est détaillée. L'origine, le nombre et le format (surligné en jaune) des fichiers manipulés sont symbolisés pour chaque étape. Les noms des programmes utilisés dans les traitements informatiques sont reportés dans les icônes de console d'ordinateur. L'analyse converge vers la discussion des résultats avec les biologistes et les cliniciens.

Les fichiers BAM sont ensuite filtrés de façon à éliminer les alignements orphelins (une seule lecture d'extrémité alignée) ou de mauvaise qualité (la qualité d'alignement est distincte de la qualité de séquence). Les lectures qui comportent des insertions ou des délétions *(indels)* de motifs répétés peuvent parfois s'aligner de plusieurs façons, bien qu'elles correspondent au même haplotype. Pour éviter ce problème, on les réaligne à la position permise la plus à gauche sur le génome, à l'aide de l'outil *bamleftalign*. Cette opération est suivie du calcul du tag "MD" avec l'outil samtools-calmd, qui décrit dans le fichier BAM la structure des alignements dans une syntaxe spécifique interprétée par certains *variant callers*. Finalement, on peut effectuer une dernière filtration des alignements de qualité indéterminée (valeur de 255), qui correspondent généralement à des alignements discontinus sur des régions non contiguës du génome. Cependant, cette opération ne doit pas être réalisée pour les outils qui tirent parti des alignements discontinus pour détecter des remaniements génomiques.
Ces opérations de filtration utilisent beaucoup de ressources de calcul et de stockage de fichiers intermédiaires. Afin de les optimiser, nous les avons chaînés dans un seul script *artbio_bam_cleaning*.

Détection des variants SNV et indels

La sélection des variants *(variant calling)* avec substitutions de nucléides uniques (*single nucleotide variation*, SNV) et insertions/délétions *(indels)* de moins de 50 nucléotides, s'effectue en compilant pour chaque position nucléotidique les anomalies des alignements qui la recouvrent dans le fichier BAM. Les programmes pour la sélection des SNV et *indels [7]* effectuent cette opération de *pileup* à partir des fichiers BAM et retournent tous leurs résultats dans un fichier au format standard VCF, mais ils sont bâtis sur des algorithmes variés faisant appel à quantité de paramètres et fichiers annexes différents. Ils calculent également de manière différente les scores de confiance attribués aux variants détectés et il est difficile d'accéder à la formule exacte utilisée. Ainsi, la maîtrise d'un seul *variant caller* demande beaucoup de temps, rendant difficile l'évaluation comparée de plusieurs programmes. Nous avons choisi d'utiliser Varscan, dont les paramètres d'utilisation sont détaillés dans le glossaire.
Il peut être utile de combiner les résultats de différents *variant callers*, soit en recherchant un consensus, soit en les intégrant dans une approche d'apprentissage machine pour améliorer la performance globale de détection de variants *[8]*.

Détection des variants structuraux et des variations du nombre de copies

Les variants structuraux (*structural variations*, SV) *[9]* incluent les remaniements équilibrés – inversions et translocations – ainsi que des remaniements regroupés sous le terme de CNV ou CNA *(copy number variation/alteration)* conduisant à un gain ou une perte de séquences génomiques : duplications, insertions et délétions de plus de 50 nucléotides. Un 1er type d'outils détecte les SV à partir des artefacts qu'ils induisent dans les fichiers BAM : alignements discordants de paires de lecture *(discordant pairs)* et alignements fractionnés *(split reads)*. Il faut donc avoir pris soin de ne pas éliminer ces artefacts au cours des étapes précédentes de nettoyage. D'un autre côté, les CNV ne génèrent pas nécessairement des artefacts d'alignement (c'est le cas, par exemple, des pertes d'hétérozygotie (*loss of heterozygozity*, LOH) à copie neutre). Ainsi, un autre type d'outils pour la détection de CNV allie la recherche des anomalies de couverture en lecture et celle des variations abruptes des fréquences des allèles B dans le génome tumoral.
Pour la détection des SV par anomalie d'alignement de lecture, nous utilisons le programme Lumpy-Smoove qui encapsule lui-même le *variant caller* Lumpy-SV, tandis que le programme Sequenza nous sert à identifier les CNV associés à des variations de couverture et/ou de fréquence des allèles B.

