Rapport de stage de deuxième année de Master

Transposition de l'index génomique de la méthodologie Agilent/SurePrint G3 vers Affymetrix/Oncoscan CNV : comparaison d'outils bioinformatiques

Auteur :

Bordron Élie
Étudiant en Master de Bioinformatique
Parcours biologie computationnelle
Promotion 2020-2021
Université de Bordeaux

Directrices de stage :

Darbo Elodie, PhD, Ingénieure de recherche
Larmonier Claire, PhD, Ingénieure biologiste

13 juin 2022

Sommaire

Tables des figures

I Introduction

I.1 Contexte

Le diagnostic et le pronostic d'une lésion cancéreuse reposent sur des critères cliniques et anatomopathologiques. Pour le diagnostic, dans les cas difficiles, le laboratoire dispose de technologies de screening moléculaire, permettant l'identification d'altérations moléculaires spécifiques ou caractéristiques du diagnostic. Ces données moléculaires peuvent permettre également d'orienter les choix thérapeutiques, en particulier vers des thérapies ciblées.

L'unité de Pathologie moléculaire est dotée d'un panel de technologies de screening de biologie moléculaire et de cytogénétique, permettant l'identification d'altérations génétiques spécifiques contribuant au diagnostic, au pronostic ou l'identification de cibles thérapeutiques. Grâce à son expertise et l'accès à des technologies innovantes (figure 1), l'unité participe activement aux activités de routine et de recherche.

- Genomic material extraction from FFPE tissues as well as liquid biopsies
- SANGER/ NGS Sequencing technique
- QPCR/RT-PCR / Droplet Digital PCR
- Molecular signature; EndoPredict (Myriad); CINSARC signature (Affymetrix, Clariom chip)
- Idylla technology (Biocartis)
- CGH Array (Affymetrix Oncoscan and Cytoscan, Agilent CGH technology)
- NGS RNASeq ciblé ARCHERDX; RNAseq total Trusight Affymetrix
- FISH: Fluorescent *in situ* hybridization

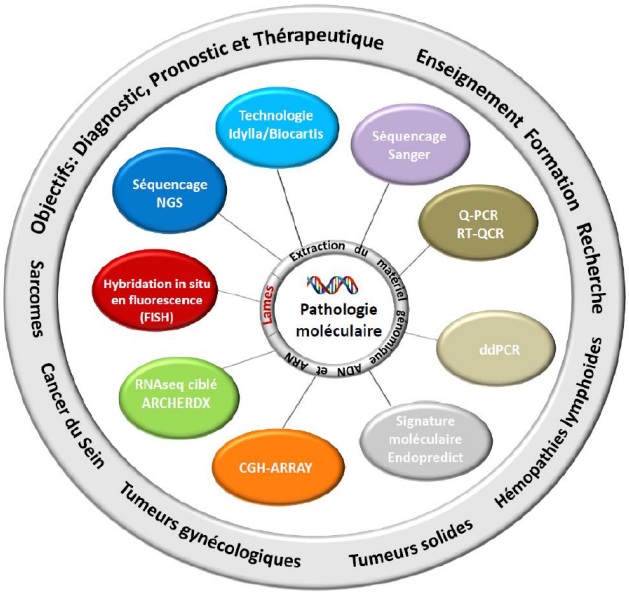

FIGURE 1 – Représentation schématique des technologies présentes dans l'unité de pathologie moléculaire, au sein du département de biopathologie de l'institut bergonié (source : support de communication de l'institut Bergonié)

I.2 Stratégie

Les génomes des cellules cancéreuses contiennent souvent des altérations chromosomiques complexes telles que des variations du nombre de copies (pertes, gains, points de cassures) qui peuvent conduire à la formation et à la progression des tumeurs. L'étude de ces anomalies chromosomiques et de l'instabilité génomique fournit des informations précieuses sur la biologie de ces tumeurs, leur évolution et leur résistance aux traitements. Parmi les technologies citées en figure 1, l'analyse globale du génome à la recherche d'anomalies de nombre de copies utilise la technique de CGH-array (array-based Comparative Genomic Hybridization).

Plus précisément, deux méthodologies sont utilisées : SurePrint G3 d'Agilent et OncoScan CNV d'Affymetrix. Le principe repose sur la capacité de 2 séquences d'acides nucléiques, complémentaires et antiparallèles, à s'apparier entre elles. L'ADN extrait du tissu à analyser est fragmenté à l'aide d'enzymes de restriction. Les fragments sont marqués à l'aide d'un traceur (biotine ou fluorochrome) puis hybridés sur un support. Les signaux quantifiés par un scanner sont comparés à ceux d'un ADN normal, utilisé comme référence. Les signaux sont analysés après calcul des Log2 ratios relatifs (LRR). Pour mettre en évidence ces relations et identifier des anomalies de continuité, les LRR sont ordonnés selon leur position génomique, puis analysés à l'aide d'algorithmes de segmentation. Leur principe est d'identifier des points de cassure témoignant de changements de niveau dans les signaux. Ces points de cassure délimitent alors des segments à l'intérieur desquels le signal peut être résumé par la moyenne des LRR des sondes qu'ils contiennent.

Les régions anormales sont qualifiées de gagnées, ou amplifiées, lorsque leur nombre de copies est supérieur à 2 (nombre de copies attendu dans un ADN normal), et «délétées» lorsqu'il est inférieur à 2. Après calcul des ratios sonde-à sonde, Log transformation et segmentation, les régions d'intérêt seront donc celles présentant un LRR > 0 pour les régions en gain, ou < 0 pour celles en perte [1].

À titre d'exemple, la plateforme Affymetrix propose une hybridation simple : seul l'ADN analysé est couplé à un fluorochrome, puis hybridé. L'analyse se fera par comparaison à un ADN de contrôle hybridé relativement à une base d'ADN virtuelle. Basée sur la technologie de la sonde d'inversion moléculaire (MIP) [2], OncoScan CNV Assay fournit une couverture du génome entier avec une résolution accrue dans tous types de cancers [3] [4] [5] [6]. Elle constitue la plateforme de choix pour la détection des variantes structurelles (VS) de l'ADN telles que les insertions, duplications et délétions chromosomiques. Outre ces variations du nombre de copies (CNV), des puces de l'ensemble du génome qui couvrent non seulement les régions polymorphes (SNP), mais également des régions non polymorphes peuvent détecter des déséquilibres chromosomiques et un déséquilibre allélique indiquant l'absence d'hétérozygotie (AOH), la perte d'hétérozygotie (LOH), ou de longues extensions contiguës d'homozygotie (LCSH). L'avantage de cette technique est la faible quantité d'ADN nécessaire ($\geq$ 80ng d'ADN), contrairement à la technologie Agilent ($\geq 1\mu g$) ainsi que l'utilisation de sondes de petites tailles permettant l'application de cette technique à des ADN partiellement dégradés (échantillons FPPE ; formalin fixed paraffin embedded tissus). L'analyse d'un profil moléculaire par CGH permet globalement :

- Le ratio en log2
- Le ratio en log2 lissé
- La fréquence allélique de l'allèle B (BAF : B allele frequency)
- Le nombre de copies pour chaque région de la puce (copy number)
- La perte d'hétérozygotie par région (LOH)
- La différence allélique ; proportion de l'allèle B par rapport à l'allèle A
- Détail des gènes présents dans chaque région

Cette analyse peut être réalisée à l'aide d'un logiciel dédié (figure 2).

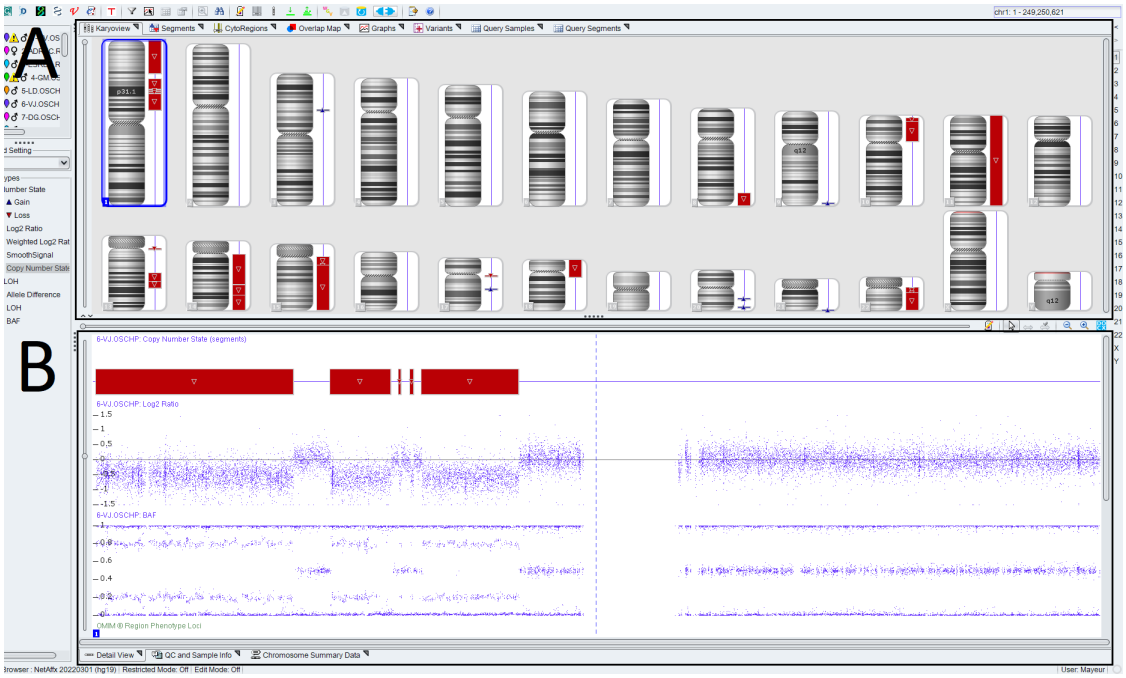

FIGURE 2 – Représentation d'un profil CGH sur le logiciel ChAS (Affymetrix, actuellement utilisé pour l'analyse des profils Oncoscan).
A : idéogrammes représentant les chromosomes humains et les régions altérées (en rouge : segments de perte, en bleu : segments de gain). B : Pistes de trois mesures de l'échantillon. En haut : nombre de copies estimé, au milieu : log Ratio, en bas : BAF. Chaque point correspond à une valeur pour une position sur le génome. Dans le panneau A, le rectangle bleu indique la région du génome que l'utilisateur souhaite agrandir pour obtenir une vue détaillée présentée dans le panneau B. Dans ce panneau, les points bleus représentent les positions génomiques mises en évidence par les sondes sur la CGH. Les segments altérés représentés dans le panneau A sont indiqués sur la vue détaillée du panneau B. Au milieu du panneau, la région dépourvue de points correspond au centromère, le point de jonction des bras chromosomiques, que la technologie oncoScan CNV ne couvre pas.

En parallèle de signatures moléculaires plus complexes, le laboratoire a précédemment démontré l'intérêt diagnostique et pronostique de l'index génomique (GI) par CGH-array dans différents types de sarcomes comme les GIST [7] [8] , les synovialosarcomes [9] et les tumeurs musculaires lisses de l'utérus [10]. Le GI est défini par le nombre d'altérations de nombre de copies[2] sur le nombre de chromosomes qui les portent. Ce « genomic index » est le reflet direct du degré de complexité moléculaire et d'instabilité génomique de la tumeur et s'avère être un puissant prédicteur de l'agressivité tumorale et de la rechute métastatique des tumeurs. Ces études ont été réalisées avec les microarrays 8x60 K whole genome d'Agilent, à partir de matériel fixé en formol et inclus en paraffine ce qui rend pertinent et accessible leur utilisation en clinique

I.3 Objectifs

Dans le but d'étendre l'utilisation de l'évaluation de l'index génomique (GI) par CGH, le laboratoire souhaiterait transposer l'approche validée précédemment sur puces AGILENT/SurePrint G3 8x60 K à la technologie Affymetrix/Oncoscan plus récente, plus résolutive et demandant moins de matériel moléculaire, également déjà utilisée au laboratoire (figure 3). Les 2 technologies n'ayant pas du tout la même couverture du génome, il est nécessaire de les comparer sur une série de tumeurs parmi les cas précédemment publiés et pour lesquels nous disposons de toutes les données cliniques et biologiques.

Fournisseur	**AGILENT**	**AFFYMETRIX**
Intitulé	SurePrint G3 Human CGH Microarray Kit, 8x60K	Affymetrix Kit, OncoScan CNV FFPE Assay Bundle
Référence	REF: G4450A	REF: 902695
Spécificité	**Hybridation génomique comparative entre l'échantillon patient et 1 ADN de référence** • coupure enzymatique 50 à 400 pb • 60 000 sondes • Sondes de 60 pb • 8 échantillons / lames • Hybridation d'une sondes tous les 33-41 kb	**Hybridation de 2 puces (AT et GC) par ADN tumoral et Comparaison à une librairie de référence** • 217 454 sondes • Sondes de 25pb • Résolution de 50kb à 125 kb sur 900 gènes • Résolution des LOH à travers le génome ≤10MB • Définition des CN
Analyse	Cytogenomics Software	ChAs software

FIGURE 3 – Spécificités des technologies OncoScan et Agilent comparées.

Pour répondre à cette problématique, nous avons 1/ Evalué des outils bioinformatiques permettant de calculer le GI à partir des données OncoScan, et 2/ Comparé leurs spécificités et leurs performances. Cela inclut le temps de calcul, mais aussi la sensibilité et la spécificité des outils dans l'identification d'altérations de nombre de copies. Enfin, nous avons 3/ corrélé les valeurs de GI définies par un outil et les valeurs obtenues par la méthodologie validée.

I.4 Spécificité de mon stage : Interaction Bioinformatique-Biologie

Ce travail s'inscrit dans le projet GIRONDE «Génomique Index Resolution by ONcoscan Definition and Expertise », pour lequel le laboratoire a obtenu un financement industriel. Ce projet a un caractère concret grâce à mon intégration dans l'unité de pathologie moléculaire. J'ai ainsi pu suivre le parcours des échantillons, de leur arrivée au laboratoire, au suivi d'une technique de CGH et à l'analyse des résultats. J'ai à la fois interagi avec l'équipe de biologistes pour mieux cerner leur préoccupation quotidienne et avec l'équipe bioinformatique du departement de Biopathologie pour comprendre les enjeux de la mise en place de nouveaux outils bioinformatiques destinés à l'analyse de routine. Lors de réunions de présentation hebdomadaires, un travail de vulgarisation des outils bioinformatiques testés a ainsi été fait pour souligner les points d'intérêt de chacun

d'eux. J'ai également participé avec l'équipe à la formation à l'utilisation du logiciel CHAS, dispensée par l'ingénieur d'application Affymetrix en mai.

