\chapter{Matériel et Méthodes}

\section{Matériel}

Les données utilisées sont des échantillons de tumeurs Gastrointestinales (GIST). Les biopsies ont été transférées en bloc FFPE, et l'ADN en a été extrait pour être utilisé par la technique OncoScan CNV.

% préciser que je travaille sur des données brutes/traitées, quels fichiers.

% ADN tumoral, FFPE, tumeurs GIST (biopsie), patients... quelques phrases pour tout ça

% \emph{tableau comparatif par étape: l'outil fait cette étape ou non.}

Les logiciels utilisés sont ChAS version 4.3; Rstudio version 2021.9.2.382\cite{rStudio}; et R version 4.1.2\cite{R}.\\

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 .

\section{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 compréhension détaillée de leurs différences est nécessaire pour apporter des réponses à ce projet. Une recherche bibliographique a été menée pour établir cette compréhension.

%%%% ChAS

\subsection{Chromosome Analysis Suite}

Le logiciel propriétaire d'Affymetrix, Chromosome Analysis Suite (ChAS), détermine les altérations de nombre de copies (fig.\ref{fig:chas\_pipe}) et permet de les visualiser sur le génome de manière interactive grâce à une interface graphique. Cette dernière est représentée par la figure \ref{fig:chas\_visu\_profile}.

idéogrammes représentant humainsrégionse : :Le rectangle bleu indique la région du génome que l’utilisateur souhaite agrandir pour obtenir une vue détaillée présentée dans le p. Dans ce panneau, le segment rouge correspond au segment perdu (en rouge) du chromosome 1 et les points bleus représentent les positions génomiques capturées par les sondes sur la CGH. L’axe y indique le log ratio et l’axe des abscisses correspond aux positions génomiques

Le panneau du bas représente le log Ratio (points en bleu) et le nombre de copies estimé du chromosome entouré en bleu dans le panneau du haut. Le profil log Ratio peut être séparé en deux régions correspondant aux bras du chromosome 1 séparées par une région intermédiaire dépourvue de points correspondant au centromère, le point de jonction des bras chromosomiques, que la technologie oncoScan CNV ne couvre pas. Les régions des bras ne présentent pas le même log Ratio moyen. Le logiciel ChAS interprète la valeur plus faible du bras p (région de gauche) par la perte d'une copie chromosomique et déclare la région comme étant un segment altéré. Il est aussi possible de manipuler le profil en le recentrant ou en fusionnant des segments, ou de rechercher des gènes spécifiques 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, sont également des étapes manuelles. Plus nombreuses sont les interventions humaines, plus grand est le risque d'erreur humaine, ce qui souligne l'importance de l'automatisation\cite{christinat2021automated}.

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 libre de droits.\\

% pas de possibilité de calculer \*directement\* ce score ni d'autres: si on veut, plus tard, ajouter des calculs de scores, on pourra plus facilement le faire avec un package R.

%%%% oncoscanR merged

\subsection{oncoscanR}

OncoscanR\cite{christinat2021automated} est un package R qui détermine les altérations de nombre de copies à l'échelle des bras chromosomiques en deux étapes (fig.\ref{fig:O\_pipe}).

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. \ref{fig:oncosR\_trim}A), puis les segments séparés de moins de 300 Kbp sont fusionnés (fig. \ref{fig:oncosR\_trim}B).

(rouge foncé) (jaune)(bleu) (?)

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 vide de moins de 300 000 paires de bases a été créée entre deux segments de même altération. Il est raisonnable de penser que ces deux segments représentent la même altération. Ils sont alors fusionnés, ce qui évite que le pourcentage de bras altéré ne soit, cette fois, sous-estimé.

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

% pour trancher sur un cas difficile, calculer le GI sur des altérations de bras permet de passer outre des altérations de petite taille qui n'ont pas forcément de d'utilité pour l'aide au diagnostic. \\

%% détermination des 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 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 litérature: généralement, on considère qu'elle correspond à une unique altération couvrant une grande part du bras\cite{ALA\_definition\_longestSegment}, mais le seuil qui détermine à quel pourcentage d'altération le bras est considéré comme altéré varie selon les études\cite{ALA\_definition\_longestSegment},\cite{mermel2011gistic2}.\\

% Un seuil de 70\%\cite{ALA\_definition\_longestSegment} tout comme un seuil de 98\%\cite{mermel2011gistic2} peuvent être utilisés selon le contexte de l'étude.

