université de BORDEAUX

Rapport de stage de deuxième année de Master

Comparaison de l'évaluation de l'index génomique en méthodologie Agilent/SurePrint G3 versus Affymetrix/Oncoscan CNV

Auteur :

Bordron Elie
Etudiant en Master 2 de bio-informatique
Université de Bordeaux

Directrices de stage :

Larmonier Claire, PhD
Darbo Elodie, PhD

23 mai 2022

Sommaire

Tables des figures

Introduction

0.1 Contexte sur la cancérologie

Dans le traitement de lésions cancéreuses, le diagnostic et le pronostic se basent sur des critères cliniques, anatomiques et pathologiques. Quand de tels critères ne suffisent pas à caractériser un cas difficile, des technologies de screening moléculaire sont disponibles, 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.
Au-delà du diagnostic, il existe actuellement un intérêt croissant pour l'identification de signatures moléculaires d'intérêt pronostique et/ou prédictif, dans l'optique d'une prise de décision thérapeutique éclairée et d'un suivi adapté. C'est le cas d'un certain nombre de signatures d'expression génique par exemple dans le cancer du sein [1], actuellement réalisées en routine et d'autres signatures en cours d'évaluation dans des essais cliniques [2].

0.2 La cancérologie à l'institut Bergonié

À l'institut Bergonié, différentes technologies sont utilisées en routine pour répondre à ces problématiques dans l'unité de pathologie moléculaire (fig.1). Parmi elles, on peut citer l'hybridation in situ en fluorescence (FISH), qui permet de détecter des anomalies au niveau de séquences d'ADN cible, ou le séquencage qui, en analysant la séquence d'un gène, permet d'évaluer le risque d'une maladie génétique. La CGH-array est utilisée pour analyser les variations de nombre de copies d'un génome. Les altérations détectées caractérisent le gain ou la perte d'ADN, ce qui permet de connaître le statut de gènes liés au développement de tumeurs et d'adapter la prise en charge thérapeutique. Plus précisément, deux méthodologies sont utilisées : SurePrint G3 d'Agilent et OncoScan CNV d'Affymetrix.

0.3 Le calcul de l'index génomique par CGH

L'intérêt diagnostique et pronostique de l'index génomique (GI) par CGH-array dans différents types de sarcomes comme les tumeurs stromales gastro-intestinales (GIST)[3,5 gironde], les synovialosarcomes [4 gironde] et les tumeurs musculaires lisses de l'utérus [6 gironde] a été précédemment démontré. Cette « signature génomique » ou « genomic index » est le reflet direct du degré de complexité moléculaire et d'instabilité génomique

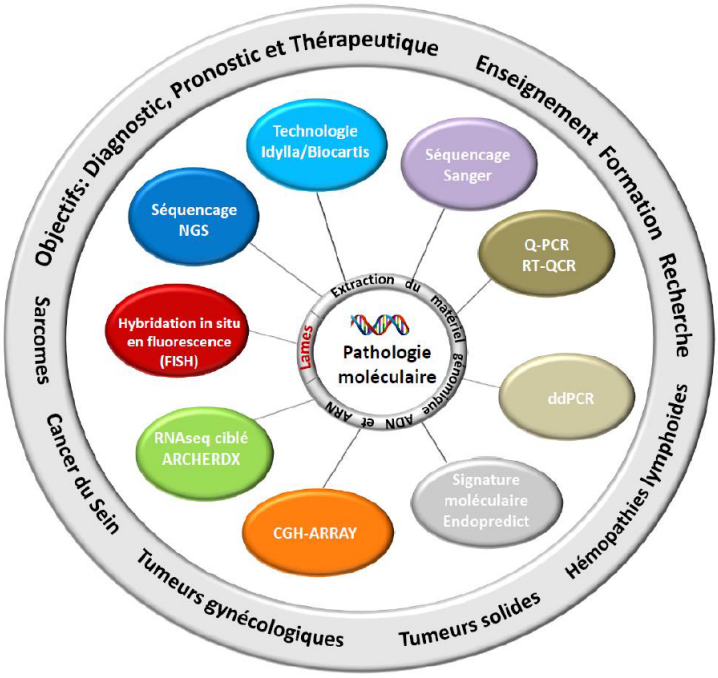

FIGURE 1 – Vue d'ensemble des technologies utilisées dans l'unité de pathologie moléculaire à l'institut Bergonié

de la tumeur et s'avère être un puissant prédicteur de l'agressivité tumorale et de la rechute métastatique des tumeurs. Ce score est utilisé en routine dans l'unité de pathologie moléculaire. Afin d'étendre l'utilisation de l'évaluation de ce score à la méthodologie OncoScan CNV, plus récente et plus résolutive, on se propose de transposer le calcul de l'index génomique validé sur puces AGILENT/SurePrint G3 8x60 K.

0.4 Transposer le calcul du GI d'Agilent à OncoScan

La couverture du génome étant différente entre les deux technologies, il est nécessaire de les comparer sur une cohorte de cas précédemment caractérisés et sur lesquels toutes les données sont disponibles. Pour mettre en place cette comparaison, on va identifier un outil bioinformatique qui permet de calculer le GI à partir des données OncoScan. On répondra alors à la question suivante : Peut-on établir la correspondance du calcul de l'index génomique et la détermination du seuil de classification des tumeurs de la technologie Agilent à la technologie Affymetrix pour pouvoir en faire bénéficier les patients analysés au sein du laboratoire dans le cadre diagnostic ou thérapeutique ?

0.5 Communication Bioinformatique-Biologie

Ce travail s'inscrit dans le projet GIRONDE, qui s'étend au-delà du stage. Un échange biologie-bioinformatique a été entretenu pour que les biologistes puissent suivre les avancées du projet, de sorte à ce que l'équipe ait la meilleure vision possible du travail effectué

et choisisse l'outil le plus pertinent. Cet échange a été facilité par l'immersion dans le côté biologique du projet, car j'ai suivi la technique à partir de l'ADN extrait jusqu'à la production des fichiers de données. Un travail de revue d'outils bioinformatiques a fait l'objet d'une vulgarisation auprès de l'équipe pour présenter les options disponibles.

Chapitre 1

Etat de l'art

1.1 Calcul du GI par Agilent

Le GI est défini ainsi : $GI = \frac{A^2}{C}$, où A est le nombre d'altérations de nombre de copies et C le nombre de chromosomes qui les portent. La détermination des altérations se base sur les données de CGH-array, notamment le log Ratio, qui témoigne des variations de nombre de copies du génome étudié. La technologie Agilent Sureprint couvre ainsi le génome sur plus de 50000 SNP[3], indiquant pour chaque SNP le nombre de copies relatif à une valeur de référence. Sur un profil, les régions de log Ratio supérieur ou inférieur au niveau normal de deux copies témoignent respectivement de gains et de pertes alléliques (fig.1.1).