II Matériel et Méthodes

II.1 Matériel

Sur la base des travaux précédemment publiés [7] [8], le laboratoire souhaite valider l'index génomique défini par la technique Oncoscan, dans un 1er temps dans les tumeurs stromales gastro-intestinales (GIST). En effet, le laboratoire est impliqué dans un protocole de recherche clinique ayant pour objectif l'évaluation de l'efficacité d'un traitement adjuvant pour les tumeurs GIST de risque intermédiaire présentant un GI de mauvais prognostic [7]. Pour ces tumeurs et sur la base du calcul de l'index génomique, les patients présentant un GI$\geq$10 peuvent bénéficier d'un traitement adjuvant. Un protocole de surveillance est mis en place pour les patients avec un GI < 10.

Les échantillons utilisés sont récents. Ils ont été choisis pour représenter la variabilité de ce sous-type tumoral.

Les logiciels utilisés sont :

— ChAS version 4.3

— R version 4.1.2[11]

— Rstudio version 2021.09.2+382 "Ghost Orchid" [12]

Les packages R utilisés sont :

— oncoscanR version 0.1.1 [13]

— CGHcall version 2.56.0 [14]

— ASCAT version 3.0.0 [15]

— rCGH version 1.24.0 [16].

II.2 Méthodes

Dans le but de calculer le GI de manière automatisée à partir de la technologie OncoScan, on cherche un outil pouvant déterminer les altérations de nombre de copies à partir des données log ratio et BAF. Les outils sélectionnés permettent le calcul des altérations de nombre de copies, et certains utilisent les mêmes méthodes. Une comparaison détaillée de leurs différences est nécessaire pour apporter des réponses à ce projet. Une recherche bibliographique a été menée pour établir cette comparaison

■ Chromosome Analysis Suite

Le logiciel propriétaire d'Affymetrix, Chromosome Analysis Suite (ChAS), détermine les altérations de nombre de copies (fig.4) et permet de les visualiser sur le génome de manière interactive grâce à une interface graphique. Cette dernière est représentée par la figure 2 de l'introduction. Le profil log Ratio peut être séparé en plusieurs régions de valeur 0 et -0,5. Les valeurs de BAF varient sur les mêmes positions.

Le logiciel ChAS interprète les régions de log Ratio négatif par la perte d'une copie chromosomique et déclare ces régions comme étant des segments altérés (en rouge dans la piste de nombre de copies). Il est possible de manipuler le profil en le recentrant ou en fusionnant ces segments, ou de rechercher des gènes spécifiques

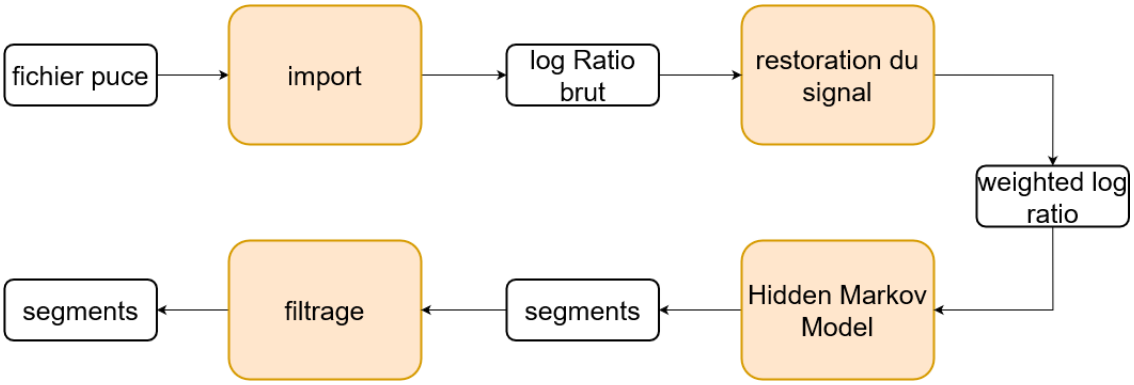

FIGURE 4 – Le Pipeline du logiciel ChAS qui détermine les altérations de nombre de copies.
Boîtes blanches, données. Boîtes orange, étapes de traitement des données. Hidden Markov Model : Modèle de Markov caché. Weighted log Ratio : log Ratio pondéré.

dans des segments altérés. Ces fonctionnalités du logiciel ChAS ont un intérêt pour une analyse optimale des altérations et sont utilisées en routine dans l'unité de pathologie moléculaire.

Pour plusieurs raisons, le logiciel ChAS ne peut pas entièrement répondre aux besoins de ce projet. D'abord, le calcul du GI ne peut pas être automatisé avec ChAS. Le recentrage et la fusion des segments, s'ils ont lieu d'être effectués, restent des étapes manuelles. Or, l'automatisation permettrait de réduire la part de subjectivité dans ce calcul. Pour un score qui peut influencer la prise de décision, cela peut avoir un intérêt.

Ensuite, le logiciel n'offre pas une vue en détail sur les méthodes de calcul qu'il emploie et les traitements qu'il applique aux données.

Les fonctionnalités utiles de ChAS en font un outil largement utilisé en routine, mais la problématique de ce travail nécessite un outil complémentaire. On étudie quatre alternatives libres de droits qui sont tous des packages R.

■ oncoscanR

OncoscanR est un package R qui détermine les altérations de nombre de copies à l'échelle des bras chromosomiques en deux étapes (fig.5).

D'abord, les segments altérés déterminés par le logiciel ChAS subissent un nettoyage : les segments de moins de 300 Kbp (milliers de paires de bases) sont supprimés (fig. 6A), puis les segments séparés de moins de 300 Kbp sont fusionnés (fig. 6B). L'intérêt du filtrage est de supprimer les segments artefacts du calcul afin d'éviter que le pourcentage de bras altéré ne soit sur-estimé. L'avantage du lissage s'exprime dans le cas particulier où, après filtrage, un artefact a été supprimé et une région non altérée de moins de 300 000 paires de bases a été créée entre deux segments de même altération, comme les deux segments de gain les plus à gauche de la figure 6B.

Comme ils sont séparés de moins de 300Kbp, OncoscanR considère que ces deux segments représentent la même altération, donc que la région qui les sépare appartient à cette altération. Ils sont alors fusionnés et la surface que recouvrait la région vide

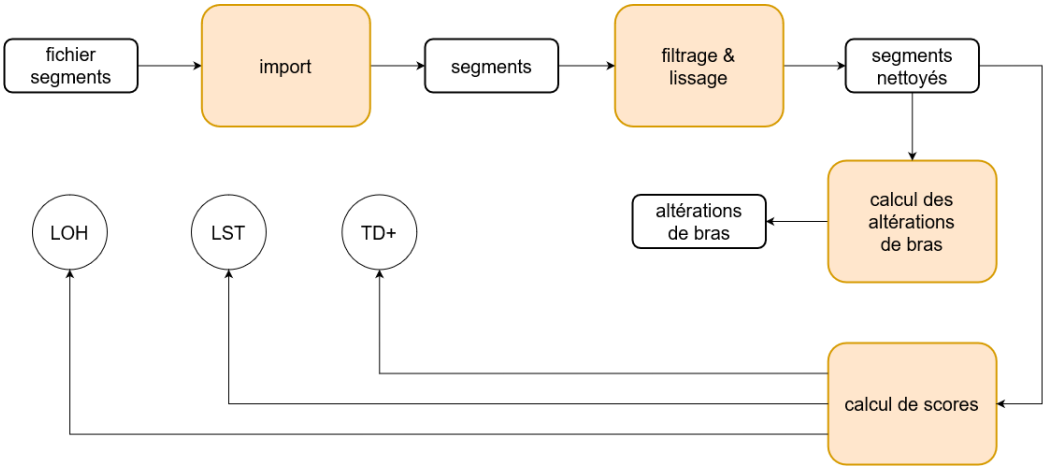

FIGURE 5 – Le Pipeline d'OncoscanR qui détermine les altérations de nombre de copies. Boîtes blanches, fichiers de données ou objets R. Boîtes orange, étapes de traitement des données. LST, Large-Scale state Transition. LOH, Loss Of Heterozygosity. TD+, Tandem Duplication positive.

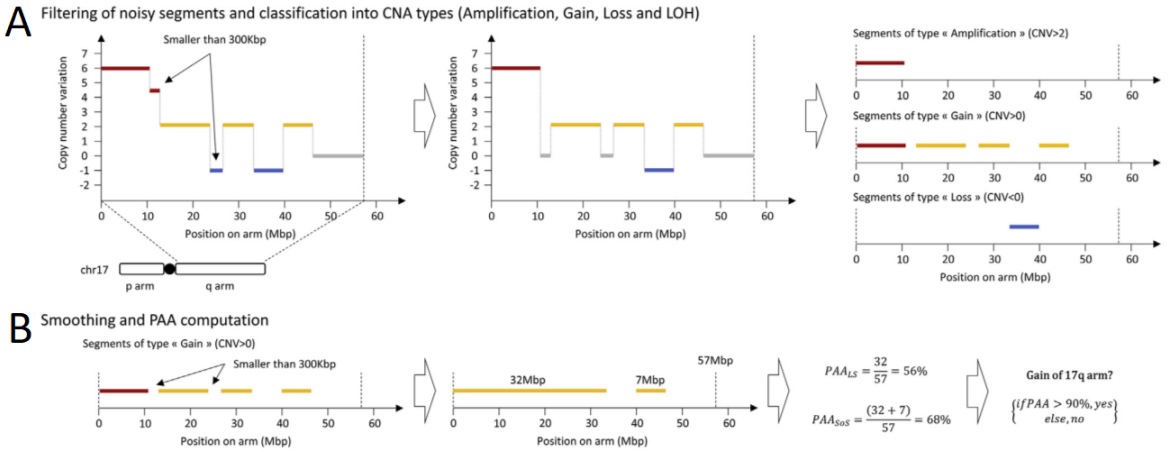

FIGURE 6 – Nettoyage des données et calcul des altérations par oncoscanR.
A : filtrage des segments de bruit de fond et classification en types d'altérations de nombre de copies (Amplification (rouge foncé), Gain (jaune), Perte (bleu)). Copy number variation (CNV) : Variation du nombre de copies. Position on arm (Mbp) : Position sur le bras en millions de paires de bases. chr17, p arm, q arm : chromosome 17, petit bras, grand bras. B : lissage des segments et calcul du pourcentage de bras altéré (PAA). LS, Longest Segment (segment le plus long). SoS, Sum of Segments (somme des segments). Image extraite de Christinat et al [13]

est bien prise en compte comme une altération, ce qui augmente le pourcentage de bras altéré. Sans cette étape, le pourcentage de bras altéré aurait été (légèrement) sous-estimé.

Ensuite, les segments altérés sont utilisés pour calculer les altérations de bras.

Les altérations de nombre de copies sont généralement classées, selon la longueur du segment altéré, en altérations focales ou de bras chromosomiques. Une altération

focale est courte et peut être, par exemple, liée à la perte de gènes suppresseurs de tumeurs ou au gain d'oncogènes, tandis qu'une altération de bras est plus large et contient des centaines de gènes. La définition d'une altération de bras ne fait pas consensus au sein de la littérature : généralement, on considère qu'elle correspond à une unique altération couvrant une grande part du bras[17], mais le seuil qui détermine à quel pourcentage d'altération le bras est considéré comme altéré varie selon les études[17],[18]. La procédure implémentée dans oncoscanR se veut applicable dans le contexte clinique au cas par cas : la somme des segments altérés est utilisée pour calculer le pourcentage de bras altéré (Percentage of Arm Altered, PAA), et le bras est dit altéré si le PAA est supérieur à 90%.

En parallèle, oncoscanR fournit le calcul de scores moléculaires d'intérêt dans la caractérisation des tumeurs. Les scores LOH et LST sont lié à la mutation des gènes BRCA[19],[20], et le score TD+ est lié à la mutation du gène CDK12[21]. La perte de ces gènes est liée au développement de tumeurs.Le score LOH correspond au nombre de segments en perte d'hétérozygotie (Loss Of Heterozygosity, LOH) de plus de 15Mbp (millions de paires de bases), en excluant les chromosomes en LOH sur toute leur longueur. Le score LST correspond au nombre de transition d'état à grande échelle (Large-scale State Transition, LST). Un LST est un point de cassure (breakpoint) entre deux régions de plus de 10 Mbp chacune. Le score TD+ correspond au nombre de segments concernés par une duplication en tandem (Tandem Duplication, TD) entre 1 et 10 Mb. Une TD est la duplication d'un exon au sein d'un gène.

■ rCGH

Le package rCGH détermine les segments d'altération en quatre étapes (fig.7) et permet une étape de visualisation interactive du profil obtenu.

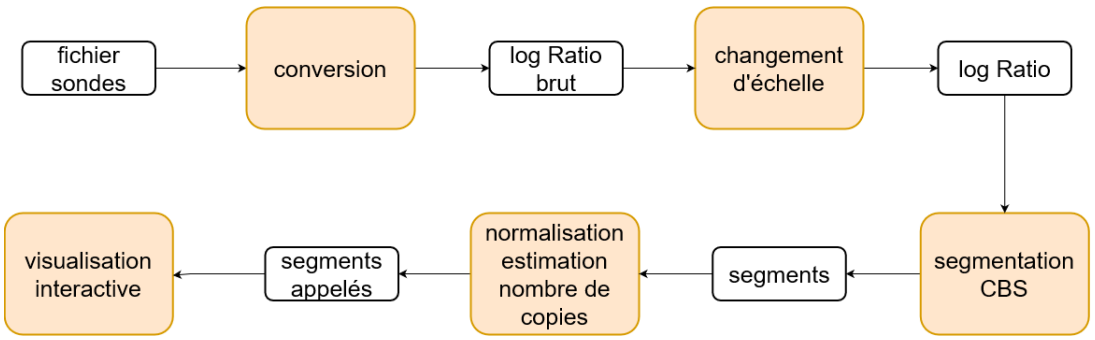

FIGURE 7 – Le Pipeline de rCGH qui détermine les altérations de nombre de copies. Boîtes blanches, fichiers de données ou objets R. Boîtes orange, étapes de traitement des données et visualisation interactive. CBS, Circular Binary Segmentation.

Un changement d'échelle est d'abord effectué sur les données de log ratio (figure 8). La dispersion des valeurs de log ratio se fait en divisant chaque valeur par la moyenne quadratique de l'ensemble du profil. Cette étape permet d'amener les données à une échelle plus facilement utilisable par les étapes suivantes.