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\%.

%Le nettoyage des segments est fait avant ce calcul.

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\cite{abkevich2012LOHscore},\cite{popova2012LST}, et le score TD+ est lié à la mutation du gène CDK12\cite{popova2016tdplus}. La perte de ces gènes est liée au développement de tumeurs.% La perte de ce gène a elle-même un lien avec l'apparition de HRD dans les tumeurs ovariennes.

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 EA

% Point principal de l'article -> prise de décision thérapeutique -> déterminer statut gènes exploitables -> aCGH -> défis techniques:

% centralisation: a un impact important car définit la baseline (état neutre de 2 copies) -> rCGH implémente une approche qui a fait ses preuves \cite{commo2015centralization}\\

\subsection{rCGH}

Le package rCGH\cite{commo2016rcgh} détermine les segments d'altération en quatre étapes (fig.\ref{fig:r\_pipe}) et permet une étape de visualisation interactive du profil obtenu.

\begin{figure}[!htb]

\centering

\includegraphics[scale=0.35]{figures/pipelines/rCGH\_pipeline(7)(2).drawio.png}

\caption{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.}

\label{fig:r\_pipe}

\end{figure}

%% changement d'échelle

Un changement d'échelle est d'abord effectué sur les données de log ratio (figure \ref{fig:rCGHadjust}). 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.

\begin{figure}[!htb]

\centerline{\includegraphics[scale=0.4]{figures/mat\_met/matmet\_rCGH\_adjust.png}}

\caption{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. }

\label{fig:rCGHadjust}

\end{figure}

%% segmentation

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\cite{venkatraman2007DNAcopy} (CBS), qui est parmi les plus performants \cite{willenbrock2005comparison}. Son principe est le suivant : Pour une région donnée (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.\ref{fig:algoCBS}B). 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. Quand 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 séparer les régions. (fig.\ref{fig:algoCBS}A). 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.

\begin{figure}[H]

% \centering

\centerline{\includegraphics[scale=0.25]{figures/mat\_met/matmet\_DNAcopy\_crop.png}}

\caption{Fonctionnement par fenêtre glissante de l'algorithme CBS sur des données de log Ratio (en noir). 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.}

\label{fig:algoCBS}

\end{figure}

%% à quoi sert undo.SD et comment il est estimé par rCGH

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 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.

%% normalisation

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 $\mu$ et son écart-type $\sigma$. 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 \ref{fig:rCGHnorm}).

\begin{figure}[!htb]

\centerline{\includegraphics[scale=0.3]{figures/mat\_met/matmet\_rCGH\_norm.png}}

\caption{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.}

\label{fig:rCGHnorm}

\end{figure}

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 sous-estimerait le nombre de copies. rCGH utilise une approche \cite{commo2015centralization} 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, 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 si il est potentiellement très altéré à l'origine.

%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.

%% estimation CN

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 est nécessaire de renseigner en entrée du programme.

%% data visualisation

% Il n'y a pas consensus sur les seuils à utiliser, donc on doit pouvoir manipuler et visualiser le profil -> visualisation interactive

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 quant au diagnostic (fig.\ref{fig:rCGHvisu}).

\begin{figure}[H]

\centerline{\includegraphics[scale=0.3]{figures/mat\_met/matmet\_rCGH\_visu.png}}

\caption{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.}

\label{fig:rCGHvisu}

\end{figure}

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 d'allèles A multiplié par 0,5 auquel on soustrait le nombre d'allèles 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 \ref{fig:rCGHvisu} pour le chromosome 3 notamment. % 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 -> dire lesquelles.

Il est aussi possible de trouver à quel segment appartiennent des gènes spécifiés. %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 Résultats].

%%%% CGHcall EA

\subsection{CGHcall}

CGHcall\cite{van2007cghcall} 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.\ref{fig:C\_pipe}).

\begin{figure}[!htb]

\centering

\includegraphics[scale=0.4]{figures/pipelines/CGHcall\_pipeline.drawio.png}

\caption{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.}

\label{fig:C\_pipe}

\end{figure}

%% preprocess

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.