1.2 Recherche d'un outil permettant ce calcul

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 de log ratio et de B Allele Frequency (BAF) générées. Pour un SNP donné, le log Ratio est calculé selon la formule suivante : $log_2(I_{echantillon}) - log_2(I_{reference})$ où $I_{echantillon}$ est l'intensité lumineuse perçue par le scanner qui lit la puce à hybridation et $I_{reference}$ est la valeur d'intensité lumineuse de référence. Le BAF correspond à la proportion de l'allèle B par rapport à l'allèle A : $BAF = \frac{B}{A+B}$. À l'état normal de 2 copies, un BAF de 0.5 sera observé, indiquant que l'allèle B représente 50% des copies du SNP en question. Un BAF de 0.67 indique que deux tiers des copies de ce SNP sont l'allèle B, ce qui révèle que le nombre de copies total sous-jacent est multiple de 3. Traiter ces informations permet donc de déterminer le nombre de copies, il est maintenant nécessaire de trouver l'outil adéquat parmi les suivants :

1.2.1 Chromosome Analysis Suite

Le logiciel propriétaire d'Affymetrix, Chromosome Analysis Suite (ChAS), détermine les altérations de nombre de copies (fig.1.3) ChAS permet de les visualiser sur le génome (fig.1.2) de manière interactive. Il est ainsi possible de manipuler le profil en le recentrant ou en fusionnant des segments, ce qui a un intérêt pour une analyse optimale des altérations.

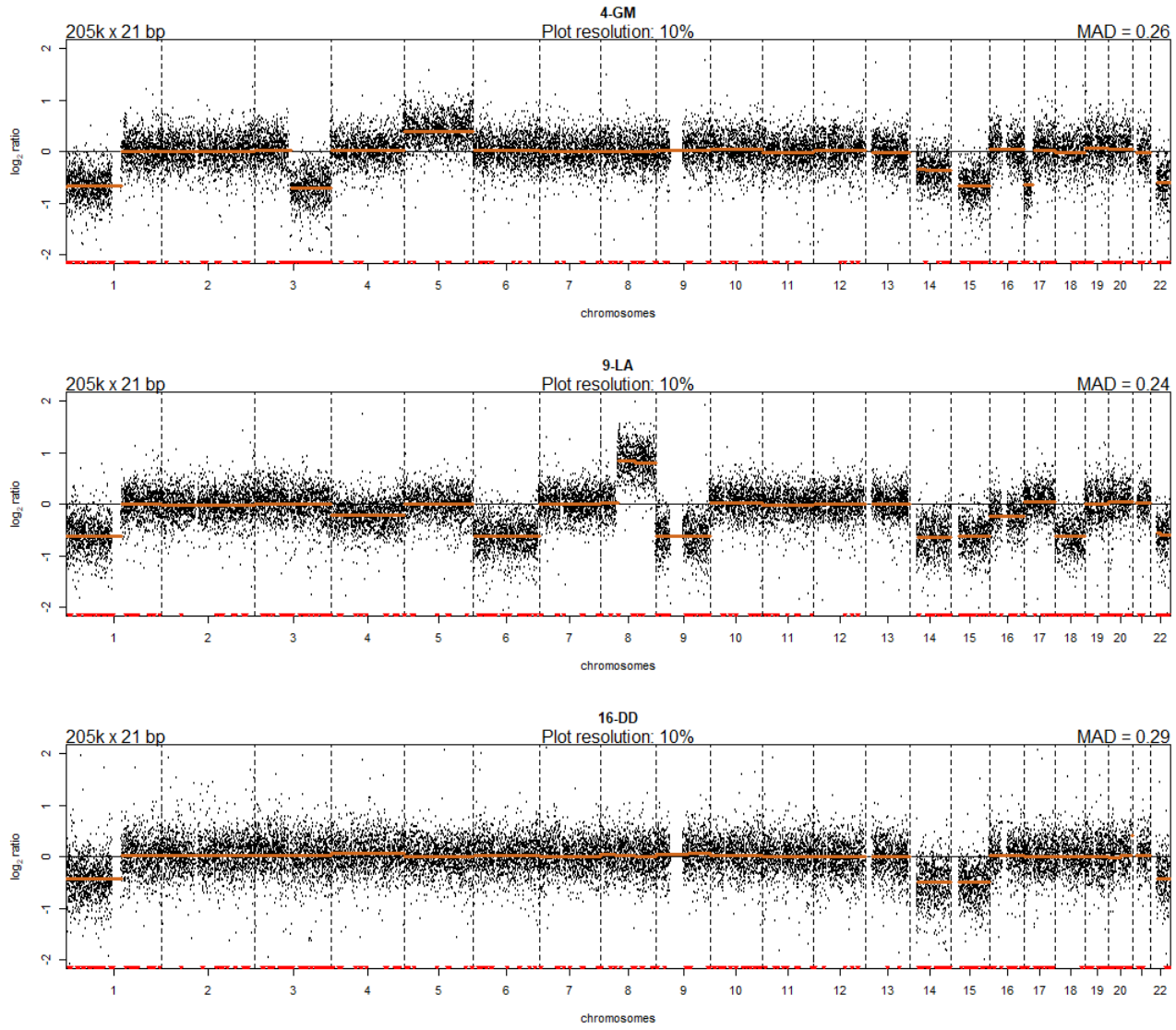

FIGURE 1.1 – Profils CGH. En noir : valeurs brutes de log Ratio. En orange : segments d'altération déterminés à partir de ces valeurs. Les trois profils présentent entre autres des pertes sur les chromosomes 1,14 et 15, et les profils 4-GM et 9-LA présentent des gains sur les chromosomes 5 et 8, respectivement.

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, si ils ont lieu d'être effectués, sont également des étapes manuelles. Plus nombreuses sont les interventions humaines, plus grand est le risque d'erreur humaine, ce qui est un argument en faveur de l'automatisation[4]. 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. D'autre part, on ne connaît pas les biais inhérents à ce logiciel, car ses résultats n'ont jamais été validés à l'aide d'une autre technologie (c'est le cas pour la plupart des logiciels propriétaires utilisés dans le domaine du diagnostic cancer). Enfin, la reproductibilité des résultats est limitée par le fait que ChAS n'est pas

