Méthodes bioinformatiques pour l'analyse de données de séquençage dans le contexte du cancer

Justine Rudewicz

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

par **Justine Rudewicz**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

Méthodes bioinformatiques pour l'analyse de données de séquençage dans le contexte du cancer

---

Date de soutenance : 30 juin 2017

Devant la commission d'examen composée de :

Valentina Boeva	Chargé de recherche ..	Rapporteur	INSERM
Jacques Colinge	Professeur	Rapporteur	Université de Montpellier I
Sylvain Marchand ..	Professeur	Examinateur	Université de La Rochelle
Jean-Philippe Merlio	Professeur	Président du jury	Université de Bordeaux
Macha Nikolski	Directeur de recherche	Directeur de thèse	CNRS
David Santamaria ..	Directeur de recherche	Examinateur	INSERM

– 2017 –

Titre Méthodes bioinformatiques pour l'analyse de données de séquençage dans le contexte du cancer

Résumé Le cancer résulte de la prolifération excessive de cellules qui dérivent toutes de la même cellule initiatrice et suivent un processus Darwinien de diversification et de sélection. Ce processus est défini par l'accumulation d'altérations génétiques et épigénétiques dont la caractérisation est un élément majeur pour pouvoir proposer une thérapie ciblant spécifiquement les cellules tumorales. L'avènement des nouvelles technologies de séquençage haut débit permet cette caractérisation à un niveau moléculaire. Cette révolution technologique a entraîné le développement de nombreuses méthodes bioinformatiques.

Dans cette thèse, nous nous intéressons particulièrement au développement de nouvelles méthodes computationnelles d'analyse de données de séquençage d'échantillons tumoraux permettant une identification précise d'altérations spécifiques aux tumeurs et une description fine des sous populations tumorales.

Dans le premier chapitre, il s'agît d'étudier des méthodes d'identification d'altérations ponctuelles dans le cadre de séquençage ciblé, appliquées à une cohorte de patientes atteintes du cancer du sein. Nous décrivons deux nouvelles méthodes d'analyse, chacune adaptée à une technologie de séquençage, spécifiquement Roche 454 et Pacifique Biosciences. Dans le premier cas, nous avons adapté des approches existantes au cas particulier de séquences de transcrits. Dans le second cas, nous avons été confronté à un bruit de fond élevé entraînant un fort taux de faux positifs lors de l'utilisation d'approches classiques. Nous avons développé une nouvelle méthode, MICADo, basée sur les graphes de De Bruijn et permettant une distinction efficace entre les altérations spécifiques aux patients et les altérations communes à la cohorte, ce qui rend les résultats exploitables dans un contexte clinique.

Le second chapitre aborde l'identification d'altérations de nombre de copies. Nous décrivons l'approche mise en place pour leur identification efficace à partir de données de très faible couverture. L'apport principal de ce travail consiste en l'élaboration d'une stratégie d'analyse statistique afin de mettre en évidence des changements locaux et globaux au niveau du génome survenus durant le traitement administré à des patientes atteintes de cancer du sein. Notre méthode repose sur la construction d'un modèle linéaire permettant d'établir des scores de différences entre les échantillons avant et après traitement.

Dans le troisième chapitre, nous nous intéressons au problème de reconstruction clonale. Cette problématique récente est actuellement en plein essor, mais manque cependant d'un cadre formel bien établi. Nous proposons d'abord une formalisation du problème de reconstruction clonale. Ensuite nous utilisons ce formalisme afin de mettre en place une méthode basée sur les modèles de mélanges Gaussiens. Cette méthode utilise les altérations ponctuelles et de nombre de copies - comme celles abordées dans les deux chapitres précédents - afin de caractériser et quantifier les différentes populations clonales présentes dans un échantillon tumoral.

Mots-clés cancer, bioinformatique, NGS, TGS, graphes de *de Bruijn*, modèles de mélanges

Title Bioinformatics methods for cancer sequencing data analysis

Abstract Cancer results from the excessive proliferation of cells decending from the same founder cell and following a Darwinian process of diversification and selection. This process is defined by the accumulation of genetic and epigenetic alterations whose characterization is a key element for establishing a therapy that would specifically target tumor cells. The advent of new high-throughput sequencing technologies enables this characterization at the molecular level. This technological revolution has led to the development of numerous bioinformatics methods.

In this thesis, we are particularly interested in the development of new computational methods for the analysis of sequencing data of tumor samples allowing precise identification of tumor-specific alterations and an accurate description of tumor subpopulations.

In the first chapter, we explore methods for identifying single nucleotide alterations in targeted sequencing data and apply them to a cohort of breast cancer patients. We introduce two new methods of analysis, each tailored to a particular sequencing technology, namely Roche 454 and Pacific Biosciences. In the first case, we adapted existing approaches to the particular case of transcript sequencing. In the second case, when using conventional approaches, we were confronted with a high background noise resulting in a high rate of false positives. We have developed a new method, MICADo, based on the De Bruijn graphs and making possible an effective distinction between patient-specific alterations and alterations common to the cohort, which makes the results usable in a clinical context.

Second chapter deals with the identification of copy number alterations. We describe the approach put in place for their efficient identification from very low coverage data. The main contribution of this work is the development of a strategy for statistical analysis in order to emphasise local and global changes in the genome that occurred during the treatment administered to patients with breast cancer. Our method is based on the construction of a linear model to establish scores of differences between samples before and after treatment.

In the third chapter, we focus on the problem of clonal reconstruction. This problem has recently gathered a lot of interest, but it still lacks a well-established formal framework. We first propose a formalization of the clonal reconstruction problem. Then we use this formalism to put in place a method based on Gaussian mixture models. Our method uses single nucleotide and copy number alterations - such as those discussed in the previous two chapters - to characterize and quantify different clonal populations present in a tumor sample.

Keywords cancer, bioinformatics, NGS, TGS, *de Bruijn* graphs, mixture models

Remerciements

Je ne saurais expliquer l'ensemble des bénéfices que cette thèse a eu sur ma vie professionnelle mais aussi personnelle (et je mettrais probablement beaucoup trop de temps en me perdant dans mes explications). Pour faire court, ces années ont été une chance pour moi et non un investissement, la curiosité et l'envie d'apprendre constante m'ont permis de ne pas voir défiler le temps. L'aboutissement de ce travail n'aurait pas été possible sans la rencontre ou encore la présence de personnes que je souhaite remercier plus particulièrement.

Je tiens tout d'abord à remercier ma directrice de thèse, Macha Nikolski, pour m'avoir offert la possibilité d'effectuer une thèse sous sa supervision et ainsi bénéficier de ses connaissances, sa bienveillance mais aussi son optimisme, son écoute, sa patience. Merci pour ta disponibilité, ton intérêt, ta confiance en moi (qui a su me porter plus d'une fois). Tu m'auras apporté énormément tant sur le plan scientifique que sur le plan humain durant ces 4 années de thèse.

Je tiens ensuite à remercier mes rapporteurs, Valentina Boeva et Jacques Collinge ainsi que l'ensemble des membres de mon jury, Sylvain Marchand, Jean-Philippe Merlio et David pour avoir accepté de participer à l'évaluation de ma thèse.

Un grand merci à Hayssam qui m'a énormément apporté statistiquement, machine learniguement, pythoneusement, ggplotement etc. Merci pour ta pédagogie, ta patiente et ta sympathie résistante à toute épreuve !

Merci beaucoup à Raluca qui m'aura transmis la connaissance, puis passion du graph de Bruijn, quand on commence, on en voit et veut partout.. Merci aussi pour ta gentillesse, ta sympathie, ta bonne humeur et ton sourire toujours présent.

Merci à l'ensemble des membres du CBiB. L'ambiance qui règne au laboratoire donne une impression quotidienne d'être en famille ou entre amis qui, doublée de professionnalisme scientifique, rend les journées de travail enrichissantes et passionnantes. Merci à toi Ben pour toutes ces histoires qui me font tellement rire et ta capacité à finir mes phrases, Émeric pour ton optimisme et tes explications parfois encore plus longues que les miennes qui me conforte que je ne suis pas la seule, Tristan pour tes conseils culinaires et séristiques mais aussi pour ces explications qui sont toujours plus longues que les miennes, merci à Aurélien pour ces partages de plats et tes petits pics/blagues toujours biens placés, Alexis pour tes conseils professionnels ou non ("prend le rouge, le rouge c'est bien"), Jeff pour tous tes conseils musicaux et toutes autres sortes de partages, Katia pour tes conseils sur toutes sortes de molécules pour bloquer le stress, June pour tes blagues dont le bide fait toujours rire, Marie pour ta joie de vivre inconditionnelle, merci à David pour ta disponibilité, ta patience et pédagogie mais aussi tes conversations, ta curiosité et gentillesse, merci à Élo qui aura été d'un grand soutien durant ma rédaction et mes chutes de confiances, merci pour ton amitié et ces heures passées au Nansouty, merci à Manu, mon google humain, pour ne pas avoir changé de place depuis toutes ces années et accepté de m'écouter parler seule (bon, tu le fait aussi..) de prendre quelques coups de pied heb-

domadaires sous le bureau, d'écouter toutes sortes de choses plus ou moins intéressantes qui m'arrivent et de toujours prendre le temps de répondre de manière constructive et de m'instruire sur toutes sortes de choses. Merci à tous ceux passés par le CBiB durant ces années, Guillaume, Thomas, Louisa, Nicolas, Emrah et tous les autres ainsi qu'à nos voisines secrétaires, Aurore, Béa et Dalila qui m'ont toujours accueillie avec le sourire pour m'expliquer des choses qui me dépassent.

Merci aux personnes rencontrées à l'institut Bergonié et au LaBRI pour leur aide et sympathie et plus particulièrement Eve, Élodie, Stéphanie, Élodie, Noël, Romaric et Vincent.

Merci à l'ensemble de mes encadrants passés qui m'ont fait découvrir l'univers de la recherche lors de mes études. Merci à Andreï Thirkov, pour m'avoir rendu passionnée pour la génétique et cancérologie durant ses cours théoriques puis permis d'effectuer mon premier stage en laboratoire ainsi qu'à Maud Privat, Yannick Bidet et Fabrice Kwiatkowski de m'avoir encadré dans l'élaboration de mes premières analyses. Ma sincère gratitude va à Hervé Bonnefoi et Richard Iggo pour m'avoir permis d'effectuer cette thèse. Merci à l'ensemble des professeurs de mathématiques et de biologie qui ont participé à l'enrichissement de mes connaissances et du développement d'une réelle passion pour ces deux disciplines que j'ai la chance de pouvoir aujourd'hui couplée grâce à l'informatique.

Merci à mes ami(e)s d'être cette bouffée d'air frais. Charlène, qui a toujours su me guider et se rendre disponible et généreuse avec moi, Angélique dont l'altruisme, la persévérance, la motivation et le courage sont pour moi exemplaires, Alexie pour ces moments de partages donnant l'impression que rien n'a changé, à Samy pour m'avoir changé les idées avec tes appels et m'avoir toujours tenue au courant des nouveaux lieux de Vichy, maintenant ce sera Paris. Merci aux familles Gravez-Marignan, Lefèvre, Laplace et Vilain pour en avoir été une seconde pour moi. Merci à mes amies de la fac d'avoir rendu ces années si plaisantes, Charline, Pauline et Bénédicte, on se voit vite, ainsi qu'à mes copines de master Louise, Cécile, Flora et Élina, au plaisir de vous voir en conf ou ailleurs ;-). Merci à Simon, Gaëlle et Marie pour ce soutien sur place et/ou à distance, conseils et écoute qui m'apportent beaucoup, mais aussi soirées tranquilles ou rocambolesques merci d'être là, d'être comme vous êtes. Merci aussi à l'ensemble de mes colocataires, Fred, Loïc, Valentin, Coralie, Romain, Jonathan, Romain et Mathieu qui ont partagé cette aventure quotidienne à leur façon et auront tous connu une phase particulière.

Merci à ma famille pour votre affection sans failles. Merci à toi papa d'être toujours présent quoi qu'il arrive même si c'est parfois en râlant ! Julian merci d'avoir toujours été un pilier, me voilà docteur en chocolatine, Jennifer et Angélina, merci pour le bonheur que vous lui apportez. Merci à toi maman, ma bonne étoile, qui est, avec du recul, le point de départ de tout cela. Merci d'être venue me soutenir le jour J, je vous aime fort.

Merci à Vincent de m'avoir enseigné que la remise en question durant la rédaction d'une thèse n'est pas la plus dure à affronter. Merci pour tout ce que tu m'apportes, ton écoute, ta patiente, ton affection.

Merci à tous ceux qui m'ont gâté de leur présence lors de ma soutenance mais aussi couvert de présents ! Mon vélo, c'est le plus beau ! !

Un dernier merci à Maxime, enfin la reconnaissance de mon bon goût dans le choix de mes chaussettes.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Biologie du cancer

La coopération des dizaines de millions de cellules composant notre organisme est essentielle au développement et à la vie. Cette coopération est maintenue par des signaux et points de contrôles cellulaires déterminant quand la cellule se divise, meurt ou se différencie (Chin et Yeong, 2010).

Le cancer peut être considéré à plusieurs niveaux d'observation. Au niveau de l'organisme, le cancer représente l'échec du maintient de cette coopération, résultant de la croissance incontrôlée de cellules indépendantes du reste de l'organisme pouvant entraîner jusqu'à sa mort (Wodarz et Komarova, 2005). Au niveau cellulaire, il est acquis que le cancer est une "maladie de l'ADN". En effet, la prolifération incontrôlée est le résultat de l'accumulation d'altérations du matériel génétique. Cette accumulation permet à la cellule de rompre le réseau de régulation nécessaire à la coopération.

1.1.1 Définition du cancer

Le cancer est une pathologie pouvant affecter l'ensemble des tissus humains. Il est caractérisé par une augmentation de la masse cellulaire provoquant la formation d'une tumeur (Lacave *et al.*, 2005) causée par une croissance cellulaire incontrôlée avec invasion des tissus environnants dans le cas de tumeurs solides (Barillot *et al.*, 2012). Certains cancers peuvent devenir métastatiques si les cellules tumorales migrent dans un site distant du site primaire de formation de la tumeur.

Les cancers peuvent être classés selon la cellule normale dont ils sont originaires :

1. carcinomes - provenant de cellules épithéliales (sein, ovaire, prostate, poumon, pancréas, colon, etc.),
2. sarcomes - provenant de tissus conjonctifs ou de soutien (os, cartilage, muscle, vaisseau sanguin, tissu adipeux, etc.),
3. lymphomes et leucémies - provenant des tissus hématopoïétiques,
4. glioblastomes, neuroblastomes, neurinome et médulloblastome - provenant des tissus nerveux (périphériques ou centraux).

Tous les cancers sont des tumeurs solides sauf les leucémies qui sont des cellules tumorales circulantes dans le sang. Les tumeurs bénignes sont des tumeurs qui ne présentent pas de capacité invasive.

Le cancer provient de la dégénérescence d'une cellule normale en une cellule cancéreuse, phénomène appelé transformation cellulaire. Cette dégénérescence cellulaire est causée par des changements dynamiques du génome (Hanahan et Weinberg, 2000) et de l'épigénome. En effet, le cancer est le résultat de l'accumulation d'altérations génétiques et épigénétiques entraînant l'acquisition de nouveaux caractères phénotypiques. Cette accumulation est progressive dans le temps et va être soumise à une pression sélective analogue au modèle d'évolution Darwinien. Ainsi, la succession des changements génétiques durant la progression tumorale confère un ou plusieurs avantages de développement et mène ainsi à la conversion progressive des cellules normales en cellules cancéreuses (Nowell, 1976).

1.1.2 Initiation, promotion et progression tumorale

Le cancer est dû à l'accumulation d'altérations génétiques ou épigénétiques. Ces altérations peuvent être causées par des stress exogènes, tels que les habitudes alimentaires, la consommation de tabac ou d'alcool mais aussi à l'exposition à des agents chimiques, des radiations ou encore des virus ainsi qu'à des stress endogènes tels que les hormones ou encore le système immunitaire (Liu *et al.*, 2015). Il existe de plus des prédispositions au cancer qui ne sont autre que des altérations pré-existantes dans le génome germinal.

Ce processus se déroule sur une période de latence pouvant durer jusqu'à vingt ans ou plus et est constitué de 3 étapes : l'initiation, la promotion et la progression (figure 1.1).

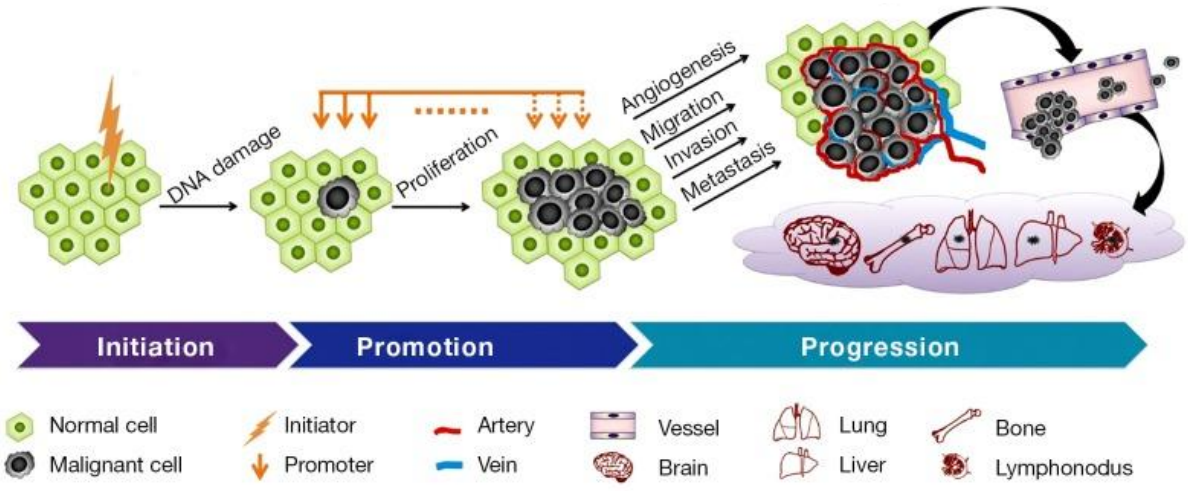


FIGURE 1.1 – *Processus de carcinogenèse (adaptation de Liu* et al. *(2013).* Le cancer est un processus se déroulant en 3 étapes : l'initiation, la promotion et la progression.

L'initiation correspond au premier phénomène de la transformation cancéreuse par l'acquisition d'une ou plusieurs altérations initiatrices. Ces premières altérations ne transforment pas la cellule en cellule cancéreuse mais réduisent le nombre d'altérations à acquérir pour cette transformation. L'initiation est suivie de la promotion des anomalies acquises avec un début de prolifération non contrôlée. Les altérations promotrices ne provoquent pas nécessairement le cancer, mais augmentent l'expansion clonale des cellules initiées, phénomène aussi appelée prolifération *in situ* ou hyperplasie. Dans certains cas, les cellules peuvent continuer de proliférer *in situ* et ne jamais infiltrer les tissus voisins. La dernière étape de transformation cellulaire correspond à la progression durant laquelle

des altérations en série permettent aux cellules tumorales d'acquérir de nouvelles propriétés impliquant un phénotype malin. Cette étape s'accompagne d'une augmentation du taux de croissance ou encore de l'invasivité liée à l'instabilité génétique permettant alors aux cellules de traverser la membrane basale et atteindre d'autres tissus pour former des métastases.

L'initiation, la promotion et la progression tumorale résultent des altérations génétiques telles que les mutations ponctuelles, les réarrangements chromosomiques et la variation du nombre de copies de gènes ou encore d'altérations épigénétiques (Lacave *et al.*, 2005).

En effet, durant leur évolution progressive, les cellules normales acquièrent donc successivement un ensemble de capacités caractéristiques (Hanahan et Weinberg, 2000, 2011). Ainsi, Hanahan et Weinberg proposent que l'ensemble des génotypes des cancers soit la manifestation de 6 capacités essentielles de la physiologie cellulaire dictant collectivement le développement tumoral (figure 1.2) :

1. l'indépendance aux signaux de croissance,
2. l'insensibilité aux signaux antiprolifératifs,
3. l'échappement à l'apoptose,
4. l'acquisition d'un potentiel de réplication illimité,
5. l'induction de l'angiogenèse,
6. l'invasion des tissues et métastases.

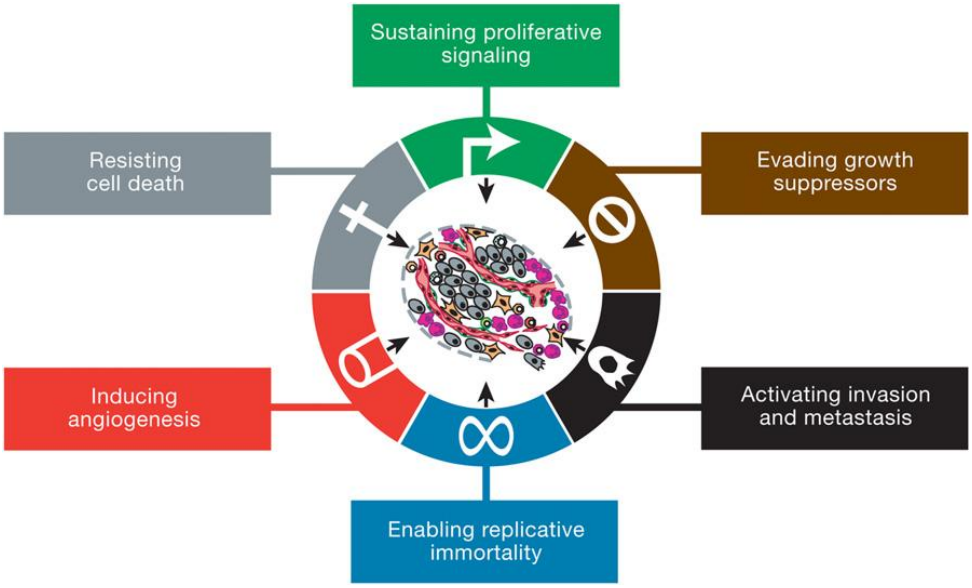


FIGURE 1.2 – *Les caractéristiques du cancer.* Les 6 capacités acquises par les cellules cancéreuses proposées par Hanahan et Weinberg (2011).

1.1.3 La génétique du cancer

Il est fréquent pour les recherches en cancérologie de se focaliser sur le point de départ de cette maladie : le génome. L'idée que le cancer soit une maladie du génome a été

proposée par Theodor Boveri en 1914 [1] avant même que le concept du gène ne fut introduit. T. Boveri a proposé que la croissance incontrôlée de la tumeur peut être une conséquence d'un nombre anormal de chromosomes (aneuploïdie). Dans les dernières décennies les technologies modernes de séquençage ont permit de démontrer que le cancer est en effet causé par des altérations génétiques contrôlant les mécanismes de croissance et de division cellulaire (Vogelstein et Kinzler, 2004).

En plus de mutations qui affectent le séquence génétique en elle-même, les altérations épigénétiques jouent aussi un rôle important dans le cancer. Ces altérations qui concernent la méthylation de l'ADN ou encore la modification des histones peuvent contribuer au processus de tumérigénisation en influant sur le niveau de l'expression de gènes (Gronbaek *et al.*, 2007). Enfin, un faisceau de preuves converge pour supposer que l'organisation tridimensionnelle du génome dans le noyau ait une influence sur l'initiation du processus de cancérisation (Reddy et Feinberg, 2013).

Variations somatiques et germinales

La recherche dans le cancer s'intéresse souvent à l'établissement du lien entre les variations génétiques et le processus tumorigenèse. Ces variations peuvent être transmises par voie héréditaire à partir des gamètes (ovules et spermatozoïdes) ou encore apparaître *de novo* par mutation dans ces dernières et devenir alors héréditaires. L'ensemble de ces variations constituent ce que l'on appelle les variations germinales. Ainsi, deux humains pris au hasard ont leurs séquences ADN identiques à 99.9%, les 0.1% restant constituent des variations génétiques responsables de différences phénotypiques mais aussi de la susceptibilité à certaines maladies (Feuk *et al.*, 2006). La base de données dbSNP est une base de dépôt central et public pour les variations apparaissant dans $> 1\%$ de la population aussi appelés SNPs (*Single-Nucleotide Polymorphisms*).

Au contraire, les variations somatiques peuvent apparaître dans n'importe quelle cellule à tout moment de la vie. Le recensement des gènes impliqués dans le cancer indique qu'une large part de ces gènes (80%) ne portent que des variations somatiques, 10% ne portent que des variations germinales et 10% portent les deux types de variations (Stratton *et al.*, 2009).

L'association de mutations germinales avec le cancer est le sujet principal des études d'association pangénomique (*Genome Wide Association Studies*, GWAS) (Easton et Eeles, 2008). Néanmoins, la plupart des recherches sur le cancer se concentrent sur l'analyse des altérations somatiques. Les exemples les plus connus de ce type d'études sont le TCGA, *The Cancer Genome Atlas* (Weinstein *et al.*, 2013) et l'ICGC, *International Cancer Genome Consortium* (Hudson *et al.*, 2010). Dans les dernières décennies, les recherches sur les mutations somatiques ont permit l'identification de plus de 100000 mutations somatiques différentes (Stratton *et al.*, 2009), la généralisation du séquençage entraînant une augmentation constante de ce nombre.