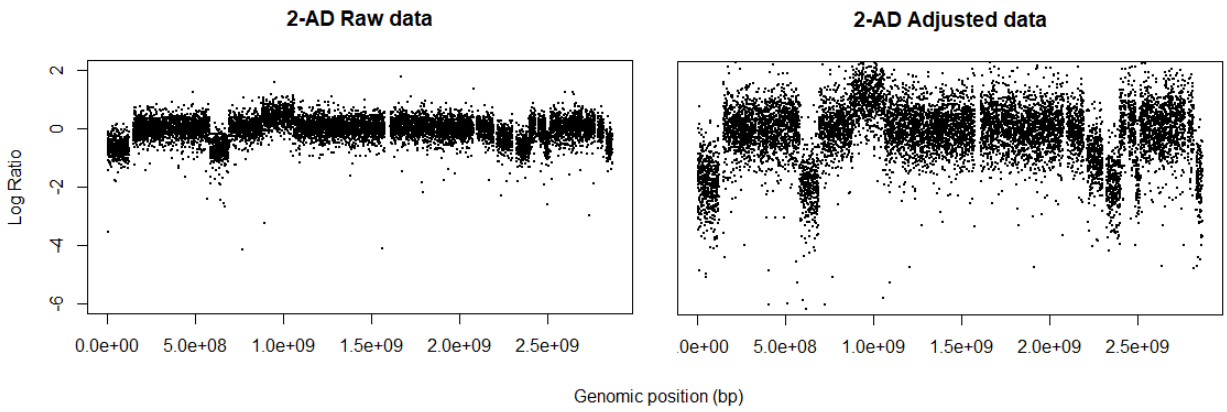

FIGURE 8 – Changement d'échelle des valeurs de log Ratio par rCGH.
À gauche, le profil brut. À droite, le profil mis à l'échelle. Genomic position (bp) : Position génomique en paires de bases. 5-LD : nom de l'échantillon. Raw data : Données brutes. Adjusted Data : Données ajustées.

Ces données sont ensuite groupées en régions de log ratio similaire appelées segments. Cette étape de segmentation, dans rCGH, utilise l'algorithme Circular Binary Segmentation[22] (CBS), qui est parmi les plus performants [23]. Son principe est le suivant : Pour une région donnée (ici de 140Kbp à 200Kbp), une fenêtre glissante (en vert) parcourt toutes les positions et cherche la sous-région au log Ratio moyen le plus différent possible du reste de la région (fig.9B).

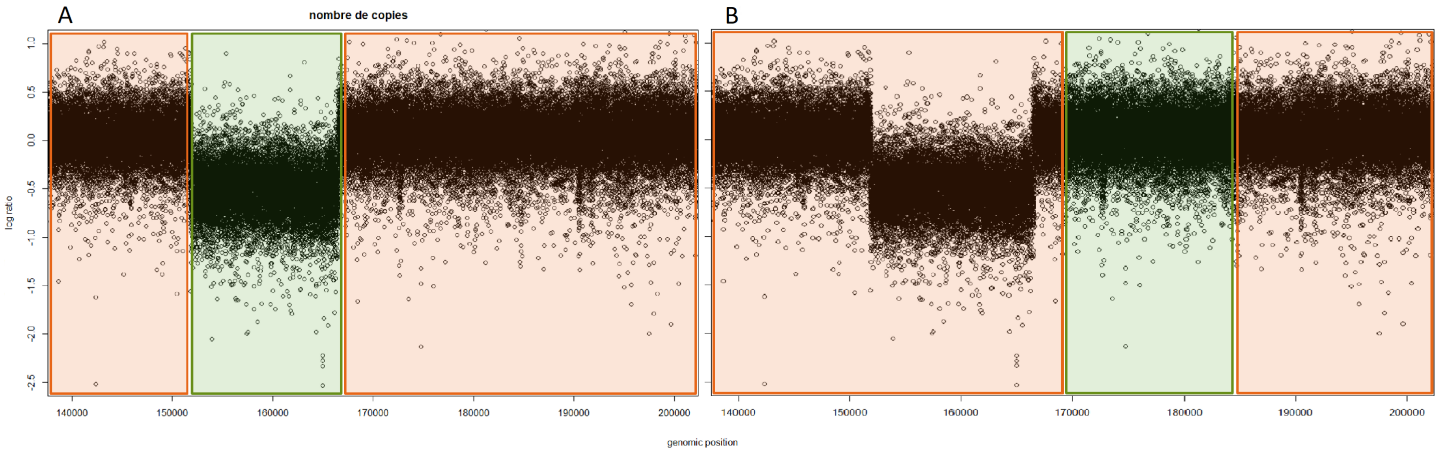

FIGURE 9 – Fonctionnement par fenêtre glissante de l'algorithme CBS.
En vert : fenêtre coulissante. En orange : le reste de la région. A et B correspondent à la même région avec des positions différentes de fenêtre coulissante. En noir : log Ratio.

Cette opération est répétée pour toutes les tailles de fenêtre glissante de 1 à l, où l est la taille de la région en paires de bases. Là où la différence de log Ratio entre la sous-région et le reste de la région est la plus grande (fig.9A), des points de coupure sont créés pour séparer les régions. Trois segments sont alors créés. L'opération est répétée récursivement sur ces derniers jusqu'à ce qu'aucun segment ne puisse

plus être créé. Les segments ont une valeur unique de log Ratio qui correspond à la médiane des valeurs que leur région contient. Les premières régions traitées sont des chromosomes ou des bras chromosomiques, choisis selon la résolution d'oncoScan CNV.

Certains points de cassure correspondent à des tendances locales qui ne reflètent pas la segmentation réelle du profil. Pour produire la meilleure segmentation, ces points de cassure sont annulés selon certaines règles. Premièrement, si un segment court est suffisamment proche d'un autre segment, alors ils sont fusionnés. Deuxièmement, si deux segments longs sont suffisamment proches, ils sont également fusionnés. La longueur d'un segment court est par défaut de 10 points. Pour la première règle, undo.SD est la distance qui définit "suffisamment proche" : si undo.SD vaut 3, la fusion est effectuée quand les valeurs des segments sont séparées de moins de trois fois l'écart-type calculé sur les deux segments. Pour deux segments longs, cette distance correspond à undo.SD divisé par un autre paramètre, relSDlong. Augmenter la longueur qui définit un segment court peut faire s'appliquer l'une des deux règles plus souvent que l'autre, ce qui change la segmentation. Augmenter undo.SD laisse plus de points de cassure, et faire varier relSDlong permet d'orienter leur annulation : ils sont annulés plus souvent soit pour les petits segments, soit pour les grands. Utiliser les valeurs par défaut de cet algorithme peut être limitant dans l'étude de cas hautement remaniés ou qui présentent beaucoup de bruit de fond. rCGH estime la valeur du paramètre undo.SD à partir du bruit des données afin de mieux s'adapter à chaque profil.

Les données de segmentation sont ensuite normalisées. Cette étape met en oeuvre un modèle statistique appelé modèle de mélange gaussien pour déterminer le niveau normal de deux copies et recentrer le profil sur ce niveau. Pour cela, les segments sont considérés comme les individus d'une population de distributions gaussiennes. Une distribution gaussienne, ou normale, suit une loi de probabilité continue qui dépend de deux paramètres : sa moyenne μ et son écart-type σ. Le modèle de mélange classe les segments dans ces populations selon leurs valeurs, créant par exemple trois groupes dont les pics sont visibles sur la figure 10).

Ici, le plus grand pic correspond au plus grand groupe de segments de même nombre de copies. Autrement dit, il s'agit du groupe qui a le plus de chances de correspondre au niveau normal. rCGH va donc soustraire sa valeur (-0.029) à l'ensemble du profil pour que ce groupe corresponde au niveau normal diploïde, dont le log Ratio est zéro. Cependant, pour un échantillon présentant un gain allélique sur plus de la moitié de ses positions, le plus grand pic rassemble les segments de gain. Normaliser par le pic le plus grand définit donc un niveau de gain comme le niveau normal de deux copies, et les segments du niveau normal sont alors considérés comme des altérations de perte. rCGH utilise une approche [24] qui vise à empêcher ce cas de figure d'arriver. Parmi les pics trouvés, seuls ceux dont la hauteur dépasse la moitié du plus grand pic sont retenus. De ces pics, celui qui se trouve le plus à gauche sera utilisé pour recentrer le profil. Cette méthode vise à centrer le profil plus fidèlement que par une normalisation basée sur la médiane ou la moyenne, même s'il est potentiellement très altéré à l'origine.

Le nombre de copies de chaque segment est estimé à partir du log Ratio après la normalisation. Cette estimation dépend de la ploïdie *a priori* du profil, qu'il

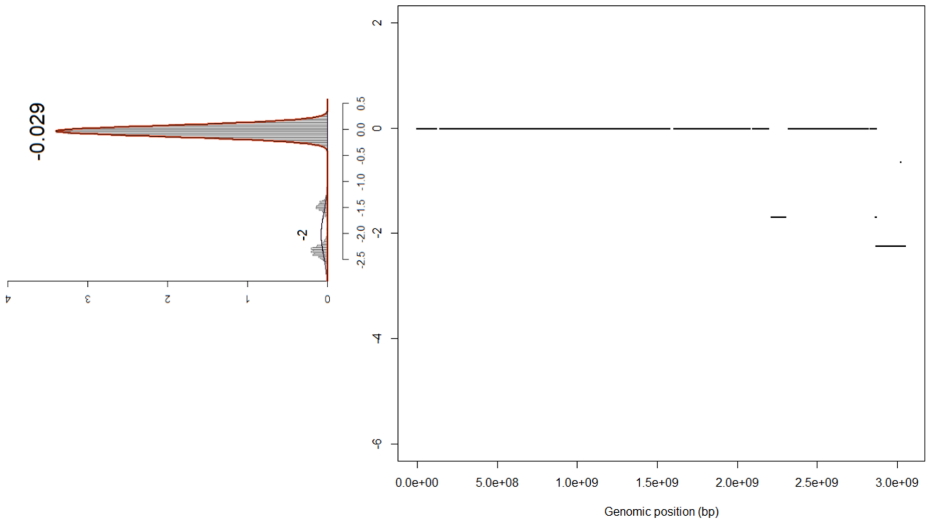

FIGURE 10 – Détermination du niveau normal par modèle de mélange gaussien de rCGH. À droite, les données segmentées. à gauche, les pics de densité leur correspondant et la valeur moyenne des segments qui les composent. La valeur en gras indique le pic qui correspond le plus au niveau normal. Genomic position (bp) : Position génomique en paires de bases.

est nécessaire de renseigner en entrée du programme. Additionnellement au calcul des altérations de nombre de copies, rCGH met à disposition une interface de visualisation interactive dans le but de pouvoir éditer un profil et utiliser différents indicateurs visuels afin de prendre les meilleures décisions quand au diagnostic (fig.11).

Cette interface permet de recentrer un profil et d'éditer les segments à l'aide du panneau de contrôle. La visualisation de plusieurs paramètres à la fois est possible, notamment la différence allélique. La différence allélique est définie, pour une position génomique qui peut présenter deux allèles A et B, par le nombre de copies de l'allèle A multiplié par 0,5 auquel on soustrait le nombre de copies de l'allèle B multiplié par 0,5. Pour une position hétérozygote à deux copies (un allèle de chaque), la différence allélique est de 0, mais une population homozygote à deux copies aura une valeur de -1 pour BB et 1 pour AA. Cela s'illustre en bas dans la figure 11 pour le chromosome 3 notamment. Il est aussi possible de trouver à quel segment appartiennent des gènes spécifiés.

■ CGHcall

CGHcall est un outil qui détermine les segments d'altération et leur attribue un nombre de copies à l'aide d'un modèle statistique, et ce, en cinq étapes (fig.12).

CGHcall prend en entrée les données de log Ratio par sonde, et commence par leur appliquer plusieurs traitements lors de l'étape de pré-traitement, qui ont pour but de préparer les données aux étapes suivantes. L'étape de pré-traitement prépare les données pour que les autres étapes du pipeline se déroulent sans problème. Ainsi, pour une cohorte de plusieurs échantillons, les données manquantes pour plus de

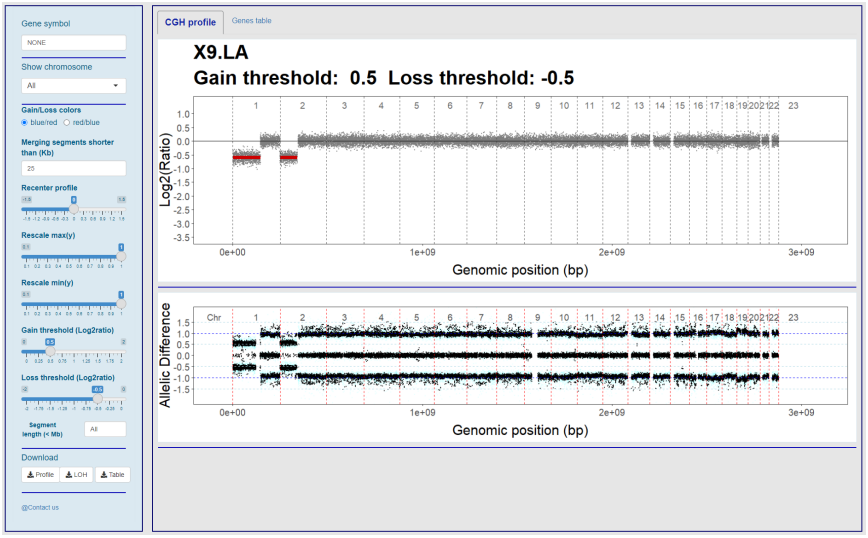

FIGURE 11 – Interface graphique de visualisation des données dans rCGH. Panneau du haut : piste log Ratio. En rouge, segments altérés ; "X9.LA", nom de l'échantillon ; Gain threshold, seuil de gain ; Loss threshold, seuil de perte ; Genomic position, position génomique. Panneau du bas : piste différence allélique (Allelic difference). À gauche : panneau de contrôle du profil.