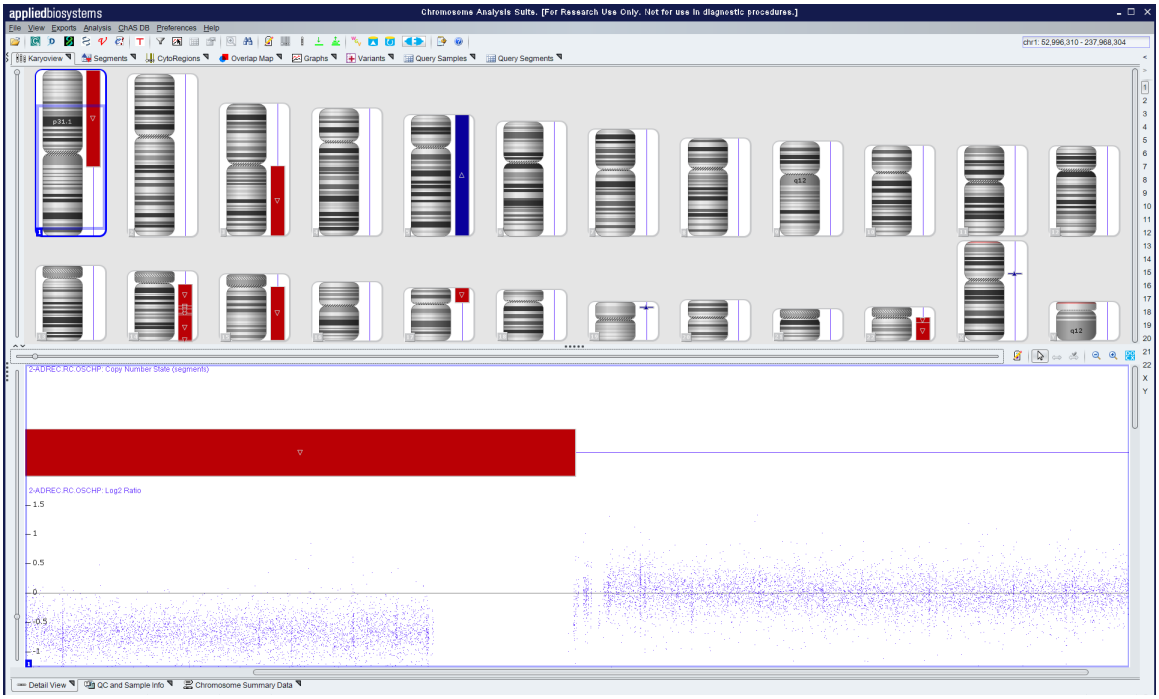

FIGURE 1.2 – Interface graphique du logiciel ChAS. Les segments d'altération (rouge, bleu foncé) sont représentés le long des chromosomes concernés.

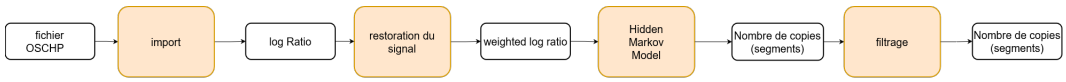

FIGURE 1.3 – Le Pipeline du logiciel ChAS qui détermine les altérations de nombre de copies.

open source.
On étudie ici quatre outils open source qui sont tous des packages R.

1.2.2 oncoscanR

OncoscanR[4] est un package R qui détermine les altérations de nombre de copies à l'échelle des bras chromosomiques en deux étapes (fig.1.4). Les segments d'altération déterminés par le logiciel ChAS subissent un nettoyage visant à fusionner certains segments et supprimer les artefacts, puis ils sont utilisés pour évaluer le pourcentage d'altération de chaque bras. Si un bras donné est altéré au-dessus du seuil défini, l'altération est validée ; dans le cas contraire, le bras est considéré comme non altéré. Un état binaire est donc obtenu pour chaque bras, ce qui limite le nombre d'altérations pouvant être trouvées à 46, le nombre de bras chromosomiques. En parallèle, oncoscanR permet le calcul de scores moléculaires d'intérêt dans la caractérisation des tumeurs.

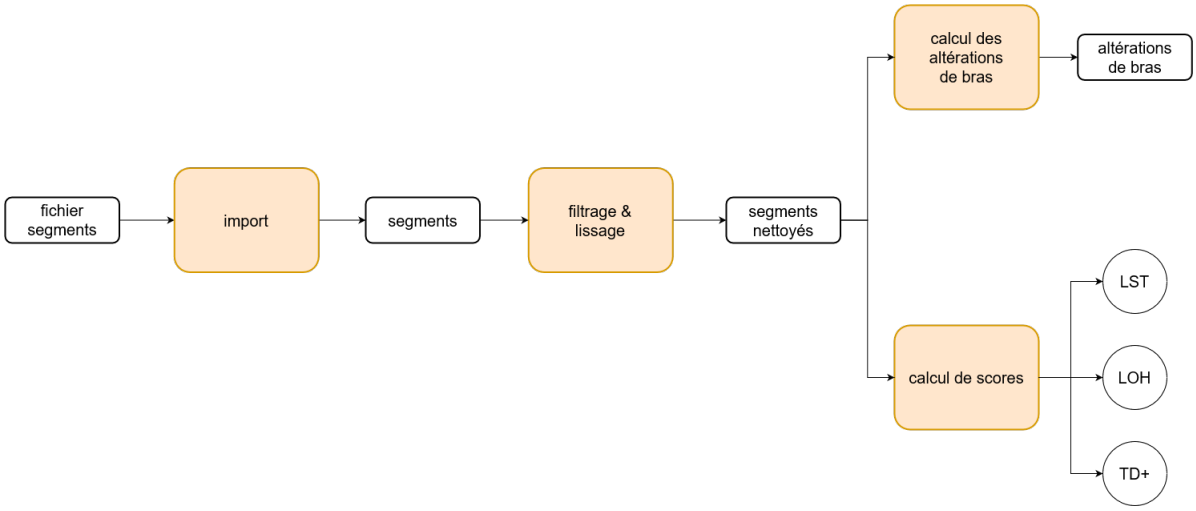

FIGURE 1.4 – Le Pipeline d'OncoscanR qui détermine les altérations de nombre de copies.

1.2.3 rCGH

Le package rCGH[5] détermine les segments d'altération en quatre étapes (fig.1.5). Un changement d'échelle (rescaling) est d'abord effectué sur les données de log ratio par SNP. Ces données sont ensuites segmentées en régions de log ratio similaire par l'algorithme Circular Binary Segmentation (CBS). Les données de segmentation sont ensuite normalisées par un modèle de mélange, qui permet de trouver le groupe de segments étant le plus probablement au niveau normal et ce, même si la majorité du profil est altérée. En parallèle de la normalisation, le nombre de copies de chaque segment est estimé, ce qui permet d'identifier les segments d'altération. rCGH met à disposition un système de visualisation interactive des données générées par son pipeline qui permet de retravailler un profil et d'en faire ressortir des gènes d'intérêt [note : La comparaison entre ChAS et cet aspect de rCGH sera faite en partie Méthodes].

1.2.4 CGHcall

CGHcall[6] 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.1.6). CGHcall prend en entrée les données de log Ratio par SNP, et commence par leur appliquer plusieurs traitements lors de l'étape de preprocess, qui ont pour but de préparer les données aux étapes suivantes. Les données sont ensuite normalisées par la médiane ou le mode, et subissent un lissage des outliers (points dont la valeur est très différente de ceux qui les entourent). L'étape suivante est la segmentation par l'algorithme CBS. Une deuxième normalisation est appliquée, cette fois sur les données segmentées, puis le nombre de copies de chaque segment est estimé lors du calling. Cette étape est effectuée par un modèle de mélange gaussien. L'intérêt d'utiliser un tel modèle est qu'il cherche à classer les segments en catégories représentant des statuts biologiques (gain, perte, normal), en comparaison d'estimations différentes qui ignorent cet aspect.