Annotation et priorisation des variants

La dernière étape consiste à annoter les variants et à les associer à des scores de prédiction d'effets, en lien avec le modèle clinique étudié. Une 1re tâche est d'identifier les variants qui affectent une séquence codant une protéine, des sites d'épissage ou d'autres éléments génomiques fonctionnels. Ensuite, on peut ne sélectionner que ces variants, mais cela dépend des objectifs de l'étude. Les variants doivent aussi être comparés avec ceux des bases de données telles que dbSNP, 1 000 Genomes ou gnomAD, qui répertorient les SNP humains. On peut ainsi éliminer les variants dont la fréquence observée dans les populations est supérieure à 1 %, car leur implication dans l'oncogenèse est peu vraisemblable.
Les variants restants peuvent enfin être annotés avec les bases ClinVar ou COSMIC, qui référencent les variants déjà associés à des cancers, ainsi que des bases de score de prédictions fonctionnelles telles que SIFT, PolyPhen-2, etc., dont le choix se fait en fonction des questions biologiques et cliniques ayant motivé l'analyse. L'annotation peut aussi prendre en compte à ce stade les connaissances de l'équipe, préalablement compilées et structurées de façon appropriée. Les outils

d'annotation capables d'interagir avec toutes ces bases de données incluent Variant Effect Predictor d'Ensembl, Funcotator de GATK et beaucoup d'autres discutés par F. Borchert et al. *[10]*. Notre *workflow* d'annotation à ARTbio utilise séquentiellement SnpSift (pour identifier les SNV répertoriés dans dbsnp), SnpEff (annotations de génomique fonctionnelle) et ANNOVAR qui nous permet d'annoter les variants avec les bases de données de prédictions COSMIC, ClinVar et dbNSFP *[11]*, qui agrège elle-même les prédictions d'un grand nombre de bases (SIFT, PolyPhen2, LRT, MutationTaster, FATHMM, etc.)

Si l'on dispose d'une cohorte, il est encore possible de réduire de façon drastique la liste de variants candidats en croisant les résultats observés pour différents patients. Les variants apparaissant dans plus de 15 à 20 % des échantillons de la cohorte sans être référencés dans ClinVar ou COSMIC doivent d'abord être vérifiés : si leur VAF est systématiquement faible (< 0,05), ils peuvent avoir été induits par des artefacts de séquençage, tandis que si elle est systématiquement élevée (> 0,3 ou 0,4), ils peuvent correspondre à des SNP non répertoriés ou mal annotés lors des étapes précédentes. Ces vérifications faites, les variants avec des annotations de haut score et surreprésentés dans la cohorte sont potentiellement impliqués dans la biologie du cancer étudié. Il conviendra de les soumettre à une analyse plus ciblée et, si possible, à une procédure de confirmation avec des données ou des expériences indépendantes.

Conclusion

Nous avons présenté ici un *workflow* simplifié d'identification de mutations somatiques dans un échantillon de tissu tumoral. Notre objectif était de permettre aux biologistes et aux cliniciens non spécialistes de mettre un pied à l'étrier de l'analyse bioinformatique à travers un exemple pratique. S'ils le souhaitent, ils peuvent également reproduire eux-mêmes cette analyse dans un serveur Galaxy en s'aidant d'un tutoriel disponible en suivant le lien https://artbio.github.io/startbio/Run-COH/

En l'absence actuelle de méthodes de référence pour identifier les mutations associées au cancer, les équipes, y compris la nôtre, mettent en œuvre des *workflows* variés, obtenant des résultats qui convergent souvent pauvrement. Il est donc urgent, en particulier pour la pratique clinique, d'établir des méthodes d'analyse reproductibles et consensuelles. La reproductibilité demande de connaître dans les moindres détails les scripts, transformations manuelles et lignes de commandes simples utilisés pour lier les étapes computationnelles entre elles. Nous capturons cette "matière noire de la bioinformatique" à l'aide du système Galaxy et de son langage formel de *workflows*. Ainsi, le fichier Galaxy-Workflow-COH.ga (disponible dans le tutoriel mentionné ci-dessus) renferme l'intégralité des informations nécessaires pour reproduire à l'identique dans un serveur Galaxy l'analyse présentée ici. Le consensus appelle quant à lui une meilleure coordination internationale et des études comparatives, tout comme une montée en compétence et la contribution des biologistes et des cliniciens.

L'annotation des variants est essentielle pour parvenir à des conclusions pertinentes. Elle requiert des bases de données fiables, si possible non redondantes et aux formats compatibles, voire identiques. Des progrès considérables sont encore possibles dans ces directions. On peut aussi espérer que des approches d'intelligence artificielle aideront dans un avenir proche à améliorer les prédictions des bases de données, par exemple celles concernant les traitements à utiliser en fonction des profils génomiques observés chez les patients. Mais ces approches nécessiteront l'accès à de très grands jeux de données sur les patients atteints de cancer, alliant séquences génomiques, observations cliniques et réponses thérapeutiques, tout en garantissant l'éthique de leur utilisation. ■

C. Antoniewski déclare ne pas avoir de liens d'intérêts en relation avec cet article. L. Bellenger et N. Naouari n'ont pas précisé leurs éventuels liens d'intérêts.