Les altérations somatiques peuvent être classées en deux catégories en fonction de leur impact sur le développement et progression de la maladie. Les mutations *driver* sont celles qui confèrent un avantage aux cellules qui les portent et sont donc sélectionnées lors du processus de l'évolution du cancer. Par définition elles résident dans les gènes associés au cancer. Au contraire, les mutations n'ayant pas d'incidence fonctionnelle sont appelées des mutations *passagères* (Stratton *et al.*, 2009). Pouvoir efficacement distinguer entre les

1. L'ouvrage de T. Boveri *Concerning the Origin of Malignant Tumours* cité dans (Weinberg, 2013) et récemment traduit de l'allemand par l'auteur.

mutations drivers et passagères est essentiel pour l'interprétation des génomes tumoraux.

Gènes impliqués

Comme nous l'avons vu, le cancer est une maladie du génome entraînant la prolifération anarchique due au dérèglement du système de contrôle du cycle cellulaire garantissant le maintien de l'intégrité des cellules et le contrôle de leur croissance. Ainsi, les gènes impliqués dans le développement tumoral peuvent être différenciés selon leur rôle dans le cycle cellulaire (Lacave *et al.*, 2005). Quand ces gènes sont altérés, les cellules deviennent susceptibles de développer un phénotype cancéreux (Wodarz et Komarova, 2005).

Ainsi, les *proto-oncogènes* sont des gènes qui favorisent la prolifération et/ou inhibent l'apoptose tandis que les *gènes suppresseurs de tumeurs*, ou onco-supresseurs, favorisent l'apoptose et/ou inhibent la prolifération cellulaire. Les altérations impliquées dans le cancer sont activatrices pour les proto-oncogènes ce qui donne des oncogènes avec une augmentation de leurs fonctions et inhibitrices pour les gènes suppresseurs de tumeurs avec une diminution ou une perte de leurs fonctions. L'altération d'un seul allèle pour un proto-oncogène suffit pour entraîner une stimulation exagérée de la prolifération cellulaire. Au contraire, d'après l'hypothèse de Knudson, lorsque les gènes suppresseurs de tumeurs sont altérés, ils se comportent comme des allèles récessifs (Knudson, 1971). L'inactivation implique alors que les deux allèles soient altérés, le premier allèle étant le plus souvent inactivé par une mutation ponctuelle tandis que le second l'est par une délétion ou une insertion.

1.2 Analyse de séquences pour le cancer

Le séquençage nouvelle génération (NGS) est un outil d'acquisition rapide de données génomiques (Metzker, 2010). Dans le contexte de la recherche contre le cancer, les technologies NGS ont été très rapidement appliquées à l'étude de ce que l'on appelle communément la maladie du génome. Ainsi, après la publication du génome humain complet obtenu par le séquençage nouvelle génération en 2008 (Wheeler *et al.*, 2008), le premier génome du cancer séquencé et la première analyse des données générées ont été publiés dans la même année (Ley *et al.*, 2008) en application à la leucémie myeloïde aiguë.

Aujourd'hui le séquençage est un outil standard en cancérologie, et ce autant dans le contexte clinique que dans la recherche. Dans le premier cas il est plus fréquent de faire appel au *séquençage ciblé* d'un panel de gènes, tandis que dans le deuxième cas le recours au séquençage de génome complet (*Whole Genome Sequencing*, WGS) est très répandu (LeBlanc et Marra, 2015).

Dans cette section, nous allons voir qu'elles sont les altérations du génomes observables et leur implications en terme de méthodologies bioinformatiques à partir des séquences issues de séquençage aussi appelées *reads*.

1.2.1 Variations génomiques : définitions

Dans notre contexte l'objectif du séquençage concerne principalement l'identification de variations. Une *variation* est une différence présente dans une séquence par rapport à un génome de référence. Comme illustré dans la figure 1.3.A, un génome de référence peut être représenté comme une chaîne de caractères que sont les nucléotides. Un locus va être

caractérisé par une coordonnée de début c_1 et de fin c_2 dans ce génome de référence. Ainsi, une variation est une différence à un locus donné par rapport au génome de référence.

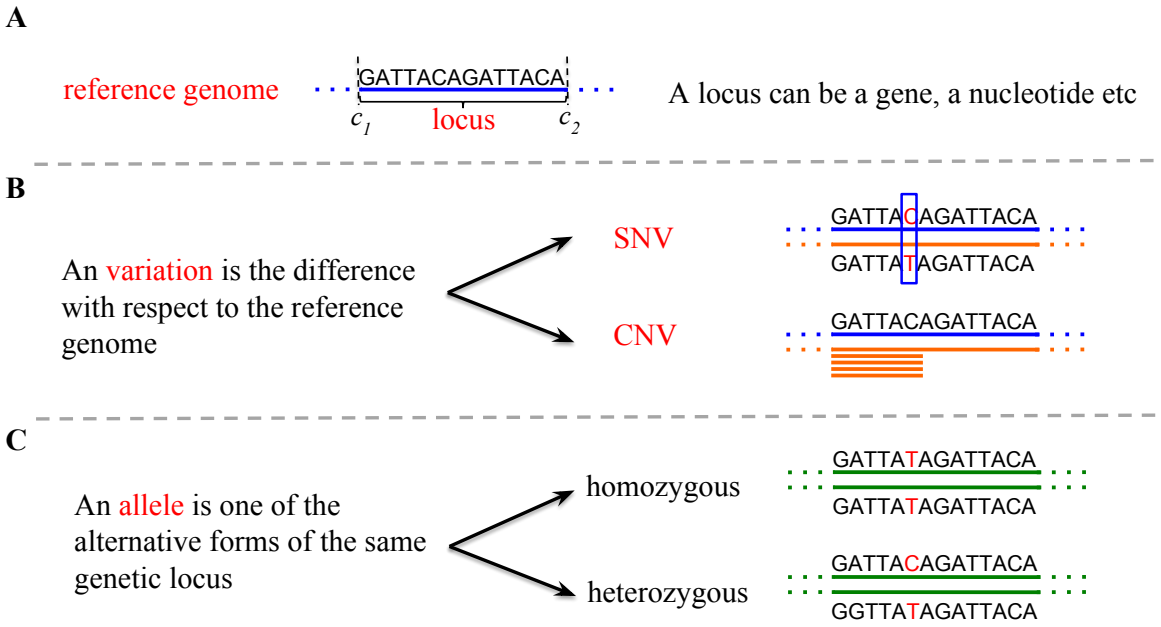


FIGURE 1.3 – *Définitions.* Illustration de (A) un locus génomique, (B) un SNV et d'un CNV et (C) d'un allèle.

Dans le génome humain normal, les sources de variations peuvent être décomposées en deux grandes classes selon la taille du locus.

Les variations de petites tailles, aussi appelées ponctuelles, correspondant aux *Single Nucleotide Variations* (SNVs) et aux insertions ou délétions de quelques nucléotides aussi appelées *indels*. Dans cette thèse, nous incluront les indels dans le terme SNVs et la différence sera explicitée si nécessaire. Ces variations constituent une différence de caractère à un locus de un à quelques nucléotides dans le cas des indels (exemple d'une substitution du nucléotide C par le nucléotide T dans la figure 1.3.B).

Les variations de grandes taille ou *Structural Variations* (SVs) correspondent à de larges duplications, délétions, insertions, inversions et translocations. Dans certains cas, ces variations entraînent des différences en terme de nombre de copies du locus impliqué ou *Copy Number Variations* (CNVs) dont la taille est supérieure à 1 kilobase (surreprésentation de la première partie de la séquence dans la figure 1.3.B).

Le génome humain normal est diploïde, ce qui se traduit par la présence de chaque portion génomique au nombre de 2. Un *allèle* est une forme alternative à un locus génomique donné. Ainsi, pour un génome humain normal diploïde, chaque locus génomique présente deux allèles parentaux différents : si les deux allèles sont les mêmes, ils sont dits *homozygotes*, s'ils sont différents, ils sont dits *hétérozygotes*.

Dans le contexte d'échantillons tumoraux, en plus des variations présentes dans le génome sain, appelé aussi normal, de l'individu, des variations vont être acquises durant le processus de la tumorigenèse. Ces variations sont en fait variations somatiques ou encore altérations et sont propres au génome tumoral (figure 1.4).

La figure 1.5 issue de (Beerenwinkel *et al.*, 2015), reporte les altérations génomiques

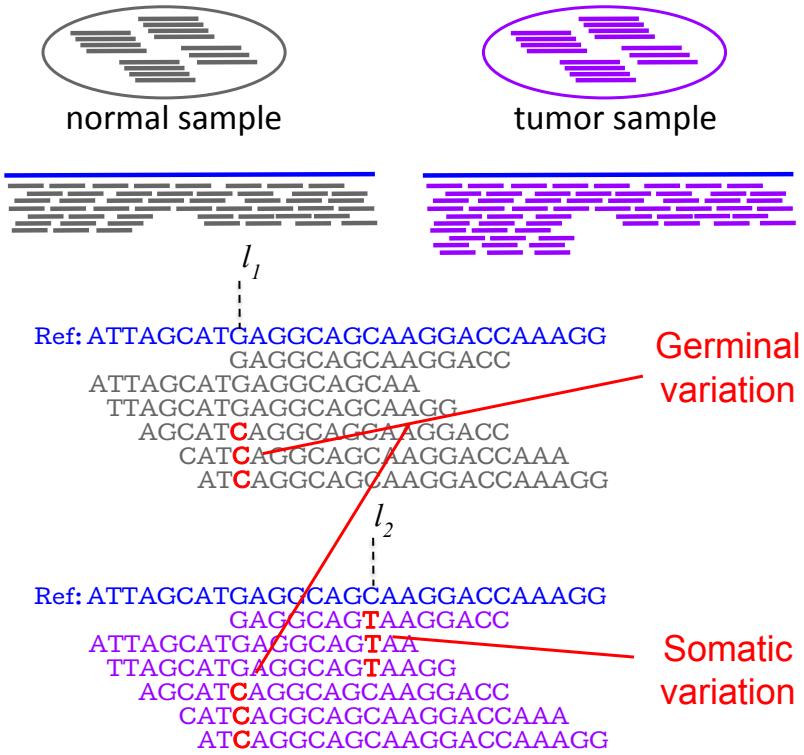


FIGURE 1.4 – *Variations germinale et somatique.* Une variation est dite germinale lorsqu'elle est présente dans le génome normal, comme illustré au locus l_1. Une variation germinale sera aussi présente dans le génome tumoral si le locus y est conservé. Une variation est dite somatique lorsqu'elle est présente seulement dans l'échantillon tumoral, comme illustré au locus l_2.

classiques retrouvées dans les cancers.

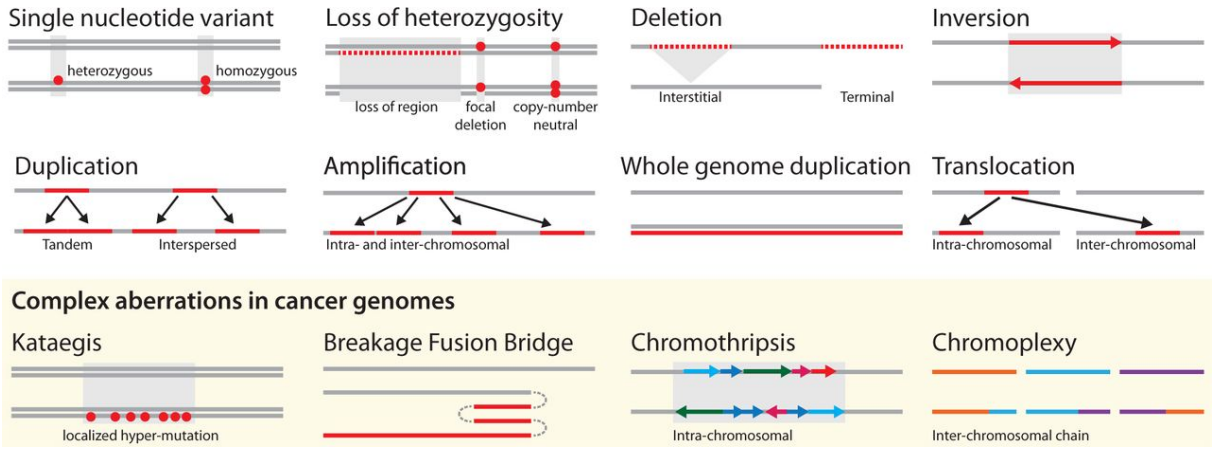


FIGURE 1.5 – *Altérations classiques des génomes tumoraux (Beerenwinkel* et al.*, 2015).* Les lignes indiquent les différents génomes : le génome germinal en haut et le génome tumoral avec des altérations somatiques en bas. Les lignes doubles sont utilisées lorsque la différentiation des modifications hétérozygotes et homozygotes est utile. Les points représentent des SNVs, tandis que les lignes et les flèches représentent des changements structurels.

Dans cette thèse, nous nous intéresserons plus particulièrement aux variations de types SNVs et CNVs. Ainsi, les SNVs et les CNVs somatiques représentent des altérations acquises durant le processus de tumorigénisation et peuvent aussi être appelés dans la littérature SNAs (pour *Single-Nucleotide Alteration*) et CNAs (*Copy Number Alteration*), respectivement.

1.2.2 Pipeline standard d'analyse

De façon classique l'analyse de données de séquençage pour le cancer comporte des étapes de pré-traitement de données, l'identification des altérations et leur mise en contexte et interprétation comme montré sur le schéma de la figure 1.6.

Chacune des ces étapes constitue en soi un sujet de recherche et beaucoup de méthodes algorithmiques et d'outils qui les implémentent ont été mis en œuvre, comme décrit par exemple dans la revue (Ding *et al.*, 2014).

Deux points essentiels de la recherche sont la reproductibilité et la réplicabilité. La reproductibilité consiste à obtenir les mêmes résultats en appliquant les mêmes méthodes aux mêmes données. En bioinformatique, si les méthodes ne font pas intervenir de solutions heuristiques, la reproductibilité n'est pas à mettre en doute et est essentielle (Peng, 2011). De plus de nombreux outils permettent de s'en assurer (Piccolo et Frampton, 2016). La réplicabilité quant à elle consiste à confirmer les résultats obtenus en variants les données et/ou les méthodes. Ici, même si l'on se restreint seulement à faire varier les méthodes bioinformatiques, ces dernières étant basée sur des méthodes algorithmiques différentes produisent des résultats parfois très variables. Ce phénomène met en évidence la nécessité d'amélioration des méthodes actuelles.

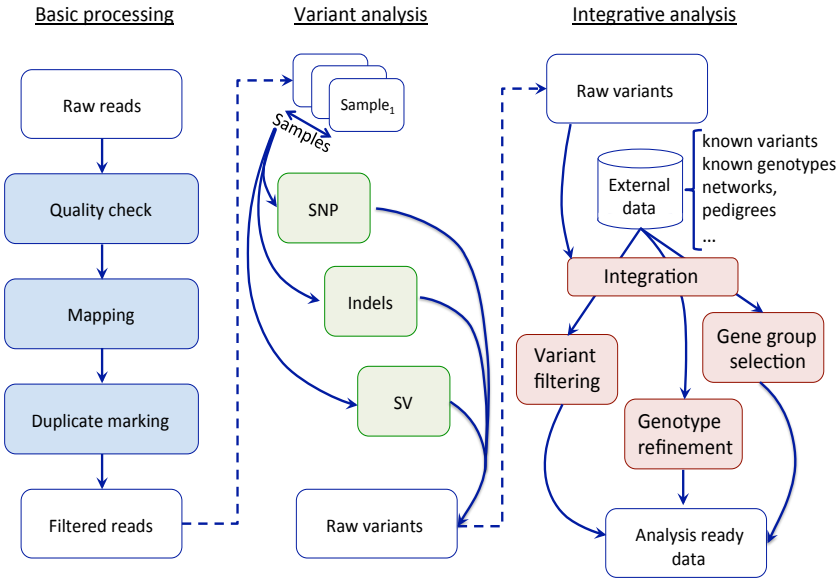


FIGURE 1.6 – *Pipeline standard.* Illustration des différentes étapes standard d'analyse de données de séquençage pour le cancer (adapté de Baker (2011)).

Par ailleurs, de nombreux biais influencent les analyses. Ainsi, des biais biologiques peuvent être engendrés par les étapes de réactions par PCR lors de l'amplification de la séquence d'intérêt pour son séquençage en créant par exemple des erreurs contexte-spécifiques (Robasky *et al.*, 2014) ou encore par la dégradation de la qualité de la séquence ou sa sous représentativité dans le cas d'échantillons tumoraux (Peng *et al.*, 2015).

De plus, des biais technologiques peuvent aussi entraîner l'apparition d'erreurs dîtes de séquençage, par exemple, lors de l'étape d'élongation des nucléotides durant le séquençage, si l'élongation n'est pas effectuée jusqu'au bout, le signal permettant la lecture du nucléotide souffre d'interférences (Schirmer *et al.*, 2015). Une autre erreur de séquençage bien connue est celle liée aux homopolymères qui en saturant le signal créent des erreurs d'estimation du nombre de nucléotides le composant (Robasky *et al.*, 2014). Lors du séquençage, le séquenceur attribut un score de qualité à chaque base nucléotidique correspondant au niveau de confiance de sa lecture. Ce score de qualité permet ainsi de déterminer la qualité d'une séquence. Le plus utilisé est le score phred Q qui est lié à la probabilité d'erreur d'identification du nucléotide de manière logarithmique : $Q = -10log_{10}P$. Ainsi, un score de 10 correspond à une probabilité d'identification incorrecte de 1 pour 10, soit une précision d'identification de 90%, 20 de 1 pour 100 (précision de 99%), 30 de 1 pour 1000 (précision de 99.9%) etc.

Enfin, de multiples biais bioinformatiques peuvent être rencontrés lors des différentes étapes d'analyse des données de séquences. Par exemple, lors de l'alignement, certains *reads* peuvent être assignés à différentes régions du génome s'ils proviennent de régions répétées ou de faible complexité et sont dans ce cas éliminés de l'analyse, créant une baisse de couverture pour ces régions (Sims *et al.*, 2014). Ce biais, dit de mappabilité, est augmenté par les erreurs introduites lors des étapes de préparation ou du séquençage mais aussi par la présence de variations dans la séquence biologique de départ par rapport au génome de référence (Miller *et al.*, 2011). De même, il influencera les étapes situées en

aval lors de la recherche de variations. Par ailleurs, de nombreuses études mettent en corrélation le contenu en GC local et la faible profondeur de couverture pour de nombreuses technologie de séquençage (Harismendy *et al.*, 2009). Sur le même principe que le score d'identification de nucléotide, un score de *mapping* peut être calculé lors de l'alignement en prenant en compte la qualité de séquençage.

Dans la suite de cette section nous allons aborder plus en détail les éléments nécessaires pour la lecture des chapitres suivants notamment l'alignement des données sur le génome de référence, la recherche de variations de type SNV et CNV dans le contexte du cancer et enfin la mise en commun de ces deux types d'altérations pour la reconstruction de l'hétérogénéité intra-tumorale.

1.2.3 Alignement et mapping

La détection de variants prend souvent comme point d'entrée les *reads* alignés sur un génome de référence, souvent représenté sous forme de chromosomes ou de contigs. Par conséquent, la qualité du *mapping* est essentielle pour une détection précise de variants.

Les algorithmes d'*alignement* de *reads* sur une séquence de référence sont des algorithmes qui mettent en correspondance un ensemble de *reads* $\mathcal{R}$ et un ensemble de contigs $\mathcal{C}$ correspondant à des ensembles de séquences (chaînes de caractères) : $\mathcal{R} = \{r_1, \ldots, r_{|R|}\}$ et $\mathcal{C} = \{c_1, \ldots, c_{|C|}\}$. Cette mise en correspondance passe par la définition de deux fonctions :

1. $\mathcal{M} : \mathcal{R} \to \mathcal{C} \cup \{c_0\}$ où $|c_0| = 0$ qui indique sur quel contig s'aligne un *read* (avec la possibilité pour un *read* de ne pas être aligné),
2. $\mathcal{A}_\mathcal{R} : \mathcal{R} \to \mathbb{N} \cup \{-1\}$ qui indique pour chaque base d'un *read* sur quelle base du contig elle s'aligne (avec la possibilité de gaps représenté par -1).

Ces concepts ont été introduits en application au séquençage Sanger dans le papier Gordon *et al.* (1998) présentant la méthode Consed en faisant référence à l'outil d'alignement multiple phrap. On peut voir la distinction faite entre la notion d'alignement lui-même $\mathcal{A}_R$ et le *mapping* $\mathcal{M}$. Depuis, les mêmes concepts sont appliqués aux *reads* issus du séquençage NGS et les deux termes d'alignement et de *mapping* de *reads* sont utilisés de façon interchangeable dans la littérature et nous les utiliseront sans faire de distinction.

Depuis 1998 de très nombreux algorithmes et outils logiciels de *mapping* de *reads* ont été développés. Ainsi, la revue de Fonseca *et al.* (2012) cite 60 méthodes différentes.

Les algorithmes modernes d'alignement de *reads* sur le génome de référence procèdent en deux étapes : l'indexation et le *mapping* lui-même. La première étape consiste à calculer un index pour le génome de référence, pour les *reads* ou pour les deux. Deux types principaux types d'algorithmes existants sont basés soit sur des tries de suffixes / préfixes, soit sur des tables de hachage (Li et Homer, 2010; Mielczarek et Szyda, 2016). Un trie de suffixes / préfixes est une structure de données qui étant donné une chaîne de caractères permet un *matching* efficace, c'est-à-dire dans notre cas la mise en correspondance d'un *read* et d'une sous-séquence sur le génome de référence. Dans le cas des données NGS cette approche générale n'est pas suffisamment efficace en termes de temps d'exécution et des approches de type transformation de Burrows–Wheeler (BWT) sont la plupart de temps utilisées. Les aligneurs les plus utilisés en routine sont BWA (Li et Durbin, 2009), Bowtie and Bowtie2 (Langmead *et al.*, 2009) et permettent l'alignement de séquences génomiques ou transcriptomiques.

1.2.4 Détection des SNVs

Étant donné l'alignement de *reads*, le but des algorithmes d'identification des SNVs est d'estimer le génotype, c'est à dire la composition allélique, à un locus donné pour un échantillon séquencé. De manière classique, l'allèle de référence, c'est à dire identique au génome de référence pour un locus donné, est dénoté A et l'allèle alternatif est dénoté B.

Soit le séquençage d'un génome diploïde sans altération et dans un contexte idéal dépourvu de tout biais biologiques, technologiques ou bioinformatiques (figure 1.7). Soit les loci l_1, l_2 et l_3 représentant les 3 alternatives possibles du génotype $G = \{AA, AB, BB\}$ d'un locus.

Le nombre d'allèles alternatifs c^B et total c^T pour chacun des loci seront alors de :

— $c^B/c^T = 0/2$ pour le locus l_1 de génotype AA

— $c^B/c^T = 1/2$ pour le locus l_2 de génotype AB

— $c^B/c^T = 2/2$ pour le locus l_3 de génotype BB

Soit 10 *reads* couvrant l'ensemble des loci l_1, l_2 et l_3. La VAF pour chacun des loci sera alors de :

— $0/10 = 0$ pour le locus l_1 de génotype AA

— $5/10 = 0.5$ pour le locus l_2 de génotype AB

— $10/10 = 1$ pour le locus l_3 de génotype BB

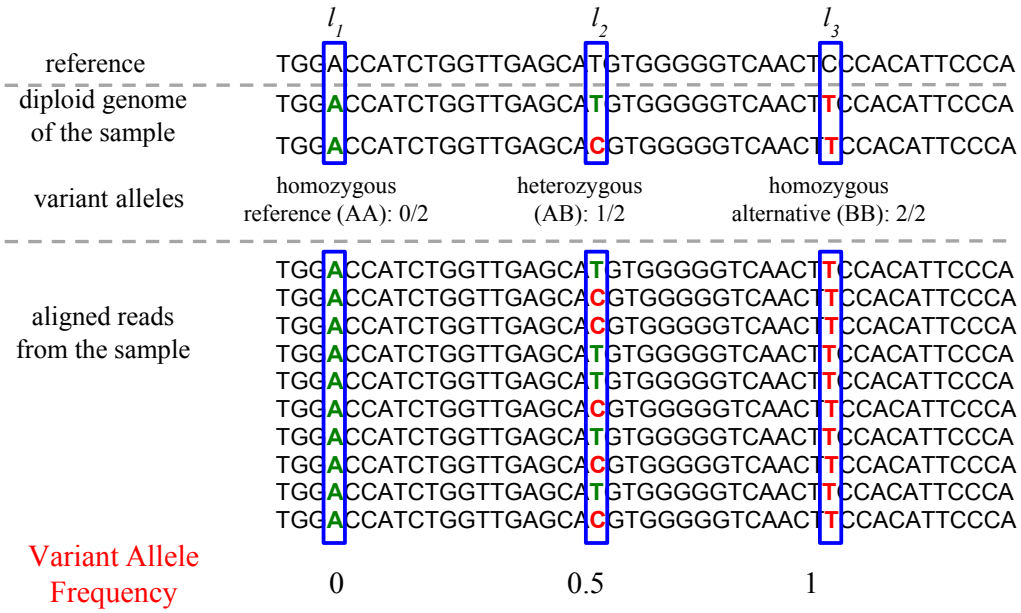


FIGURE 1.7 – *Illustration du calcul des VAFs.* Illustration d'un locus homozygote AA (l_1), hétérozygote AB (l_2) et homozygote BB (l_3) et des fréquences alléliques théoriques correspondantes.