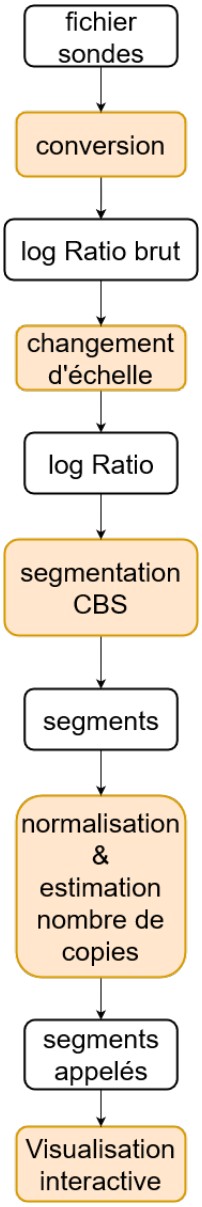

FIGURE 1.5 – Le Pipeline de rCGH qui détermine les altérations de nombre de copies.

1.2.5 ASCAT

Le package R ASCAT[7] détermine les altérations de nombre de copies en deux étapes. Les données log Ratio et BAF par SNP sont extraites du fichier puce (.OSCHP) généré par le logiciel ChAS et données en entrée à ASCAT. La segmentation Allele-Specific Piecewise Constant Fitting (ASPCF) utilise ces deux valeurs en parallèle pour déterminer les segments de même log Ratio. Le calling Allele-Specific Copy number Analysis of Tumors (ASCAT) attribue ensuite un nombre de copies à chaque segment en estimant la ploïdie et la cellularité du profil. Utiliser le BAF 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. pipeline 1.7

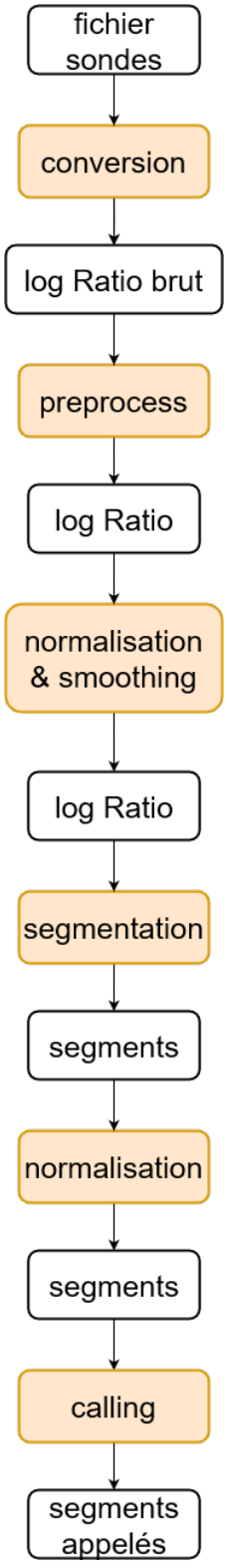

FIGURE 1.6 – Pipeline de détermination du nombre de copies du package CGHcall

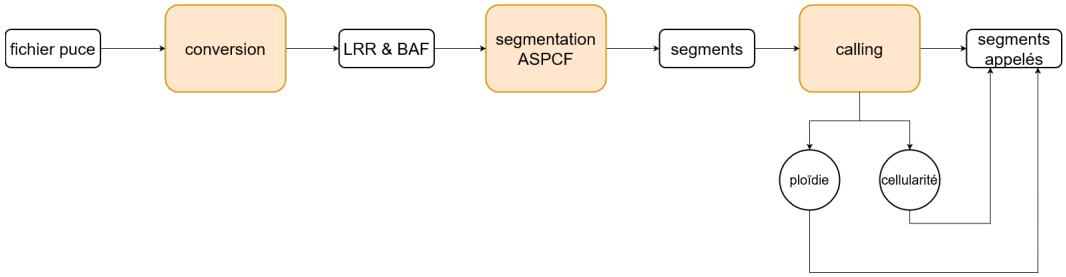

FIGURE 1.7 – Pipeline d'ASCAT aboutissant au nombre d'altérations de nombre de copies

Outil	oncoscanR	rCGH	CGHcall	ASCAT
input	Fichier Segments	Fichier Sondes	Fichier Sondes	Fichier puce
nettoyage des données	oui	oui	deux	non
Normalisation/recentrage	non	recentrage	deux	non
Segmentation	non	CBS	CBS	ASPCF
Calling	non	non (estimation)	Modèle de mélange	ASCAT
autre	calcul des altérations de bras calcul de scores	Visualisation interactive	non	estimation ploïdie et cellularité
output permettant le calul du GI	altérations de bras scores HRD	Nombre de copies par segment	Nombre de copies par sonde	Nombre de copies par segment

FIGURE 1.8 – Comparaison des apports de chaque outil dans le calcul du GI

1.2.6 Comparaison des pipelines

Les outils ne font pas tous les mêmes étapes pour déterminer le nombre d'altérations (fig.1.8). Il est à noter que d'autres outils auraient eu leur pertinence dans ce travail comme GISTIC[8] et oncoSNP[9], mais par manque de temps, ils n'ont pas été retenus. D'autre part, ils sont écrits en matlab, et la comparaison d'outils basés sur le langage R a été favorisée.

Chapitre 2

Matériel et Méthodes

2.1 Matériel

Les données utilisées sont des échantillons de tumeurs Gastrointestinales (GIST). L'ADN a été extrait de biopsies et transféré en bloc FFPE, puis a été utilisé par par la technique OncoScan CNV.

Les logiciels utilisés sont ChAS version 4.3 ; Rstudio version 2021.9.2.382[10] ; et R version 4.1.2[11].
Les packages R utilisés sont : oncoscanR version 0.1.1 ; CGHcall version 2.56.0 ; ASCAT version 3.0.0 ; et rCGH version 1.24.0 .

2.2 Méthodes

Les outils sélectionnés permettent le calcul des altérations de nombre de copies, mais ils y parviennent sans recourir aux mêmes méthodes.

2.2.1 oncoscanR

OncoscanR détermine les altérations de bras chromosomique. 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 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[12], mais le seuil qui détermine à quel pourcentage d'altération le bras est considéré comme altéré varie selon les études[12],[8].
La procédure validée par Christinat et al. décrit une définition qui 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, et le bras est dit altéré si ce pourcentage est supérieur à 90%. Le nettoyage des segments est fait avant ce calcul. Les segments de moins de 300 000 paires de bases sont supprimés (fig. 2.1A), puis les segments séparés de moins de 300 000 paires de bases sont fusionnés (fig. 2.1B).