Références

1. *Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. Genomics 2016;107(1):1-8.*

2. *Schwaederle M et al. Impact of precision medicine in diverse cancers: a meta-analysis of phase II clinical trials. J Clin Oncol 2015;33(32):3817-25.*

3. *Meric-Bernstam F et al. Feasibility of large-scale genomic testing to facilitate enrollment onto genomically matched clinical trials. J Clin Oncol 2015;33(25):2753-62.*

4. *Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 2010;26(5):589-95.*

5. *Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN]. 2013. http://arxiv.org/abs/1303.3997*

6. *Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9(4):357-9.*

7. *Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J 2018;16:15-24.*

8. *Wang M et al. SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. Sci Rep 2020;10(1):12898.*

9. *Sudmant PH et al. An integrated map of structural variation in 2,504 human genomes. Nature 2015;526(7561):75-81.*

10. *Borchert F et al. Knowledge bases and software support for variant interpretation in precision oncology. Brief Bioinform 2021;22(6):bbab246. doi:10.1093/bib/bbab246*

11. *Liu X et al. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med 2020;12(1):103.*

Glossaire

"Mal nommer un objet, c'est ajouter au malheur de ce monde"

Sur une philosophie de l'expression, 1944. Albert Camus

1 000 Genomes : base de donnée des variations génétiques identifiées dans le projet "1 000 Genomes".

Allèle B (fréquence) : en analyse NGS, la fréquence d'allèle B (BAF) d'une position rapportée comme SNV est le rapport du nombre de lectures de l'allèle alternatif (colonne Alt du VCF) sur la somme des nombres de lectures de cet allèle alternatif et de l'allèle de référence (Alt/(Alt + Ref). Si l'on parcourt une région normale diploïde, la BAF des SNV rencontrés oscille ainsi entre des valeurs proches de 0, 0,5 et 1. Dans une région triploïde, par exemple, la BAF oscillera entre les valeurs 0, 0,33, 0,66 et 1 (pour les configurations AAA, AAB, ABB et BBB).

Aligneurs : programmes prenant en entrée des fichiers de lecture de séquences et un génome de référence (sous une forme comprimée et indexée), et assignant une position chromosomique à ces lectures. Les aligneurs calculent en plus différentes métriques essentielles pour évaluer le type et la qualité de chaque alignement. Ces informations sont produites en sortie dans un fichier SAM, le plus souvent compressé à la volée en un fichier BAM.

ANNOVAR : ANNOVAR *[1]* est un logiciel qui utilise un grand nombre de bases de données pour annoter de manière fonctionnelle les variants génétiques détectés à partir de divers génomes (y compris les génomes humains de référence hg18, hg19, hg38). Son développement n'a pas été soutenu depuis 2013, mais il reste fonctionnel et il est capable d'interagir avec les versions récentes des bases de données d'annotations et de prédictions. Nous utilisons une version d'ANNOVAR (Table_Annovar) développée pour fonctionner dans Galaxy.

artbio_bam_cleaning : outil Galaxy développé par la plateforme ARTbio pour gérer en un seul traitement toutes les opérations de filtration des alignements dans un fichier BAM. Pour les détails du code, voir artbio_bam_cleaning dans le Galaxy Toolshed.

BAM : format de compression des fichiers SAM.

bamleftalign : *bamleftalign* est un utilitaire du programme freebayes dont la fonction est d'aligner à gauche quand c'est possible toutes les insertions et délétions identifiées dans un alignement au format SAM/BAM, et de les fusionner.

BWA-mem : programme d'alignement *[2]* utilisé par ARTbio pour les lectures de séquences génomiques. Notre ligne de commande indicative est :

```
bwa mem -t 16 -M -Y
-R "@RG\tID:uniq_ident\tPL:Illumina\tPU:flowcell&lane&sampleid\tetc…
  LB:flowcell&lane&sampleid\tSM:sample_name\tCN:seq_center_name"
<chemin de l'index du génome> <chemin des séquences R1> <chemin des séquences R2>
```

où l'option -t 16 indique que le programme utilisera 16 processus en parallèle, -M et -Y sont des options de mise en forme de fichier BAM, et l'option -R sert à renseigner les informations de "*read group*" qui sont ajoutées pour chaque alignement. Les *read group* (@RG) simplifient l'utilisation des fichiers BAM car ils leur associent des informations facilitant leurs comparaisons ou leurs fusions. Bowtie2 offre des performances en pratique égales à celle de BWA-mem, mais nécessite un ajustement plus complexe des paramètres de fonctionnement *[3, 4]*.