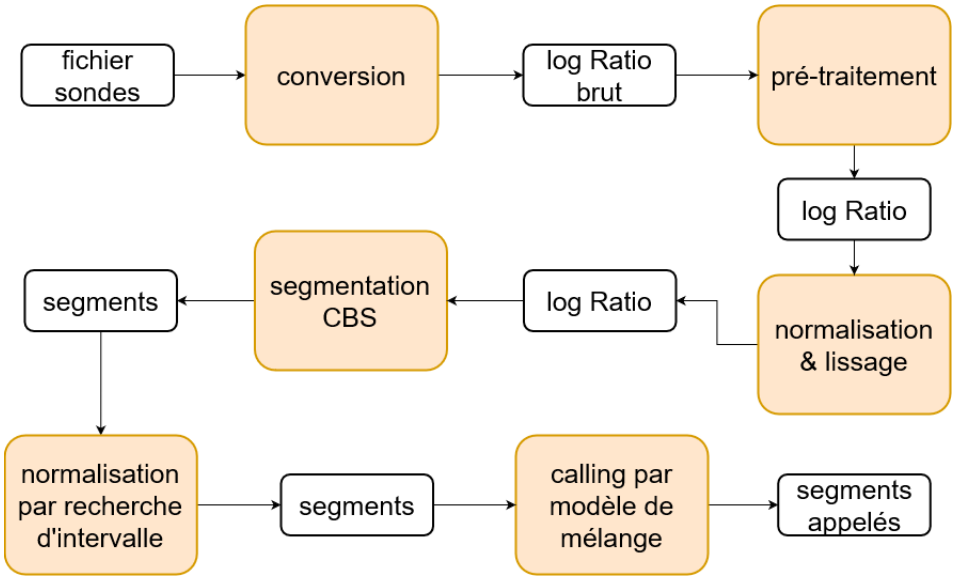

FIGURE 12 – Pipeline de détermination du nombre de copies du package CGHcall. Boîtes blanches, fichiers de données ou objets R. Boîtes orange, étapes de traitement des données. CBS, Circular Binary Segmentation.

30% des échantillons sont supprimées pour tous. (fig.13). Les données manquantes restantes sont estimées à l'aide du package R impute[25]. Les étapes suivantes n'ont donc pas à traiter des données manquantes.

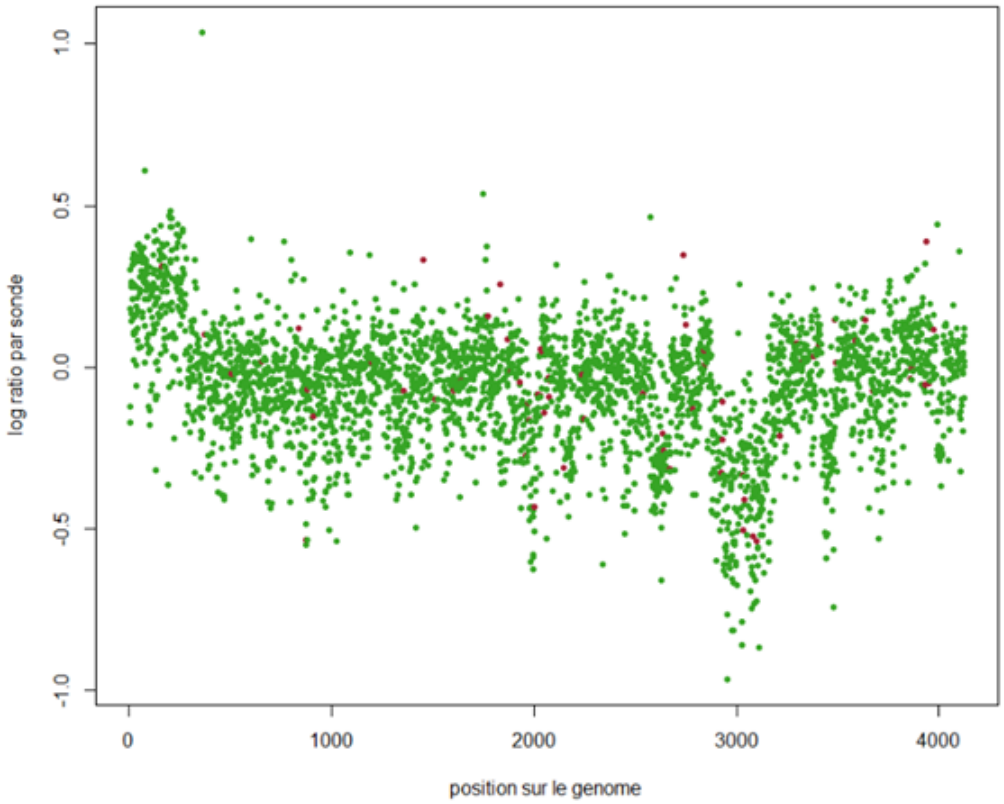

FIGURE 13 – Effet du pré-traitement sur des données de log Ratio.
En rouge, points supprimés des données. En vert, points conservés.

Les données sont ensuite normalisées par la médiane ou le mode, et subissent un lissage des «outliers »(valeurs aberrantes ; points dont la valeur est significativement différente des autres). Une telle valeur aberrante peut être définie à partir de l'écart interquartile[26].

L'étape suivante est la segmentation par l'algorithme CBS. CGHcall n'implémente pas de fonctionnalité additionnelle et utilise les paramètres par défaut.

Une deuxième normalisation est ensuite appliquée (fig.14). Elle cherche le niveau zéro de manière plus avancée que la première. Dans les données de log Ratio, l'intervalle contenant les données les plus segmentées est recherché de manière récursive . L'intervalle contenant tous les points est séparé en quatre zones, et la zone comprenant le plus de segments est à son tour séparée en quatre zones. Après cinq cycles, la valeur centrale du dernier intervalle trouvé (en rouge sur la figure) est soustraite au profil pour le centraliser.

Le nombre de copies de chaque segment est estimé lors du calling (estimation du nombre de copies). Le calling de CGHcall peut être lancé sur un échantillon individuel, mais aussi sur une cohorte d'échantillons. Si les échantillons ont une ploïdie différente, cela peut amener à sous-estimer ou sur-estimer le nombre de copies. Cette étape est effectuée par un modèle de mélange gaussien. L'intérêt d'utiliser un

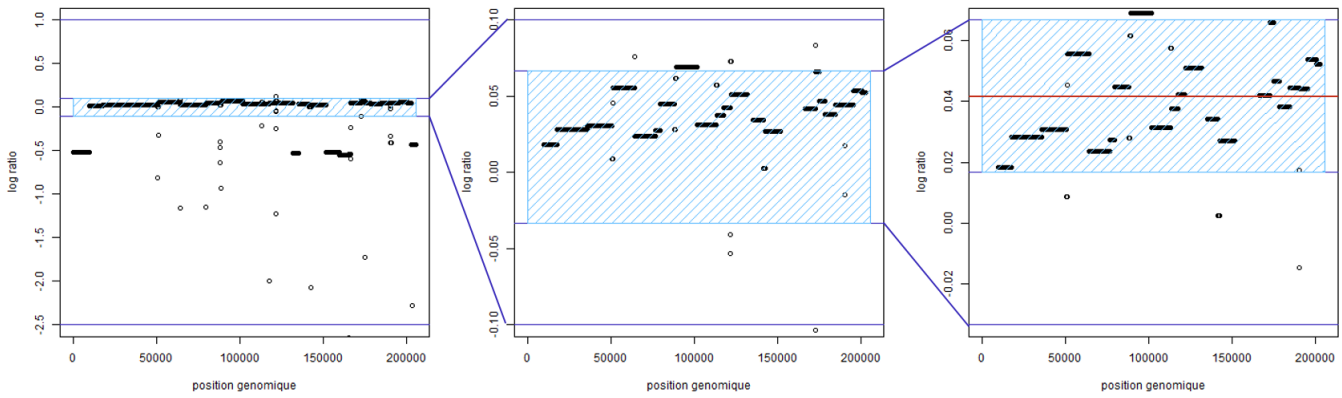

FIGURE 14 – Recherche du niveau zéro des données segmentées par la normalisation de CGHcall.
À gauche, les données de log Ratio d'un échantillon. Au milieu, agrandissement sur le meilleur intervalle trouvé. À droite, deuxième agrandissement sur le meilleur intervalle trouvé. Les zones hachurées en bleu correspondent au meilleur intervalle trouvé à chaque étape.

tel modèle est qu'il cherche à classer les segments en catégories représentant des statuts biologiques (gain, perte, normal). En effet, par rapport à une classification qui attribue à chaque segment la valeur du nombre entier le plus proche, l'estimation CGHcall conserve l'écart entre les valeurs. Par exemple, pour un nombre de copies estimé pour la plupart des segments près de 0.5, la classification de l'entier le plus proche va répartir les valeurs entre 1 et 0, et CGHcall va considérer les segments autour de 0.5 comme une même altération et les classer soit tous à 1, soit tous à 0.

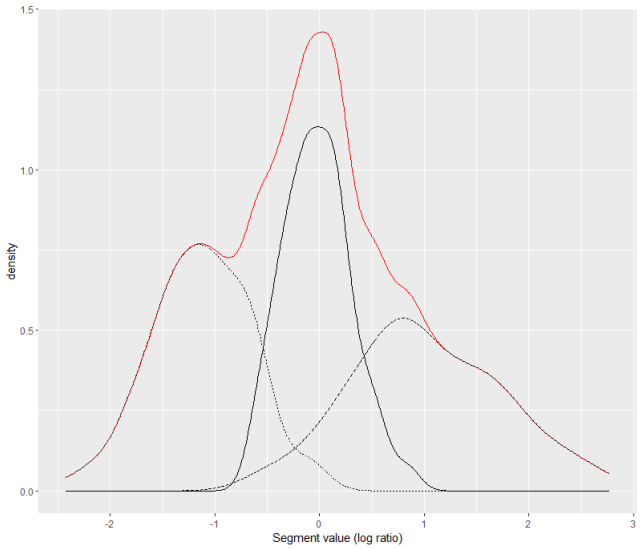

FIGURE 15 – Le calling par modèle de mélange de CGHcall.
En rouge, la distribution des segments. En noir, les groupes (déterminés par le calling) qui composent cette distribution. Ce graphe est construit à partir de données virtuelles.

Pour ce faire, les segments d' un échantillon sont mélangés pour former une unique population (courbe rouge, fig.15), et en trouvant les distributions gaussiennes sous-jacentes, le modèle classifie les segments en groupes.

Ces groupes correspondent au statut de segments : ici, trois groupes sont trouvés, dont les moyennes respectives sont -1, 0 et 1, c'est-à-dire les statuts de perte, normal et de gain, respectivement. Cette étape peut intégrer le pourcentage de cellules tumorales dans le calcul s'il est renseigné.

■ ASCAT

Le package R ASCAT détermine les altérations de nombre de copies en deux étapes (fig.16).

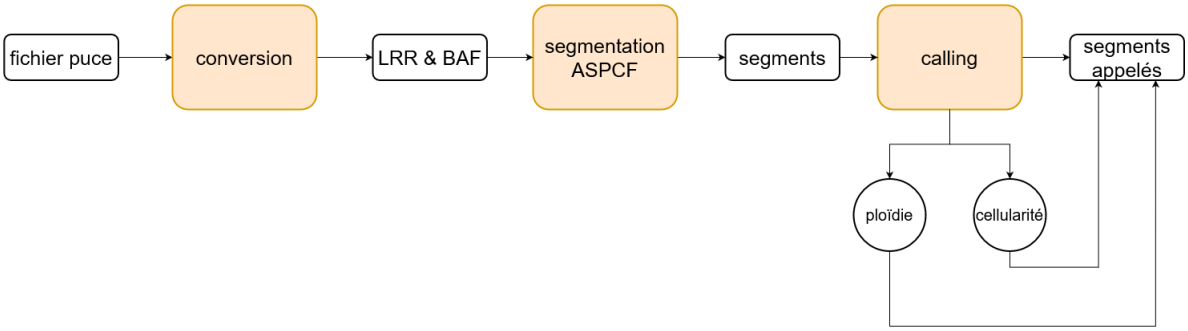

FIGURE 16 – Pipeline d'ASCAT aboutissant au nombre d'altérations de nombre de copies. Boîtes blanches, fichiers de données ou objets R. Boîtes orange, étapes de traitement des données. LRR, log Ratio. BAF, B allele frequency. ASPCF, Allele-Specific Piecewise Constant Fitting. Calling, estimation du nombre de copies.

Les données log Ratio et BAF par sonde sont d'abord extraites du fichier puce (.OSCHP) généré par le logiciel ChAS et données en entrée à ASCAT. La segmentation utilisée par ASCAT est effectuée via l'algorithme Allele-Specific Piecewise Constant Fitting (ASPCF). ASPCF ajuste des fonctions constantes sur les données log Ratio et BAF, et force les points de coupure à être présents aux mêmes positions sur les deux pistes. Utiliser le BAF additionnellement au log Ratio pour segmenter les données permet de détecter les altérations qui aboutissent à un nombre de copies normal (par exemple, la perte de l'allèle A puis le gain de l'allèle B amènent à un nombre de copies normal) et peuvent rester invisibles si seul le log Ratio est observé.

ASPCF cherche le partitionnement optimal de la région génomique en segments. Un partitionnement optimal minimise le critère d'optimisation indiqué en figure 17.