A partir des données de séquençage, nous avons accès aux nombres de *reads* qui supportent chaque nucléotide du génome de référence.

Ainsi, étant donné le séquençage d'un échantillon, la *fréquence allélique* d'un variant (VAF) pour un locus l correspond au nombre de *reads* r^B supportant l'allèle alternatif B

relatif au nombre total de *reads* $r^T = r^A + r^B$, où r^A est le nombre de *reads* supportant l'allèle de référence A, tel que :

$$VAF_l = \frac{r^B}{r^A + r^B} = \frac{r^B}{r^T}$$

Cependant, un séquençage n'est pas parfait, de nombreuses méthodes ont ainsi été développées. Les plus utilisées étant VarScan, samtools ou encore GATK. De nombreuses études comparatives ont établi les performances particulièrement intéressantes du pipeline GATK (DePristo *et al.*, 2011) pour l'identification de variants (voir par exemple (Pirooznia *et al.*, 2014; O'Rawe *et al.*, 2013) entre autres). De plus, dans une étude récente, (Warden *et al.*, 2014) montrent une bonne performance de qualité comparable pour les méthodes GATK (UnifiedGenotyper et HaplotypeCaller) et VarScan.

VarScan (Koboldt *et al.*, 2009) est une approche robuste basée sur l'application de filtres tels que la profondeur de couverture, la qualité de séquençage et la qualité de *mapping*, la fréquence du variant et un test statistique permettant une détection des variations efficace.

La méthode d'origine d'identification de SNVs dans GATK, le UnifedGenotyper (Nielsen *et al.*, 2011) est la même que celle de samtools (Li *et al.*, 2009; Li, 2011) et est basée sur une approche Bayesienne de calcul de la vraisemblance de génotypes (*genotype likelihood*). Il s'agit d'estimer la vraisemblance des *reads* étant donné les génotypes possibles et la convertir ensuite en probabilité de génotypes en utilisant la règle de Bayes. Ce type d'approches fonctionne bien dans le cas de forte couverture de séquençage, mais donne un taux de faux positifs entre 4% et 11% pour les données de faible couverture (4x) (Le et Durbin, 2011).

L'algorithme HaplotypeCaller de GATK est basé sur une méthode différente et procède par la construction d'un graphe combiné avec un modèle de Markov caché. Cette méthode procède en deux étapes : elle identifie d'abord des régions génomiques appelés des *sites actifs* ayant une forte probabilité que les *reads* présentent des variations, ou allèles alternatifs, et les analyse ensuite en détail. Pour cela, l'algorithme procède à la construction d'un graphe dans la région en amont et en aval du site. Le graphe permet de définir les haplotypes possibles, c'est à dire les possibilités de succession entre les différents allèles alternatifs situés dans un même site actif. Ces haplotypes sont ensuite évalués en alignant l'ensemble des *reads* couvrant la région en question à chaque haplotype. L'alignement et le calcul d'un score de vraisemblance sont effectués grâce à un modèle de Markov caché prenant en compte la qualité du séquençage. Ainsi, ce score est utilisé pour calculer un score pour chacun des allèles alternatifs individuels composant les sites actifs. Cette méthode produit des résultats d'une meilleure précision que l'ancienne méthode UnifedGenotyper (Pirooznia *et al.*, 2014).

Bien que les méthodes classiques puissent être utilisées avec des paramètres appropriés, de nombreux outils ont été développés spécifiquement pour la recherche de variations somatiques ou SNAs. La plupart d'entre elles requièrent un échantillon sain pairé à l'échantillon tumoral ce qui est parfois une limitation. Par exemple Mutect (Cibulskis *et al.*, 2013), détecte les variants somatiques en utilisant un classifieur Bayesien. DeepSNV (Gerstung *et al.*, 2012), quant à lui, teste si la fréquence d'un variant a été modifiée par rapport à l'échantillon de tissu normal ou modélise sa fréquence par une distribution bêta lors de l'absence d'échantillon contrôle. La version améliorée de VarScan (Koboldt *et al.*, 2012) (parfois appelée VarScan2), permet désormais la détection de SNAs.

De nombreuses études ont mis en avant la faible réplicabilité lors de l'utilisation de différents outils de détection de SNAs. Par exemple, Pabinger *et al.* (2014) comparant entre autres méthodes GATK et VarScan partagent seulement 50% des variants détectés. Wang *et al.* (2013) comparent quant à eux 6 outils dédiés à la détection de SNAs reportent que Mutect et VarScan sont les méthodes les plus efficaces. La publication la plus récente (Hofmann *et al.*, 2017) comparant 11 outils spécifiques ou non aux SNVs somatiques, dont ceux mentionner précédemment GATK (UnifiedGenotyper et HaplotypeCaller), VarScan, Mutect et deepSNV, révèle que deepSNV et JointSNVMix2 présentent les meilleurs performances. De plus, les auteurs mettent en avant que la qualité de l'alignement est importante pour obtenir une bonne sensibilité de détection.

1.2.5 Détection des CNVs

Les variations de type *copy number* (CNVs) correspondent au gain et à la perte de larges segments génomiques pouvant correspondre à des loci d'une taille allant de 1 kilobase jusqu'à des chromosomes entiers. Ce type de variation est souvent assimilé aux variants structuraux (SVs) car il en découle directement. En effet, le changement du nombre de copies d'un locus est la conséquence directe des SVs tel que les larges insertions, délétions, translocations etc.

Il existe 4 grandes méthodes de détection de SVs à partir du séquençage WGS d'un échantillon (Medvedev *et al.*, 2009; Teo *et al.*, 2012; Pirooznia *et al.*, 2015) (figure 1.8) :

1. La méthode *Read-Pair* ou *reads* pairés. Cette méthode s'appuie sur le fait que lorsque l'on effectue un séquençage *paired-end*, la distance entre les *reads* alignés présente une distribution autour de la taille moyenne de l'insert pour l'ensemble des paires de *reads*. Ainsi, si la distance est significativement différente de la moyenne attendue pour une paire de *reads*, ces derniers sont considérés comme appartenant à un locus présentant une SVs (figure 1.8.1).
2. La méthode *Split Read*. Cette méthode utilise les *reads* pairés qui n'ont pas été alignés correctement afin d'identifié des zones de cassures de l'ADN suite à des événement de SVs (figure 1.8.2).
3. La méthode *Read Depth*, *read count* ou encore de profondeur de couverture. Cette méthode est basé sur l'hypothèse qu'il existe une corrélation entre la couverture d'un locus, en matière de nombre de *reads* mappés, et le nombre de copies de ce locus. Cette méthode permet la détection de variations de type *copy number* (figure 1.8.3).
4. La méthode d'assemblage *de novo*. Cette méthode va comparer les contigs obtenus à partir de l'assemblage des *reads* au génome de référence afin d'identifier de potentielles SVs (figure 1.8.4).

L'ensemble de ces méthodes est complémentaire. En effet, chacune d'entre elles est efficace pour la détection de SVs particuliers (Alkan *et al.*, 2011) bien qu'aucune ne permet l'identification simultanée des différents types de SVs (Abel et Duncavage, 2013). Dans cette thèse, nous allons nous intéresser plus particulièrement aux méthodes destinées à l'identification de variations en matière de *copy number* dans le cadre de séquençage WGS.

Les méthodes basées sur la profondeur de couverture sont les approches les plus utilisées pour la détection de CNVs. En effet, comparées aux autres méthodes qui reportent seulement les coordonnées génomiques des loci présentant un CNVs, ces dernières permettent d'estimer le nombre de copies exactes de ces loci. De nombreux outils ont été

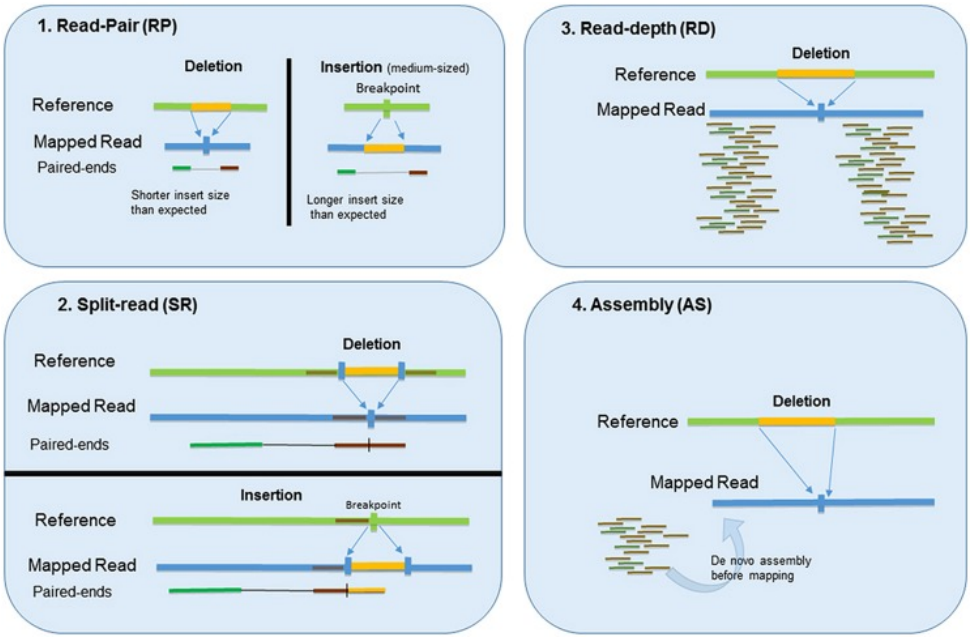


FIGURE 1.8 – *Méthodes de détection des SVs à partir de données NGS (Pirooznia* et al.*, 2015).* Il existe 4 grandes méthodes de détection de CNVs : *read-pair* (1), *split-read* (2), *read-depth* (3) et les méthodes basée sur l'assemblage *de novo* (4)

développés ces dernières années, la revue de Zhao *et al.* (2013) présente 15 méthodes comportant des étapes similaires. Tout d'abord, les *reads* sont alignés sur le génome de référence. Le génome de référence est alors partitionné en fenêtres dans lesquelles sont comptés le nombre de *reads*. Ces comptes sont alors normalisés afin de supprimer certains biais potentiels. Par exemple, un contenu extrême en GC (bas ou élevé) diminue la profondeur de couverture ou encore la présence de régions répétées du génome implique une baisse de mappabilité (Wineinger *et al.*, 2008). A partir de ces fenêtres de comptes normalisés, un algorithme de segmentation peut-être appliqué afin d'identifier les régions contiguës présentant le même *copy number* et ainsi estimer si elles correspondent à une perte, un gain ou aucune variation.

Dans le cas d'échantillons tumoraux, il est classique de rechercher les altérations de *copy number* ou CNAs. En effet, en plus des altérations de type SNA, les cellules tumorales possèdent très souvent un nombre anormal de chromosomes, phénomène aussi appelé aneuploïdie. En plus des CNAs, certaines méthodes permettent l'identification de loci qui sont représentés plus que par seul un allèle parental, dupliqué ou non, l'autre ayant été supprimé, phénomène appelé *Loss Of Heterozygosity* (LOH).

De nombreuses méthodes ont été développées pour la détection de CNAs dans les génomes tumoraux, les programmes de détection classiques n'étant pas adaptés. En effet, en plus des biais tels que le contenu en GC, l'identification de CNAs dans des échantillons tumoraux est complexifiée par leur quantité et leur diversité en matière de taille et de niveaux de *copy number*, la contamination de l'échantillon par le tissus sain, la présence de différents populations tumorales ou encore par le fait que le niveau de ploïdie moyen (nombre moyen de chromosomes pour chaque autosome de la tumeur) ne soit pas connu. La revue de Liu *et al.* (2013) explicite les caractéristiques d'une dizaine de méthodes spécifiques de détection de CNAs.

Ainsi, certaines méthodes reportent seulement les segments avec des pertes ou gain de

copy number. Par exemple, Gusnanto *et al.* (2012) permet d'estimer le nombre de copies dans des échantillons tumoraux après une correction prenant en compte la contamination par le tissu sain et la ploïdie tumorale. Cette méthode développée pour l'analyse de données de très faible couverture est décrite plus en détail dans le chapitre 2.

D'autres méthodes incorporent des informations supplémentaires contenues dans les *reads* telles que la fraction allélique de l'allèle alternatif présent aux loci de SNPs hétérozygotes aussi appelé BAF pour *B-allele fraction*. La valeur de BAF attendue à ces loci étant de 0.5 pour un génome diploïde, toute déviation de 0.5 peut être considérée comme le résultat d'une altération de *copy number* ou encore de LOH. Par exemple, Control-FREEC (Boeva *et al.*, 2012) segmente les profils de *copy number* et de BAF tandis que CLImAT (Yu *et al.*, 2014) les modélise par un modèle de Markov caché et assigne un génotype à chaque segment identifié (aussi appelé profil allélique ou encore état de *copy number*).

Une manière d'encoder le profil allélique peut être de le représenter sous la forme de 2 entiers aussi appelés *allele specific copy number* (ASCN) (Wang *et al.*, 2015). Ainsi, l'ASCN correspond aux nombre de copies de chacun des deux chromosomes parentaux sous la forme du nombre d'allèle majeur et du nombre d'allèle mineur. Par exemple, l'ASCN correspondant à la duplication d'un des deux chromosomes parentaux est $(1, 2)$. La somme de ces deux nombres est égale au nombre total de copies du segment chromosomique.

1.2.6 Reconstruction de l'hétérogénéité intra-tumorale

Il existe ainsi de nombreuses méthodes dédiées à la caractérisation des altérations de type SNAs et CNVs. Comme vu dans la section 1.1.2, la progression tumorale est due à l'accumulation progressive de ces altérations dans le temps conduisant à la formation de différentes populations de cellules tumorales, ou clones, dérivant tous d'un clone ancestral. Ce phénomène d'évolution clonale suit un processus Darwinien où les cellules se diversifient et sont sélectionnées impliquant ce que l'on appelle l'hétérogénéité intra-tumorale. Reconstruire la composition d'un échantillon tumoral est essentiel pour proposer une thérapie ciblant l'ensemble des populations composant la tumeur.

Ainsi se pose le problème de reconstruction clonale qui consiste en l'identification et la caractérisation des différents clones composant un échantillon tumoral à partir de son séquençage. En effet, il est possible de reconstruire de manière probabiliste la proportion des populations d'un échantillon à partir de son profil transcriptomique ou génomique en utilisant les propriétés statistiques de mélanges de populations (Yadav et De, 2014). De nombreuses méthodes ont été développées pour la déconvolution d'échantillons de tissus hétérogènes, nous nous concentrerons sur celles basées sur le WGS qui peuvent être regroupées en trois grandes classes selon les informations utilisées :

1. *Les CNAs.* Ces approches tentent ainsi de résoudre ce problème de reconstruction via le *copy number*. L'algorithme THeta a été le premier a identifier de manière automatique les altérations de type CNA à partir de données de WGS d'échantillons tumoraux composés de plus de deux populations présentant des différences pour ce type d'altération (Oesper *et al.*, 2013). Ainsi, à partir des données de séquençage, cette méthode identifie la contamination par le tissu sain ainsi que le nombre de clones et leurs proportions à partir des altérations de *copy number* qu'ils présentent en séléctionnant la solution de composition qui explique le mieux les données c'est à dire qui maximise la vraisemblance. Theta2 (Oesper *et al.*, 2014), nouvel algorithme basé sur le précédent inclus l'utilisation de BAF afin d'augmenter sa spécificité de

reconstruction clonal. Cependant, cet algorithme ne renseigne pas sur le génotype associé aux CNAs. Par ailleurs, les méthodes basées sur la seule identification de CNA ne sont pas capables d'identifier les populations qui ne présentent pas ce type d'altération.

2. *Les SNAs.* Ces approches s'appuient sur les fréquences alléliques des SNAs afin d'identifier le nombre et la composition des clones de l'échantillon tumoral. Ces méthodes font l'hypothèse que les SNAs présentes dans les mêmes proportions cellulaires possèdent les mêmes valeurs de VAFs. Par exemple, l'algorithme Sciclone (Miller *et al.*, 2014) clusterise les VAFs des SNAs de loci ne présentant ni de variation de *copy number* ni de LOH en proposant différents types de modèles de mélanges Bayésien : bêta, binomial et Gaussien. La méthode Clomial (Zare *et al.*, 2014) quant à elle décompose les VAFs en produit de deux matrices : l'une code les proportions des différentes populations, la seconde codant les génotypes. Ensuite, elle recherche la solution la plus vraisemblable via un algorithme d'espérance-maximisation (EM). Cependant, ces méthodes considèrent seulement les altérations présentes à des loci non affectés par des altérations de *copy number* ou LOH ce qui entraîne une perte d'information considérable.
3. *La combinaison SNAs et CNAs.* Cette troisième classe de méthodes prend en compte ces deux types d'information afin de reconstruire la composition de l'échantillon tumoral. ABSOLUTE (Carter *et al.*, 2012), première méthode a avoir couplé les SNAs et les CNAs, détermine si ces derniers sont clonaux ou sous-clonaux mais ne permet cependant pas de quantifier les proportions des populations tumorales. De nombreuses autres méthodes telles que SCHIME, PhyloSub ou encore BitPhyloggeny ont ainsi été développées et s'attardent sur la reconstruction phylogénique après avoir identifié les différentes populations tumorales. Cependant l'ensemble de ces méthodes ne prend pas en compte l'hétérogénéité possible en matière de *copy number*. En effet, seules quelques méthodes prennent en compte l'hétérogénéité intra-tumorale à la fois en matière de SNAs et de CNAs comme par exemple la méthode Canopy (Jiang *et al.*, 2016). Cette méthode prend en entrée les SNAs et les CNAs identifiées de manières séparées par des algorithmes de détections tels que décrits dans les précédentes sous-sections et estime les différentes proportions de cellules tumorales par un *Markov Chain Monte Carlo*. Néanmoins, cette méthode a besoin d'échantillons multiples pour une même tumeur afin de résoudre le problème de reconstruction clonale.

De même, certaines méthodes telles que SciClon, Pyclone, Clomial ou encore Phylosub, réparties dans les trois classes définies précédemment, ont été développées pour la reconstruction de la composition tumorale à partir de multiples biopsies différentes en terme d'espace ou de temps. Cependant, une grande partie des études de séquençage de génomes tumoraux est établie avec un seul échantillon provenant du même patient, complexifiant la déconvolution des mélanges de populations tumorales.

Enfin, certaines méthodes, dont une grande partie de celles présentées précédemment, poursuivent la reconstruction de l'échantillon tumoral par la reconstruction de l'histoire phylogénique entre les différentes populations identifiées. La revue (Beerenwinkel *et al.*, 2015) présent 12 de ces méthodes.

Chapitre 2

Identification de *Single Nucleotide Variations*

Dans ce chapitre, nous allons nous intéresser à la détection de SNVs germinales et somatiques dans le cas de séquençage ciblé à partir d'échantillons tumoraux. La détection de SNVs consiste à identifier à partir des *reads* les positions dans lesquelles au moins une des bases diffère du génome de référence.

Comme présenté dans l'introduction (sous-section 1.2.4), un grand nombre de méthodes de détection de SNVs ont été développées ces dernières années. Cependant, l'identification de SNVs dans des données de séquençage issues d'échantillons tumoraux reste difficile en raison d'un certain nombre de facteurs tels que les biais biologiques, technologiques et bioinformatiques (voir la sous-section de l'introduction 1.2.2) ou encore la contamination par le tissus sain et l'hétérogénéité intra-tumorale.

Cette difficulté est accentuée lorsque l'on s'intéresse au séquençage à partir de transcrits. En effet, bien que le séquençage à partir de la rétrotranscription d'ARNm facilite la détection d'épissages pathologiques, ce paramètre doit être pris en compte. De plus, la complexité de détection est augmentée lorsque le gène cible est un gène suppresseur de tumeur. En effet, la perte de l'activité de cette classe de gènes est cruciale pour la progression tumorale et peut être satisfaite par des mécanismes tels que les altérations de type SNVs ou encore par la perte d'hétérozygotie (LOH) ou la dégradation post-traductionnelle. Dans ces deux derniers cas, la détection des altérations est d'autant plus difficile puisque le nombre de transcrits présents dans l'échantillon tumoral est diminué. Par conséquent, les méthodes classiques de détection de SNVs doivent être adaptées pour ce type d'échantillons.

Par ailleurs, cette difficulté de détection est d'autant plus importante lorsque le séquençage est effectué via une technologie dite de troisième génération (Third Generation Technology ; TGS) telle que Pacifique Biosciences (PacBio). En effet, cette technologie présente un taux d'erreurs 2.5% (Wang *et al.*, 2013) dont la majorité sont des insertions et des délétions (Eid *et al.*, 2009; Chin *et al.*, 2011) ce qui engendre des difficultés pour les méthodes de détection de SNVs classiques (Zook *et al.*, 2014). L'utilisation de cette technologie en recherche biomédicale étant un phénomène émergeant, les rares études ayant recours à PacBio sont établies avec des filtres très stringents (Shukla *et al.*, 2015). Ainsi, le développement de nouvelles méthodes de détection de SNVs adaptées aux TGS est indispensable pour la détection de SNVs dans le contexte de séquençage ciblé d'échantillons tumoraux.

L'étude de phase III multicentrique EORTC 10994 s'inscrit dans cette problématique.

L'objectif de cette étude est de déterminer si le statut mutationnel de TP53 des tumeurs du sein influence la réponse aux traitements. Pour cela le séquençage de l'ARNm de p53 provenant de biopsies de tumeurs du sein a été réalisé à partir de deux technologies de séquençage 454 Life Sciences (Roche) et PacBio.

Dans la première partie de ce chapitre, nous allons présenter le contexte de l'étude. Dans une seconde, nous verrons la mise en place d'une approche classique de détection de SNVs pour les données issues du séquençage 454. Les résultats découlant de cette analyse ont été l'objet d'une publication (Iggo *et al.*, 2013). La troisième partie décrira une nouvelle approche que nous avons mise en place, sans alignement, faisant appel à un outil informatique courant, les graphes, et permettant un compromis entre sensibilité et spécificité pour l'analyse du séquençage par la technologie PacBio. Enfin, dans la quatrième partie, nous verrons l'application de cette approche aux autres données de séquençage ciblé dans le cadre du cancer. L'algorithme et les résultats de cette nouvelle méthode ont donné lieu à une publication (Rudewicz *et al.*, 2016).

2.1 Contexte

2.1.1 Étude EORTC 10994 : motivations et challenges

L'EORTC (European Organisation for Research and Treatment of Cancer) est une organisation internationale réalisant des études cliniques au niveau européen pour tous types de cancer. A la fin des années 1990, lorsque l'étude clinique EORTC 10994/BIG1-00 débute, les résultats d'études précliniques suggèrent que les tumeurs du sein mutées pour le gène *TP53* sont résistantes aux anthracyclines et sensibles aux taxanes (Bonnefoi *et al.*, 2011).

C'est dans ce contexte que cet essai clinique de phase III multicentrique a été mis en place afin de tester si le statut de la protéine p53 influence la réponse des tumeurs mammaires humaines à la chimiothérapie. Afin de tester cette hypothèse, le traitement aux taxanes est comparé à un régime aux anthracyclines sur 1851 patients présentant une tumeur du sein. En parallèle, un test des levures est effectué afin de déterminer le statut de p53 dans 1469 cas.

Le test des levures

Le test fonctionnel des levures distingue les mutations inactivant la fonction de facteur de transcription de la protéine p53 des mutations passagères (figure 2.1).

Pour cela, les ARNm p53 sont extraits à partir des échantillons de biopsies de tumeurs avant traitement, convertis en ADNc et amplifiés par PCR. Chaque produit de PCR est alors cloné dans un vecteur d'expression puis transfecté dans une levure. La levure contient un gène rapporteur *ADE2* qui est sous le contrôle transcriptionnel d'un promoteur sensible à la protéine p53. Ainsi, la levure forme des colonies blanches lorsque p53 est de phénotype sauvage et des colonies rouges lorsque p53 n'est pas fonctionnel. Chaque colonie est dérivée d'une seule levure et donc d'une seule molécule d'ARNm p53. Par conséquent, le pourcentage des colonies rouges reflète l'abondance relative des ARNm de p53 sauvages et mutants présent dans un échantillon. Le statut de l'ensemble des tumeurs a ainsi été testé.