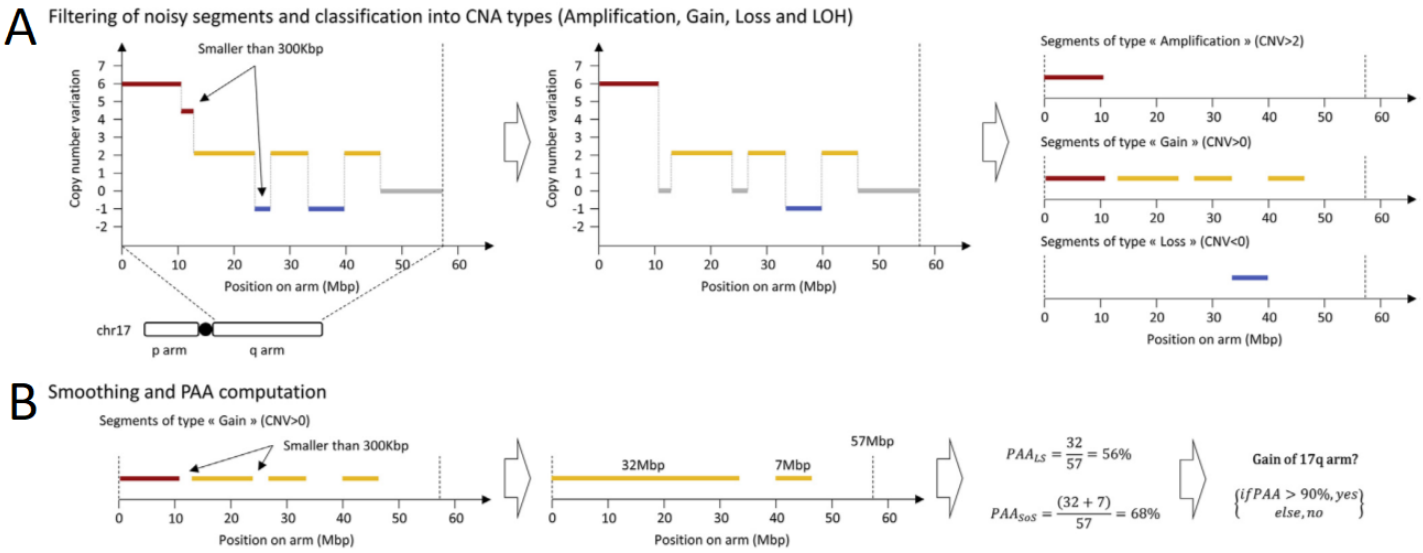

FIGURE 2.1 – Nettoyage appliqué aux données par oncoscanR. A : filtrage des segments de moins de 300 kbp. B : lissage des segments distants de moins de 300 kbp et calcul du pourcentage de bras altéré. Image extraite de Christinat et al[4]

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 un vide de moins de 300 000 paires de bases a été créé entre deux segments de même altération. Il est raisonnable de penser que ces deux segments représentent la même altération. Ils sont alors fusionnés et le vide est rempli, ce qui évite que le pourcentage de bras altéré ne soit, cette fois, sous-estimé.
Le calcul de 3 scores HRD et HRD-like est possible par oncoscanR. Le score LOH est le nombre de segments LOH >15Mbp, en excluant les chromosomes à 100% de LOH. Ce score est lié à la mutation des gènes BRCA[13]. Le score LST correspond au nombre de LST (Large-scale State Transition). Un LST est un point de cassure (breakpoint) entre deux régions de plus de 10 Mbp chacune. Ce score est lié à la mutation des gènes BRCA[14]. Le score TD+ correspond au nombre de segments TD (Tandem Duplication) entre 1 et 10 Mb. Une TD est la duplication d'un exon au sein d'un gène. Ce score est lié à la mutation du gène CDK12. La perte de ce gène a elle-même un lien avec une HRD dans les tumeurs ovariennes.

2.2.2 rCGH

La mise à l'échelle des données effectuée par rCGH est visible dans la figure 2.2. La dispersion des valeurs de log ratio se fait en divisant chaque valeur par la moyenne quadratique de l'ensemble du profil.

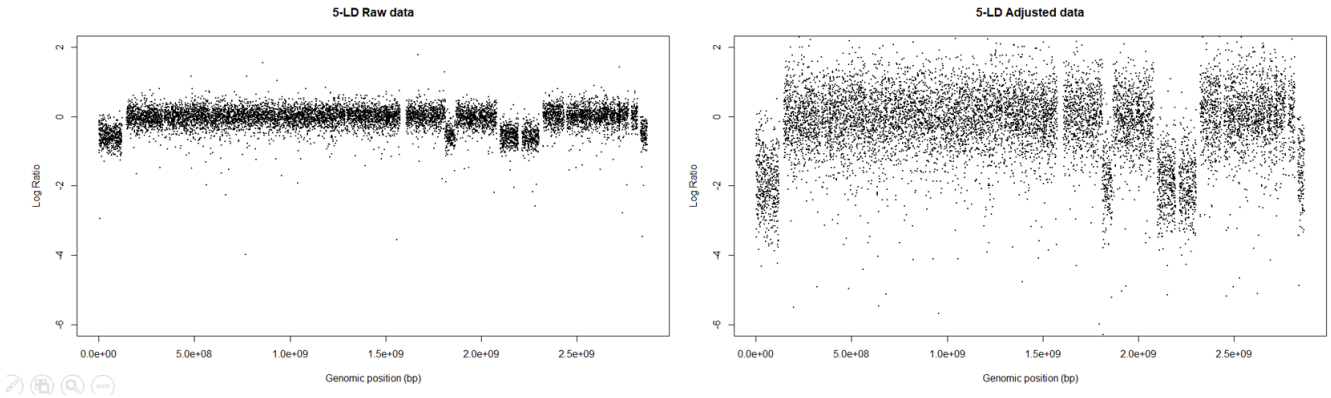

FIGURE 2.2 – Changement d'échelle des valeurs de log Ratio par rCGH. À gauche, le profil brut. À droite, le profil mis à l'échelle.

La segmentation réalisée par rCGH utilise l'algorithme Circular Binary Segmentation[15] (CBS), qui est parmi les plus performants [16]. Son principe est le suivant : Pour une région donnée, une fenêtre glissante parcourt toutes les positions et cherche la sous-région au log Ratio moyen le plus différent possible du reste de la région (fig.2.3B). Cette opération est répétée pour toutes les tailles de fenêtre glissante de 1 à l, où l est la taille de la région. Là où la différence de log Ratio entre la sous-région et le reste de la région est la plus grande, des points de coupure sont créés pour les séparer (fig.2.3A). 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.