Dans cette expression, pour un segment donné, les deux termes entre crochets sont respectivement la qualité de l'ajustement (goodness of fit) sur les données log Ratio et sur les données BAF. $ave(\{r_s\}_{s\in I_j})$ représente la moyenne des données log Ratio sur le segment I. La qualité de l'ajustement, ici, est une mesure des écarts à la moyenne : elle est élevée si cet écart est grand, ce qui indique que le segment est hétérogène (fig.18, en bas à droite) ; Si les valeurs du segment sont proches de la moyenne, la qualité de l'ajustement sera au contraire plus faible (fig.18, en bas à gauche). La qualité de l'ajustement élevée alourdit le critère d'optimisation en conséquence, et si c'est le cas de nombreux segments,

$$\sum_{j=1}^{Q} \sum_{i \in I_j} \left[w(r_i - ave(\{r_s\}_{s \in I_j}))^2 + (1-w)(b_i - ave(\{b_s\}_{s \in I_j}))^2 \right] + \lambda \cdot Q$$

FIGURE 17 – Le critère d'optimisation à minimiser pour trouver la meilleure solution. w est un poids qui permet de donner plus d'importance à la qualité de l'ajustement des données log Ratio ou BAF.

le partitionnement est moins crédible qu'un partitionnement qui séparerait les segments hétérogènes en plusieurs sous-segments. Le terme en dehors des crochets est une pénalité

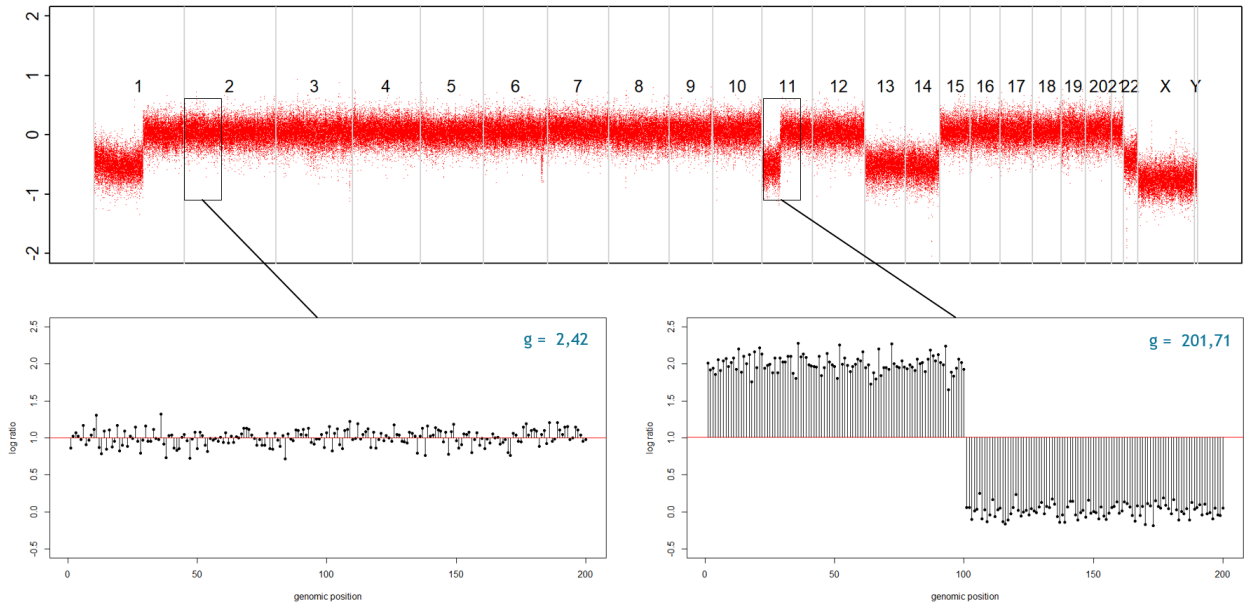

FIGURE 18 – Profil CGH à traiter par ASPCF et détail de deux de ses régions. En haut, les données de log Ratio sur le génome entier. En bas, deux régions de ce profil. La qualité de l'ajustement (g pour goodness of fit) qui serait obtenue si ces régions étaient des segments est indiquée en bleu.

qui correspond aux points de coupure trouvés : plus le nombre de segments trouvés Q est grand, plus la pénalité est grande. Ce terme empêche ASPCF de considérer que créer un segment par point de donnée est le partitionnement optimal. Le terme *lambda* permet d'ajuster l'importance de la pénalité dans l'équation. Dit autrement, la meilleure solution est celle qui génère des segments ayant chacun la plus faible variabilité possible, tout en gardant le nombre de segments modéré.

Le calling Allele-Specific Copy number Analysis of Tumors (ASCAT) de l'outil éponyme attribue ensuite un nombre de copies à chaque segment en estimant la ploïdie et la cellularité du profil. La ploïdie correspond au nombre de copies global d'un génome : normalement de 2, elle peut varier dans le cas d'un cancer. Certaines tumeurs sont ainsi triploïdes ou tétraploïdes. La cellularité tumorale correspond au pourcentage de cellules tumorales dans l'échantillon d'où les données sont extraites.

Le nombre de copies des deux allèles de chaque position est déterminé ainsi : Les

valeurs de log Ratio et BAF peuvent être exprimées en fonction du nombre de copies de chaque allèle (1,2).

$$r_i = \gamma log_2 \left(\frac{n_{A,i} + n_{B,i}}{2} \right) \tag{1}$$

$$b_i = \frac{n_{B,i}}{n_{A,i} + n_{B,i}} \tag{2}$$

Dans ces équations, r_i et b_i représentent respectivement log Ratio et BAF. Le paramètre γ est une constante qui dépend de la technologie utilisée et contrebalance la compaction que subit le log Ratio par la technique.

Pour référence, si le nombre de copies est de 2, $r_i = 0$. Aux positions hétérozygotes, $b_i = 0.5$, mais aux positions homozygotes, $b_i = 0$ (allèle A) ou $b_i = 1$ (allèle B). Dans une tumeur aneuploïde, c'est-à-dire dont la ploïdie n'est pas 2, ces équations ne sont plus vraies. Pour prendre ça en compte, la ploïdie est modélisée par le paramètre ψ, ce qui donne l'équation 3.

$$r_i = \gamma log_2 \left(\frac{n_{A,i} + n_{B,i}}{\psi} \right) \tag{3}$$

Dans le cas d'une contamination de l'échantillon par des cellules saines, le calcul du nombre de copies correspond au nombre de copies des cellules tumorales multiplié par la proportion de ces cellules, auquel s'additionne le nombre de copies des cellules saines (2, donc) multiplié par la proportion de cellules saines. Prendre en compte la cellularité (le pourcentage de cellules tumorales) à l'aide du paramètre ρ dans le calcul du log Ratio et du BAF implique donc d'adopter les équations 4 et 5.

$$r_i = \gamma log_2 \left(\frac{2(1-\rho) + (n_{A,i} + n_{B,i})}{\psi} \right) \tag{4}$$

$$b_i = \frac{1 - \rho + \rho n_{B,i}}{2 - 2\rho + (n_{A,i} + n_{B,i})} \tag{5}$$

Les équations obtenues expriment log Ratio et BAF en fonction de la cellularité ρ, de la ploïdie ψ et du nombre de copies n. À partir de cette relation, il est possible d'isoler le nombre de copies et de l'exprimer en fonction du log Ratio, du BAF, de la cellularité et de la ploïdie. (6,7).

$$\hat{n}_{A,i} = \frac{\rho - 1 + 2^{\frac{r_i}{\gamma}}(1 - b_i)(2(1-\rho) + \rho\psi)}{\rho} \tag{6}$$

$$\hat{n}_{B,i} = \frac{\rho - 1 + 2^{\frac{r_i}{\gamma}} b_i (2(1-\rho) + \rho\psi)}{\rho} \tag{7}$$

Le calling d'ASCAT utilise les équations 6 et 7 pour déterminer le nombre de copies. La cellularité et la ploïdie, qui sont des inconnues, sont déterminées itérativement : l'algorithme ASCAT les fait varier de 0.1 à 1.05 et de 1 à 6 (fig. 19), respectivement, calcule le nombre de copies pour chaque couple de valeurs, estime la qualité de l'ajustement de chaque solution, et retient la meilleure. Une solution a une bonne qualité si les valeurs de nombre de copies qu'elle calcule sont proches de nombres entiers (fig. 20).

Le nombre de copies de chaque position (et incidemment de chaque segment) est alors déterminé.

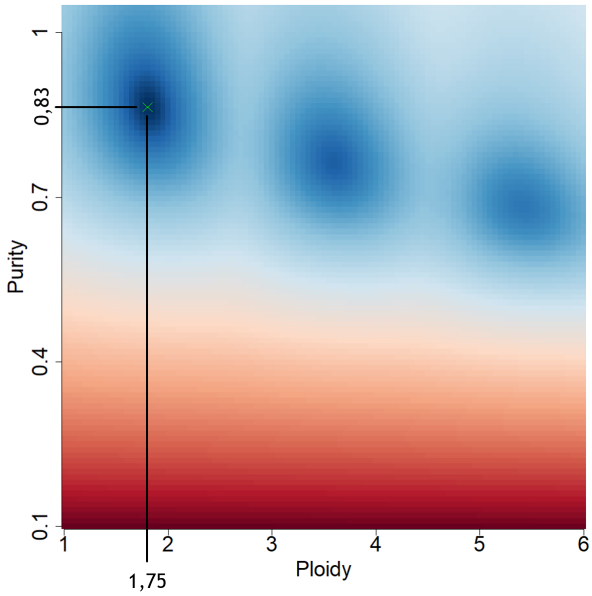

FIGURE 19 – La qualité de l'ajustement de toutes les solutions testées par le calling ASCAT.
Rouge, faible qualité, bleu, qualité élevée, blanc, intermédiaire. Ploidy : Ploïdie. Purity : Cellularité tumorale. La solution retenue est marquée d'une croix verte.

L'ADN utilisé par la technologie OncoScan CNV est extrait à partir d'échantillons tumoraux souvent contaminé par le tissu sain alentour, et identifier la séparation entre les deux peut constituer un défi. L'hypothèse de présence d'ADN issu de cellules saines ne peut pas être exclue lors de l'analyse d'une tumeur. La cellularité tumorale a pour intérêt de renseigner sur l'état de cette hétérogénéité tumorale.

Il est à noter que d'autres outils auraient eu leur pertinence dans ce travail comme GISTIC[18] et oncoSNP[27]. Ces outils sont écrits dans le langage de programmation matlab[28], et il est compliqué d'intégrer un programme écrit en matlab dans un pipeline de routine dans le contexte de ce travail. Pour cette raison, GISTIC et oncoSNP n'ont pas été retenus, et la sélection d'outils basés sur le langage R a été favorisée.

Pour comparer ces outils qui déterminent les altérations de nombre de copies différemment, le GI est calculé par chacun d'eux. Ce projet étant actuellement en phase d'exploration, la comparaison des outils est manuelle. On cherche à savoir s'il existe une proportionnalité similaire entre les GI calculés à partir des données Agilent et OncoScan. Cela va être fait à l'aide du calcul du coefficient de corrélation de Pearson

La comparaison des outils passe également par leurs performances. Leur pouvoir prédictif est évalué à l'aide de courbes ROC, et on regardera également leur temps de calcul.

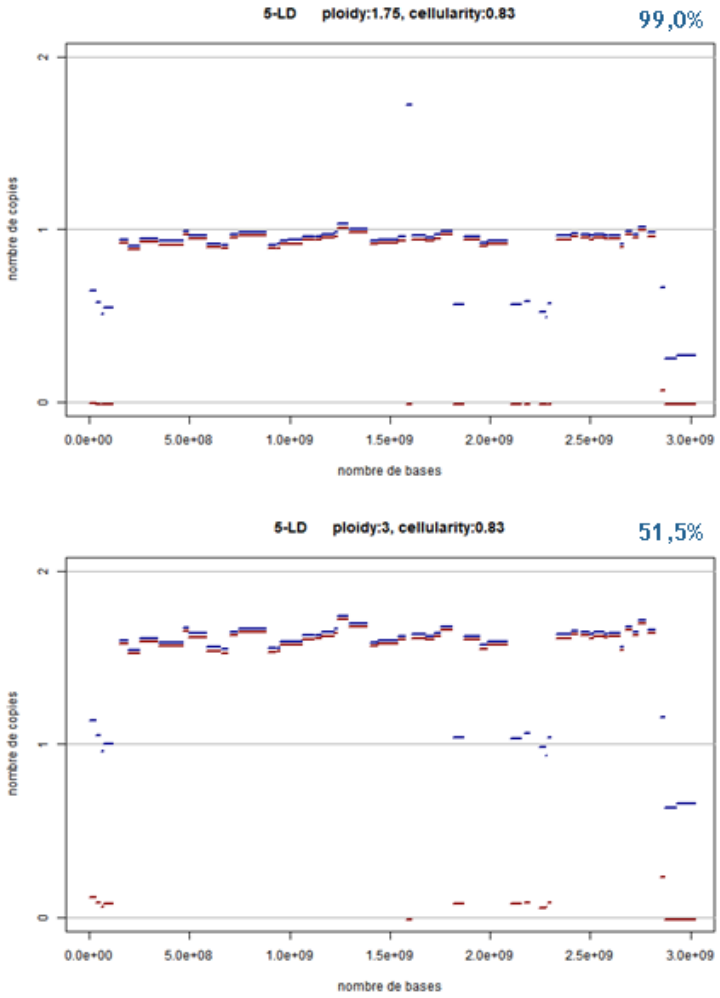

FIGURE 20 – Deux solutions et leur qualité de l'ajustement associée (note sur 100) déterminées par ASCAT pour un même profil.
Les deux allèles (bleu et rouge) de chaque SNP sont représentés légèrement décalés pour des raisons de lisibilité. En haut : solution pour une ploïdie de 1.75 et une cellularité tumorale de 0,83. En bas : solution pour une ploïdie de 3 et une cellularité tumorale de 0,83.

III Résultats et Discussion

III.1 Comparaison des pipelines

Les différences entre les outils sont résumées dans le tableau 1.

outil	input	nettoyage	recentrage	segmentation	calling	CNA	autre
oncoscanR	segments	oui	non	non	non	bras	scores
rCGH	sondes	oui	oui	CBS	oui	segments	GUI
CGHcall	sondes	oui	deux	CBS	Modèle de mélange	segments	non
ASCAT	sondes	non	non	ASPCF	ASCAT	segments	estimation ploïdie et cellularité

TABLE 1 – Comparaison des apports de chaque outil dans le calcul du GI.
Input désigne le type de données nécessaire en entrée de l'outil. La colonne autre désigne les fonctionnalités qui ne servent pas directement à déterminer les altérations, mais qui ont un intérêt pour ce type d'analyse. CBS, Circular Binary Segmentation ; ASPCF, Allele-Specific Piecewise Constant Fitting. ASCAT, Allele-Specific Copy number Analysis of Tumors .CNA : altérations de nombre de copies obtenues. GUI : interface graphique.

Les données en entrée de chaque outil diffèrent, mais toutes proviennent du fichier puce généré par OncoScan CNV, et des adaptations ont dû être faites pour chaque outil. Les données de segments nécessaires à oncoscanR ont été générées sous forme de fichier texte par le logiciel ChAS. Cette étape est manuelle et doit être faite pour chaque échantillon individuellement. Pour rCGH et CGHcall, les données par sonde ont aussi été générées par ChAS, mais il est possible de le faire en cohorte pour les fichiers sondes. Enfin, dans le cas d'ASCAT, le fichier puce a été directement lu et converti en objet R par le package EaCoN [29] Dans l'optique de mettre en place un outil automatisé, éviter cette étape fastidieuse serait intéressant. Pour cela, utiliser APT (Analysis Power Tools) [1], un outil en ligne de commande développé par Affymetrix pour traiter entre autres les données OncoScan serait intéressant, mais le calcul des altérations serait différent de celui fait par ChAS, et l'unité de pathologie moléculaire utilise le logiciel ChAS et pas APT.