ClinVar : ClinVar est une base de données du NCBI répertoriant les relations entre les variations génétiques humaines et les phénotypes ainsi que les preuves expérimentales les appuyant.

Compression des fichiers : elle est souhaitable pour soulager l'utilisation des disques durs et accélérer certains traitements. Il s'agit le plus souvent d'une compression gzip (extension de fichier ".gz"). Dans les systèmes Linux, l'outil de compression d'un fichier est gzip et l'outil de décompression d'un fichier compressé est gunzip. Le taux de compression est compris entre ~2 et ~10 selon la nature des données contenues dans le fichier. La plupart des outils de traitement des fichiers fastq acceptent également les fichiers fastq.gz, sans décompression préalable. Il existe d'autres procédés de compression, par exemple zip, ou bzip. À noter que le format BAM est aussi un format de compression spécifique des fichiers SAM.

Copy number variation/alteration **(CNV/CNA)** : remaniements conduisant à un gain ou perte de séquences génomiques : duplications, insertions et délétions de plus de 50 nucléotides (nt) (voir *[5]* pour une présentation simple des différents remaniements). Un 1er type d'outils, dont lumpy-SV est un exemple, détecte les SV à partir des paires d'alignements discordants *(discordant pairs)* et des alignements fractionnés *(split reads)*. Un autre type d'outil, plus adapté à la détection de CNV de taille importante, par exemple Sequenza, allie la recherche des anomalies de couverture en lecture et celle des variations abruptes de fréquence des allèles B *(B-allele frequencies)* dans le génome tumoral, en comparaison du génome normal. La couverture en lecture est affectée par divers biais, le plus courant étant le contenu en GC qui sera donc pris en compte par ces programmes de détection de CNV. Il est à noter qu'une synchronisation, même partielle, de la réplication des cellules d'un échantillon peut introduire un biais de couverture qu'il faudra aussi prendre en compte pour interpréter les CNV candidats.

COSMIC : COSMIC est une base de données du Sanger Institute répertoriant les mutations somatiques trouvées dans les cancers humains ainsi que leur impact.

Couverture (profondeur) : la profondeur ou la couverture de lecture est le nombre de lectures non dupliquées s'alignant sur une position du génome donnée et doit être calculée en fonction du fichier BAM post-traité, indépendamment pour la tumeur et le contrôle. Des outils comme *bedtools genomecov [6]*, par exemple, sont largement utilisés à cette fin. La couverture est souvent exprimée sous forme de moyenne ou de pourcentage sur un ensemble d'intervalles, tels que les exons ou uniquement les exons ciblés sur un panel. Pour le séquençage de panel ou les données WES, le ratio sur cible des lectures cartographiées doit être calculé en plus de la couverture sur cible. Un faible ratio sur la cible justifie la prudence et peut indiquer que le processus de préparation de la bibliothèque doit être répété (puisqu'une forte proportion de la séquence s'aligne en dehors des cibles prévues). Il est ainsi crucial de connaître exactement et exhaustivement les régions cibles d'un séquençage génomique. Il est à noter que ces régions diffèrent souvent selon les panels de gènes et les kits de WES.

dbNSFP : dbNSFP *[7]* est une base de données agrégeant les autres bases SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, GERP++, PhyloP and SiPhy. Pour des raisons historiques, elle est référencée sous l'acronyme LJB* pour son utilisation avec ANNOVAR.

dbSNP : base de données du NCBI répertoriant les polymorphismes de nucléotides uniques dans les populations humaines. La base compte 3 341 554 567 enregistrements.

fastq : format de fichiers de séquences incluant la qualité de séquençage. Chaque lecture de séquence est décrite par 4 lignes. La 1re ligne est un identifiant (ou entête) unique commençant par le caractère "@" (vert). Pour un séquençage "*paired-end*", les résultats sont le plus souvent fournis sous la forme d'un couple de fichiers R1 et R2. Dans ce cas, un même identifiant se retrouve à la même position dans chacun des fichiers, assorti des suffixes 1 et 2 (rouge). La 2e ligne est la séquence nucléotidique (A, T, G, C ou N), la 3e contient le plus souvent le signe unique "+" (bleu) et la 4e indique la qualité de lecture de chaque nucléotide en 2e ligne, sous la forme d'un caractère imprimable. La valeur ASCII de ces caractères est comprise entre 33 et 73, et, soustraite de 33, elle renvoie à un score de qualité compris entre 0 et 40 (voir ce tableau). Grâce à ce procédé, on peut aligner un caractère unique avec chacun des nucléotides, indiquant leur score de qualité de lecture. La qualité Q peut être reliée à la probabilité P de lecture erronée par la formule :

$$Q = -10 \log_{10}(P)$$
$$\text{soit } P = 10^{-Q/10}$$

Ainsi, une qualité de 40 est codée par le caractère "I" qui renvoie à une probabilité de lecture erronée de 1/10 000, tandis qu'une qualité de 29, codée par le caractère ">" (jaune) renvoie à une probabilité d'erreur de ~1,2/1 000.