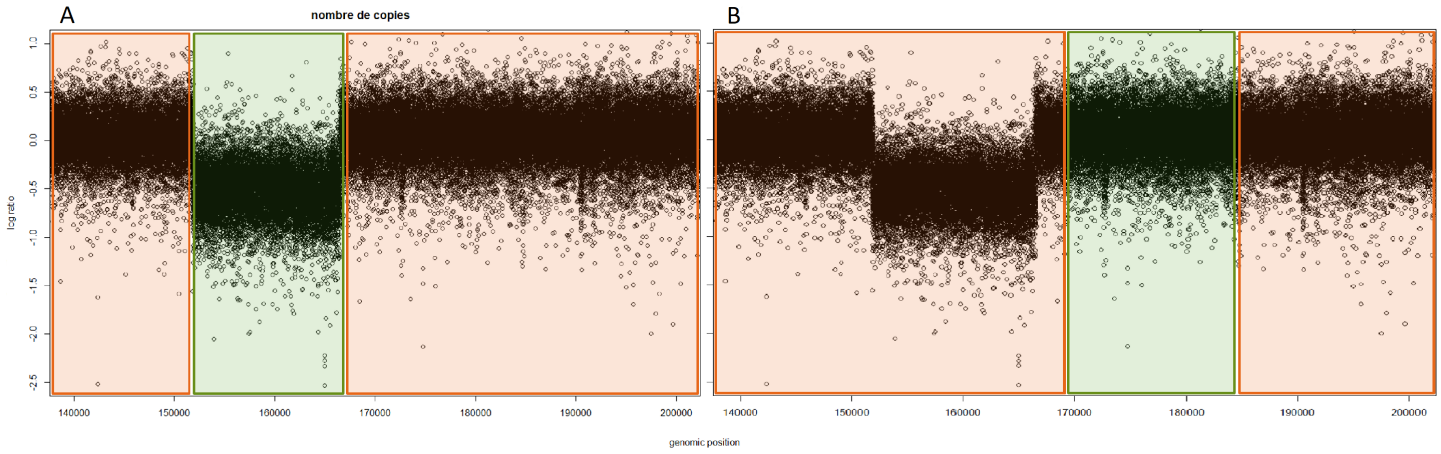

FIGURE 2.3 – Fonctionnement par fenêtre glissante de l'algorithme CBS. A : En vert : fenêtre coulissante. En orange : le reste de la région.

L'algorithme CBS annule ensuite certains points de cassure qui correspondent à des tendances locales à l'aide de trois paramètres. Cependant, 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 d'un de ces paramètres clé à partir du bruit des données dans le but de mieux s'adapter à chaque profil.

La normalisation réalisée par rCGH met en oeuvre un modèle de mélange gaussien pour trouver le niveau normal de deux copies et recentrer le profil dessus. Pour cela, les segments sont considérés comme les individus d'une population de distributions gaussiennes. Le modèle de mélange va classer les segments en groupes selon leurs valeurs, créant par exemple trois groupes dont les pics sont visibles sur la figure 2.4). Le programme pourrait alors recentrer le profil sur la valeur du plus grand pic, qui correspond au plus grand groupe de segments de même nombre de copies, autrement dit le groupe qui a le plus de chances de correspondre au niveau normal. Cependant, pour un échantillon globalement en gain (triploïde ou tétraploïde ou plus), le plus grand pic rassemble les segments de gain ; normaliser par le pic le plus grand n'est donc pas toujours fiable. Pour empêcher ça, rCGH utilise une approche qui a fait ses preuves [17]. Parmi les pics trouvés, seuls ceux dont la hauteur dépasse la moitié du plus grand pic sont retenus. De ces pics, le plus à gauche sera utilisé pour recentrer le profil. Cette normalisation assure donc que le profil est bien centré, même si il est potentiellement très altéré à l'origine.

Le nombre de copies de chaque segment est estimé par rCGH après la normalisation. Cette estimation dépend de la ploïdie "a priori" du profil, qu'il est nécessaire de renseigner en entrée du programme. Or, dans le cas d'un profil dont la ploïdie est complexe à déterminer, on ne peut intrinsèquement pas la donner à rCGH.

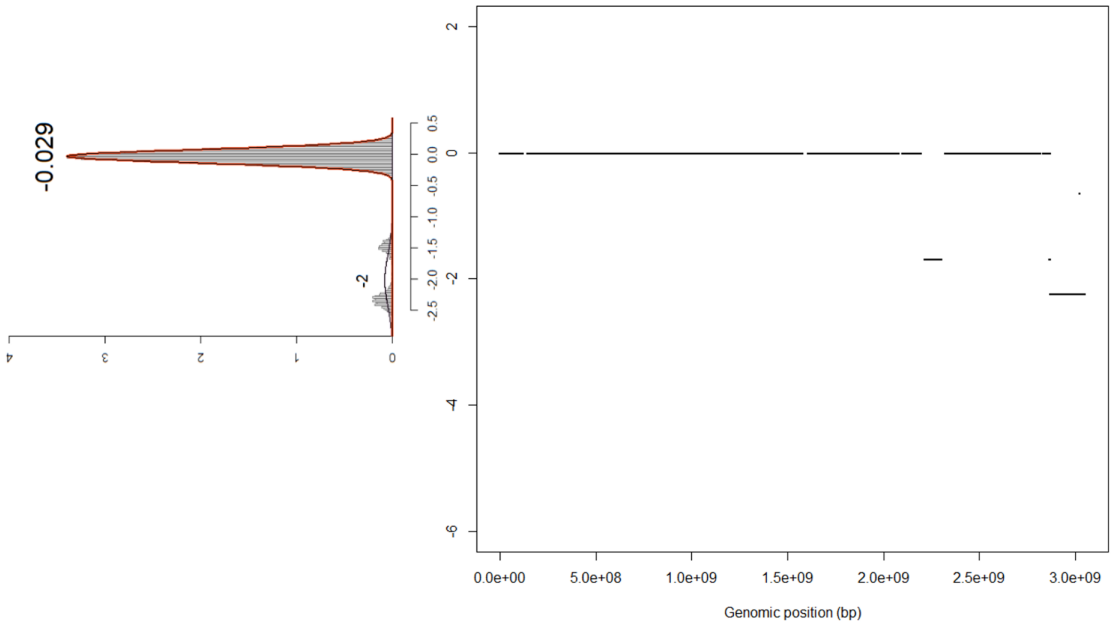

FIGURE 2.4 – Normalisation par modèle de mélange de rCGH. Les segments sont considérés comme une population de gaussiennes dont on retrouve les pics de densité à gauche. Ici, le plus grand pic est utilisé pour recentrer le profil.

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. 2.5). Cette interface permet de recentrer un profil et d'éditer les segments plus rapidement que le logiciel ChAS, mais prend plus de temps à afficher les changements. D'autre part, beaucoup de fonctionnalités de ChAS ne sont pas présentes dans cette

interface.