L'importance de l'étape de nettoyage (ou pré-traitement) varie entre les outils. Elle est primordiale pour oncoscanR : le nettoyage peut influer fortement sur le calcul des altérations en supprimant et fusionnant des segments. Pour CGHcall et rCGH, l'étape de nettoyage a pour intérêt majeur de préparer les données afin que les étapes suivantes

1. http ://media.affymetrix.com/support/developer/powertools/changelog/

n'aient pas à les prendre en compte. Cela passe par la suppression des valeurs manquantes pour CGHcall, mais aussi par leur estimation, tandis que rCGH met les données à une échelle utilisable par le reste du pipeline. ASCAT n'a pas d'étape de nettoyage.

La normalisation (ou le recentrage) est une étape importante de l'analyse des CNV, car définir le niveau normal détermine la nature des CNV qui sont sur des niveaux différents, et détermine par extension les altérations de nombre de copies. Pour rCGH, CGHcall et ASCAT, une même région est souvent marquée du même niveau de log Ratio (Annexes : figure 32). Les méthodes mises en place par rCGH et CGHcall pour cette étape ne sont donc pas forcément nécessaires pour cette analyse, mais lors du traitement de cas ayant plusieurs niveaux de log Ratio différents, l'étape de normalisation peut avoir une importance plus grande.

Les segmentations CBS et ASPCF produisent des résultats globalement similaires. En effet, pour la plupart, les segments déterminés par rCGH, CGHcall et ASCAT recouvrent fidèlement les régions de log Ratio, comme le montre la figure 32 de l'échantillon 6 en annexe. La segmentation ASPCF (du package ASCAT) crée cependant plus de segments. D'autre part, bien que le même algorithme soit utilisé pour rCGH et CGHcall, des différences de segmentation sont observables : Alors que les segments altérés déterminés par CGHcall se superposent, pour la plupart, sur les régions de log Ratio leur correspondant, les segments de rCGH sont placés aux mêmes régions génomiques, mais à des valeurs de log Ratio plus éloignées de 0. Cela est dû à l'étape de changement d'échelle, qui étale les données de log Ratio et les éloigne de 0. D'autre part, entre 1.5×10^9 et 2×10^9, les segments altérés sont un peu différents : la région au niveau de 1.75×10^9 est plus lisse pour CGHcall que pour rCGH. Cela est causé par un paramètre de la segmentation undo.SD, qui utilise la valeur par défaut de 3 pour CGHcall, mais est estimé par rCGH pour calculer la meilleure segmentation possible. Pour l'échantillon 6, la segmentation rCGH semble ainsi mieux suivre le profil de log Ratio.

L'attribution d'un nombre de copies réel (le calling) aux segments altérés est une étape importante dans l'analyse du nombre de copies, mais peut constituer une transformation très forte des données. Le sens biologique de cette estimation du nombre de copies doit être interprété en gardant un regard critique sur la méthode de calcul. Le modèle de mélange de CGHcall permet cela en calculant pour chaque segment la probabilité qu'il appartienne à chaque type d'altération : Des différences très tranchées indiquent une forte crédibilité du résultat. Pour ASCAT, pour un échantillon donné, différentes solutions de calling sont évaluées par un score sur 100 appelé qualité de l'ajustement. La crédibilité de la solution retenue est alors connue et l'interprétation des résultats peut en tenir compte. L'estimation du nombre de copies par rCGH étant documentée de manière minimale, il est compliqué de la comparer au calling d'ASCAT et CGHcall.

Au-delà du calcul des altérations de nombre de copies, OncoscanR, rCGH et ASCAT possèdent des fonctionnalités qui ont un intérêt dans le diagnostic des tumeurs cancéreuses. OncoscanR peut ainsi calculer des scores moléculaires dont le lien avec le développement de tumeurs précises a été mis en évidence [19][20][21]. Partant du principe qu'une analyse optimale des altérations détectées passe par leur lecture et leur édition à l'aide d'un outil visuel, rCGH dispose d'une interface graphique permettant la visualisation interactive et la modification d'un profil. Certaines de ses fonctionnalités sont plus faciles d'accès que celles du logiciel ChAS (notamment de par l'utilisation de curseurs) pour définir les seuils d'altération ou recentrer le profil, alors que ChAS nécessite plusieurs étapes

pour accomplir la même chose (le recentrage du profil passe par un re-chargement des données). Cependant, la plupart des changements faits dans l'interface graphique de rCGH mettent un peu de temps à être appliqués sur un ordinateur de bureau aux performances normales ; c'est un problème que n'a pas ChAS. D'autre part, ChAS dispose d'un panel de fonctionnalités plus complet, comme la visualisation de plusieurs profils à la fois, ou la conversion des données en fichiers de différents formats. L'utilisation de ChAS ne se limitant pas à l'analyse de nombre de copies, ce logiciel est voué à continuer d'être utilisé, car la méthode fait l'objet d'un dossier de validation de méthode en vue d'une accréditation selon la norme NF ISO 15189. En parallèle, l'évaluation d'un autre outil pour le calcul de l'index génomique a fait l'objet de mon projet de stage et pourra être appliqué au laboratoire.

Estimation de la cellularité par ASCAT

La cellularité tumorale correspond au pourcentage de cellules tumorales présentes dans l'échantillon dont on prélève l'ADN. Dans l'unité de pathologie moléculaire, la cellularité d'un échantillon est estimée à partir d'une lame de microscope sur laquelle une coloration HES est appliquée. Pour construire cette lame, une fine coupe de 3 μm est effectuée en prélevant la couche supérieure de l'échantillon. La coloration est ensuite appliquée : l'hématoxyline et l'éosine colorent respectivement les noyaux des cellules en bleu et le cytoplasme en rose. La lame est observée au microscope et les zones présentant le plus de cellules tumorales sont cerclées (figure 21, en haut, contour noir). La proportion de cellules tumorales est alors estimée visuellement sur lame par le pathologiste.

L'échantillon est ensuite foré (figure 21, en bas) à l'emplacement de la zone cerclée pour en extraire l'ADN qui sera utilisé dans les techniques de diagnostic (ici, la CGH-array). La cellularité est une information qui peut ensuite être prise en compte lors de l'analyse

FIGURE 21 – Représentation d'un échantillon inclus en bloc FFPE (en bas) et de la lame correspondante (en haut).

de l'échantillon. ASCAT effectue une estimation de la cellularité. Afin de savoir si cette estimation est comparable avec les valeurs estimées sur lame, on fait un graphe des valeurs de cellularité données par ces deux méthodes (fig. 22). Les valeurs de cellularité estimées

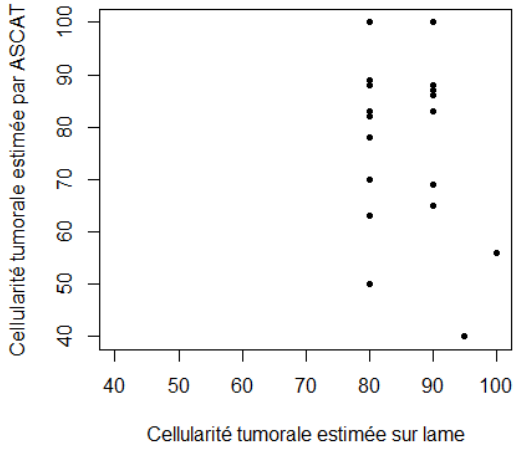

FIGURE 22 – Comparaison entre les valeurs de cellularité déterminée sur lame et estimée par ASCAT.

sur lame présentent des valeurs se trouvant entre 80 et 100, et les valeurs estimées par ASCAT vont de 40 à 100. Ce graphe ne met pas en évidence de proportionnalité dans ces valeurs. Cette discordance peut s'expliquer par le fait que l'estimation visuelle est faite sur une lame et l'estimation d'ASCAT est faite sur les données de l'ADN extrait par forage. La répartition en trois dimensions d'une tumeur au sein d'un tissu étant hétérogène, l'estimation de la cellularité peut ne pas concerner la zone forée quelques centimètres en profondeur. En résumé, il est complexe de tirer des conclusions de cette comparaison en ayant connaissance des prérequis techniques d'évaluation.

La problématique de ce projet était la transposition du calcul du GI, dont l'approche est validée sur la technologie Agilent, à la technologie OncoScan. Dans un premier temps, on peut répondre à cette problématique concernant le fonctionnement des outils uniquement.

rCGH, CGHcall, et ASCAT utilisent les données des sondes produites par la technologie Oncoscan CNV et les utilisent pour déterminer les régions (segments) du génome dont le nombre de copies est altéré. En résumé, ces trois outils déterminent les altérations de nombre de copies, ce qui permet de calculer le GI ; ils sont en mesure de répondre à cette problématique. OncoscanR détermine les altérations de nombre de copies au niveau des bras. Cette définition permet de déterminer le GI, en théorie. Cependant, les cela transforme les données : à partir des segments d'altération, on détermine un état binaire pour chaque bras chromosomique : altéré ou non. Le nombre d'altérations utilisées dans le calcul du GI correspond alors au nombre de bras altérés. un bras présentant 1 ou 8 segments d'altération peut ainsi donner le même état d'altération. Sur les profils qui présentent de nombreux segments, le GI ainsi calculé sera donc limité par le nombre de bras chromosomiques. Cela pose un problème, car le GI précédemment établi se base sur le nombre de segments d'altération : Une grande partie de cette information est perdue lors de l'estimation des altérations de bras. Pour cette raison, OncoscanR ne semble pas apte à être utilisé pour répondre à la problématique.

■ **Distribution des valeurs de GI**

La figure 23 présente la distribution des valeurs de GI pour les outils testés et pour le GI calculé manuellement à partir des données Agilent, pour référence.

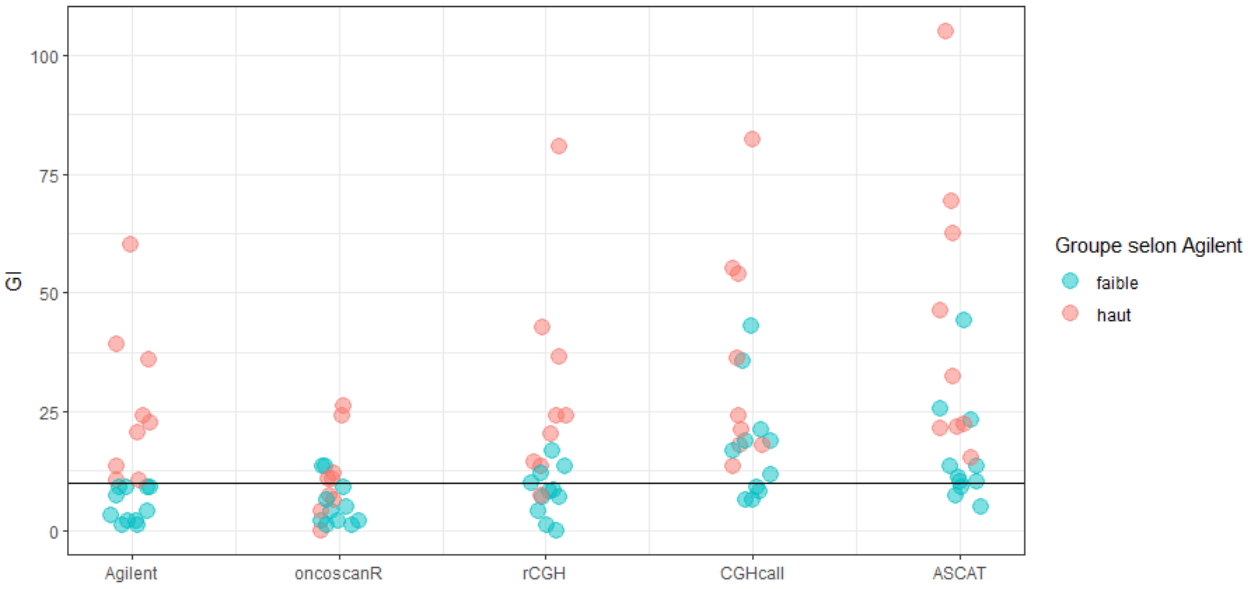

FIGURE 23 – La répartition des valeurs de GI (index génomique) par outil. Les échantillons dont le GI calculé à partir des données Agilent (à gauche) est supérieur ou égal à 10 sont représentés en rouge, et ceux dont le GI est inférieur à 10 sont en bleu pour tous les groupes. La valeur de 10 sur l'axe y est indiquée par une ligne noire. Un bruit de fond a été ajouté aux données sur l'axe X pour éviter la superposition des points.

Ce graphe permet :

1. de visualiser la distribution des valeurs de GI
2. d'identifier l'éventuelle répartition qui en découle

Les échantillons ont été classés en deux groupes selon leurs valeurs de GI calculées par Agilent par un seuil de 10 défini précédemment [7] [8]. On cherche à savoir si, pour l'un des outils testés, le calcul du GI classe aussi bien les échantillons en deux groupes. On retrouve une répartition similaire pour rCGH, CGHcall et ASCAT : Malgré le recouvrement de ces deux groupes, le groupe de GI haut présente en effet des valeurs globalement supérieures à celles du groupe de GI faible.

Ces résultats indiquent que parmi les outils testés, rCGH, CGHcall et ASCAT déterminent des altérations qui amènent à des écarts de GI similaires à ceux de la procédure manuelle appliquée aux données Agilent. Dans le cas d'OncoscanR, les écarts de GI ne présentent pas de similitudes notables avec ceux du GI agilent. Cela s'explique par le fait qu'OncoscanR, pour un bras chromosomique donné, convertit les segments altérés en une unique altération, ce qui supprime un nombre d'altérations parfois conséquent. Le GI qui en découle peut alors avoir une valeur inférieure à celle d'un autre outil. Cela ne fait pas d'OncoscanR un bon candidat.

■ **Corrélations** Pour avoir plus d'informations sur les similarités de classement des échantillons entre chaque outil et le GI déterminé par Agilent, des corrélations sont calculées (fig.24). Des corrélations fortes (r>0.75, coefficient de corrélation de

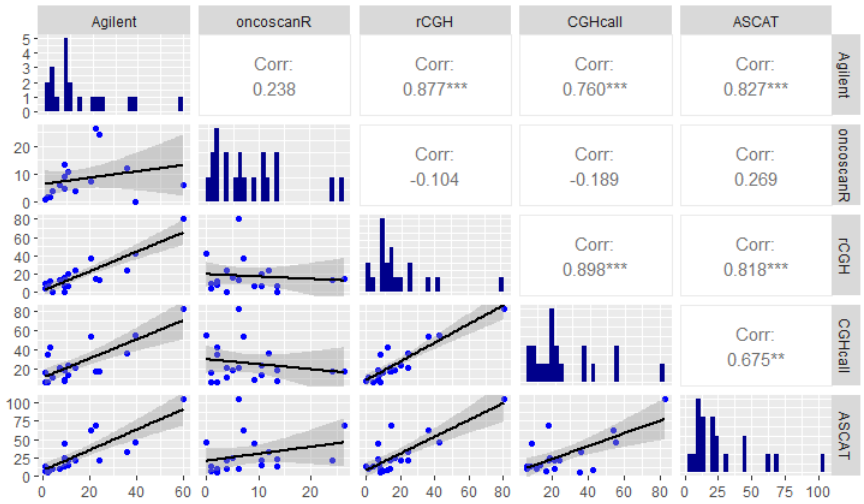

FIGURE 24 – Corrélations entre les valeurs obtenues par les outils testés et sur les données Agilent.