Ainsi, pour une cohorte de plusieurs échantillons, les données manquantes chez plus de 30\% des échantillons sont supprimées pour tous. (fig.\ref{fig:cghcall\_prepro}. Les données manquantes restantes sont estimées à l'aide du package R impute\cite{hastie2011Rpkg\_impute}

\begin{figure}[H]

\centerline{\includegraphics[scale=0.9]{figures/mat\_met/matmet\_cghcall\_preprocess.png}}

\caption{Effet du pré-traitement sur des données de log Ratio. En rouge, points supprimés des données. En vert, points conservés.}

\label{fig:cghcall\_prepro}

\end{figure}

%% normalisation 1

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\cite{tukey1977outliers}.

%% segmentation CBS

L'étape suivante est la segmentation par l'algorithme CBS. CGHcall n'implémente pas de fonctionnalité additionnelle.

%% normalisation 2

Une deuxième normalisation est ensuite appliquée. 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 (fig.\ref{fig:cghcall\_norm2}). 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.

\begin{figure}[!h] %%%%%%% l'image apparaît au bon endroit quand elle est appelée ici, ne pas y toucher

\centerline{\includegraphics[scale=0.4]{figures/mat\_met/matmet\_cghcall\_norm2.png}}

\caption{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, agrandissement sur un intervalle plus restreint. Les zones hachurées en bleu correspondent au meilleur intervalle trouvé à chaque étape. }

\label{fig:cghcall\_norm2}

\end{figure}

%% Calling par modèle de mélange

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 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.

Pour ce faire, les segments d'un échantillon sont mélangés pour former une unique population (courbe rouge, fig.\ref{fig:cghcall\_call}), 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.

%%%% ASCAT EA

\subsection{ASCAT}

Le package R ASCAT\cite{van2010ascat} détermine les altérations de nombre de copies en deux étapes (fig.\ref{fig:A\_pipe}).

\begin{figure}[!hb]

\centering

\includegraphics[scale=0.3]{figures/pipelines/ASCAT\_pipeline(3).drawio.png}

\caption{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.}

\label{fig:A\_pipe}

\end{figure}

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.

%% Segmentation ASPCF

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.

%Sur la figure \ref{fig:ascatseg\_breakpoints} notamment, cela s'illustre sur les chromosomes 1 et 11. %% cette figure a été déplacée dans l'intro pour définir les concepts de logRatio et BAF.

ASPCF cherche le partitionnement optimal de la région génomique en segments. Un partitionnement optimal minimise le critère d'optimisation indiqué en figure \ref{fig:ascatseg\_equa}.

\begin{figure}[H]

\centerline{\includegraphics[scale=0.7]{figures/mat\_met/matmet\_ascat\_seg\_equation.png}}

\caption{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.}

\label{fig:ascatseg\_equa}

\end{figure}

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.\ref{fig:ascatseg\_gofs}, 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.\ref{fig:ascatseg\_gofs}, 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, le partitionnement est moins crédible qu'un partitionnement qui séparerait les segments hétérogènes en plusieurs sous-segments.

\begin{figure}

\centerline{\includegraphics[scale=0.35]{figures/mat\_met/matmet\_ASCAT\_seg\_deux\_segments.png}}

\caption{En haut, les données de log Ratio d'un profil à traiter par ASPCF. 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.}

\label{fig:ascatseg\_gofs}

\end{figure}

Le terme en dehors des crochets est une pénalité 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.

%%%% ASCAT matmet

La méthode de calling de l’outil ASCAT 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. En effet, 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.

% 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.

%Ensuite, le calling ASCAT détermine le nombre de copies des deux allèles de chaque SNP à partir des données de segmentation et de l'estimation de deux caractéristiques du profil: la cellularité et la ploïdie.

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 (\ref{eq:logR},\ref{eq:BAF}).

\begin{equation} \label{eq:logR}

% \[

r\_i = \gamma log\_2\left(\frac{n\_{A,i} + n\_{B,i}}{2}\right)

% \]

\end{equation}

\begin{equation} \label{eq:BAF}

% \[

b\_i = \frac{n\_{B,i}}{n\_{A,i} + n\_{B,i}}

% \]

\end{equation}

Dans ces équations, $r\_i$ et $b\_i$ représentent respectivement log Ratio et BAF. Le paramètre $\gamma$ 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 $\psi$, ce qui donne l'équation en \ref{eq:logR\_psi}.

\begin{equation} \label{eq:logR\_psi}

% \[

r\_i = \gamma log\_2\left(\frac{n\_{A,i} + n\_{B,i}}{\psi}\right)

% \]

\end{equation}

Dans le cas d'une contamination de l'échantillon par des cellules non aberrantes, 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) multiplié par leur proportion % (figure \ref{fig:ascat\_call\_equations}C, $\rho$ représente le pourcentage de cellules tumorales).