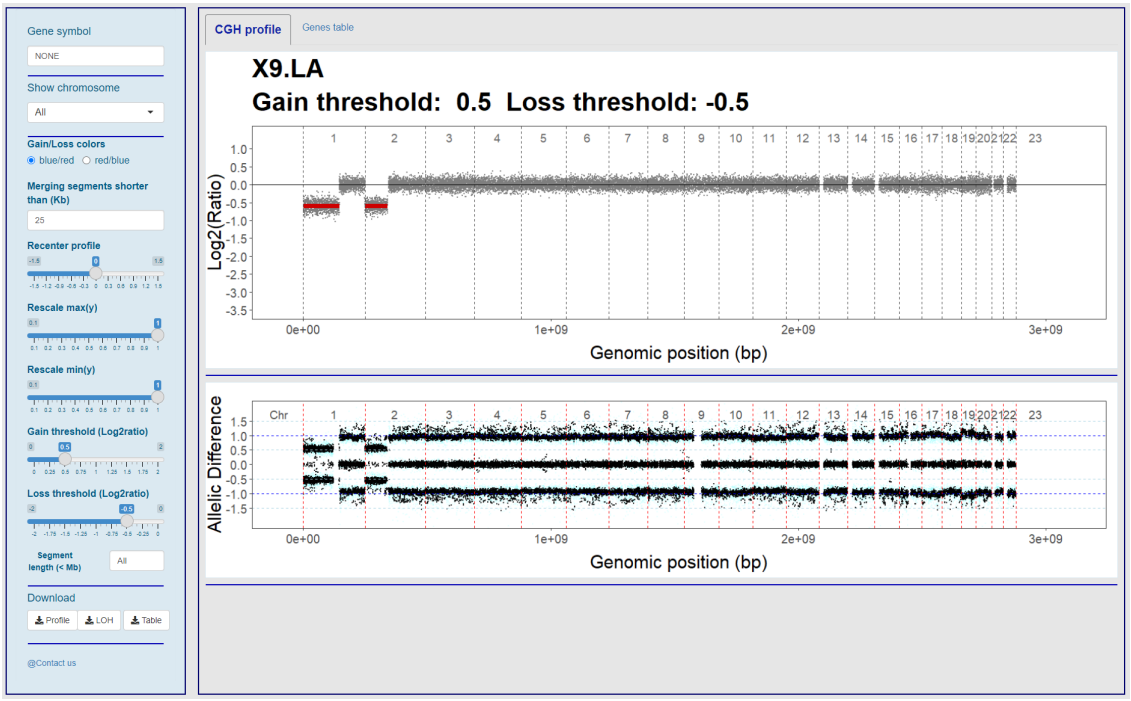

FIGURE 2.5 – Interface graphique de visualisation des données dans rCGH.

2.2.3 CGHcall

Pour le package CGHcall, les étapes de détermination du nombre de copies procèdent comme suit.

L'étape de preprocess prépare les données pour que les autres étapes du pipeline se déroulent sans problème. Ainsi, pour une cohorte de plusieurs échantillons, les données manquantes chez plusieurs échantillons sont supprimées pour tous (fig.2.6, et remplace les données manquantes qu'il reste par des valeurs estimées.

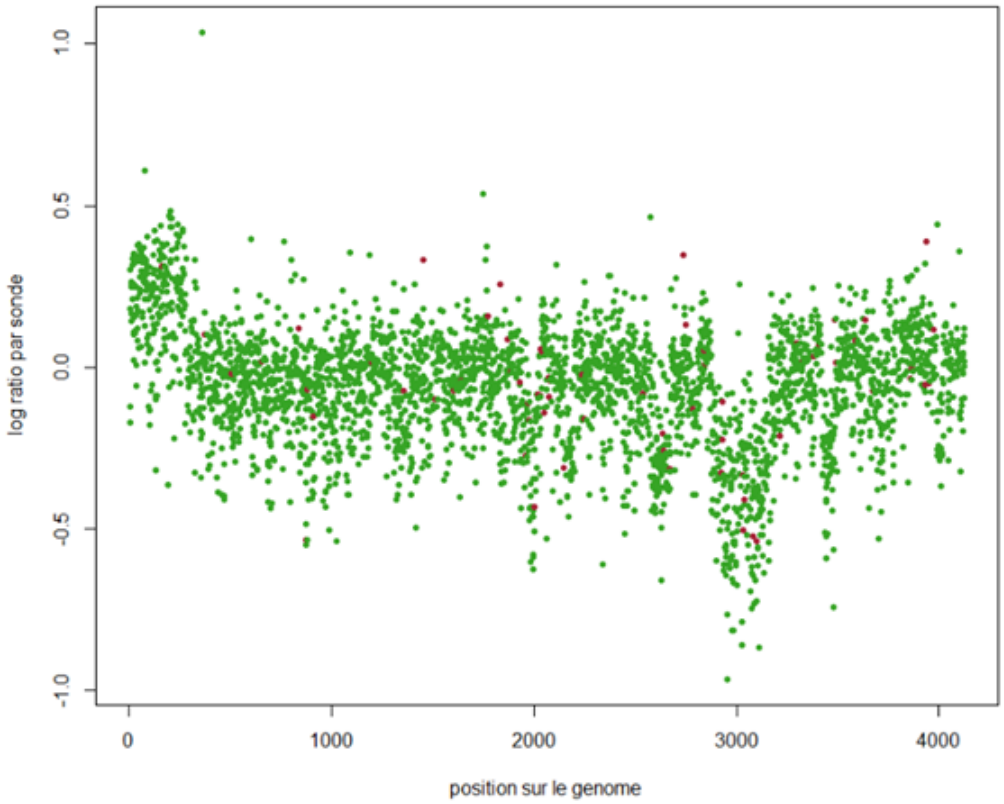

FIGURE 2.6 – Effet du preprocess sur un échantillon. Des points de données ont été retirés (en rouge) car sur l'ensemble de la cohorte, un nombre important d'échantillons ont des valeurs manquantes pour ces points. Les autres points (en vert) ont été conservés.

La première normalisation centre le profil par la médiane ou le mode des données et lisse les outliers.

La segmentation est effectuée par l'algorithme CBS. Les valeurs par défaut sont utilisées.

La deuxième normalisation trouve le niveau zéro de manière plus poussée que la première. Dans les données de log Ratio, l'intervalle contenant les données les plus segmentées est recherché de manière récursive (fig.2.7). 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 5 cycles, la valeur centrale du dernier intervalle trouvé est soustraite au profil pour le centraliser.