Pearson) sont observables entre la méthode validée sur Agilent et ASCAT, CGHcall et rCGH, tandis que la corrélation entre Agilent et OncoscanR n'est pas significative .

Cela indique que pour ces trois outils, les valeurs de GI calculées sont proportionnelles entre l'outil et le GI de référence, Agilent. En d'autres termes, les écarts relatifs de deux points sont similaires entre l'outil et la référence, ce qui indique l'intérêt d'ASCAT, CGHcall et rCGH pour répondre à la problématique. Ce n'est pas le cas d'OncoscanR : Dans le cas d'une proportionnalité parfaite entre deux séries de valeurs, tous les points passent par une droite d'équation y=x ; certains points de la comparaison OncoscanR - Agilent en sont très éloignés. À titre d'exemple, sur la figure 25, la valeur de l'échantillon 12-BC est d'environ 40 pour Agilent, et proche de 0 pour oncoscanR.

Le profil de cet échantillon présente de nombreuses petites altérations notamment sur les chromosomes 1 et 18, qui sont vraisemblablement la cause du GI élevé pour Agilent (fig. 26). Le GI faible de cet échantillon pour oncoscanR est causé par sa méthode de calcul des altérations : la surface recouverte par les segments altérés n'est pas suffisante pour qu'oncoscanR déclare des bras comme altérés.

■ **performances**

On a maintenant une mesure de la capacité des outils à classer les échantillons. Afin de connaître en détail leur performances sur ce point, des courbes ROC ont été construites. Pour un outil donné, une telle courbe est construite à partir des valeurs de GI trouvées et du groupe (déterminé par Agilent) auquel appartiennent les échantillons. Ces courbes sont très peu lisses de par le faible nombre d'échantillons utilisés (fig.27, fig. 28).

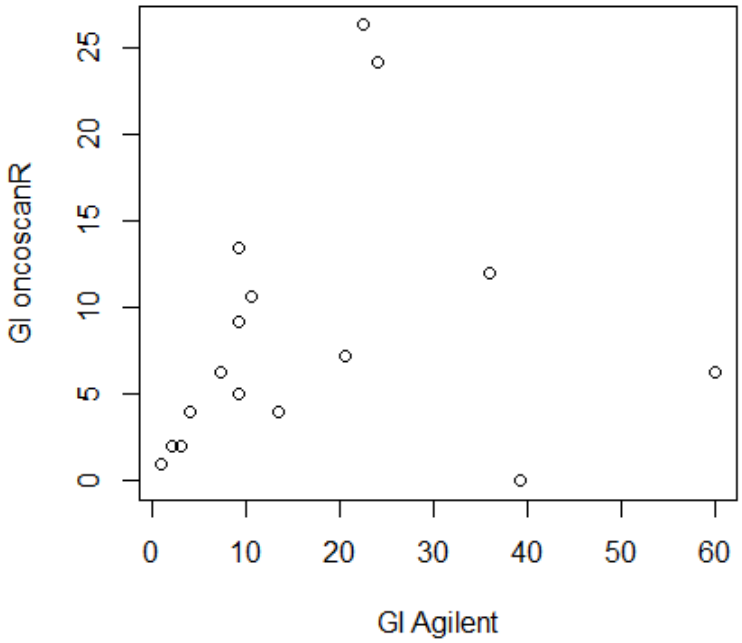

FIGURE 25 – Valeurs de GI d'oncoscanR par rapport à Agilent.
GI : index génomique.

FIGURE 26 – Les données de log Ratio (en haut), différence allélique (au milieu) et BAF (en bas) de l'échantillon 12-BC.

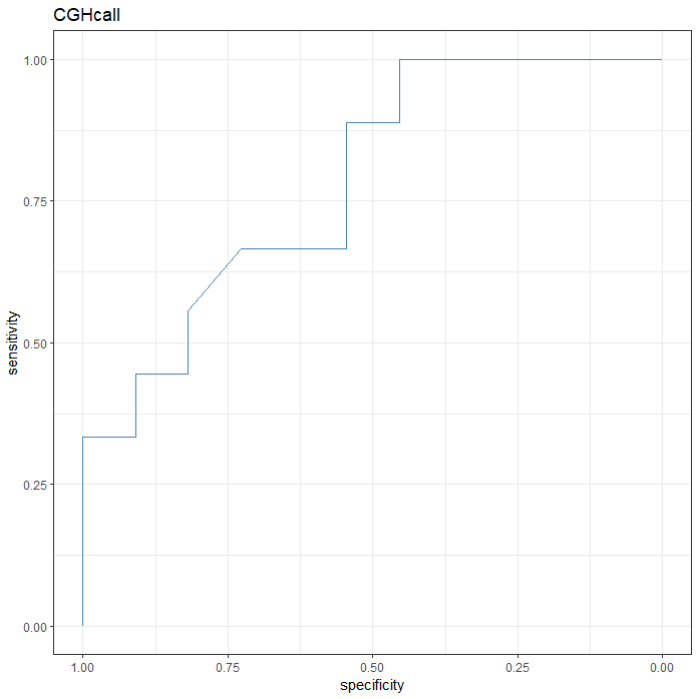

FIGURE 27 – Courbe ROC des donnés CGHcall.

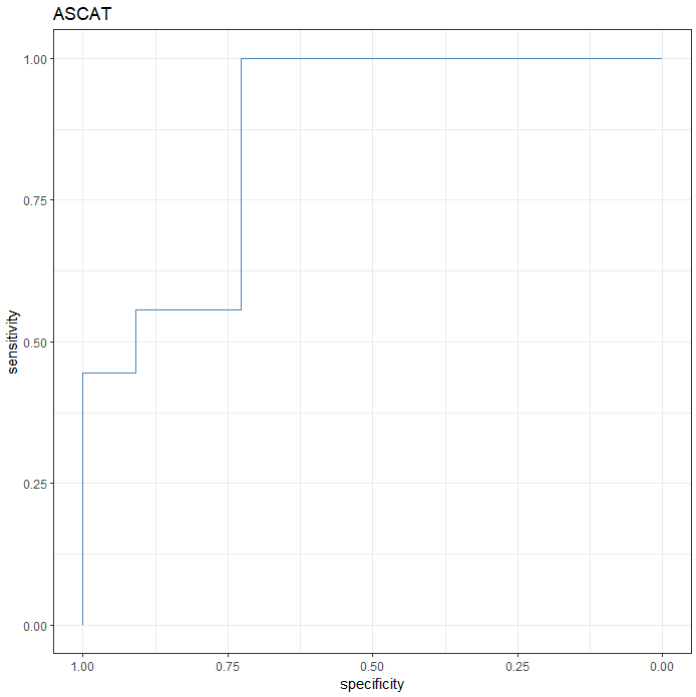

FIGURE 28 – Courbe ROC des données ASCAT.

L'aire sous la courbe ROC (area under curve, AUC) correspond à la probabilité qu'un échantillon donné soit classé dans la bonne catégorie par l'outil. Les valeurs d'AUC sont représentées dans le tableau 2.

	oncoscanR	rCGH	CGHcall	ASCAT
AUC	0.71	0.91	0.78	0.87
ρ	0.24 (p=0.3)	0.88 (p<$3e^{-7}$)	0.76 (p<$1e^{-4}$)	0.83 (p<$7e^{-6}$)

TABLE 2 – Métriques sur les outils.
AUC : aire sous la courbe. ρ : Coefficient de corrélation de Pearson et les p-values associées.

Les quatre outils présentent des valeurs relativement élevées (>0.7), rCGH ayant l'AUC la plus élevée de 0.91. Ces résultats indiquent que les quatre outils ont une meilleure performance prédictive qu'une classification aléatoire, et en général, ASCAT et rCGH classifient les échantillons correctement près de 90% du temps, ce qui en fait des outils dont la performance est intéressante pour répondre à la problématique. OncoscanR et CGHcall sont moins prometteurs en termes de performance mais ne sont pas à exclure par ce résultat seul.

Dans la figure 29, le temps de calcul du GI par échantillon est représenté pour chaque outil. OncoscanR est l'outil le plus rapide, avec un temps de calcul médian inférieur à 10

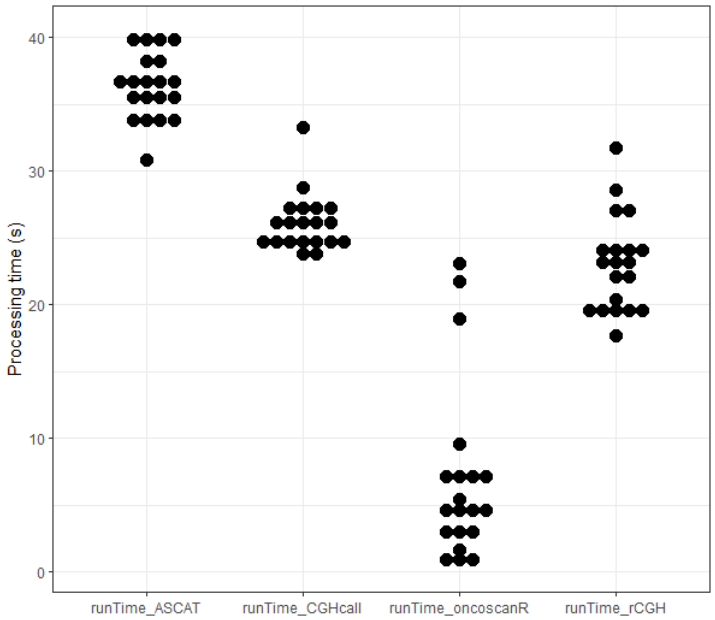

FIGURE 29 – Temps de calcul pour chaque outil.

secondes. Pour rCGH et CGHcall, ce temps de calcul se situe entre 20 et 30 secondes, et

il avoisine les 35 secondes pour ASCAT. Ces résultats s'expliquent par le faible nombre de computations que cet outil effectue par rapport à rCGH, CGHcall ou ASCAT dont les étapes de segmentation et de calling sont plus lourdes algorithmiquement, ce qui se traduit par des temps de calcul plus élevés. Le temps de calcul peut être réduit en allégeant la charge computationnelle d'un outil. Cela peut être accompli en modifiant des paramètres, voire en retirant des étapes dans le processus de détermination des altérations utilisé par l'outil. Le temps de calcul des outils utilisés dans le diagnostic peut avoir une importance non négligeable dans un contexte de routine où de nombreux échantillons sont traités en un temps donné. Cependant, le temps de calcul doit ici être considéré comme une caractéristique de l'outil plutôt qu'un critère décisionnel pour le choix d'un outil pouvant répondre à la problématique.

La définition de l'index génomique constitue un outil critique pour la classification des tumeurs, préciser un diagnostic et assigner un traitement pertinent. Utilisé couramment au laboratoire sur la technologie Oncoscan, il permettrait de communiquer des données moléculaires complexes, en termes quantitatifs et soutiendra l'application réussie de la génomique dans les soins aux patients.

Le nombre d'échantillons utilisés ici ne permet pas de définir un outil de choix pour répondre à la problématique. Néanmoins, sur la base des résultats, tous les outils présentent un intérêt, à l'exception d'OnoscanR.

Dans des travaux futurs, utiliser un plus grand nombre d'échantillons permettrait de mieux définir l'outil le plus adapté.

Bibliographie

[1] F. Commo, « Analyse génomique en médecine de précision : Optimisations et outils de visualisation, » thèse de doct., Université Paris Saclay (COmUE), 2015.

[2] Y. Wang, M. Cottman et J. D. Schiffman, « Molecular inversion probes : a novel microarray technology and its application in cancer research, » *Cancer genetics*, t. 205, n° 7-8, p. 341-355, 2012.

[3] Y. Chekaluk, C.-L. Wu, J. Rosenberg et al., « Identification of nine genomic regions of amplification in urothelial carcinoma, correlation with stage, and potential prognostic and therapeutic value, » *PloS one*, t. 8, n° 4, e60927, 2013.

[4] T. Xie, G. d'Ario, J. R. Lamb et al., « A comprehensive characterization of genome-wide copy number aberrations in colorectal cancer reveals novel oncogenes and patterns of alterations, » 2012.

[5] I. Skirnisdottir, M. Mayrhofer, M. Rydåker, H. Åkerud et A. Isaksson, « Loss-of-heterozygosity on chromosome 19q in early-stage serous ovarian cancer is associated with recurrent disease, » *BMC cancer*, t. 12, n° 1, p. 1-9, 2012.

[6] F. Bellido, M. Pineda, R. Sanz-Pamplona et al., « Comprehensive molecular characterisation of hereditary non-polyposis colorectal tumours with mismatch repair proficiency, » *European Journal of Cancer*, t. 50, n° 11, p. 1964-1972, 2014.

[7] P. Lagarde, G. Pérot, A. Kauffmann et al., « Mitotic checkpoints and chromosome instability are strong predictors of clinical outcome in gastrointestinal stromal tumors, » *Clinical cancer research*, t. 18, n° 3, p. 826-838, 2012.

[8] P. Lagarde, J. Przybyl, C. Brulard et al., « Chromosome instability accounts for reverse metastatic outcomes of pediatric and adult synovial sarcomas, » *Journal of clinical oncology*, t. 31, n° 5, p. 608-615, 2013.

[9] L. Lartigue, A. Neuville, P. Lagarde et al., « Genomic index predicts clinical outcome of intermediate-risk gastrointestinal stromal tumours, providing a new inclusion criterion for imatinib adjuvant therapy, » *European Journal of Cancer*, t. 51, n° 1, p. 75-83, 2015.

[10] S. Croce, A. Ribeiro, C. Brulard et al., « Uterine smooth muscle tumor analysis by comparative genomic hybridization : a useful diagnostic tool in challenging lesions, » *Modern Pathology*, t. 28, n° 7, p. 1001-1010, 2015.

[11] R Core Team, *R : A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. adresse : `https://www.R-project.org/`.

[12] RStudio Team, *RStudio : Integrated Development Environment for R*, RStudio, PBC, Boston, MA, 2022. adresse : http://www.rstudio.com/.