Fichier des lectures "R1"	Fichier des lectures "R2"
@M00860:26:000000000-A6UGV:1:1101:18278:1676 1:N:0:3 TTGAACCTGGGAGGCAGAGGTTGCAGTGAGCCCAGATTGTG + >3>>ABAB4FAF2EAFGFGE4E4CFGGCHHHFCHG2FHHHHHHHH HEBHHHHH	@M00860:26:000000000-A6UGV:1:1101:18278:1676 2:N:0:3 CTATTCTTCATATTTTACCTCTAATTGAAAGGGATTATATTCAG + 1111133B3333AAAG31A1ABD13D31111000001B33D2DA 2111B1AF

FastQC : logiciel de contrôle de qualité des données de séquençage. Le code est développé en langage JAVA ce qui le rend compatible avec les plateformes Windows/MacOS/Linux. Il propose une interface graphique autonome, mais il peut aussi être utilisé avec des lignes de commandes. Un outil Galaxy FastQC est également disponible dans Galaxy.

Filtration des fichiers fastq : élimination des lectures de mauvaise qualité pouvant induire l'identification de faux variants. Elle peut être effectuée à l'aide des programmes trimmomatic *[8]*, cutadapt *[9]* ou SOAPnuke *[10]*.

Fréquence des allèles B *(B-allele frequency)* : fréquences des allèles "B" *(B-allele frequencies)* des polymorphismes de nucléotides (SNP).

Galaxy : Galaxy est une application *open source* (distribuée sous la licence gratuite académique permissive) qui s'installe sur un serveur de calcul bio-informatique. Galaxy offre un système (souvent appelé "*framework*") qui permet aux chercheurs sans expertise en informatique d'accéder à la plupart des outils d'analyses utilisés par les bio-informaticiens à travers une interface graphique conviviale, plutôt qu'en utilisant des lignes de commandes. Un utilisateur interagit avec Galaxy via le web en téléchargeant et en analysant les données. Galaxy se charge

d'interagir avec l'infrastructure de calcul sous-jacente (serveurs qui exécutent les outils et les *workflows*, et disques qui stockent les données) sans l'exposer à l'utilisateur.

Génome de référence : 3 références de génome humain sont disponibles. GRCh37/hg19 date de juillet 2007 et ne devrait être utilisé que dans les rares cas où les résultats attendus doivent être intégrés avec des résultats plus anciens obtenus avec cette référence, et impossibles à convertir dans une version plus récente. L'utilisation de GRCh38/hg38 (décembre 2011) lui sera toujours préférée. GRCh39/hg39 est également disponible depuis juin 2020. Beaucoup de zones jusqu'alors non séquencées ou non associées à des positions chromosomiques y sont maintenant cartographiées, en particulier les zones hétérochromatiques péricentromériques ou télomériques. On devrait donc la privilégier dès à présent.

gnomAD : la base de données d'agrégation du génome (gnomAD) est hébergée par le Broad Institute et elle est construite par un consortium de chercheurs dans le but d'agréger et d'harmoniser les données de séquençage de l'exome et du génome humain à partir de différents projets de séquençage à grande échelle, et de rendre les données collectées disponibles à la plus large communauté scientifique possible.

Haplotype : un haplotype est un groupe de variants génomiques (ou polymorphismes) qui se localisent à proximité les uns des autres sur le même chromosome et qui ont donc tendance à être transmis ensemble.

Indel : *indel* est un terme qui désigne une insertion et/ou une délétion de bases dans le génome. En évolution moléculaire, on parle d'*indel* pour les délétions/insertions de 1 à 10 000 paires de bases de longueur, et de *microindel* si cette longueur est inférieure ou égale à 50 nt. Cependant, en analyse bioinformatique des variants génomiques, l'usage est plutôt de désigner les *microindels* par le terme *(small) indels*, et de classer les insertions ou délétions de plus grande taille dans la catégorie des (larges) variations structurales (SV). Pour une étude intéressante, voir *[11]*.

LOH : les pertes d'hétérozygotie "copie-neutre" *(copy-neutral LOH)* sont classées avec les CNV. En effet, même si le nombre global de copies ne varie pas dans une LOH, elle implique bien la duplication d'un segment de chromosome accompagnée d'une perte de la région homologue correspondante, de sorte que la cellule conserve 2 copies mais provenant d'un seul parent.