Prendre en compte la cellularité (le pourcentage de cellules tumorales) dans le calcul du log Ratio et du BAF implique donc d'adopter les équations \ref{eq:logR\_psi\_rho} et s\ref{eq:BAF\_psi\_rho}.

\begin{equation} \label{eq:logR\_psi\_rho}

% \[

r\_i = \gamma log\_2\left(\frac{2(1-\rho) + (n\_{A,i} + n\_{B,i})}{\psi}\right)

% \]

\end{equation}

\begin{equation} \label{eq:BAF\_psi\_rho}

% \[

b\_i = \frac{1-\rho + \rho n\_{B,i}}{2 - 2\rho + (n\_{A,i} + n\_{B,i})}

% \]

\end{equation}

Les équations obtenues expriment log Ratio et BAF en fonction de la cellularité $\rho$, de la ploïdie $\psi$ 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. (\ref{eq:nCopyA},\ref{eq:nCopyB}).

\begin{equation} \label{eq:nCopyA}

% \[

\hat{n}\_{A,i} = \frac{ \rho - 1 + 2^{ \frac{r\_i}{\gamma} } (1-b\_i) ( 2 ( 1-\rho) + \rho\psi) }{\rho}

% \]

\end{equation}

\begin{equation} \label{eq:nCopyB}

% \[

\hat{n}\_{B,i} = \frac{ \rho - 1 + 2^{ \frac{r\_i}{\gamma} } b\_i ( 2 ( 1-\rho) + \rho\psi) }{\rho}

% \]

\end{equation}

Le calling d'ASCAT utilise ces dernières équations 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. \ref{fig:ascat\_call\_sunrise}), 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. \ref{fig:ascat\_call\_bestSol}).

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

% \begin{figure}

% \centerline{\includegraphics[scale=0.2]{figures/mat\_met/matmet\_ascat\_calling\_equations.png}}

% \caption{Comment ASCAT exprime le nombre de copies en fonction de la ploïdie, de la cellularité, du log Ratio et du BAF.}

% \label{fig:ascat\_call\_equations}

% \end{figure}

\begin{figure}

\centerline{\includegraphics[scale=0.75]{figures/mat\_met/matmet\_ascat\_calling\_bestSolution.png}}

\caption{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.}

\label{fig:ascat\_call\_bestSol}

\end{figure}

\begin{figure}

\centerline{\includegraphics[scale=0.25]{figures/mat\_met/matmet\_ascat\_calling\_sunrise.png}}

\caption{La qualité de l'ajustement de toutes les solutions testées par le calling ASCAT (rouge, faible qualité, bleu, qualité élevée, blanc, intermédiaire). Ploidy: Ploïdie. Putity: Cellularité tumorale. La solution retenue est marquée d'une croix verte.}

\label{fig:ascat\_call\_sunrise}

\end{figure}

%% le début de la partie méthodes dit:

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

\subsection{Comparaison des pipelines}

Les différences entre les outils sont résumées dans la figure\ref{fig:tabPipes}.

Il est à noter que d'autres outils auraient eu leur pertinence dans ce travail comme GISTIC\cite{mermel2011gistic2} et oncoSNP\cite{yau2010oncoSNP}. Ces outils sont écrits dans le langage de programmation matlab\cite{MATLABR2021a}, 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.

\begin{figure}

\centering

\includegraphics[scale=0.6]{figures/pipelines/tableauPipelines.png}

\caption{Comparaison des apports de chaque outil dans le calcul du GI. Input, type de données nécessaire en entrée de l'outil. CBS, Circular Binary Segmentation; ASPCF, Allele-Specific Piecewise Constant Fitting. ASCAT, Allele-Specific Copy number Analysis of Tumors }

\label{fig:tabPipes}

\end{figure}

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. %% dire que pour oncoscanR, une altération de bras = une altération %% pas besoin, déjà dit, ainsi que le fait qu'on a 46 altérations max. 42 vu qu'on exclut les bras des chr X et Y.

% § 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 -> on l'a dit individuellement.

On cherche à savoir si 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 Spearman %(ici, on annonce que cette question est posée dans la partie résultats). utiliser Spearman, plus fiable.

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.