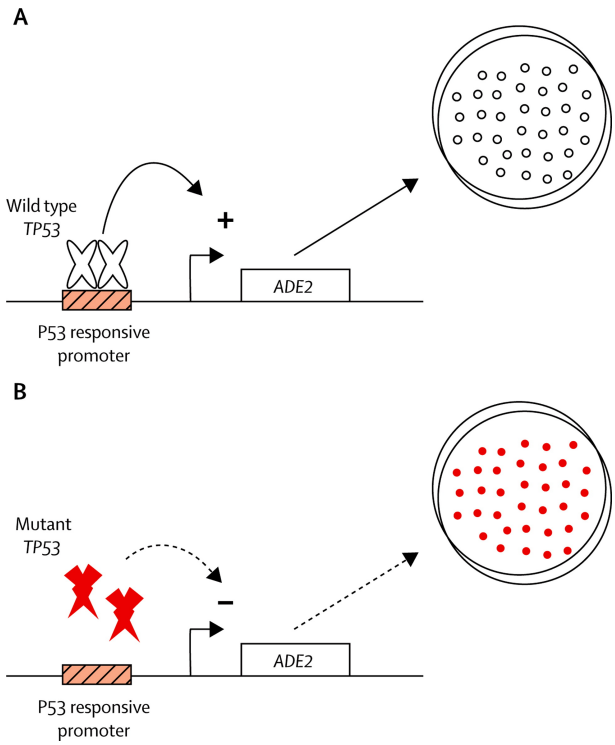


FIGURE 2.1 – *Test fonctionnel des levures (Bonnefoi* et al., *2011).* Lorsque la protéine p53 de type sauvage est exprimée dans la levure, elle se lie au site de liaison du promoteur du gène rapporteur *ADE2* et active son expression via son activité de facteur de transcription. La protéine p53 mutée est incapable de se lier à l'ADN et d'induire l'expression de *ADE2*. Ainsi, les levures contenant des protéines p53 sauvages forment des colonies blanches (A) tandis que les levures contenant des protéines p53 mutées forment des colonies rouges (B).

Résultats de l'étude du statut de p53 par le test des levures

La distribution du pourcentage de colonies rouges des patients révèle la présence de trois pics (Figure 2.2) :

— un pic sauvage situé à 11% de colonies rouges qui correspond au pic des phénotypes sauvages et peut être interprété comme étant le niveau du bruit de fond du test des levures.

— un pic mutant situé à 29% de colonies rouges qui représente des mutations soit hétérozygotes soit qui entraînent la dégradation de l'ARNm par un processus de Nonsens-Mediated Decay (NMD).

— un pic mutant situé à 76% de colonies rouges qui correspond aux mutations de type "classique"

Le pourcentage de colonies rouges est relatif à la quantité d'ARNm mutés présente dans les cellules tumorales des biopsies mais aussi à la quantité d'ARNm de type sauvage dans les cellules saines de la biopsie et au bruit de fond. L'ensemble de ces biais entraîne le chevauchement du pic d'échantillons de type sauvage et du pic d'échantillon mutés présentant une faible concentration d'ARNm muté et/ou fortement dilué dans celui du tissu sain.

A partir de cette distribution, le seuil du pourcentage de colonies rouges sélectionné pour différencier les tumeurs sauvages des mutantes pour p53 a été fixé à 20%, afin de minimiser le nombre de faux positifs.

Les transcrits des 50 premières tumeurs classées comme mutées (avec un pourcentage de colonies rouges supérieur à 20) ont été séquencés à partir des plasmides contenus dans 4 à 5 colonies. Le résultat a permis de confirmer que les séquences plasmidiques présentent dans différentes colonies contiennent bien des mutations identiques.

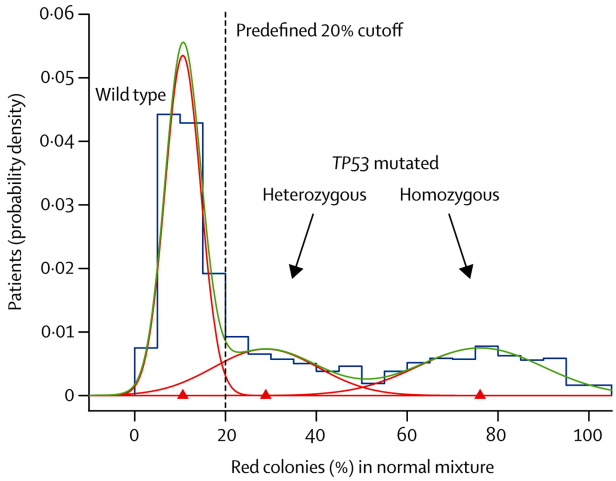


FIGURE 2.2 – *Distribution du pourcentage des colonies rouges (Bonnefoi* et al.*, 2011).* L'histogramme de densité en bleu montre les résultats de l'analyse des pourcentage de colonies rouges retrouvé dans le test des levures. La courbe verte correspond à l'ajustement estimé de la densité normalisé. Les courbes rouges représentent la répartition théorique des résultats du dosage de levure après décomposition en trois pics. Les triangles rouges indiquent la valeur médiane des colonies rouges pour chaque pic (de gauche à droite) : 11%, 29% et 76%.

Conclusion de l'étude du statut de p53 par le test des levures

Les résultats de cette étude n'ont pas permis de mettre en évidence une corrélation entre le statut de *TP53*, le type de chimiothérapie et la réponse au traitement. En effet, un simple test des levures ne permet pas la sélection des patients pour le choix du traitement. Ceci peut être dû à une mauvaise évaluation du statut de *TP53* en raison de la proximité du pic sauvage et du pic des mutations ayant un faible pourcentage de colonies rouges ou encore en raison de la binarité du test. En effet, les tumeurs sont classées soit en tant que sauvages soit en tant que mutantes, or la réponse au traitement peut dépendre du type de la mutation.

Suite à cette observation, un séquençage Sanger direct des produits de RT-PCR utilisés pour la transfection des levures a alors été effectué afin d'identifier les mutations du gène *TP53*. Les échantillons utilisés présentent des mutations déjà définies (voir 2.1.1). Au total, 74 biopsies ont été séquencées par séquençage Sanger à partir de 51 tumeurs. Lorsque le pourcentage de colonies rouges est supérieur à 40%, les variations identifiées par séquençage direct correspondent avec le séquençage des plasmides. Cependant, aucune mutation n'a pu être identifiée pour 15 cas mutés dont le pourcentage est proche du bruit de fond dans le test des levures.

Au vu de ces résultats, il peut être admis que le séquençage Sanger direct des produits PCR échoue pour les échantillon dont le pourcentage de colonies rouges est proche de celui du bruit de fond. Il a alors été proposé de tester si les technologies de séquençage haut débit permettent l'identification des mutations dans de tels échantillons. La figure 2.3 résume l'ensemble des données disponibles pour cette étude.

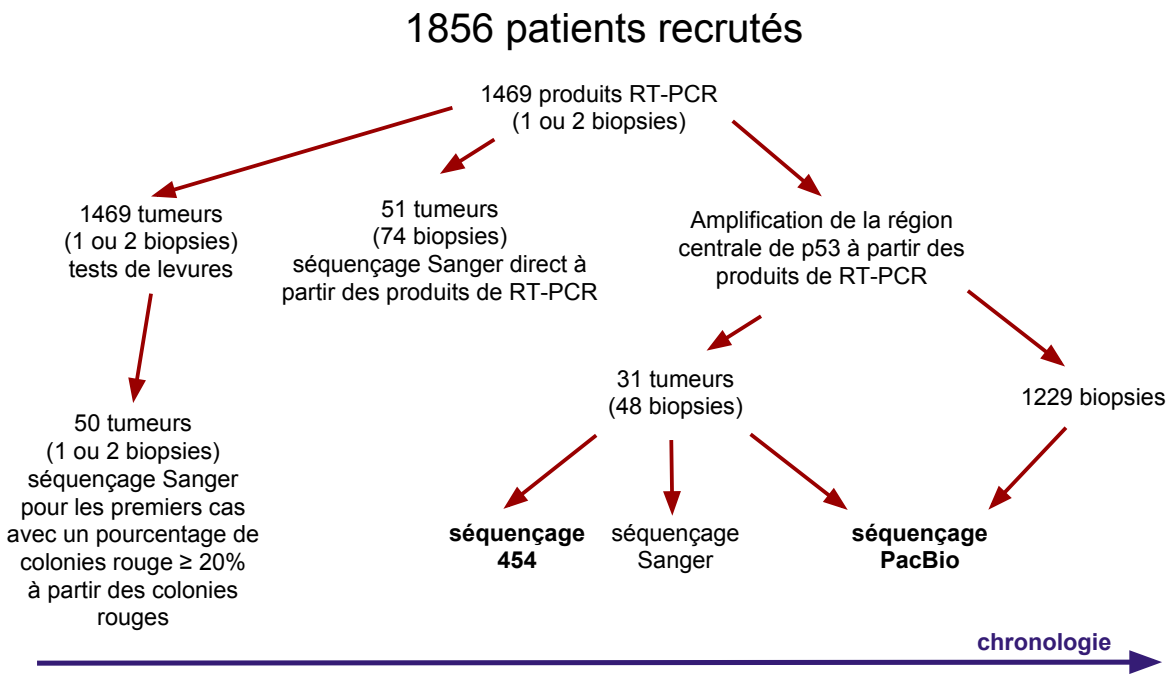


FIGURE 2.3 – *Résumé des données disponibles pour l'étude EORTC 10994.* Le séquençage 454 et le séquençage PacBio (en gras) sont les séquences dont l'analyse fait l'objet de ce chapitre.

2.1.2 Matériel biologique et séquençage

Sélection des échantillons

Afin d'établir la faisabilité du séquençage par NGS et TGS de l'ensemble des échantillons de l'étude EORTC, un ensemble d'échantillons à été sélectionné. Ainsi, 31 tumeurs ont été choisies afin de constituer trois groupes distincts. Ces trois groupes ne sont autres que des sous-ensembles représentatifs des trois pics précédemment décris dans la figure 2.2 :

1. *contrôles négatifs* : 12 échantillons considérés comme sauvages avec un taux d'ARNm mutés nul : ces biopsies sont classées comme négatives au test des levures puisqu'elles présentent un pourcentage de colonies rouges inférieur à 13%.
2. *contrôles positifs* : 18 échantillons considérés comme mutés avec un taux d'ARNm mutés important : ces biopsies sont classées comme positives au test des levures puisqu'elles présentent un pourcentage de colonies rouges compris entre 59 et 92%.
3. *cas difficiles* : 18 échantillons considérés comme mutés avec un taux d'ARNm mutés faible : ces biopsies sont classées comme positives au test des levures et présentent un pourcentage de colonies rouges compris entre 22 et 48%.

Afin de tester la reproductibilité du séquençage, 2 biopsies par tumeur ont été séquencées pour 17 tumeurs appartenant aux groupes des contrôles positifs et des cas difficiles, ramenant le nombre d'échantillons séquencés à 48.

Préparation de la librairie

Le séquençage est effectué à partir des produits de PCR utilisés dans l'étude EORTC lors des tests des levures. La taille de l'ARNm du variant prédominant de p53 est de 2591 nucléotides mais les mutations entraînant une perte de fonction ont lieu principalement dans sa région centrale (Walerych *et al.*, 2012). Ainsi, seule la région d'environ 1 kb comprise entre les nucléotides 159 et 1045, soit la région correspondant aux codons 54 à 348 de p53, est séquencée.

Afin de pouvoir réaliser le séquençage de cette région en une seule fois, cette dernière est amplifiée de façon à être séparée en deux séquences par une PCR qui permet aussi de marquer l'ADN pour son séquençage en multiplexage. Ainsi, les amorces F159 et R642 amplifient la partie C-terminale et les amorces F590 et R1045 la partie N-terminale du fragment à séquencer (figure 2.4). Ceci permet l'obtention de deux fragments de longueur de 454 nucléotides pour le fragment C et 482 nucléotides pour le fragment N. La librairie finale est composée de 192 fragments marqués différemment correspondant au brin sens et anti-sens des 96 fragments d'environ 500 pb issues des 48 biopsies.

2.2 Analyses des données 454

Dans cette section, nous présentons une méthode basée sur l'adaptation de l'approche classique de détection de SNVs à partir de séquences NGS ciblées. Cette méthode a été évaluée sur les données du séquençage des produits de RT-PCR provenant de l'ARNm p53. Elle a du répondre à différentes problématiques : la validation de la spécificité des amorces PCR, la détection les épissages pathologiques mais aussi la considération de l'hétérogénéité des échantillons et plus précisément, la contamination par le tissus sain. Le séquençage NGS a été effectué sur un sous ensemble des 48 biopsies par la technologie

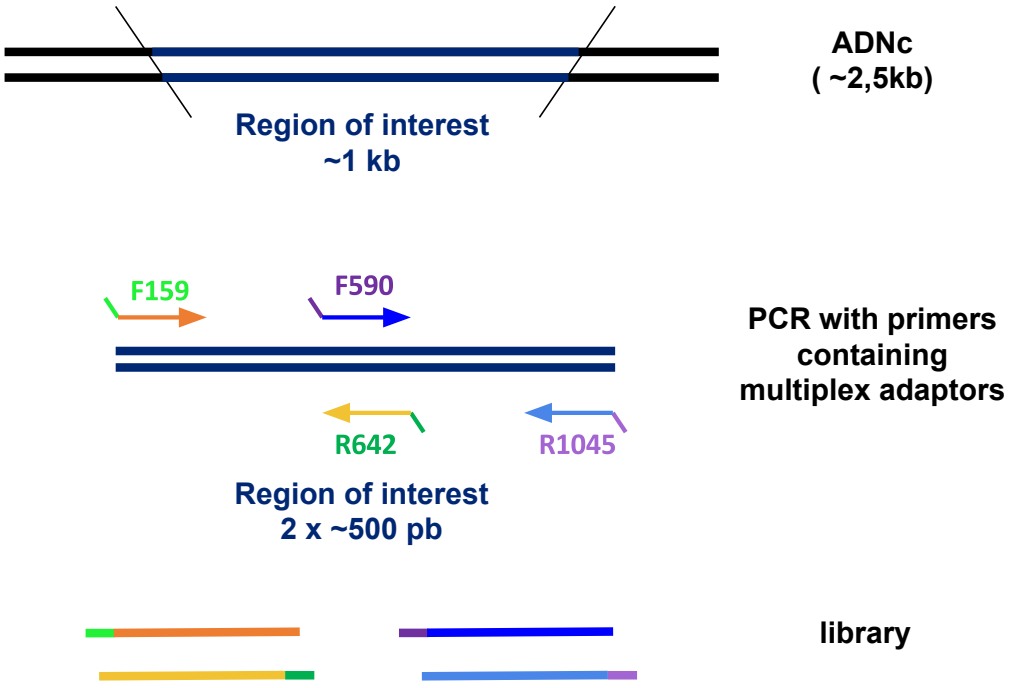


FIGURE 2.4 – *Stratégie de séquençage.* Seule la région centrale d'environ 1 kb est séquencée. Cette région centrale est amplifiée de façon à être séparée en deux séquences d'environ 500 kb par une PCR qui permettant le multiplexage. Ainsi, les amorces F159 et R642 amplifient la partie C-terminale et les amorces F590 et R1045 la partie N-terminale du fragment à séquencer.

454 GS FLX Titanium (Roche). Ces données ont été rendues disponibles sur la base de donnée SRA du NCBI sous le numéro d'accession SRP020456 (BioProject PRJNA193388). Un fichier qualité et un fichier fasta pour chaque amorce sont disponibles pour l'ensemble des biopsies.

2.2.1 Description des données 454

Le nombre de *reads* par fragment est compris entre 84 et 1757, avec seulement 2 inférieurs à 100. La faible profondeur peut être due au fait que la quantification avant la mise en commun des produits de PCR n'ai pas été suffisamment précise. Une manière usuelle d'évaluer la qualité d'un séquence est d'observer la qualité de chaque base de chaque *read* ainsi que la qualité moyenne par *read*. La qualité et la profondeur de chaque base en fonction de sa position sur le *read* avant et après nettoyage des séquences est décrit dans la figure 2.5 (A et B). La qualité des *reads* est satisfaisante (QPhred compris entre 30 et 40) en début de séquence mais tend à diminuer considérablement à la fin (figure 2.5.A). Un nettoyage des séquences est alors effectué en utilisant l'outil FASTX toolkit (Gordon et Hannon, 2010) (version 0.0.13.1) qui permet la suppression des nucléotides séquencés dont la qualité est inférieur à 30 en fin de séquence.

Le nettoyage des *reads* permet un bon compromis entre un niveau de qualité et une profondeur acceptable (figure 2.5.B). Ceci est aussi observable dans les figures 2.5 (C et D) qui représentent le nombre de *reads* en fonction de la qualité moyenne. En effet, après nettoyage, la qualité moyenne des *reads* est améliorée, la majorité présentant une qualité moyenne supérieure ou égale à 30. Le nettoyage diminue ainsi l'identification de mutations qui seraient dues à une erreur de séquençage. Par ailleurs, nous observons la présence de régions où la qualité chute considérablement (QPhred 20). Ce phénomène provient de la présence d'homopolymères, la technologie 454 étant en effet connue pour produire des erreurs liées à leur présence.

2.2.2 Méthode

La détection de SNVs repose sur l'identification de variations entre un ensemble de *reads* et le génome de référence correspondant (sous-section 1.2.4 de l'introduction). A ce jour, l'approche classique est de procéder au *mapping* de chacun des *reads* en identifiant la position du *read* sur le génome de référence et en opérant à son alignement sur ce *locus* puis de rechercher les différences. Dans notre cas, cette approche classique a été adaptée à la recherche de SNVs dans des séquences NGS générées à partir du séquençage de transcrits d'un gène suppresseur de tumeur (figure 2.6).

Dans le cas de séquençage de transcrits, l'alignement avec un algorithme adapté à ce type de séquençage sur le génome de référence peut entraîner des faux négatifs. En effet, les algorithmes d'alignement sont capable d'identifier des sites d'épissages utilisés pour la suppression des introns lors de l'étape de maturation de l'ARNm. Cependant, l'utilisation de certains sites d'épissage entraîne la formation d'ARNm puis de protéines pathologiques. Ainsi, si l'alignement est effectué uniquement sur la séquence génomique, ces altérations liées à un épissage pathologique ne sont pas identifiées. Afin de contourner ce problème, nous avons adapté la méthode classique. Cette adaptation est constituée des étapes suivantes :

1. L'alignement des *reads* sur :

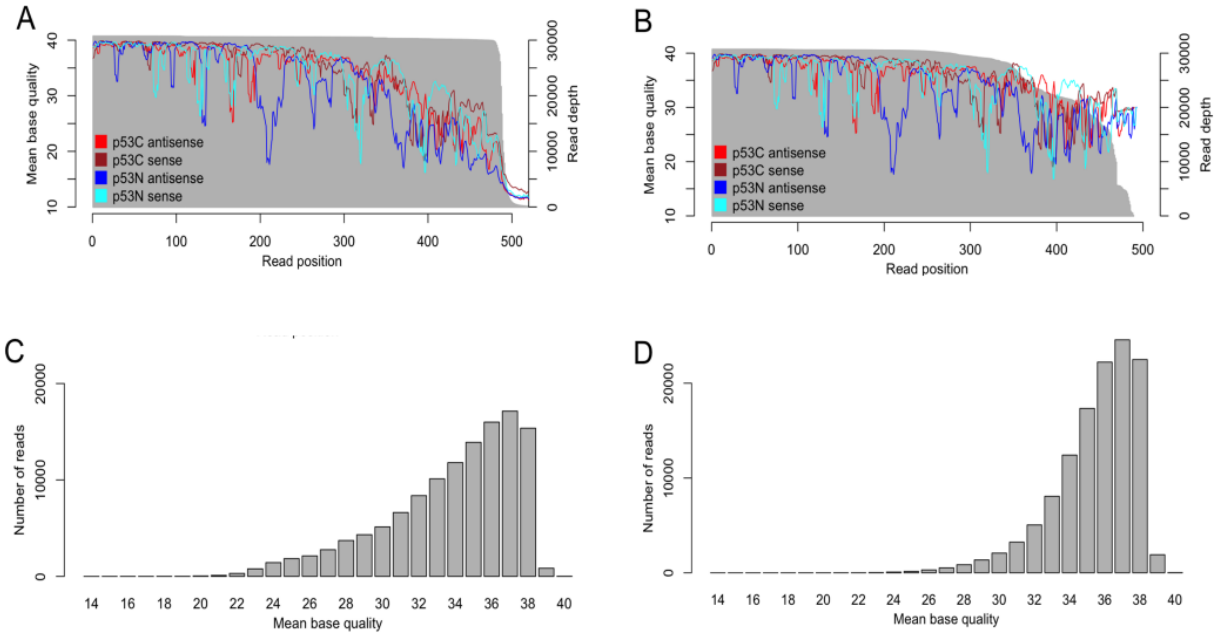
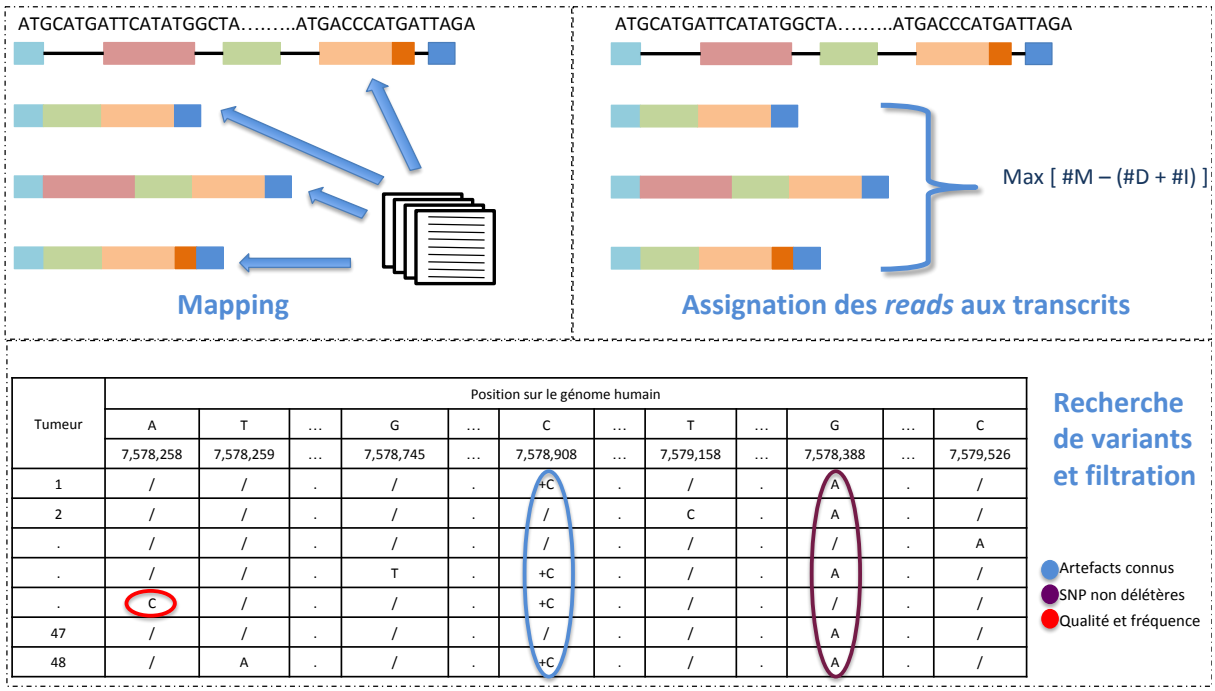


FIGURE 2.5 – *Qualité des séquences NGS avant et après nettoyage.* Les figures A et B représentent le score de qualité (QPhred) moyen pour chaque position sur les *reads*. Les fragments C-terminaux sont représentés en bordeaux (fragment sens) et rouge (fragment antisens) et les fragments N-terminaux en cyan (fragment sens) et bleu (fragment antisens). La silhouette grise représente la profondeur de lecture. Les figures C et D représentent la distribution du nombre de *reads* en fonction de la qualité moyenne sur l'ensemble du fragment. Les données brutes sont présentées dans A et C et les données nettoyées sont présentées dans B et D.



Tumeur	Position sur le génome humain											
	A	T	…	G	…	C	…	T	…	G	…	C
	7,578,258	7,578,259	…	7,578,745	…	7,578,908	…	7,579,158	…	7,578,388	…	7,579,526
1	/	/	.	/	.	+C	.	/	.	A	.	/
2	/	/	.	/	.	/	.	C	.	A	.	/
.	/	/	.	/	.	/	.	/	.	/	.	A
.	/	/	.	T	.	+C	.	/	.	A	.	/
.	C	/	.	/	.	+C	.	/	.	/	.	/
47	/	/	.	/	.	/	.	/	.	A	.	/
48	/	A	.	/	.	+C	.	/	.	A	.	/

FIGURE 2.6 – *Identification de SNVs par alignement.* L'alignement des *reads* est effectué sur le génome de référence ainsi que sur un ensemble de transcrits alternatifs. L'ensemble des *reads* est alors soumis à un tri permettant l'assignation de chacun d'entre eux au transcrit dont il est le plus proche en terme de distance. Une approche classique de recherche de variants permet alors l'identification des mutations dans l'ensemble des échantillons.

   - une référence génomique. Cet alignement permet de valider la stratégie d'amplification et la spécificité des amorces PCR. En effet, cela permet de vérifier que l'ensemble des *reads* soient *mappés* au *locus* du gène dont l'ARNm est amplifié et exclure toute contamination aspécifique potentielle lors des différentes PCR.
   - un ensemble de références transcriptomiques. Cet alignement permet la détection d'épissages alternatifs pathologiques. Les transcrits utilisés sont restreints à ceux présentant une différence de séquence pour la région séquencée.