Lumpy-Smoove : nous avons développé cet outil Galaxy basé sur le variant *caller Lumpy-SV [12]*, lui-même encapsulé dans l'outil Smoove qui ajoute une filtration interne des alignements discordants ou fractionnés afin d'accélérer la sélection et le génotypage des variants structuraux. Une ligne de commande indicative de l'outil est :

```
smoove call --name output --exclude exclude.cnvnator_100bp.GRCh38.20170403.bed
            --fasta reference.fa --processes 4 --genotype --removepr *.bam
```

pileup : il décrit les informations sur les paires de bases à chaque position chromosomique. Ce format facilite l'inspection visuelle et la sélection des SNP/indel par les variant callers.
Le format pileup ressemble à ceci :

```
seq1 272 T 24  ,.$.....,,.,.,...,,,,.,..^+. <<<+;<<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,,.,.,...,,,,.,..A <<<;<<<<<<<<<3<=<<<;<<+
seq1 274 T 23  ,.$....,,.,.,...,,,,.,....  7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23  ,$....,,.,.,...,,,,.,...^!.  <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22  ...T,,.,.,...,,,,.,....  33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22  .....,.,.,.C.,,,.,..G.  +7<;<<<<<<<&<=<<:;<<&<
```

```
seq1 278 G 23  ....,,.,.,...,,,,.,.....^k.  %38*<<;<7<<7<=<<<;<<<<<
seq1 279 C 23  A..T,,.,.,...,,,,.,..... ;75&<<<<<<<<<=<<<9<<:<<
```

où chaque ligne est constituée du chromosome, de la coordonnée (basée à 1), de la base de référence, du nombre de lectures couvrant la position nucléotidique, des nucléotides lus et de leur qualité de séquence. Dans la colonne des nucléotides lus, un point représente une correspondance avec le nucléotide de référence sur le brin sens, une virgule une correspondance sur le brin antisens, "ACGTN" pour une non-concordance sur le brin sens et "acgtn" pour une non-concordance sur le brin antisens.
Voici un exemple d'insertions de 2bp communes à 3 lectures :

```
seq2 156 A 11  .$.....+2AG.+2AG.+2AGGG   <975;:<<<<<
```

Voici un exemple de suppression de 4 pb de la référence dans deux lectures :

```
seq3 200 A 20 ,,,,,,.,.-4CACC.-4CACC.....,,,.^~. ==<<<<<<<<<<<::<;2<<
```

Pour les détails de la syntaxe du format pileup, voir la documentation de l'outil samtools.

Outil : terme générique en vigueur en bioinformatique pour désigner tout programme ou script réalisant une tâche d'analyse.

Paired-end sequencing : à l'heure actuelle, la plupart des séquençages génomiques en oncologie sont générés avec des appareils Illumina en lisant 100 à 150 nt des 2 extrémités de fragments génomiques de ~500 nt contenus dans une librairie de séquençage. Le séquençage d'un échantillon produit ainsi une paire de fichiers compressés, respectivement R1 et R2 pour les séquences des extrémités "gauche" et "droite".

Picard MarkDuplicates : élément de la boîte à outils Picard développée par le Broad Institute en langage JAVA. MarkDuplicate permet de marquer et, si besoin, de retirer, les alignements SAM résultant d'une duplication PCR des fragments dans les librairies.

PolyPhen-2 : PolyPhen-2 est une base de données répertoriant les impacts possibles de substitutions d'acides aminés sur la structure et la fonction des protéines humaines. Elle est construite avec l'outil PolyPhen-2 *[13]* qui utilise des caractéristiques physiques et comparatives des séquences protéiques.

Qualité du séquençage : voir fastq.

Qualité d'alignement : reportée dans la 5e colonne "MAPQ" d'un fichier SAM, elle est égale à −10 * log10(P) arrondi à l'entier le plus proche, où P est la probabilité que la position d'alignement soit erronée. Une valeur 255 indique que la qualité d'alignement n'est pas disponible. Il s'avère que différents programmes d'alignement calculent différemment les scores MAPQ. Il n'est donc pas possible de comparer directement ces scores, à moins qu'ils ne proviennent du même programme.