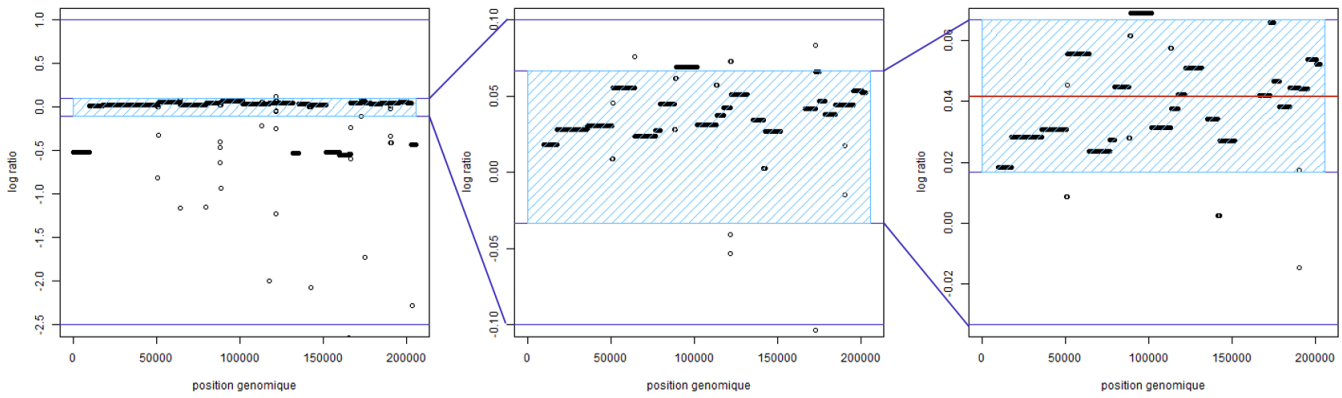

FIGURE 2.7 – Mode de recherche du niveau zéro de CGHcall.

Lors du calling, à chaque segment, un nombre de copies ayant un sens biologique est estimé à l'aide d'un modèle statistique de mélange. Les segments d' un échantillon sont mélangés pour former une unique population (courbe rouge, fig.2.8), 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 si il est renseigné. D'autre part, les segments de *tous* les échantillons d'une cohorte peuvent être utilisés pour ce calcul. Le statut des segments est donc déterminé pour tous les échantillons à la fois, mais chaque échantillon garde ses segments à la fin du processus.

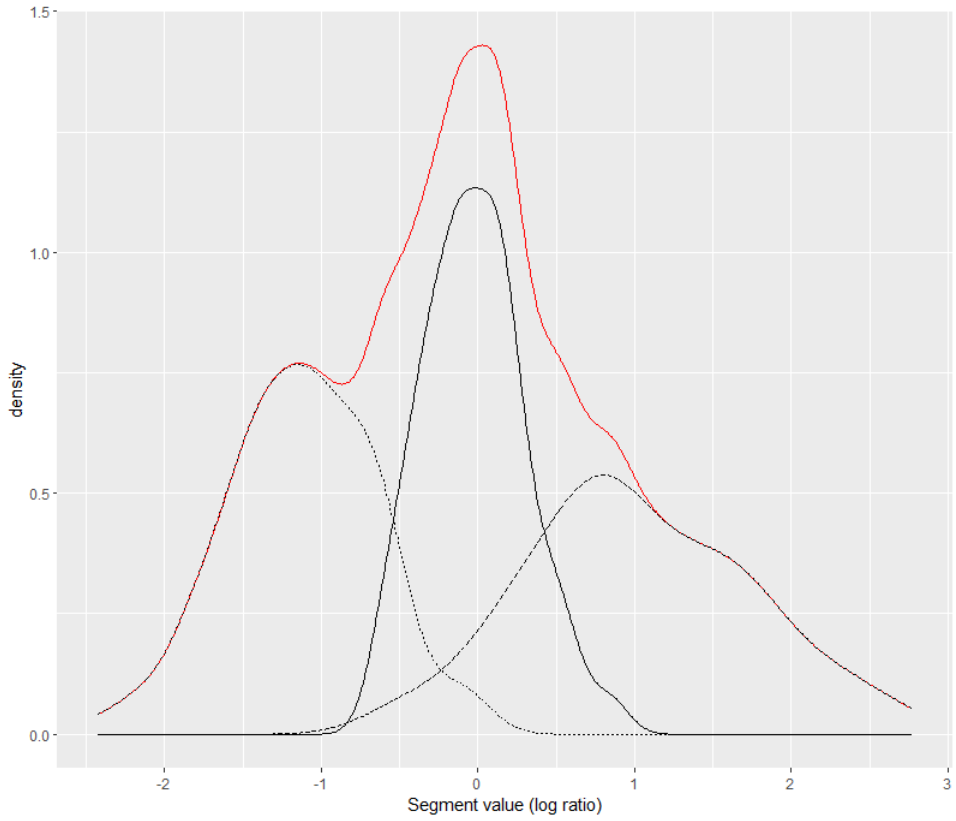

FIGURE 2.8 –

2.2.4 ASCAT

Comment compare-t-on les outils ? § GI comment il est calculé : à partir des segments d'altération ce qui a dû être adapté pour que les 4 outils fassent ce calcul Y a-t-il une proportionnalité similaire entre Agilent et OncoScan ? -> graphe de corrélation (ici, on annonce que cette question est posée dans la partie résultats). Discuter l'amplitude, le cutoff, a-t-on 2 ou 3 groupes...

GGally

§ performance vitesse précision courbes ROC

Chapitre 3

Résultats et Discussion

3.1 Résultats

Sachant les plus et moins de chaque outil, on fait une comparaison des outils :

GI Y a-t-il une proportionnalité similaire entre Agilent et OncoScan ? -> graphe de corrélation (ici, on répond à cette question). Discuter l'amplitude, le cutoff, a-t-on 2 ou 3 groupes... Y a-t-il des outils qui n'en expriment pas ?

performance vitesse : quel est l'outil le plus rapide précision : quel est l'outil le plus précis ? (On prend aussi en compte la spécificité, et d'autres paramètres...) courbes ROC Sur cette comparaison, quel est le meilleur compromis vitesse/précision ?

3.2 Discussion

+ et - de chaque outil oncoscanR : L'intérêt d'utiliser les ALA pour le calcul du GI : ça reste une façon de déterminer les altérations de nombre de copies, même si le nombre d'altérations en elle-mêmes est réduit. Cependant, les données sont transformées : à partir des segments d'altération, on détermine un état binaire pour chaque bras chromosomique : altéré ou non. Le nombre d'altérations utilisé 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 GI. Sur les profils qui présentent de nombreux segments, le GI ainsi calculé sera donc limité par le nombre de bras chromosomiques

Au vu de ce compromis, du GI obtenu et des points + et - de chaque outil, quel est l'outil retenu ?

Conclusion sur la question

Bibliographie

[1] S. Lal, A. E. M. Reed, X. M. de Luca et P. T. Simpson, « Molecular signatures in breast cancer, » *Methods*, t. 131, p. 135-146, 2017.

[2] F. Chibon, P. Lagarde, S. Salas et al., « Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity, » *Nature medicine*, t. 16, n° 7, p. 781-787, 2010.

[3] Agilent. « SurePrint G3 Human CGH Microarray 8x60K. » (), adresse : `https://www.agilent.com/en/product/cgh-cgh-snp-microarray-platform/cgh-cgh-snp-microarrays/human-microarrays/sureprint-g3-human-cgh-microarray-8x60k-228417`. (accessed : 20.05.2022).

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[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] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.