2. L'assignation des *reads* à un transcrit de référence. Nous avons pour cela développé un algorithme de tri des *reads* en fonction du transcrit alternatif sur lequel il s'aligne le mieux. Cet algorithme calcule un score de correspondance s pour chaque paire *read* / transcrit grâce à la décomposition du CIGAR contenu dans les fichiers d'alignements, tel que :

$$s = M - (I + D)$$

   où M est le nombre de *matches*, I est le nombre d'insertions et D est le nombre de délétions. Le *read* est alors assigné au transcrit qui maximise le score de correspondance.

3. la recherche de variants.

2.2.3 Application aux données 454

L'alignement à partir des *reads* nettoyés (voir 2.2.1) a été réalisé sur deux types de séquences de référence comme décri dans la sous-section précédente :

- *une référence génomique :* le génome de référence humain GRCh37.
- *trois références transcriptomiques :* les isoformes $NM\_000546.5$ (α), $NM\_001126114.2$ (β) et $NM\_001126113.2$ (γ). L'isoforme α correspond au transcrit fonctionnel de p53 tandis que les isoformes β et γ correspondent aux transcrits présentant une variation de séquence dans la région séquencée. Seul un l'isoforme α permet de produire une protéine entièrement fonctionnelle, tandis que les isoformes β et γ incluent un exon alternatif dans l'intron 9.

L'outil GMAP (version 2012-11-09) a été sélectionné pour son efficacité à aligner des séquences transcriptomiques contenant des mutations susceptibles d'entraîner des variations d'épissage ou d'être situées à proximité ou à l'intérieur de sites d'épissage (Wu et Watanabe, 2005). En effet, ce type de mutations peut être à l'origine de mauvais alignement entraînant la détection erronée de SNVs (Mielczarek et Szyda, 2016). Par ailleurs, bien que que cet outil soit spécialement efficace pour l'alignement sur une référence génomique, nous avons choisi de le conserver pour celui sur les transcrits alternatifs afin de s'affranchir de biais d'alignement éventuels lors de la mise en commun des résultats. Pour cela, la fixation d'un seuil élevé pour la taille des introns dans GMAP ("min-intronlength=15000") facilite l'identification des délétions. En effet, si ce paramètre n'est pas renseigné, GMAP identifie un épissage ce qui exclue la détection de mutation bien que cet épissage soit pathologique et provoque une délétion.

Suite à l'alignement, les données sont alors converties en format pileup grâce au logiciel Samtools (Li *et al.*, 2009). Les variants sont alors recherchées avec Varscan (version 2.3.4). Les seuils fixés pour admettre une mutation sont de 5 pour la profondeur de lecture, 30 pour la qualité moyenne et de 4,5% pour la fréquence de l'allèle mutée.

2.2.4 Résultats

Le tableau 2.1 décrit pour l'ensemble des biopsies, le génome de référence à partir duquel la mutation a été identifié, sa position sur ce génome, l'allèle de référence et de l'allèle muté ainsi que leurs profondeur de lecture et qualité respectives. La figure 2.7 quant à elle présente les résultats des mutations identifiées dans l'ensemble des biopsies comparées à celles identifiées par séquençage Sanger direct des produits RT-PCR.

S	B	Genomes	Position	F	RD	AD	RQ	AQ	Ref	Alt
83	1	GRCh37	7576525	100	0	5	0	38	A	C
169	1	GRCh37,NM_000546.5	7578442	84	162	824	38	38	T	C
169	2	GRCh37,NM_000546.5	7578442	74	264	741	38	38	T	C
183	1	NM_000546.5	7578554	20	862	222	38	38	[...]	A
183	1	NM_000546.5	7579313	9	715	74	38	37	[...]	G
183	2	NM_000546.5	7578554	21	570	149	38	38	[...]	A
183	2	NM_000546.5	7579313	10	536	67	37	37	[...]	G
192	1	GRCh37,NM_000546.5	7579479	18	623	140	39	39	[...]	C
192	2	GRCh37,NM_000546.5	7579479	24	174	54	39	39	[...]	C
207	1	GRCh37,NM_000546.5	7578210	17	1263	260	39	38	T	C
207	1	GRCh37,NM_000546.5	7579358	57	254	340	38	38	C	G
207	2	GRCh37,NM_000546.5	7578210	7	1116	83	38	38	T	C
207	2	GRCh37,NM_000546.5	7579358	69	172	391	38	37	C	G
215	1	GRCh37,NM_000546.5	7577557	22	1081	297	39	38	AG	A
215	2	GRCh37,NM_000546.5	7577557	14	1562	253	39	37	AG	A
221	1	GRCh37,NM_000546.5	7578190	71	232	570	39	39	T	C
221	2	GRCh37,NM_000546.5	7578190	65	318	601	39	39	T	C
256	1	GRCh37,NM_000546.5	7578184	81	289	1257	39	38	G	A
269	1	GRCh37	7578176	100	0	363	0	38	C	T
269	2	GRCh37	7578176	100	0	298	0	38	C	T
272	1	NM_000546.5	7578177	13	756	112	39	35	C	CAGACT
272	1	GRCh37	7578177	12	750	101	39	38	C	T
272	2	NM_000546.5	7578177	7	632	47	39	36	C	CAGACT
272	2	GRCh37	7578177	7	632	44	39	38	C	T
276	1	GRCh37,NM_000546.5	7577545	5	1144	59	38	37	TGCC	T
276	1	GRCh37,NM_000546.5	7578278	11	303	38	38	35	G	C
279	1	GRCh37,NM_000546.5	7577115	71	348	855	38	38	A	G
279	2	GRCh37,NM_000546.5	7577115	70	367	856	38	38	A	G
290	1	GRCh37,NM_000546.5	7578433	10	471	54	38	39	G	GACTGC
290	2	GRCh37,NM_000546.5	7578433	8	805	74	38	38	G	GACTGC
316	1	GRCh37,NM_000546.5	7578528	92	61	731	32	37	A	T
316	2	GRCh37,NM_000546.5	7578528	93	43	568	32	37	A	T
318	1	GRCh37,NM_000546.5	7577094	51	551	570	38	38	G	A
318	2	GRCh37,NM_000546.5	7577094	31	1051	475	38	38	G	A
320	1	GRCh37,NM_000546.5	7579347	56	314	396	37	36	[...]	A
320	1	GRCh37	7579349	42	162	115	38	35	[...]	A
320	2	GRCh37,NM_000546.5	7579347	54	381	452	37	36	[...]	A
320	2	GRCh37	7579349	37	224	134	38	36	[...]	A
323	1	GRCh37,NM_000546.5	7578437	5	970	51	38	37	G	A
326	1	GRCh37,NM_000546.5	7578495	15	732	132	38	34	CA	C
326	2	GRCh37,NM_000546.5	7578495	13	452	69	38	32	CA	C
340	1	GRCh37,NM_000546.5	7578265	65	253	465	36	36	A	C
340	2	GRCh37,NM_000546.5	7578265	71	104	256	36	37	A	C
341	1	GRCh37,NM_000546.5	7577120	68	352	753	38	38	C	T
341	2	GRCh37,NM_000546.5	7577120	80	343	1341	38	38	C	T

TABLE 2.1 – *Tableau des SNVs identifiées par séquençage NGS.* Pour chaque biopsie (B) de chaque échantillon (S), la mutation identifiée est décrite par le(s) génome(s) de référence (Genomes) à partir duquel elle a été identifiée, sa position sur ce génome (Position), sa fréquence relative(F), l'allèle de référence (Ref) et de l'allèle alternative (Alt) ainsi que leurs profondeur de lecture (RD / AD) et qualité (RQ / AQ) respectives. Les délétions de taille supérieure à 4 nucléotides sont remplacé par [...].

EORTC ID	Red yeast colonies biopsy 1 (%)	Red yeast colonies biopsy 2 (%)	No of plasmids with the indicated mutation	Sanger plasmid mutations	No of biopsies tested by NGS/Sanger cDNA pool sequencing	Sanger cDNA pool mutations	NGS mutations	Mutation type
169	85	77	3/3	Y163C	2	Y163C	Y163C	Missense
207	62	73	2/2	R110P	2	R110P	R110P	Missense
221	84	67	4/4	Y220C	2	Y220C	Y220C	Missense
279	79	72	4/4	C275R	2	C275R	C275R	Missense
316	92	92	4/4	F134L	2	F134L	F134L	Missense
318	67	33	nt	na	2	R282W	R282W	Missense
320	83	73	4/4	R110ΔRLGF	2	R110ΔRLGF	R110ΔRLGF	In-frame deletion
340	66	77	4/4	I195S	2	I195S	I195S	Missense
341	59	86	3/3	R273H	2	R273H	R273H	Missense
83	37	22	4/4	p53-beta	2	p53-beta*	p53-beta	Splicing
183	48	45	4/4	intron 4 sd¶	2	Intron 4 sd¶	Intron 4 sd¶	Splicing
192	23	29	1/4	P64fs	2	P64fs	P64fs	Frameshift
215	38	25	2/3	S241fs	2	S241fs†	S241fs	Frameshift
269	40	43	3/4	Intron 6 sd	2	Intron 6 sd	Intron 6 sd	Splicing
272	36	25	2/3†	Intron 6 sd	2	Intron 6 sd	Intron 6 sd	Splicing
276	27	na	2/4	G245ΔG	1	wt	G245ΔG§	In-frame deletion
290	22	21	2/4	S166fs	2	S166fs	S166fs	Frameshift
323	22	na	2/4	Q165Z	1	wt	Q165Z	Nonsense
326	22	23	1/4	L145fs	2	L145fs	L145fs	Frameshift
158	7	na	na	na	1	wt	wt	na
193	7	na	na	na	1	wt	wt	na
256	8	na	na	na	1	P222L	P222L	Missense (SNP)
267	8	na	na	na	1	wt	wt	na
284	5	na	na	na	1	wt	wt	na
285	13	na	na	na	1	wt	wt	na
288	10	na	na	na	1	wt	wt	na
306	6	na	na	na	1	wt	wt	na
311	10	na	na	na	1	wt	wt	na
312	7	na	na	na	1	wt	wt	na
319	10	na	na	na	1	wt	wt	na
322	8	na	na	na	1	wt	wt	na

Full details of the mutations are given in Supplementary Tables 3 and 4. ¶The plasmids contained a mixture of abnormally spliced products; the same splice variant was seen in one plasmid as in the pooled cDNA and NGS reads (the causal mutation was only visible in the plasmid sequences).
*The causal mutation was not identified.
†A probable PCR splicing artefact was also identified in biopsy 1.
‡A splice donor mutation was present in one plasmid and intron retention in the other.
§A P191A mutation encoding functional p53 was also present in this case (see text for details).
nt = not tested; na = not applicable; wt = no mutations identified; fs = frameshift (the mutant codon is indicated); sd = splice donor.

FIGURE 2.7 – *Résumé des SNVs identifiées par séquençage NGS et comparées à celles identifiées par séquençage Sanger.* Ce tableau décrit pour l'ensemble des biopsies, leur pourcentage de colonies rouges, le nombre de plasmides contenant la mutation identifiée par séquençage Sanger ainsi que sa description, le nombre de biopsies testé via le séquençage Sanger direct des produits de RT-PCR et la caractérisation des mutations identifiées.

Contrôles positifs

Dans le groupe des échantillons présentant un pourcentage élevé de colonies rouges dans les tests des levures, les mutations identifiées à partir des séquences NGS sont les mêmes que celles présentes dans les séquences plasmidiques avec un score QPhred moyen de l'allèle mutant compris entre 32 et 39. De plus, lorsque deux biopsies d'une même tumeur ont été séquencées, la mutation retrouvée est identique (figure 2.7 et tableau REF).

Le mutant pour lequel le séquençage du plasmide n'a pu être effectué (cas 318) contient une mutation R282W dans les deux biopsies. Avec les seuils permissifs de Varscan, le cas 276 présente deux mutations. La seconde mutation (P191A) est présente dans 11% des reads mais correspond probablement à une mutation passagère. Ceci est soutenu par le fait que la protéine mutante est de type sauvage dans la base de données de tests fonctionnels (Kato *et al.*, 2003).

Comme attendu, lorsque le pourcentage de colonies rouges est élevé dans le test de levure, c'est à dire lorsque la quantité d'ARNm mutés est importante, le séquençage NGS détecte efficacement les mutations.

Contrôles négatifs

Le groupe des échantillons présentant un faible pourcentage de colonies rouges dans les tests des levures ne présente aucune mutation dans 11 des 12 échantillons.

L'échantillon restant (cas 256) possède une mutation P222L dans 81% des *reads*. La qualité moyenne de l'allèle mutante est de 39 indiquant une grande confiance dans ce résultat. Le test des levures a été utilisé dans l'étude EORTC puisqu'il permet la distinction entre les mutations entraînant une perte de fonction de la protéine p53 en tant que facteur de transcription et les mutations passagères. Au contraire, le séquençage NGS permet l'identification de l'ensemble des altérations.

Cas difficiles

Les échantillons présentant un pourcentage de colonies rouges au test de levure compris entre 22 et 48% présentent des mutations non-sens ou décalant le cadre de lecture.

Dans un cas (cas 192), aucune mutation n'est retrouvée à partir du séquençage des plasmides lors d'un premier séquençage. Les séquences NGS quant à elles, révèlent une mutation entraînant un décalage du cadre de lecture dans une région extérieure à la région séquencée dans le premier séquençage à partir des plasmides. Un nouveau séquençage incluant cette région a permis de confirmer la présence de cette mutation.

Lors de l'alignement des séquences d'ADNc sur le génome, si la mutation entraîne une variation d'épissage (*splicing type*) cette anomalie n'est pas détectée. En effet, le logiciel d'alignement ne permet pas la différenciation entre les épissages alternatifs "normaux" et les épissages pathologiques. L'inclusion de cas "difficiles" dans cette étude de faisabilité permet d'appréhender ce type de mutations. Il n'y a pas de contrainte sur la taille de l'insert dans la stratégie de clonage par réparation homologue dans le test des levures. En revanche, tout ce qui augmente la taille d'un fragment est susceptible d'entraîner sa perte lors du séquençage NGS. Or, les mutations d'épissage mènent généralement à un mélange de rétention d'introns et de suppression d'exons, la mutation étant souvent plus facilement détectable dans le second cas. C'est exactement ce qui a été vu dans les données NGS.

En effet, alors que la mutation réelle est visible dans les séquences plasmidiques dans les quatre cas présentant des défauts d'épissage, seulement trois sont présentes dans les données NGS. Dans le quatrième cas (cas 183) la mutation est situé à cinq nucléotides après le site d'épissage (cg^gtca**g** > cg^gtca**c**). Aucune des séquences NGS ne contient l'intron conservé (il augmente la taille du fragment de 757 pb). Cependant, plusieurs *reads* correspondent à l'ARNm épissé à partir d'un site donneur cryptique dans l'exon 4, cet événement a également été retrouvé dans un plasmide. Le donneur cryptique a la séquence ag gt et correspond à un donneur d'épissage normal lors de l'alignement sur la référence génomique. En revanche, la délétion est bien détectée à partir des alignements sur les références transcriptomiques.

L'épissage alternatif dans le cas 83 est causé par une mutation dans le site donneur à la fin de l'exon alternatif présent dans l'isoforme beta. Cette mutation améliore le site donneur (t**t**^gt > t**g**^gt) et conduit à l'inclusion plus fréquente de l'exon alternatif. Les quatre plasmides séquencés pour ce cas contiennent l'exon supplémentaire et le donneur d'épissage muté. La mutation dans le site donneur est facilement identifiée dans une des deux biopsies de cette tumeur à partir des données NGS car la base mutée présente un score QPhred élevé (38). La principale difficulté à identifier la mutation est la faible profondeur de l'allèle muté empêchant la détection dans la seconde biopsie. Ceci s'explique probablement par l'augmentation de la taille des séquences (+132 pb) entraînant leur perte. De plus, dans le cas 83, le pourcentage de colonies rouges dans le test des levures est de 37 et 22%, pour les biopsies 1 et 2 respectivement, révélant une faible concentration en ARNm mutants.

Les mutations responsables des défauts d'épissage dans les deux derniers cas 269 et 272 ont été facilement identifiées. Dans ces deux cas, la mutation affecte le site donneur d'épissage de l'intron 6, ce qui entraîne l'utilisation d'un donneur cryptique (tg^gt) cinq nucléotides après le donneur normal.

Lors de l'alignement sur le génome, les formes pathologiques d'épissage n'ont pas été identifiées en tant que telles mais la mutation elle-même est bien identifiée. L'alignement sur les variants permet quant à lui d'identifier correctement la rétention de cinq nucléotides de l'intron. Ainsi, l'ensemble des mutations du groupe "difficile" est correctement identifié par séquençage 454.

La stratégie d'identification des mutations par l'adaptation de l'approche classique de détection de SNVs est efficace pour les données issues du séquençage des ARNm de p53 via la technologie 454.

2.2.5 Qualité des données 454

Biais biologiques

Un phénomène d'épissage par PCR peut se produire lorsque la reverse transcription se termine avant que la polymérase n'atteigne le site d'amorce opposé. Les extrémités des fragments obtenus s'apparient les uns aux autres dans les cycles suivants au niveau de séquences présentant quelques bases d'homologie. Bien que la taille et la position de ces délétions soient très variables, ces dernières peuvent être amplifiées si elles arrivent dans un cycle de PCR précoce et entraîner un bruit de fond élevé. L'utilisation d'amorces spécifique de p53 dans l'étude EORTC 10994 permet de diminuer sensiblement ce phénomène. Cependant, de tels épissages sont tout de même visibles dans les séquences NGS et se traduisent par la présence de petites délétions (figure 2.8).

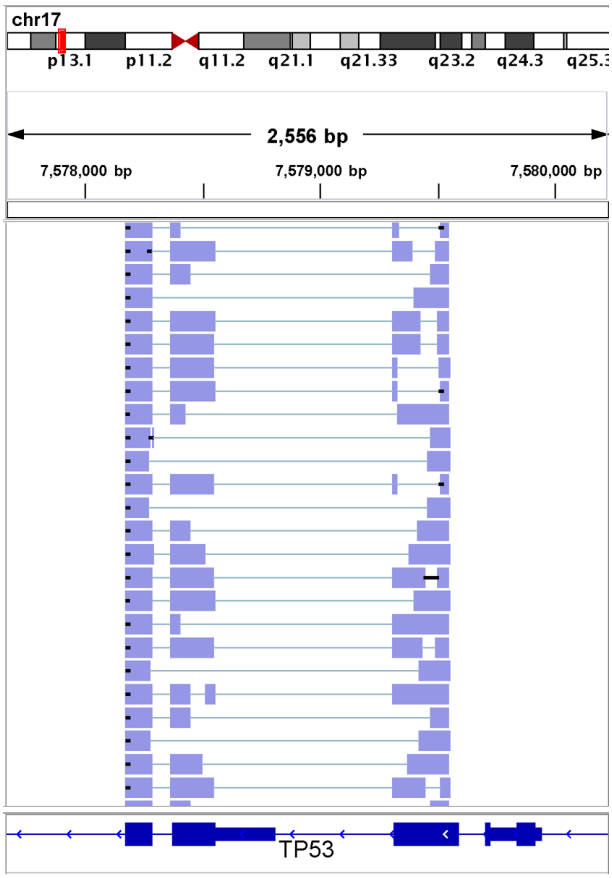


FIGURE 2.8 – *Visualisation du phénomène d'épissage par PCR dans une tumeur sauvage pour p53.* La visualisation (IGV) du phénomène d'épissage par PCR dans une tumeur classée comme type sauvage (échantillon 158) montre que ces artefacts entraînent de petites délétions dans les exons 4 à 6. Les barres épaisses bleues foncées dans le cadre du bas montrent les exons codants du gène p53. Les barres bleues claires dans le cadre central représentent les *reads*. Les délétions dans les *reads* qui ne commencent pas aux bornes des exons sont causées par des artefacts d'épissage par PCR.

Le compte du nombre de *reads* contenant des délétions de taille supérieure à 5 nucléotides et présentes dans au moins 2 biopsies permet de quantifier l'impact de ce phénomène. La figure 2.9 permet de visualiser les délétions en fonction de leur position sur l'ADNc de p53 et du nombre de *reads* qui les supportent. Ainsi, le pourcentage de *reads* comportant de telles délétions varie de 0.5 à 8.7 avec une moyenne sur l'ensemble des biopsies de 3.21%. Afin de ne pas générer de faux positifs, ces artefacts ont été exclus des résultats finaux.

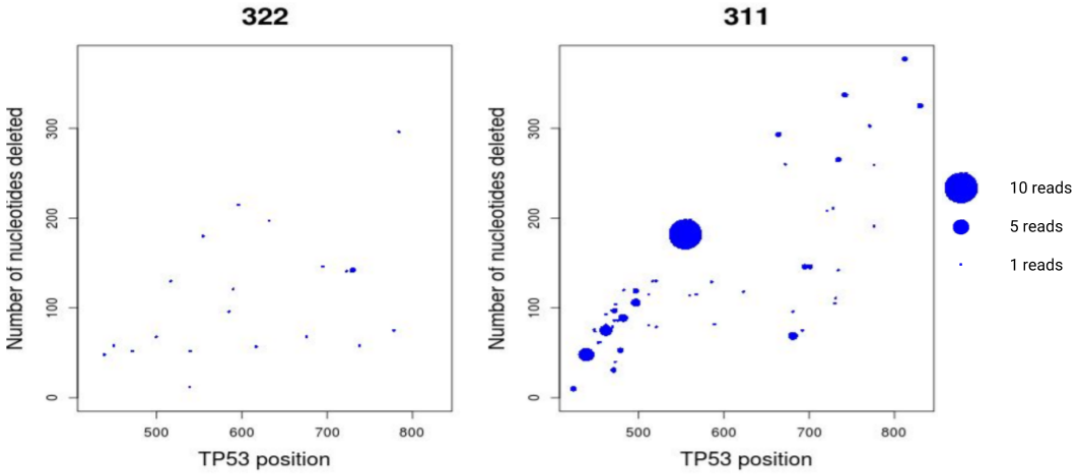


FIGURE 2.9 – *Nombre de nucléotides délétés en fonction de leur position sur le fragment N-terminal de p53* Les délétions présentes dans les tumeurs 322 et 311 sont représentées en fonction de leur taille et de leur position sur le transcrit du gène *TP53*. La taille des points est proportionnelle aux nombre de *reads* présentant la délétion. La tumeur 322 (A) présente quelque délétions tandis que la tumeur 311 (B) en présente de nombreuses.

Biais technologiques

La qualité pour chaque nucléotide et la couverture de p53 de l'ensemble des séquences NGS sont résumées dans la figure 2.10. La profondeur moyenne est plus élevée pour le fragment 3' que le pour le fragment 5'. Après nettoyage des séquences, le pourcentage de *reads* s'alignant au locus p53 est compris entre 96 et 100%. Ceci indique qu'il n'y a pas eu de contamination des librairies et que GMAP a fonctionné correctement. Par ailleurs, une faible qualité est observée principalement aux positions de la séquence de l'ARNm de p53 présentant des homopolymères. Les sites les plus fréquemment touchés sont indiqués par des points rouges et correspondent à un score de qualité inférieur à 25.

Corrélation entre le test des levures et le séquençage 454

La figure 2.11 représente le pourcentage de *reads* altérés présents dans les données par rapport au pourcentage de colonies rouges obtenus lors des tests des levures. Les cas 83 et 183 ont été exclus de l'analyse, les événements d'épissages alternatifs empêchant toute comparaison. Pour les cas restants, la corrélation entre le nombre de *reads* altérés et le pourcentage de colonies rouges est très significative (coefficient de corrélation de 0.97). Le

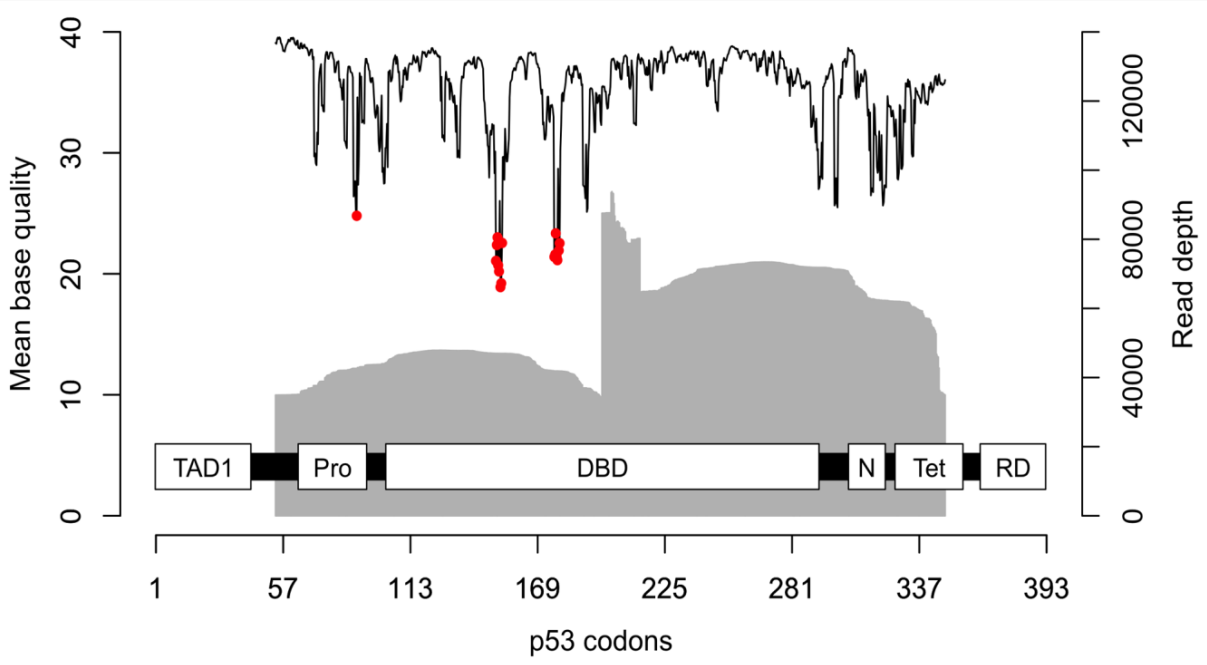


FIGURE 2.10 – *Qualité des séquences NGS sur p53.* La qualité moyenne des bases (QPhred) et la profondeur de lecture sont reportées en fonction de la position sur p53. La ligne continue représente la qualité des bases et la silhouette grise la profondeur de lecture. L'ADNc est séquencé entre les codons 54 et 348. Les points rouges indiquent 16 bases intégrant des homopolymères où la qualité moyenne retrouvée à ces positions est inférieure à 25. Cette baisse de qualité affecte les codons 89, 151, 152, 153, 176, 177 et 178.

point d'intersection de la droite de régression linéaire avec l'axe des ordonnées est 15.1% ce qui représente le bruit de fond du test des levures, c'est à dire le pourcentage de colonies rouges attendu pour un échantillon sain.