SAM : les fichiers SAM sont des fichiers TSV *(tabulation-separated values)*, qui commencent par un en-tête facultatif, et décrivent des alignements de séquences. Un exemple simplifié de fichier SAM est donné ci-dessous :

```
@HD VN:1.3  SO:coordinate
@SQ SN:chr1  LN:45
@SQ SN:chr2 LN:40
r001   163 chr1 7   30  8M4I4M1D3M  =   37  39  TTAGATAAAGAGGATACTG *   XX:B:S,12561,2,20,112
r002   0   chr1 9   30  1S2I6M1P1I1P1I4M2I *   0   0   AAAAGATAAGGGATAAA   *
r003   0   chr1 9   30  5H6M    *   0   0   AGCTAA  *
r004   0   chr1 16  30  6M14N1I5M  *   0   0   ATAGCTCTCAGC    *
r003   16  chr1 29  30  6H5M    *   0   0   TAGGC   *
r001   83  chr1 37  30  9M  =   7   -39 CAGCGCCAT   *
```

L'en-tête se compose de plusieurs lignes, commençant par un caractère '@', chaque ligne décrivant une métadonnée d'alignement. Elle commence par son identifiant, suivi de balises séparées par des tabulations. Chaque balise se compose d'un identifiant à 2 caractères, suivi de ':', et de la valeur assignée. Si elle est présente, la métadonnée @HD vient en premier et indique la version SAM (balise VN) utilisée dans le fichier et l'ordre de tri (SO) des alignements. Les métadonnées facultatives @SQ indiquent les noms des séquences de référence (balise SN) et leur longueur (balise LN). Il existe bien d'autres types de métadonnées qui peuvent être renseignées dans l'en-tête du fichier SAM, comme par exemple @PG ID:bwa PN:bwa VN:0.7.15-r1140 CL:bwa mem -t 10 -M -Y -R @RG\tID qui spécifie une commande du programme pour générer l'alignement.

La section d'en-tête facultative est suivie des enregistrements d'alignement, qui sont à nouveau séparés par des tabulations, définissant 11 colonnes obligatoires.

Col	Champ	Type	Valeur (défaut)	Description
1	QNAME	string	obligatoire	Identifiant de la lecture
2	FLAG	int	obligatoire	"Flag" de l'alignement
3	RNAME	string	*	Nom de la référence de l'alignement
4	POS	32-bit int	0	Position sur la référence (première position = 1)
5	MAPQ	8-bit int	255	Qualité de l'alignement
6	CIGAR	string	*	Chaîne CIGAR de l'alignement
7	RNEXT	string	*	Référence de l'alignement de la lecture appariée
8	PNEXT	string	0	Position de l'alignement de la lecture appariée
9	TLEN	string	0	Longueur calculée du fragment séquencé
10	SEQ	string	*	Séquence de la lecture
11	QUAL	string	*	Valeur ASCII de la qualité (codée au format PHRED)

Quelques remarques :

✓ la standard SAM évoque des "*queries*". Dans un contexte d'alignement, les *queries* sont des lectures ;

✓ la standard SAM évoque des "*templates*" et "segments". Dans un contexte d'alignement de lectures "*paired-end*", un *template* se compose de 2 segments, chacun correspondant à une lecture. La longueur du *template* correspond à la taille calculée du fragment séquencé ;

✓ les lectures appariées sont stockées sous la forme de 2 alignements avec le même QNAME. Les lectures d'une paire sont alors discriminées par leur valeur de FLAG ;

✓ lorsque le FLAG indique que SEQ est "inverse-complémentée", alors QUAL est aussi inversé ;

✓ les positions dans le fichier SAM sont basées à 1 (premier nucléotide = 1). Dans un fichier BAM, les positions deviennent basées à 0 (premier nucléotide = 0).

Les qualités doivent être stockées sous un format ASCII PRED. Les noms identifiants de lectures et de références ne doivent pas contenir d'espaces. Il est courant de couper les identifiants de lectures ou de références au 1er espace rencontré.

Les 11 colonnes obligatoires sont suivies d'un nombre arbitraire de tags facultatifs. Les tags ont un identifiant à 2 caractères suivi de ":${TYPE}:", et de la valeur du tag. Par exemple, MD:Z:12A130A6 indique que le tag MD, de type Z (chaîne de caractères), prend la valeur 12A130A6.

Le format SAM code de nombreuses autres informations que nous ne détaillons pas ici. Étant donné qu'il est une référence centrale pour les outils de recherche de variants,

nous recommandons vivement de se référer au document de spécification du format SAM aussi souvent que nécessaire.

samtools-calmd : calmd est un outil de la suite samtools *[14]* qui calcule le tag MD dans la 12e colonne des fichiers SAM/BAM. Si la balise MD est déjà présente dans le fichier SAM/BAM et que calmd calcule un tag MD différent de celui existant, l'outil renverra un avertissement.

samtools-markdup : élément de la boîte à outils samtools *[14]* permettant de marquer et, si besoin, de retirer les alignements SAM résultant d'une duplication PCR des fragments dans les librairies.

Sequenza : Sequenza *[15]* est un logiciel de détection de CNV basé sur la détection des variations de couverture en lecture et de fréquence des allèles-B. Nous l'avons encapsulé dans l'outil Galaxy snvtocnv.