[13] Y. Christinat, P. Chaskar, S. Clément et al., « Automated Detection of Arm-Level Alterations for Individual Cancer Patients in the Clinical Setting, » *The Journal of Molecular Diagnostics*, t. 23, n° 12, p. 1722-1731, 2021.

[14] M. A. Van De Wiel, K. I. Kim, S. J. Vosse, W. N. Van Wieringen, S. M. Wilting et B. Ylstra, « CGHcall : calling aberrations for array CGH tumor profiles, » *Bioinformatics*, t. 23, n° 7, p. 892-894, 2007.

[15] P. Van Loo, S. H. Nordgard, O. C. Lingjærde et al., « Allele-specific copy number analysis of tumors, » *Proceedings of the National Academy of Sciences*, t. 107, n° 39, p. 16 910-16 915, 2010.

[16] F. Commo, J. Guinney, C. Ferte et al., « rCGH : a comprehensive array-based genomic profile platform for precision medicine, » *Bioinformatics*, t. 32, n° 9, p. 1402-1404, 2016.

[17] D. M. Roy, L. A. Walsh, A. Desrichard et al., « Integrated Genomics for Pinpointing Survival Loci within Arm-Level Somatic Copy Number Alterations, » *Cancer Cell*, t. 29, n° 5, p. 737-750, 2016, issn : 1535-6108. doi : https://doi.org/10.1016/j.ccell.2016.03.025. adresse : https://www.sciencedirect.com/science/article/pii/S1535610816301088.

[18] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim et G. Getz, « GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers, » *Genome biology*, t. 12, n° 4, p. 1-14, 2011.

[19] V. Abkevich, K. Timms, B. Hennessy et al., « Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer, » *British journal of cancer*, t. 107, n° 10, p. 1776-1782, 2012.

[20] T. Popova, E. Manié, G. Rieunier et al., « Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation, » *Cancer research*, t. 72, n° 21, p. 5454-5462, 2012.

[21] T. Popova, E. Manié, V. Boeva et al., « Ovarian cancers harboring inactivating mutations in CDK12 display a distinct genomic instability pattern characterized by large tandem duplications, » *Cancer research*, t. 76, n° 7, p. 1882-1891, 2016.

[22] E. Venkatraman et A. B. Olshen, « A faster circular binary segmentation algorithm for the analysis of array CGH data, » *Bioinformatics*, t. 23, n° 6, p. 657-663, 2007.

[23] H. Willenbrock et J. Fridlyand, « A comparison study : applying segmentation to array CGH data for downstream analyses, » *Bioinformatics*, t. 21, n° 22, p. 4084-4091, 2005.

[24] F. Commo, C. Ferte, J. Soria, S. Friend, F. Andre et J. Guinney, « Impact of centralization on aCGH-based genomic profiles for precision medicine in oncology, » *Annals of Oncology*, t. 26, n° 3, p. 582-588, 2015.

[25] T. Hastie, R. Tibshirani, B. Narasimhan, G. Chu, M. B. Narasimhan et M. biocViews Bioinformatics, « Package 'impute', » 2011.

[26] J. W. Tukey et al., *Exploratory data analysis*. Reading, MA, 1977, t. 2.

[27] C. Yau, D. Mouradov, R. N. Jorissen et al., « A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data, » *Genome biology*, t. 11, n° 9, p. 1-15, 2010.

[28] *MATLAB version 9.10.0.1613233 (R2021a)*, The Mathworks, Inc., Natick, Massachusetts, 2021.

[29] B. Job, *EaCoN : EaCoN : Easy Copy Number !* R package version 0.3.6-2, 2021.

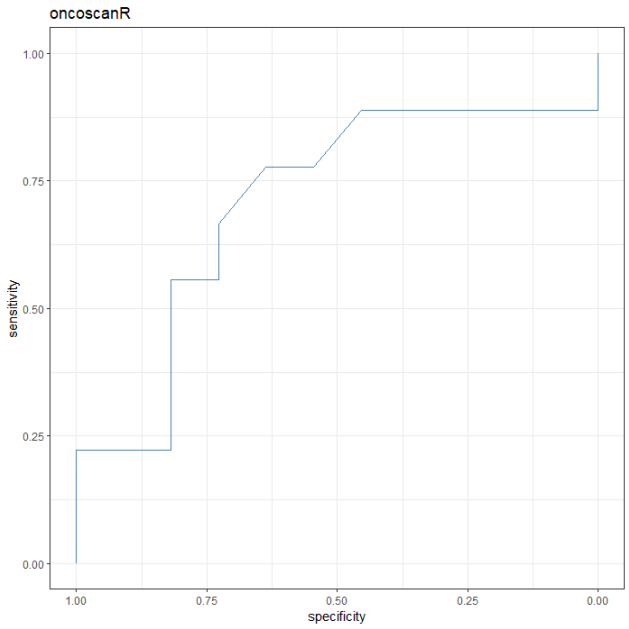

FIGURE 30 – Courbe ROC issue des résultats d'oncoscanR.

Annexes

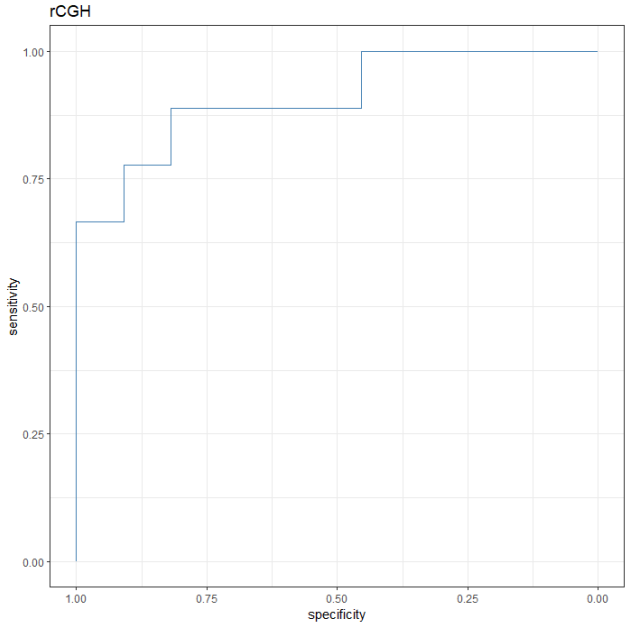

FIGURE 31 – Courbe ROC issue des résultats de rCGH.

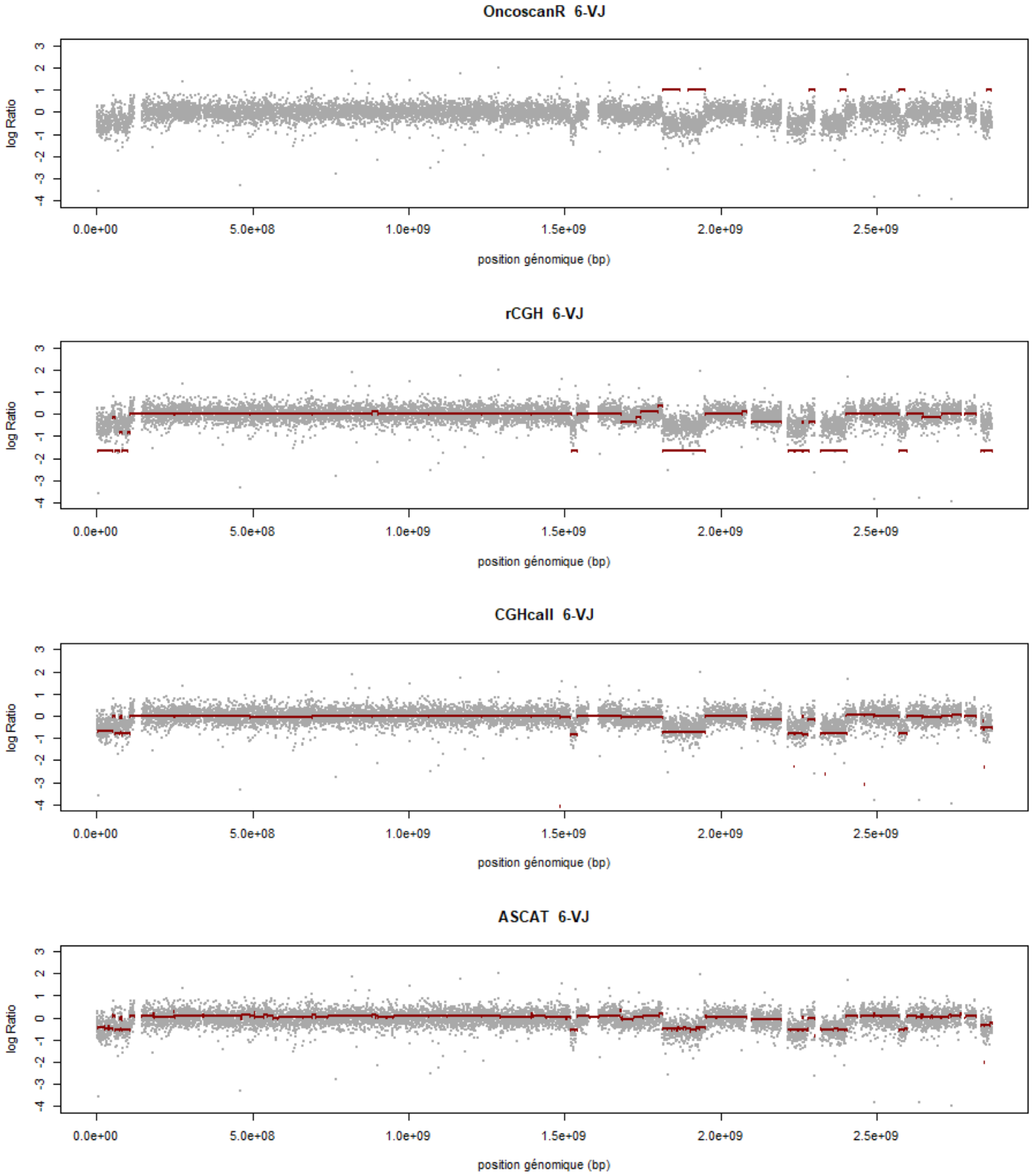

FIGURE 32 – altérations trouvées par les quatre outils.
En gris : log ratio. En rouge : segments altérés calculés par les outils. Les quatre panneaux correspondent de haut en bas à oncoscanR, rCGH, CGHcall et ASCAT.

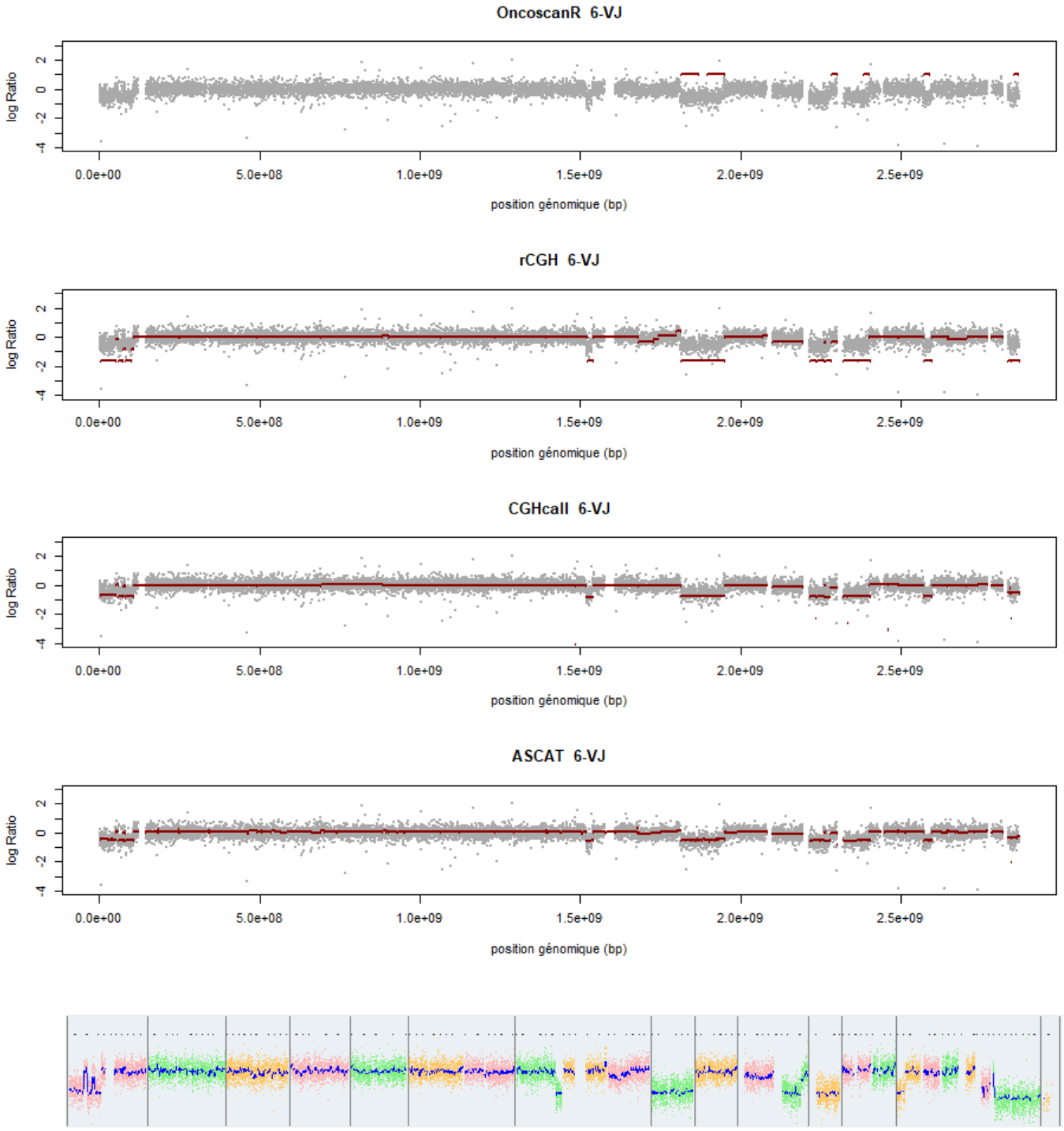

FIGURE 33 – altérations trouvées par les quatre outils et le logiciel ChAS. En gris : log ratio. En rouge : segments altérés calculés par les outils. Les quatre premiers panneaux correspondent de haut en bas à oncoscanR, rCGH, CGHcall et ASCAT. Le cinquième représente le profil déterminé par ChAS. En rose, jaune et vert : le log Ratio, coloré pour mettre en évidence les séparations entre chromosomes. En bleu : le signal de log Ratio lissé.