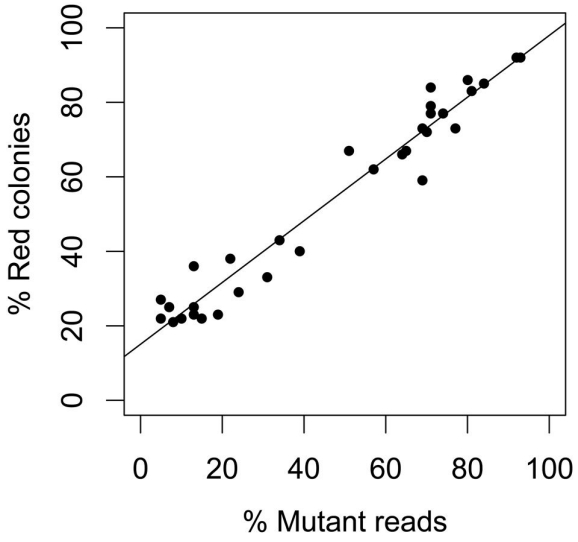


FIGURE 2.11 – *Pourcentage de colonies rouges en fonction du pourcentage de* reads *altérés.* Le pourcentage de *reads* contenant des mutations est représenté en fonction du pourcentage de colonies rouges.

Dans l'étude EORTC, le bruit de fond de colonies rouges était estimé à 11 ± 4% à partir des biopsies sauvages. Ainsi, le bruit de fond calculé à partir des biopsies de l'étude pilote est supérieur à l'estimation initiale faite à partir de l'ensemble des échantillons.

2.3 Analyses des données PacBio

Dans cette section, nous allons voir que l'adaptation de l'approche classique de détection de SNVs présentée dans la sous-section 2.2.2 est inefficace lors de son application aux données PacBio. Après l'identification de caractéristiques liées aux séquences générées par la technologie PacBio et les altérations qu'elles comprennent, nous proposons une nouvelle méthode de détection de SNVs : MICADo. Cette méthode, basée sur les graphes de *de Bruijn*, permet une identification efficace des SNVs en différenciant les altérations spécifiques aux échantillons des altérations liées aux différents biais biologiques ou technologiques. L'algorithme MICADo est développé en langage Python. Le code source est disponible sur github `http://github.com/cbib/MICADo`.

2.3.1 Description des données PacBio

Le séquençage a été réalisé sur 1277 biopsies de l'étude clinique EORTC 10994 via la technologie de séquençage circulaire PacBio (PacBio RSII ; P4-C2 chemistry). Ces données ont été rendues disponible sur la base de donnée SRA du NCBI sous le numéro d'accession SRP064161 (BioProject PRJNA290142). Dans ce travail, nous nous concentrerons sur la recherche de SNVs dans les 48 échantillons précédemment décrits (voir sous-section 2.1.2) pour lesquels nous disposons de données de référence (séquençages Sanger et 454).

Après démultiplexage et suppression des adaptateurs, le nombre de *reads* varie de 89 à 8771 par fragment.

2.3.2 Recherche de variants par l'adaptation de l'approche classique (2.2.2)

L'utilisation de la méthode par alignement ayant fournis des résultats très satisfaisants sur les données issus du séquençage 454, nous avons conservé la même approche pour la recherche de variants dans les données PacBio. Une fois les données converties en format pileup, les variants sont alors recherchées avec Varscan (version 2.4.0). Les seuils fixés pour admettre une mutation sont de 5 pour la profondeur de lecture, 60 pour la qualité moyenne et de 5% pour la fréquence de l'allèle muté et un seuil de 0,001 pour la p-valeur.

L'ensemble des altérations attendues a été retrouvé à partir de ce pipeline (tableau 2.2). Cependant, VarScan présente un taux élevé de faux positifs dont le nombre semble être indépendant de la catégorie des échantillons (figure 2.18). Dans le contexte d'une étude clinique où la distinction entre un SNV réel et un artefact introduit par des erreurs de séquençage est essentiel, ces résultats sont inexploitables.

2.3.3 Observations des données PacBio

Afin de comprendre pourquoi le nombre de faux positifs est si élevé bien que les paramètres de détection des SNVs choisis soient très stringents, nous avons recherché si ces altérations présentent des caractéristiques particulières.

Lorsque l'on regarde les séquences alignées d'un échantillon, nous observons en effet, un nombre très important d'altérations (figure 2.12.A). Les erreurs de séquençage de la technologie PacBio sont connues pour être majoritairement des insertions et des délétions. En effet, nous observons un nombre moyen d'insertions et de délétions par *read* de 2 et un nombre de substitutions, ou mismatches, de 1 dans l'ensemble des échantillons séquencés (figue 2.12.B). Si l'on s'intéresse au nombre d'altérations uniques par échantillon, nous observons que leur distribution est autour de 700 par échantillon (figure 2.12.C).

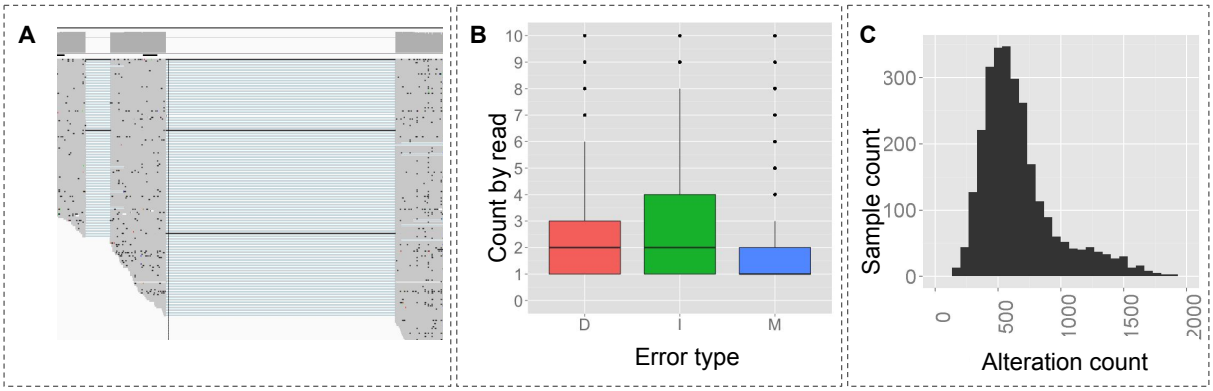


FIGURE 2.12 – *Observation des séquences alignées issue de la technologie PacBio.* A. Visualisation de l'alignement des séquences d'un échantillon sur 3 exons de p53 à partir de l'outil de visualisation IGV. B. Nombre d'altération de type insertion (I), délétion (D) et mismatch (M) par *read*. C. Histogramme du nombre d'altérations uniques par échantillon.

A partir des données d'alignement, deux caractéristiques liées aux séquences générées par la technologie PacBio sont identifiées :

1. la qualité des bases altérées n'est pas informative. En effet, lorsque l'on s'intéresse aux altérations retrouvées dans les échantillons ne présentant pas de mutations (échantillons présentant un pourcentage de colonies rouges inférieur à 10), la qualité des bases qui les supportent semble inutile pour leur filtrage (figure 2.13) puisqu'un grand nombre des altérations présentent des qualités supérieures à 40.
2. les altérations sont contexte-spécifiques. En effet, certaines altérations sont récurrentes, voire systématiques, dans l'ensemble des échantillons et sembleraient être dépendantes du contexte dans lequel se trouve le nucléotide altéré (figure2.14). De plus, les motifs sont majoritairement constitués d'homopolymères.

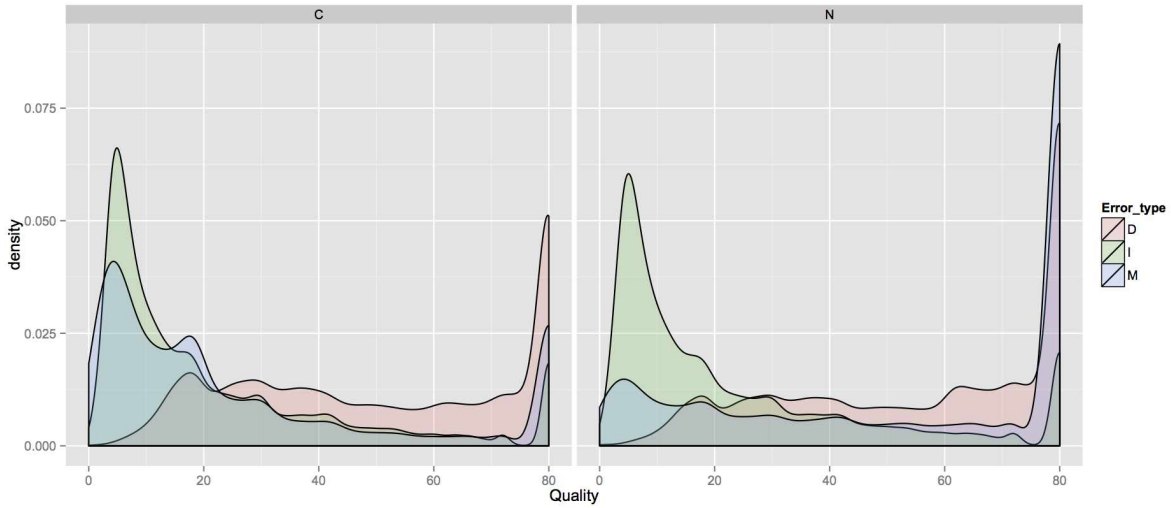


FIGURE 2.13 – *Distribution de la densité de qualité des altérations.* La distribution de qualité des altérations de type insertion (I), délétion (D) et mismatch (M) pour les *reads* issus du séquençage de l'ARNm p53 séparés par fragments (N et C). Seuls des échantillons de types sauvage ont été sélectionnés (pourcentage de colonies rouges < 10).

2.3.4 Développement d'une nouvelle méthode : MICADo

MICADo, pour *Mutation In CAncer Data*, prend en entrée les séquences de référence pour un gène d'intérêt et plusieurs ensembles de *reads* correspondant au séquençage d'un panel de gènes d'une cohorte de patients. Cette méthode permet de détecter avec précision des mutations dans des données de séquençage ciblées et de les distinguer des autres altérations dues à différents biais biologiques ou technologiques. MICADo est composé de trois étapes principales (figure 2.15).

Tout d'abord, la séquence de référence et les *reads* provenant du séquençage de la cohorte sont représentés avec un graphe de *de Bruijn* (GDB). Un GDB est un graphe orienté largement utilisé dans les méthodes de traitement de données NGS (Pevzner *et al.*, 2001; Compeau *et al.*, 2011). Cet objet représente la redondance rencontrée dans une ou plusieurs séquences. Les sommets du graphe correspondent à l'ensemble des mots de longueur k contenus dans les séquences, appelés k-mers, tandis que les arrêtes dans le graphe correspondent à l'ensemble des $k-1$ chevauchements entre les k-mers.

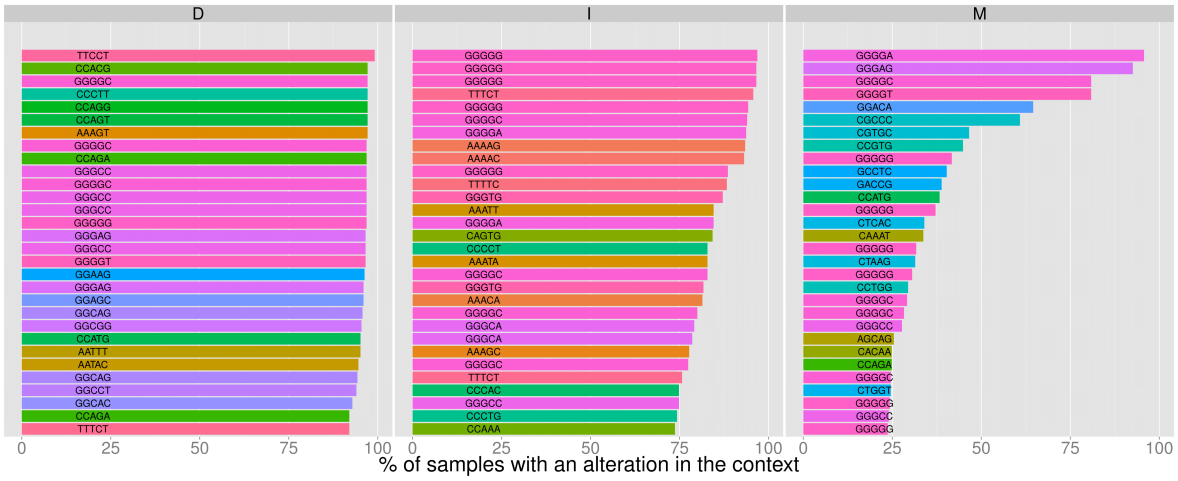


FIGURE 2.14 – *Contexte des altérations les plus fréquentes dans l'ensemble des échantillons.* Le pourcentage d'échantillons présentant l'altération et son contexte (après altération) est représenté pour les insertions (I), délétions (D) et mismatches (M) les plus fréquentes dans l'ensemble des échantillons.

Ici, nous utilisons une extension du GDB classique, le GDB *coloré* (Iqbal *et al.*, 2012). Un GDB coloré comprend plusieurs ensembles de *reads* et possède des sommets multicolores destinées à garder une trace des ensembles auxquels les k-mers appartiennent. Nous construisons un premier GDB coloré, le *graphe de référence*, pour les séquences de référence. Ces séquences de références sont composées des séquences ciblées par le séquençage et des SNPs connus. Les sommets de ce graphe sont colorés en fonction des séquences de référence dont ils proviennent. Nous construisons ensuite un second GDB coloré, le *graphe échantillon*, pour chacun des échantillons, ou ensemble de *reads*, des différents patients. Les sommets du graphe échantillon sont colorés en une seule couleur représentant chaque échantillon.

Ensuite, nous comparons le graphe de référence pour chaque graphe échantillon afin de mettre en évidence les *chemins alternatifs* correspondant aux chemins dans le graphe échantillon qui ne sont pas présents dans le graphe de référence (figure 2.15.A).

Enfin, pour chaque chemin alternatif, nous vérifions si ce dernier est spécifique à l'échantillon ou commun à la cohorte. Ceci est réalisé en générant des échantillons aléatoires construits à partir de *reads* provenant de l'ensemble de la cohorte et en vérifiant la présence de chaque chemin alternatif précédemment identifié au sein de ces échantillons aléatoires (figure 2.15.B).

2.3.5 Méthodologie

Construction des graphes de *de Bruijn*

Un GDB est un graphe orienté dont les sommets sont des mots d'une longueur donnée k, appelée k-mers, et dont les arrêtes connectent les mots se chevauchant par $k-1$ lettres. Deux définitions de graphes de *de Brujn* sont rencontrés dans la littérature en fonction de la façon dont les arrêtes sont traités. La définition implicite infère les arrêtes des sommets, c'est à dire que chaque chevauchement $k-1$ est comptabilisée pour savoir si les k-mers correspondants sont consécutifs ou non dans le reads (Uricaru *et al.*, 2014). A l'inverse, la

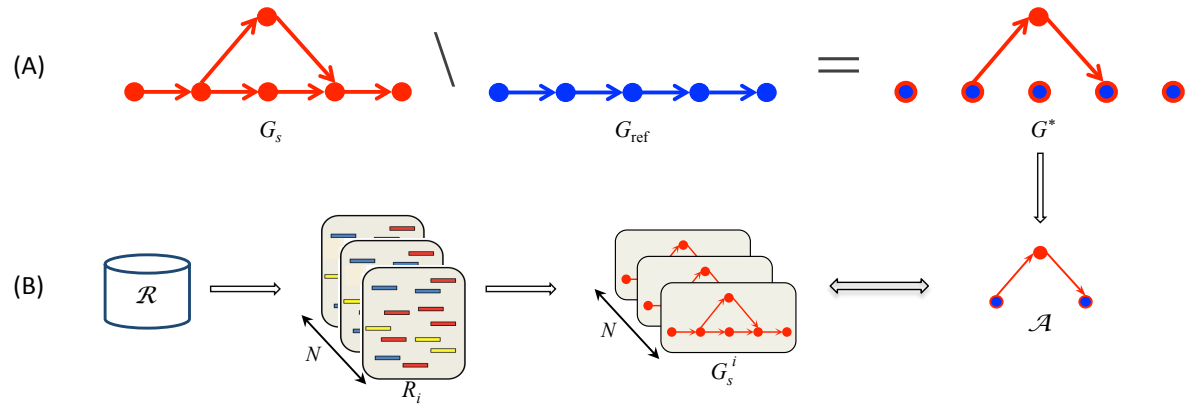


FIGURE 2.15 – *MICADo.* G_{ref} et G_s sont respectivement construits à partir de plusieurs séquences de référence et des *reads* de l'échantillon analysé. Après avoir détecté les différences entre G_s et G_{ref} en construisant G^*, l'objectif est de décider si elles sont spécifiques à l'échantillon en vérifiant leur présence dans des échantillons aléatoires R_i construits à partir de l'ensemble $\mathcal{R}$ des *reads* de la cohorte.

formulation explicite représente spécifiquement les arrêtes, en plus des sommets, auquel cas seul les chevauchements de $k-1$ lettres entre les k-mers consécutifs présents dans les *reads* vont être considérés (Iqbal *et al.*, 2012).

Pour l'approche MICADo, la formulation explicite est utilisée et il est communément admis qu'un sommet dans un GDB représente à la fois le k-mer et son reverse-complément. Ici, les *reads* reverse-complément sont reverse-complémentés avant la construction du graphe, nous représentons ainsi uniquement k-mers (et non pas leur reverse-complément). De plus, nous utilisons une version étendue du GDB, un graphe *coloré* (Iqbal *et al.*, 2012), mieux adapté au traitement de plusieurs échantillons. Un GDB coloré prend plusieurs ensemble de *reads* correspondant à plusieurs échantillons et les agrègent dans un seul GDB en colorant les sommets avec des étiquettes propres aux échantillons dans lesquels ils sont présents. Une définition plus formelle suit ci-dessous.

Définition 2.1. Pour n ensembles de séquences donnés S_1, ..., S_n et $k \geq 2$, les graphes de De Bruijn $G = \langle V, E, l \rangle$ correspondants sont définis par :
V, ensemble de sommets représentant tous les k-mers de S_1, ..., S_n
E, ensemble d'arêtes tel que :

$$E = \{(v,w) : v,w \in V \text{ et } v_2 \dots v_k = w_1 \dots w_{k-1} \text{ et } \exists\, i \in [1,n] \text{ t.q. } \exists\, s \in S_i, v_1 v_2 \dots v_k w_k \subseteq s\},$$

l, correspond à la fonction d'étiquette tel que $l : V \to \mathcal{P}(L)$ renvoie les couleurs des sommets avec L représentant les n étiquettes (couleurs) correspondant aux n ensembles de séquences.

Lorsque le contexte ne sera pas ambiguë, la notation $v = w$ sera utilisée pour l'égalité des k-mers qui sont encodés par les sommets correspondants sans associer la fonction étiquette l.

Pour une séquence d'intérêt, son *graphe de référence*, noté $G_{ref} = \langle V_{\text{ref}}, E_{\text{ref}}, l \rangle$, est un GDB n-coloré construit à partir de la décomposition en k-mers des n séquences correspondant à sa séquence génomique (ou ses variants d'épissages dans le cas d'une étude

transcriptomique) et ses SNP connus. Un *graphe échantillon*, noté $G_S = \langle V_S, E_s \rangle$, est un GDB 1-coloré, c'est à dire un GDB simple, construit à partir de l'ensemble des *reads* d'un seul échantillon. Pour le graphe échantillon, nous définissons le *support en reads* d'un sommet v, noté $r(v)$, comme le nombre de *reads* dans lesquels le k-mer correspondant apparaît. Afin d'écarter les erreurs de séquençage identifiables dès à présent, les sommets ayant un support de *reads* en dessous d'un seuil fixé sont retirés de G_S.

Recherche de chemins alternatifs

Étant donné les séquences de référence et les *reads* d'un échantillon modélisés par G_{ref} et G_S, la détection d'altérations de séquences (insertions, délétions et substitutions) revient à capturer les différences entre ces deux graphes. Cela signifie que nous devons identifier les chemins, ou suites de sommets, présents dans G_S, mais absents dans G_{ref}, nommés *chemins alternatifs*. Fondamentalement, un chemin alternatif correspond à une suite de k-mers (sommets) dans l'échantillon qui diffère de celle présente dans la séquence de référence, sauf pour les deux k-mers d'ancrage qui sont communs à la fois à l'échantillon et à la référence (figure 2.16.A).

Définition 2.2. Le *graphe de différences* $G^* = \langle V^*, E^* \rangle$ est défini tel que $V^* = V_s \cup V_{\text{ref}}$ et $E^* = E_s \setminus E_{\text{ref}}$. Un chemin $p_a = (v_1^* \dots v_n^*)$ dans G^* est appelé *chemin alternatif* s'il existe un chemin $p_r = (v_1 \dots v_m)$ dans G_{ref} tel que $v_1^* = v_1$ et $v_n^* = v_m$.

La notion de *support en reads*, défini à l'origine sur les sommets, est généralisée aux chemins dans le graphe échantillon : le *support en reads* d'un chemin p est le nombre $r(p)$ de *reads* de l'échantillon apparaissant dans la succession de k-mers composant ce chemin p. Dans ce cas, un *read* supporte un chemin si et seulement si il supporte l'ensemble des k-mers composant ce chemin. Un chemin alternatif p_a peut comporter $n \geq 1$ modifications par rapport à son chemin de référence p_r. La distance de Levenshtein $\text{lev}(p_a, p_r)$ entre les séquences définies par p_a et p_r, permet de déterminer l'ensemble minimum δ_i (avec $1 \leq i \leq n$) d'opérations d'édition (insertions, délétions et substitutions) qui transforment p_a en p_r. Cela définit l'ensemble des *chemins alternatifs atomiques* correspondant à $\{\delta_i(p_a)\}$ (figure 2.16.B). Le support en *reads* de ces chemins alternatifs atomiques est égal à $r(p_a)$.

Par la suite, l'ensemble des tuples correspondant aux modifications calculées pour un échantillon donné est noté $A = \{\langle p_a, p_r, c(p_a) \rangle\}$, où p_a est un chemin alternatif atomique dans G^*, p_r son chemin de référence correspondant dans G_{ref}, et $c(p_a)$ son *ratio de comptes* avec $c(p_a) = r(p_a)/(r(p_a) + r(p_r))$, où $r(p_r)$ est calculé dans le graphe échantillon. Par exemple, dans la figure 2.16.A, le *ratio de comptes* pour le chemin alternatif p_a est calculé tel que $c(p_a) = 8/(8+3) = 0.72$. p_a est atomique car il comporte une seule altération (la substitution `G` vs. `T`).

De plus, un chemin alternatif doit être *simple*, c'est à dire composé d'au plus deux sommets de V_{ref}. En effet, les chemins composés ne sont pas considérés. L'existence de plusieurs séquences de référence peut engendrer une explosion combinatoire des chemins de référence pour un chemin alternatif donné. Par conséquent, pour un p_a donné, un seul chemin de référence est conservé et correspond à celui qui maximise l'intersection, en termes de support de *reads*, entre le chemin alternatif et les différents chemins de référence possible.

Pour construire l'ensemble A, l'algorithme de recherche de chemin alternatif (algorithme 1) est appliqué.