SIFT : SIFT *(sorting inrolerant from tolerant)* est une base de données de prédictions de l'effet des substitutions d'acides aminés sur les fonctions des protéines. Elle est associée à l'outil informatique Oncotator *[16]* utilisé pour effectuer ces prédictions, et améliorée plus récemment avec l'outil SIFT 4G *[17]* qui tire parti de processeurs graphiques.

SnpEff : SnpEff *[18]* est un outil d'annotation des variants génétiques et de prédiction de leurs effets. Il a été adapté pour une utilisation dans Galaxy.

SnpSift : SnpSift est un outil permettant de filtrer et de manipuler des bases de données d'annotations ou de prédictions. Il est inclus comme utilitaire dans le logiciel principal SnpEff, et une version de l'outil est disponible pour Galaxy.

SNV : *single nucleotide variants*.

Stockage : les exigences de stockage pour les données NGS et leurs traitements dépendent de la couverture de séquençage et du nombre de traitements effectués pour détecter des variants. Par exemple, un fichier BAM pour un génome entier couvert 30 fois (WGS) a une taille d'environ 90 gigaoctets (Go). Pour une paire d'échantillons tumeur (couv ×30)/normal (couv ×90), 2 fichiers BAM de 90 et 220 Go doivent être stockés, auxquels s'ajoutent plusieurs fichiers VCF générés par l'analyse, soit environ 320 Go par patient.

Variant calling : opération effectuée par les programmes "*variant callers*".

Variant callers : programme de détection de variants à partir d'un fichier d'alignement BAM. Pour une revue, voir *[19]*.

VCF *(variant Calling Format)* : les outils de détection de variants retournent un fichier au format standard VCF, décrivant les conditions du variant *calling*, les positions des variants et leurs caractéristiques génomiques et génétiques. Si les cliniciens ne devaient maîtriser parfaitement qu'un seul format de fichier d'analyse bioinformatique, ce serait indubitablement le VCF. Voir le document de référence décrivant le standard du VCF.

Varscan : Varscan *[20]* est un *variant caller* adapté à la recherche de variants somatiques à partir d'alignements BAM d'échantillons normal et tumoral. Il a été adapté à l'environnement Galaxy. Notre ligne de commande indicative pour paramétrer Varscan est :

```
varscan.py --normal 'normal.bam' --tumor 'tumor.bam' --normal-purity 1.0 --tumor-purity 0.32 --ofile 'snv_and_indels.vcf' --threads 4 --verbose '/genomes/hg38.fa'
```

Workflow : ensemble de traitements informatiques reliés entre eux par des données en entrée et/ou en sortie, la sortie ou production d'une étape précédente devenant l'entrée d'une étape suivante. Un *workflow* peut être linéaire ou ramifié, ouvert (une donnée en entrée générant plusieurs sorties finales) ou fermé (plusieurs données en entrées se combinant pour une sortie finale). Pour être réutilisable, un *workflow* est formalisé sous la forme d'un script composé avec un code standard et décrivant la totalité des outils, paramètres utilisés et liens entre ses différentes étapes.

Références

1. *Wang K et al. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164.*

2. *Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN]. 2013. http://arxiv.org/abs/1303.3997*

3. *Musich R et al. Comparison of short-read sequence aligners indicates strengths and weaknesses for biologists to consider. Front Plant Sci 2021;12:657240.*

4. *Hwang S et al. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep 2015;5:17875.*

5. *Jung HS et al. Utilization of the oncoscan microarray assay in cancer diagnostics. Applied Cancer Research 2017;37:1.*

6. *Liu X et al. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med 2020;12(1):103.*

7. *Bolger AM et al. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114-20.*

8. *Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal 2011. http://journal.embnet.org/index.php/embnetjournal/article/view/200*

9. *Chen Y et al. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience 2018;7(1):1-6.*

10. *Sudmant PH et al. An integrated map of structural variation in 2,504 human genomes. Nature 2015;526(7571):75-81.*

11. *Layer RM et al. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 2014;15(6):R84.*

12. *Adzhubei I et al. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet 2013;Chapter 7: Unit7.20.*

13. *Li H et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25(16):2078-9.*

14. *Favero F et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol 2015;26(1):64-70.*

15. *Ramos AH et al. Oncotator: cancer variant annotation tool. Hum Mutat 2015;36(4):E2423-9.*

16. *Vaser R et al. SIFT missense predictions for genomes. Nat Protoc 2016;11(1):1-9.*

17. *Cingolani P et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 2012;6(2):80-92.*

18. *Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J 2018;16:15-24.*

19. *Koboldt DC et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012;22(3):568-76.*