Certains $v \in V_{\text{start}}$ peuvent ne pas appartenir à V_{ref}. Nous appelons ces sommets de départ ou de fin des *tips*. Ces tips peuvent être présents dans G^* pour trois raisons :

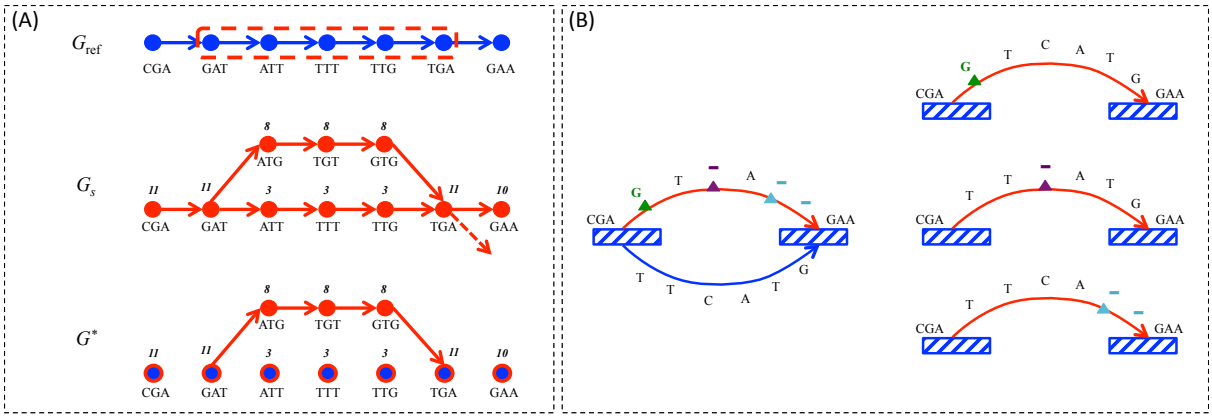


FIGURE 2.16 – *Recherche de chemins alternatifs.* A. G_{ref} encode la séquence `CGATTTGAA` pour $k = 3$. L'ensemble de séquences correspondant à cet échantillon est composé de 8 *reads* contenant la séquence `CGATGTGAA`, 2 *reads* avec `CGATTTGAA` et 1 *reads* contenant la séquence `CGATTTGA` suivis par un caractère différent de `A` ; G_s représente cet échantillon pour $k = 3$. Le graphe de différence G^* contient un chemin alternatif p_a codant la séquence `ATGTG` avec un support en *reads* $r(p_a) = 8$, ainsi que des sommets isolés (singletons) (points bleus encerclés en rouge) représentant les k-mers partagés avec G_{ref} et G_s, sans chemins alternatifs les reliant. Le chemin de référence correspondant à p_r est encerclé en rouge dans G_{ref}. B. Un chemin alternatif p_a encodant la séquence `CGAGTAGAA` est identifié avec son chemin de référence p_r encodant `CGATTCATGGAA`. Les k-mers communs à p_a et p_r sont représentés par les rectangles rayés bleu. p_a comporte 3 altérations, correspondant aux opérations d'édition suivantes : δ_1 - substitution d'un `T` par un `G` (triangle vert), δ_2 - délétion d'un `C` (triangle violet) et δ_3 - délétion d'un `TG` (triangle bleu). p_a est ainsi décomposé en trois chemins alternatifs atomiques, chacun portant une altération.

Algorithme 1 : Alternative path search

```
A = ∅
V_start = {v ∈ V* | deg+(v) > 0}                    // start vertices
V_end = {v ∈ V* | deg−(v) > 0}                      // end vertices
for each v_s in V_start do
    for each v_e in V_end do
        compute A, the set of alternative paths p_a between v_s and v_e in G*
        if A ≠ ∅ then
            retrieve the best reference path p_r between v_s and v_e in G_ref
            for each p_a ∈ A do
                compute A' = {δ_i(p_a) | p_a ∈ A} the set of atomic alternative paths
                for each p'_a ∈ A' do
                    c = r(p'_a)/(r(p'_a) + r(p_r))
                    A = A ∪ ⟨p'_a, p_r, c⟩
return A
```

1. la présence de *reads* qui ne démarrent/finissent pas au départ/fin de la zone d'intérêt,
2. la présence de *reads* qui portent des altérations au début ou à la fin de leur séquence,
3. en raison de la suppression de sommets dans G_s ayant un support en *reads* en dessous d'un seuil fixé.

Afin de permettre à l'algorithme de fonctionner dans ces cas particuliers, un chemin de référence approprié doit être défini pour les p_a débutant ou se terminant par un *tip*. Ceci est effectué par une recherche heuristique du chemin de référence le plus vraisemblable (figure 2.17) qui consiste en la recherche d'*ancres* dans G_{ref} pour chacun de ces *tips*. Les ancres sont définis comme des sommets qui appartiennent à des chemins entre le sommet de départ dans G_{ref} et les v_e correspondant (et inversement, v_s) ayant la plus petite distance de Levenshtein avec les k-mers correspondants. Si plusieurs sommets remplissent ce critère, celui définissant le p_r de longueur la plus proche de p_a est conservé.

Chemins alternatifs spécifiques

Une fois l'ensemble des chemins alternatifs atomiques A calculé, MICADo les partitionne soit en altérations spécifiques à l'échantillon (c'est à dire en mutations) soit en altérations non spécifiques récurrentes au sein de la cohorte. Pour cela, un test de permutation basé sur la construction d'échantillons aléatoires est effectué. Cette approche se justifie par les deux hypothèses suivantes :

1. Les mutations sont supposées indépendantes entre les échantillons et non-récurrents dans la cohorte. Ceci est particulièrement vrai pour les mutations perte de fonction (telles que celles conduisant à l'inactivation de fonction de p53).
2. Les erreurs de séquençage sont supposées suivre une distribution non uniforme et être récurrentes dans les échantillons (voir la sous-section 2.3.3).

Sur la base de ces hypothèses, un chemin alternatif atomique p_a peut être déterminé comme spécifique à un échantillon donné en se référant aux informations contenues dans le bruit de fond la cohorte. Plus précisément, un chemin alternatif atomique p_a est spécifique à un graphe échantillon G_s si elle est rare dans l'ensemble des *reads* de la cohorte, notée $\mathcal{R}$.

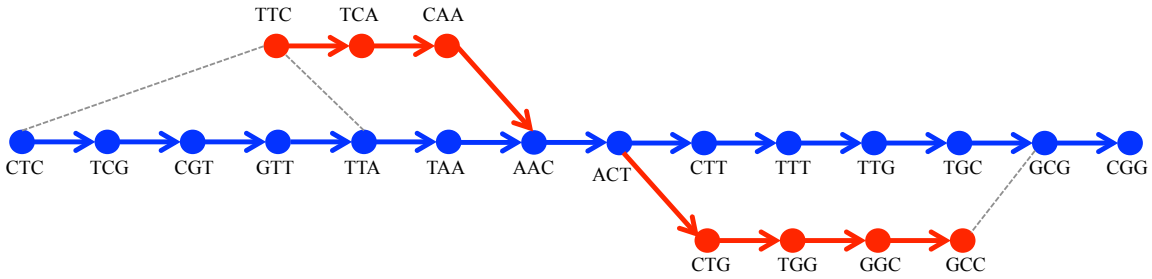


FIGURE 2.17 – *Ancrage des tips.* Considérant G_{ref} encodant la séquence `CTCGTTAACTTTGCGG` (en bleu) et G_s encodant `TTCAACTGGCC` (en rouge), il existe deux *tips* : `TTC` et `GCC`. Les deux chemins alternatifs correspondant sont p_a^1 encodant `TTCAAC` et p_1^2 encodant `ACTGGCC`. Pour `TTC` les k-mers les plus proches parmis ceux du chemin entre `CTC` et `AAC` en terme de distance de Levenshtein sont `CTC` et `TTA` (représentés par les pointillés gris). Le calcul des longueurs des deux chemins de référence possibles $|\texttt{CTC}\ldots\texttt{AAC}| = 6$ et $|\texttt{TTA}\ldots\texttt{AAC}| = 2$ permet de choisir `TTA` comme ancre. En effet, ce dernier forme le chemin ayant la longueur la plus proche de p_a^1. De même, `GCG` est choisi comme ancre pour `GCC`.

La probabilité de cette association entre p_a et G_s est mesurée en calculant le rapport de comptes $c(p_a)$ selon l'hypothèse nulle de l'absence d'association entre les *reads* et les échantillons. Pour un $c(p_a)$ observé, N échantillons aléatoires R_i sont générés par piochage sans remplacement de *reads* dans $\mathcal{R}$. Pour chacun des R_I, le ratio $c_i(p_a)$ correspondant est calculé et correspond au rapport de comptes de p_a observé dans le graphe G_s^i qui encode R_i. Cette étape est répétée afin d'obtenir une distribution de $c_i(p_a)$ et en déduire une p-valeur en comparant cette distribution et le comptes de *reads* $c(p_a)$ de l'échantillon.

Une fois les N ré-échantillonnages effectués, l'hypothèse nulle de l'absence d'association entre l'échantillon et le chemin alternatif est rejeter si α-pourcent de l'ensemble des $c_i(p_a)$ sont inférieurs à $c(p_a)$, où α est le seuil de signification requis. Dans ce cas, la $p - valeur$ associée au chemin alternatif atomique p_a est calculé telle que $p = \frac{|\{c_i(p_a) > c(p_a)\}|}{N}$. Après avoir observé que les $c_i(p_a)$ sont normalement distribués, un $z - score$ standardisé z est calculé pour $c(p_a)$. Ainsi, les altérations sont considérées comme spécifiques à l'échantillon si p est inférieur au seuil de significativité requis et z supérieur à un nombre d'écart-type de différence à la moyenne requis.

2.3.6 Évaluation sur les données PacBio

Afin de comparer les résultats de MICADo, les méthodes VarScan et GATK classiquement utilisées ont été sélectionnées.

Paramètres et références utilisées

Pour la recherche de SNVs par la méthode MICADo, l'isoforme de p53 NM_000546.5 a été utilisé comme séquence de référence ainsi que ses SNPs connus situés dans les régions ciblées : rs1042522, rs137852793, rs137852792, rs121912665 et rs1800372. SNPnexus (Chelala *et al.*, 2009) a été utilisé afin d'obtenir les positions des SNPs sur les transcrits. La recherche de SNVs s'est faite avec les paramètres suivants :

— une taille de k-mer de 18 qui correspond à la longueur minimale permettant l'obtention d'un DBG sans cycle à partir de la séquence de référence,

— 1000 permutations,

— une p-valeur inférieure à 0.01,

— un z-score supérieur à 10.

Le *mapping* préalable à GATK est le même que celui décrit dans la sous-section 2.2.3. L'utilisation de l'outil GATK (version 3.4-46-gbc02625) a été menée selon les recommandations de DePristo *et al.* (2011) et basée sur les étapes suivantes : `SplitNCigarReads`, (ii) `RealignerTargetCreator`, (iii) `IndelRealigner` and (iv) `HaplotypeCaller` (avec l'option GVCF). Seuls les mutations `NON_REF` ont été sélectionnées pour les analyses.

Les résultats de VarScan, GATK et MICADo sont comparées aux mutations identifiées précédemment dans (Iggo *et al.*, 2013). Les polymorphismes affectant la région séquencée ont été filtrés après la recherche des variants pour GATK et VarScan afin de rendre les résultats comparable à ceux de MICADo.

Résultats

Les résultats sont représentés sur la figure 2.18 et la description des mutations identifiées sont rapportées dans le tableau 2.2.

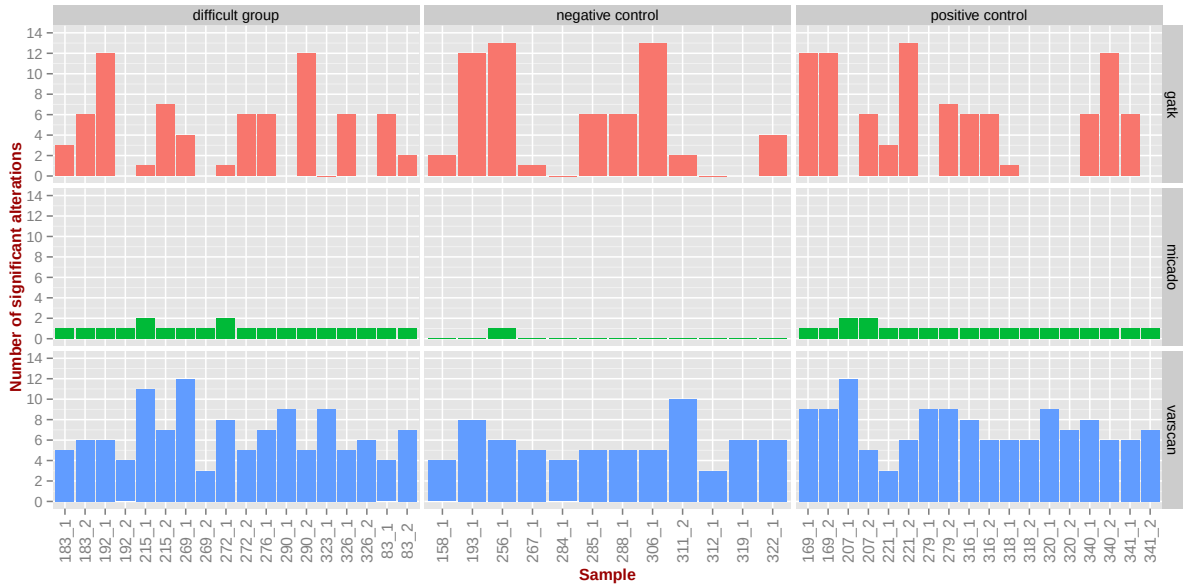


FIGURE 2.18 – *Comparaison des résultats.* Les trois outils (axe vertical) ont été testés sur les données PacBio EORTC 10994 (axe x) pour identifier des altérations dans les trois groupes (axe horizontal). Nous rapportons ici le nombre de variations identifiées (axe y) par catégorie (axe vertical) et par échantillon.

On observe dans la figure 2.18 et le tableau 2.2 que GATK et VarScan présentent un taux élevé de faux positifs, et que le nombre de mutations identifiées semble être indépendant de la catégorie des échantillons. A l'inverse, MICADo limite le taux de faux positifs et seul un échantillon de la catégorie de contrôle négatif est identifié comme étant muté. En effet, tout en étant dans le groupe des contrôles négatifs, l'échantillon 256_1

présente une mutation qui avait été précédemment identifiée dans les séquences NGS obtenues par la technologie 454 (voir la sous-section 2.2.4). Cependant, les mutations identifiées pour 83_1 et 83_2 sont des faux positifs. En effet, ces échantillons présentent une mutation différente de celle identifiée par MICADo). De même, l'échantillon 276_1 appartenant au groupe des cas difficiles, comporte deux mutations dont une seule est correctement identifiée par MICADo.

	EORTC ID	category	Exp. #	MICADo	GATK	VarScan
1	158_1	negative control	0	0/0	0/1	0/4
2	193_1	negative control	0	0/0	0/3	0/8
3	256_1	negative control	1	1/1	1/4	1/6
4	267_1	negative control	0	0/0	0/1	0/5
5	284_1	negative control	0	0/0	0/0	0/4
6	285_1	negative control	0	0/0	0/2	0/5
7	288_1	negative control	0	0/0	0/2	0/5
8	306_1	negative control	0	0/0	0/4	0/5
9	311_2	negative control	0	0/0	0/1	0/10
10	312_1	negative control	0	0/0	0/0	0/3
11	319_1	negative control	0	0/0	0/4	0/6
12	322_1	negative control	0	0/0	0/2	0/6
13	169_1	positive control	1	1/1	1/3	1/9
14	169_2	positive control	1	1/1	1/3	1/9
15	207_1	positive control	1	1/2	1/5	1/12
16	207_2	positive control	1	1/2	1/2	1/5
17	221_1	positive control	1	1/1	1/2	1/3
18	221_2	positive control	1	1/1	1/4	1/6
19	279_1	positive control	1	1/1	1/5	1/9
20	279_2	positive control	1	1/1	1/3	1/9
21	316_1	positive control	1	1/1	1/2	1/8
22	316_2	positive control	1	1/1	1/2	1/5
23	318_1	positive control	1	1/1	1/1	1/6
24	318_2	positive control	1	1/1	1/5	1/6
25	320_1	positive control	1	1/1	1/5	1/9
26	320_2	positive control	1	1/1	1/4	1/7
27	340_1	positive control	1	1/1	1/2	1/8
28	340_2	positive control	1	1/1	1/3	1/6
29	341_1	positive control	1	1/1	1/3	1/6
30	341_2	positive control	1	1/1	1/6	1/7
31	183_1	difficult group	1	1/1	0/2	1/5
32	183_2	difficult group	1	1/1	0/2	1/6
33	192_1	difficult group	1	1/1	1/3	1/6
34	192_2	difficult group	1	1/1	1/4	1/4
35	215_1	difficult group	1	1/2	1/1	1/11
36	215_2	difficult group	1	1/1	1/3	1/7
37	269_1	difficult group	1	1/1	1/2	1/12
38	269_2	difficult group	1	1/1	1/6	1/3
39	272_1	difficult group	1	1/2	1/1	1/8
40	272_2	difficult group	1	1/1	1/3	1/5
41	276_1	difficult group	2	1/2	1/2	1/7
42	290_1	difficult group	1	1/1	0/4	1/9
43	290_2	difficult group	1	1/1	0/3	1/5
44	323_1	difficult group	1	1/1	0/0	0/9
45	326_1	difficult group	1	1/1	1/2	1/5
46	326_2	difficult group	1	1/1	1/4	1/6
47	83_1	difficult group	1	0/1	0/2	0/3
48	83_2	difficult group	1	0/1	0/1	0/7

TABLE 2.2 – *Résultats de l'identification de SNVs par MICADo, GATK et VarScan pour les données PacBio.* Pour chacun des outils, la catégorie de l'échantillon *category*, son nombre de mutations attendues (exp.#), et le nombre de mutations correctes (Vrais Positifs) (c) sur le total des mutations identifiées (t) sont reportés.

Les mutations identifiées par GATK et VarScan ont été analysées afin de déterminer si certaines étaient plus fréquentes que d'autres. Le tableau 2.3 montre les 25 positions les plus fréquemment altérées, ainsi que leur contexte nucléotidique avant et après. On observe

immédiatement que les mutations identifiées par VarScan et GATK sont récurrentes. En effet, ces altérations sont essentiellement retrouvées dans un contexte d'homopolymères et correspondent dans la majeure partie des cas à la délétion d'un nucléotide de cet homopolymère bien que VarScan et GATK aient été exécutés avec un seuil stringent concernant le seuil de qualité. Cela confirme les observations mises en avant dans la sous-section 2.3.3 : le contexte des altérations est récurrent et la qualité des bases attribuée par le séquenceur n'est pas informative. Les outils de recherche de variants s'appuyant sur le score de qualité pour filtrer les faux positifs ne sont alors pas en mesure d'éviter ces erreurs.

Ainsi, ces résultats démontrent que plusieurs mutations identifiées à un niveau individuel par GATK et VarScan résultent d'altérations de séquences dues à différents biais tel que le biais de séquençage. L'algorithme MICADo quant à lui mesure la spécificité de chaque altération en la comparant au bruit de fond présent dans l'ensemble des échantillons par une méthode de ré-échantillonnage.

Pos	Before	R	A	Type	After	Count
412	AGGCTGCT	TC	T	D	CCCCCCGT	34
1099	ACGAGCTG	GC	G	D	CCCCCAGG	34
652	ATTCCACA	AC	A	D	CCCCCGCC	29
464	TGCACCAG	GC	G	D	CCCCCTCC	28
581	CACGTACT	TC	T	D	CCCCTGCC	16
729	AGGCGCTG	GC	G	D	CCCCCACC	16
1078	GCAAGAAA	AG	A	D	GGGGAGCC	13
767	TGGTCTGG	GC	G	D	CCCCTCCT	12
1147	GCTCCTCT	TC	T	D	CCCCAGCC	11
1201	AGATCCGT	TG	T	D	GGGCGTGA	11
475	CCTCCTGG	GC	G	D	CCCCTGTC	6
652	ATTCCACA	A	AC	I	CCCCCGCC	6
729	AGGCGCTG	G	GC	I	CCCCCACC	6
502	CTTCCCAG	GA	G	D	AAAACCTA	5
1126	GAGCACTG	GC	G	D	CCCAACAA	5
996	AATCTACT	TG	T	D	GGGACGGA	4

Pos	Before	R	A	Type	After	Count
452	ACCGGCGG	GC	G	D	CCCCTGCA	26
652	ATTCCACA	AC	A	D	CCCCCGCC	13
412	AGGCTGCT	TC	T	D	CCCCCCGT	11
422	CCCCGTGG	GC	G	D	CCCCTGCA	10
729	AGGCGCTG	GC	G	D	CCCCCACC	10
1099	ACGAGCTG	GC	G	D	CCCCCAGG	9
767	TGGTCTGG	GC	G	D	CCCCTCCT	7
464	TGCACCAG	GC	G	D	CCCCCTCC	5

TABLE 2.3 – *Top 25 des* hotspots *de mutations identifiées par VarScan (en haut) et GATK (en bas).* Pour chaque nucléotide (colonne *P*os, en coordonnées transcriptomiques), la colonne *C*ount présente le nombre d'altérations significatives identifiées par chaque outil, ainsi que le contexte nucléotidique avant *B*efore et après *A*fter, dans la séquence de référence *R* et alternative *A*. La colonne *T*ype représente le type des altérations : insertions (I) et délétions (D).

Corrélation entre le nombre de *reads* et le nombre de k-mers

La figure 2.19 représente la taille des graphes de *de Bruijn* d'un échantillon, mesurée par le nombre de sommets $|V_s|$. Le nombre de k-mers est ainsi corrélé avec le nombre de *reads* après la suppression des k-mers avec un support en *reads* inférieur à 3%. Cependant,

après suppression de ces sommets, la corrélation disparaît. Ainsi, cette étape de nettoyage supprime entièrement les erreurs de séquençage aléatoires et l'importance de la réduction est fonction du bruit de fond présent dans les *reads* des différents échantillons.

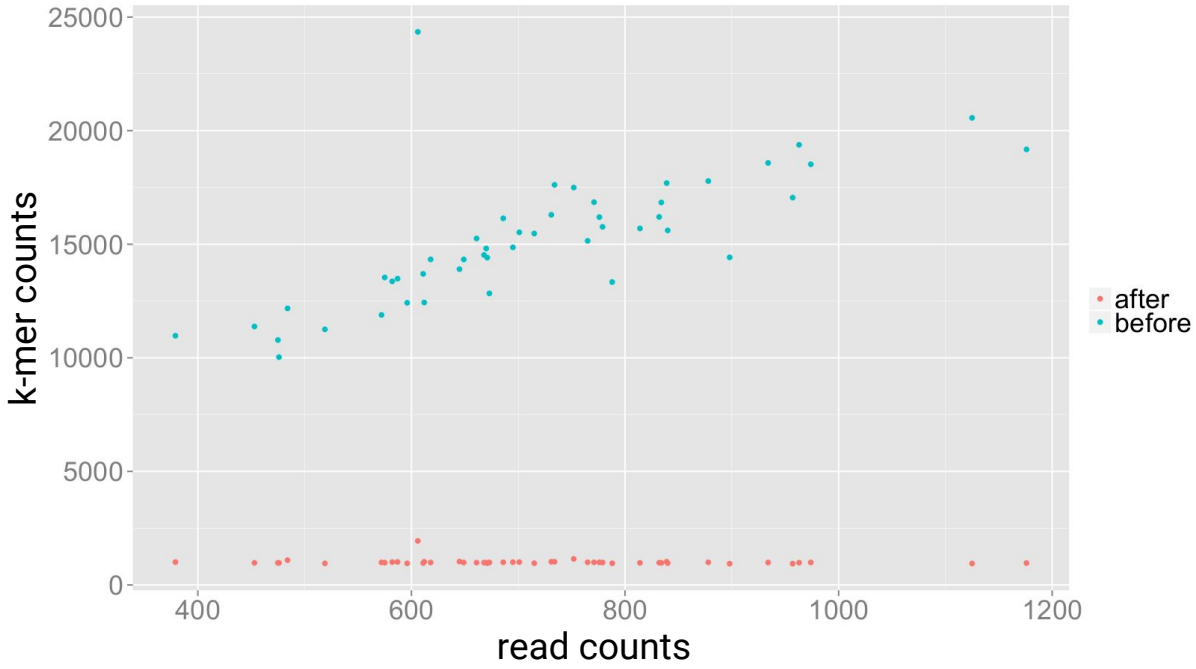


FIGURE 2.19 – *Taille des graphes de de Bruijn par échantillons.* Le nombre de k-mers par échantillon (axe y) est représenté en fonction du nombre de *reads* (axe x). Les couleurs correspondent au nombre de k-mers avant (points bleus) et après (points rouges) la suppression des k-mers dont le support en *reads* est inférieur à un seuil fixé t de 3%.

2.4 Application de MICADo

2.4.1 Séquençage PacBio du gène *FLT3*

Le récepteur tyrosine kinase FLT3 est une protéine fortement impliquée dans le développement des cellules souches et dans le système immunitaire (Gilliland et Griffin, 2002). Des mutations de FLT3 sont détectées dans environ 20% des patients atteints de leucémie myéloïde aiguë (LMA) et sont associées à un mauvais pronostic (Smith *et al.*, 2012).

Ce jeu de données est composé de 19 échantillons obtenus par séquençage PacBio des transcrits FLT3 pour 8 patients atteints d'une *AML* avant traitement et ces mêmes patients après rechute, et 3 individus sains sans antécédents de cancer. Les données sont disponibles à partir de la base de données NCBI SRA sous le numéro d'accession SRP011010 (Bioproject PRJNA85103).

Tous les patients présentent une duplication interne en tandem (ITD) qui est associée après la rechute à de nouvelles mutations ponctuelles. L'ensemble des mutations retrouvées dans chaque patient est décrit dans (Smith *et al.*, 2012). Le séquençage a été effectué sur le domaine kinase FLT3 de taille de 1346 nucléotides. La moyenne de *reads* par échantillon est de 1348 avec une variation allant de 59 à 4856).

La séquence de référence du transcrit FLT3 utilisé pour la recherche de variations correspond à son isoforme unique NM 004119 et ses 4 SNPs rs121913491, rs121913232, rs121913487 et rs147467327. Avant leurs analyses, l'ensemble des *reads* ont été orientés pour être dans le sens de la séquence de référence.

2.4.2 Résultats

Pour chaque patient, l'objectif est de vérifier l'absence de mutation dans les échantillons avant traitement et d'identifier au moins une mutation dans les échantillons après une rechute en conformité avec (Smith *et al.*, 2012). Les échantillons normaux (contrôle normal n°1, n°2 et n°3) ne présentent pas de mutation ce qui est confirmé par MICADo. La comparaison des résultats obtenus par MICADo avec ceux rapportés dans l'étude originale sont présentés dans le tableau 2.4.

Dans l'étude originale, les auteurs ont analysé manuellement une petite région limitée à seulement 4 codons en aval de la région ITD. Nous avons traité la totalité de la séquence correspondante à la région située après l'ITD. Les échantillons de avant traitement présentent tous un pourcentage de *reads* altérés inférieur à 0,43%. Nous avons par conséquent défini un nombre attendue de mutations égal à 0 (colonne exp. # dans le tableau 2.4).

Pour les échantillons après rechute, nous avons comptabilisé seulement 1 modification pour chaque *locus* modifié, prenant ainsi en compte uniquement le clone majoritaire dans le compte du nombre attendu de mutations.

Notre algorithme MICADO identifie les mutations avec une grande précision (tableau 2.4). Deux mutations non rapportées par l'étude précédente (identifiées comme faux positifs dans le tableau 2.4) et un faux négatif sont observées. Cependant, les faux positifs potentiels des échantillons avant traitement et après rechute du sujet 1005-004 sont identiques et correspondent à une substitution dans une région qui n'a pas été analysée dans l'article original. Par ailleurs, la seconde mutation dans le sujet 1011-007 (après rechute) est identifiée par MICADo, mais est filtrée par le seuil du z-score.

Pour les sujets 1009-003 (avant traitement et après rechute), 1011-007 (avant traitement et après rechute) et 1005-007 (pré-traitement) MICADo identifie de grandes insertions correspondant aux régions ITD avec $k = 30$. Pour les autres sujets, la duplication peut être identifiée avec un $k > 30$.

Subject #	Exp. pr #	MICADo pr.	Exp. pr #	MICADo rel.
1009-003	0	0/0	1	0/0
1011-006	0	0/0	1	1/1
1011-007	0	0/0	2	1/1
1005-004	0	0/1	1	1/2
1005-006	0	0/0	1	1/1
1005-007	0	0/0	1	1/1
1005-009	0	0/0	1	1/1
1005-010	0	0/0	1	1/1

TABLE 2.4 – *Résultats de MICADo pour les données FLT3.* Pour chaque échantillon, ce tableau renseigne le nombre attendu de mutations pour les prélèvements avant traitement (Exp. pr. #) et après rechute (Exp. rel. #) ainsi que le nombre de mutations correctes (Vrais Positifs) (c) sur le total des mutations identifiées (t) avant traitement et après rechute obtenues par MICADo (MICADo pr. and MICADo rel.) c/t.

Chapitre 3

Identification de *Copy Number Alterations*

Dans ce chapitre nous allons nous intéresser à la variation des altérations en terme de nombre de copies (*Copy Number Alterations* ou CNAs) à partir de données de séquençage de très faible couverture. La mise en place de cette méthodologie est motivée par une étude clinique de phase II, Horgen, comparant la réponse et la toxicité de deux traitements administrés à une cohorte de patientes.

La recherche d'altérations en terme de *copy number* repose sur la comparaison entre la couverture de séquençage du génome d'un échantillon test et d'un échantillon de référence. Cependant de nombreux biais biologiques, technologiques et bioinformatiques influencent la distribution de la couverture le long du génome séquencé c'est à dire le nombre de fois que chaque portion génomique sera observée lors du séquençage (sous-section 1.2.2 et 1.2.5). Ce phénomène de variabilité de couverture va être d'autant plus important lorsque la profondeur du séquençage est diminuée. Par ailleurs, des biais supplémentaires dus à la nature tumorale de l'échantillon tels que la contamination par le tissus sain, l'hétérogénéité intra-tumorale et la taille du génome tumoral vont s'y ajouter. Par conséquent, des méthodes de normalisation spécifiques doivent être mis en place afin de corriger ces multiples biais rencontrés lors du séquençage de faible de couverture d'un échantillon tumoral.

L'objectif de cette étude translationnelle, trans-Horgen, est de comprendre si la thérapie endocrine néo-adjuvante, administrée aux patientes atteintes de tumeurs du sein positives aux récepteurs aux œstrogènes (ERα+), induit des modifications observables en terme de CNAs. L'intérêt de la thérapie endocrine est qu'elle ne possède pas, contrairement aux chimiothérapies, d'activité clastogénique ou mutagénique conférant un cadre idéal pour explorer l'évolution des génomes tumoraux sous traitement.

Pour cela, un séquençage de très faible couverture à été réalisé avant et après traitement à partir de l'ADN de tumeurs du sein ERα+ de 20 patientes. La possibilité de multiplexage des échantillons est un avantage supplémentaire au coup faible de la génération de données à très faible couverture tout en conservant une efficacité pour l'identification des altérations de *copy number*.

La première partie de ce chapitre est consacrée à la méthode que nous avons mise en place pour la détection de CNAs à partir de ces données fortement bruitées. Dans la seconde partie, nous nous intéresserons aux analyses statistiques que nous avons appliquées aux données afin de mettre en évidence les changements intervenus durant le traitement.

Les résultats de ces travaux ont donné lieu à une publication (Quenel-Tueux *et al.*,

2015).

3.1 Contexte

3.1.1 Trans-Horgen : motivations et challenge

Les tumeurs ERα+ représentent environ 80% des cancers du sein. La thérapie administrée aux patientes atteintes de ce type de tumeurs est un traitement néo-adjuvant permettant de réduire la taille de la tumeur préalablement à une intervention chirurgicale et ainsi limiter l'étendue de l'exérèse. Dans le cas des cancers du sein ERα+, la thérapie endocrine néo-adjuvante administrée aux patientes empêche la production d'œstradiol ou interfère avec l'activation du recepteur aux oestrogènes par l'oestradiol et, dans la plupart des cas, fait régresser lentement la tumeur sur une période de plusieurs mois.

L'essai clinique de phase II, Horgen (inscrit au registre clinicaltrials.gov, numéro NCT0087858), a été mis en place afin de comparer la toxicité et la réponse à deux néo-adjuvant : l'anastrozole, qui bloque la synthèse des oestrogènes et le fulvestrant, qui bloque l'action de l'oestrogène sur son récepteur. 120 patientes ménopausées de trois centres français ont reçues aléatoirement soit 1 mg / jour d'anastrozole par voie orale durant 6 mois (61 patientes, 51%), soit 500 mg de fulvestrant par voie intramusculaire toutes les 4 semaines durant 6 mois (59 patientes, 49%). Cette étude à mis en évidence de bons taux de réponse et une faible toxicité démontrant que l'anastrozole et le fulvestrant sont des traitements hormonaux néo-adjuvants efficaces et bien tolérés chez des femmes ménopausées ayant un cancer du sein opératoire ou localement avancé.

Parallèlement, 3 biopsies ont été prélevées sur l'ensemble des patientes avant traitement puis après six mois de traitement, les tumeurs résiduelles ont été extraites chirurgicalement avec pour objectif une analyse génomique de ces tumeurs. Contrairement à la chimiothérapie, la thérapie hormonale n'est pas clastogène ou mutagène, de sorte que les changements génétiques qui se produisent pendant le traitement peuvent ne pas être considérés comme des événements secondaires provoqués par des dommages à l'ADN. Cette caractéristique est un avantage pour explorer l'évolution des génomes tumoraux induite par le traitement en permettant de considérer les changements observés comme une sélection des cellules tumorales par le traitement susceptible d'entraîner une prolifération de populations clonales résistantes.

3.1.2 Approche

L'objectif de l'approche que nous avons développée est d'identifier des variations de CNAs entre des échantillons tumoraux pairés avant et après traitement à partir de leur séquençage de très faible couverture. L'identification de CNAs consiste à déterminer le niveau de ploïdie de chaque portion génomique en comparant un échantillon test par rapport à un échantillon de référence que l'on admet diploïde. Dans le cas d'un échantillon test tumoral, l'échantillon de référence peut être soit issu du séquençage de l'ADN constitutionnel du même patient soit être construit à partir du séquençage d'un individu ou d'un pool d'individus sains.

Cependant, lors du séquençage d'un échantillon, il existe de multiples biais biologiques, technologiques et bioinformatiques (sous-section 1.2.2), qui entraînent un bruit de fond important dans les données. De plus, dans le contexte des données à très faible couverture, la variabilité aléatoire de la profondeur du séquençage constitue un biais supplémentaire

influençant la détection de CNAs. Par ailleurs, lorsque l'on séquence un échantillon tumoral, la différence de taille du génome tumoral et normal, la contamination par le tissus sain de l'échantillon tumoral et l'hétérogénéité intra-tumorale sont d'autres sources importantes de variabilité.

Afin de prendre en compte l'ensemble des caractéristiques liées à nos données, nous avons mis en place une approche permettant de réduire l'ensemble du bruit présent dans les données. Cette approche est composée d'une étape de prétraitement et de normalisation réduisant la variabilité due aux différents biais, puis d'une étape classique de segmentation permettant l'obtention de données normalisées segmentées utilisées dans l'étape finale de comparaison des paires d'échantillons avant/après traitement (figure 3.1).

L'étape de prétraitement permet d'obtenir une matrice de comptes M à partir des *reads* alignés de l'échantillon tumoral et d'un échantillon de référence. A cette étape, les valeurs aberrantes sont éliminées par un filtrage des régions présentant des valeurs extrêmes dans l'échantillon de référence. La seconde étape va permettre de calculer des ratios entre les compte de l'échantillon tumoral et de référence puis de normaliser les données. L'application successive de méthodes de normalisation va permettre de corriger :

1. la variabilité induite par la composition en GC de l'ADN induite lors du séquençage
2. la variabilité induite par l'utilisation de données très faible couverture
3. les biais liés à la nature tumorale de l'échantillon

L'ensemble de ces corrections vont alors nous permettre d'obtenir une matrice de ratios normalisés M'. A partir de la matrice M', l'étape de segmentation partitionne alors les chromosomes en segments où la valeur moyenne de ratio μ est constante. Cette matrice de ratios segmentés M'' peut-être utilisée pour effectuer des analyses statistiques telles que celles que nous avons développées dans la section 3.3.

3.1.3 Matériel biologique et séquençage

Afin d'identifier les changements survenus au cours de la thérapie endocrinienne, nous avons analysé les biopsies avant et après traitement de 20 patientes atteintes d'un cancer du sein ERα+. L'ADN des échantillons prélevés avant et après 6 mois de traitement a été extrait, préparé puis séquencé via la technologie de séquençage GAIIx (Illumina, San Diego, CA) comme décrit par Wood *et al.* (2010). Dans 5 cas, deux biopsies indépendantes ont été séquencées avant traitement et dans 4 cas 2 parties de la tumeur résiduelle après traitement ont été séquencées. Après un contrôle de la qualité du séquençage par l'outil Casava (Illumina), les *reads* ont été alignés contre le génome humain (version hg19) par l'outil BWA (version 0.5.9-r16 ; *single end mode*) en utilisant les paramètres par défaut. Les fichiers *bam* ont été rendus disponibles sur la base de données SRA du NCBI sous le numéro d'accession SRP035504 (BioProject PRJNA230247). La qualité de l'alignement a été évaluée avec qualimap (v.0.7.1.) et est résumée dans le tableau 3.1. Chaque échantillon possède entre 1 et 5 millions de *reads* correspondant à une couverture comprise entre 0.01 et 0.1.

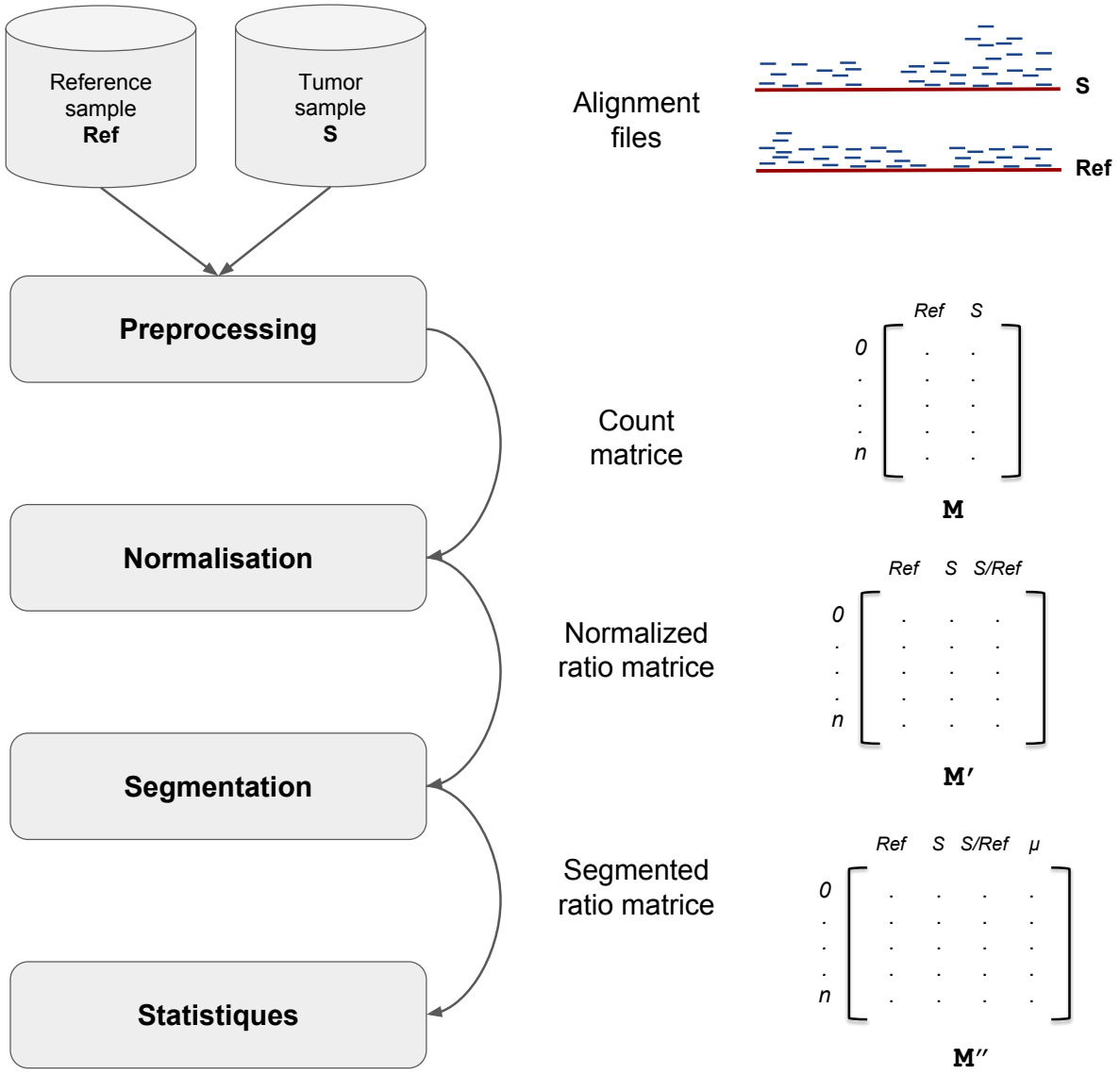


FIGURE 3.1 – *Résumé de l'approche mise en place pour la détection de changements en copy number au cours du traitement.* L'étape de pré-traitement permet d'obtenir une matrice de compte M correspondant au nombre de *reads* pour chaque intervalle génomique à partir des fichiers d'alignement de l'échantillon tumoral S et de référence Ref. L'étape de normalisation va alors transformé la matrice M en matrice M' où le ratio du nombre de compte pour S/Ref est calculé et normalisé. L'étape de segmentation permet de regrouper les intervalles en segments où la valeur moyenne de ratio μ est constante et obtenir la matrice M'' permettant d'établir des analyses statistiques.
n : nombre d'intervalles de partitionnement du génome ; S : vecteur de comptes de *reads* pour l'échantillon tumoral ; Ref : vecteur de comptes de *reads* pour l'échantillon de référence ; S/Ref : vecteur de ratios des comptes de l'échantillon tumoral par rapport à l'échantillon de référence ; μ : vecteur des moyennes de ratios segmentés.

Sample	# reads	% mapped reads	mean coverage	mean MQ	TC (histology)	TC (CNAnorm)	z (all chr)	z (1/16)	Samples compared
H01Bx1	1119422	87,88	0,02	32,26	80	57	1,7	1,5	Bx1/Ch1
H01Ch1	1268213	82,82	0,03	32,12	50	22			
H02Bx1	1254324	81,01	0,02	32,14	60	30	0,5	0,5	Bx1/Ch1
H02Ch1	1195543	81,93	0,02	32,19	25	29			
H02Bx2	2130773	93,94	0,04	31,84	60	24	1,3	0,6	Bx2/Ch2
H02Ch2	2020652	93,66	0,05	32,31	50	13			
H03Bx1	1104248	82,30	0,02	32,19	70	63			
H03Bx2	1015982	83,69	0,02	32,20	50	56	-0,3	0,6	Bx2/Ch1
H03Ch1	1251966	80,98	0,02	32,09	70	33			
H04Bx1	1310300	85,21	0,03	32,21	90	70	-0,9	-1,3	Bx1/Ch1
H04Ch1	1528554	79,35	0,03	32,12	90	71			
H04Bx2	1471926	81,04	0,03	31,96	80	81	-0,7	-0,4	Bx2/Ch2
H04Ch2	1299827	80,94	0,03	31,94	50	47			
H05Bx1	1284880	86,71	0,03	32,16	90	62	0,3	-0,3	Bx1/Ch1
H05Ch1	1821722	84,73	0,04	32,21	80	13			
H05Bx2	1254990	80,08	0,02	31,96	60	37	1,4	0,9	Bx2/Ch2
H05Ch2	1355732	78,61	0,03	31,98	70	14			
H06Bx1	1381027	87,61	0,03	32,06	70	47	0,1	-0,4	Bx1/Ch2
H06Ch1	1125906	88,33	0,02	32,18	70	52			
H06Bx2	4287306	83,94	0,09	32,23	60	55	-1,6	1,5	Bx2/Ch1
H06Ch2	1444887	88,44	0,03	31,64	40	na			
H07Bx1	852658	83,17	0,02	32,25	50	47	-0,1	-0,6	Bx1/Ch1
H07Ch1	1131168	84,94	0,02	32,26	50	50			
H08Bx1	1296674	88,05	0,03	32,23	70	38	**6,0**	**5,3**	Bx1/Ch1
H08Ch1	888927	87,19	0,02	32,18	50	60			
H09Bx1	806506	87,24	0,02	32,35	90	46	**9,4**	**7,5**	Bx1/Ch1
H09Ch1	979334	85,91	0,02	32,31	90	46			
H10Bx1	614715	89,58	0,01	32,20	80	70	**6,6**	**5,8**	Bx1/Ch1
H10Ch1	715261	90,36	0,02	32,23	60	61			
H11Bx1	1837565	88,09	0,04	32,14	70	69	-0,3	-0,4	Bx1/Ch1
H11Ch1	1664328	93,25	0,04	32,29	70	58			
H12Bx1	1441568	89,48	0,03	32,31	50	43	0,2	0,1	Bx1/Ch1
H12Ch1	749893	90,57	0,02	32,18	80	72			
H13Bx1	1836423	87,40	0,03	31,42	70	35	**6,4**	**8,5**	Bx1/Ch1
H13Ch1	1569256	85,71	0,03	31,51	60	63			
H14Bx1	2258533	88,84	0,04	31,65	80	53	**15,0**	**13,3**	Bx1/Ch1
H14Ch1	5439344	78,93	0,10	30,89	70	59			
H15Bx1	2807430	91,38	0,06	32,28	70	58	**20,8**	**27,2**	Bx1/Ch1
H15Ch1	1971001	90,44	0,04	32,14	60	68			
H16Bx1	1694192	87,65	0,03	31,66	90	42	0,9	0,8	Bx1/Ch1
H16Ch1	3603879	93,01	0,07	31,74	40	18			
H17Bx1	2092015	90,80	0,04	31,82	60	51	0,8	0,8	Bx1/Ch1
H17Ch1	1656777	87,52	0,03	31,67	90	56			
H18Bx1	2195404	93,45	0,04	31,78	60	68	0,2	0,2	Bx1/Ch1
H18Ch1	720570	85,32	0,01	32,02	40	52			
H19Bx1	1754425	82,52	0,03	32,10	80	31	**4,1**	**4,0**	Bx1/Ch1
H19Ch1	1404124	81,55	0,03	32,01	50	24			
H20Bx1	2562276	76,04	0,05	31,87	60	57	0,6	0,5	Bx1/Ch1
H20Ch1	1980348	77,21	0,04	31,96	50	60			

TABLE 3.1 – *Résumé des échantillons de l'étude trans-Horgen.* Pour chaque échantillon (*Sample*), le nombre de *reads* total (*# reads*), le pourcentage de *reads* mappés (*% mappped reads*), la couverture moyenne (*mean coverage*), la qualité de moyenne de *mapping* (*mean MQ*), le pourcentage de tissu tumoral estimé par histologie (*TC (histology)*) et par l'outil CNAnorm (*TC (CNAnorm)*) sont reportés. De même, chaque comparaison entre paire d'échantillons avant *Bx* et après *Ch* (*Samples compared*) et les z-scores associés sont reportés pour un modèle utilisant l'ensemble des chromosomes (*z (all chr)*) et utilisant les chromosomes 1 et 16 (*z (1/16)*) (voir sous-section 3.3.3.

3.2 Identification des CNAs

3.2.1 Prétraitement des données

La préparation des données est une étape qui consiste à partitionner le génome de référence et représenter chaque échantillon en vecteur de compte (figure 3.2.A). Ainsi, le nombre de *reads* est comptabilisé dans des fenêtres génomiques non-chevauchantes pour chaque échantillon tumoral indépendamment et dans échantillon de référence associé. Si le séquençage du tissu sain associé n'a pas été effectué, un échantillon de référence construit à partir d'un pool d'échantillons de femmes seines peut-être utilisé. De plus, le pourcentage en GC est calculé pour chaque fenêtre. La matrice issue du pré-traitement se constitue ainsi de 5 colonnes renseignant, le chromosome, la position de départ de la fenêtre génomique, le compte de *reads* présent dans l'échantillon tumoral, celui présent dans l'échantillon de référence et le pourcentage en GC.

Dans le but de s'affranchir des valeurs extrêmes et du biais de localisation, certaines fenêtres sont masquées pour l'ensemble des analyses (figure 3.2.B). Ainsi, nous avons exclu les valeurs des fenêtres génomiques situées à 2 Mb de chacune des extrémités des bras chromosomiques. De plus, nous avons calculé la médiane et l'écart type des valeurs de compte des fenêtres de l'ensemble des chromosomes pour l'échantillon de référence afin d'exclure les fenêtres dont les valeurs sont inférieur ou supérieur à 4 fois l'écart à la médiane. Enfin, le chromosome Y a été exclu de l'analyse, l'ensemble des patientes étant des femmes, ainsi que le chromosome mitochondrial, seul l'ADN nucléaire ayant été séquencé.

3.2.2 Normalisation et segmentation

Les méthodes CNAnorm (Gusnanto *et al.* (2012) ; version 1.8.0) et DNAcopy (Olshen *et al.* (2004) ; version 1.36.0) ont été utilisées afin d'obtenir les profils de *copy number* pour l'ensemble des échantillons (figures en annexe). CNAnorm est un algorithme permettant de normaliser les données de séquençage de très faible couverture avant d'effectuer la segmentation effectuée par DNAcopy. En effet, une étape de normalisation s'avère nécessaire lorsque la variabilité du séquençage est augmentée par la faible profondeur, la contamination par le tissus sain, les différence de taille entre le génome tumoral et de référence ou encore l'hétérogénéité intra-tumorale.

La méthode de normalisation CNAnorm procède en 4 étapes :

1. La première étape consiste à calculer le ratio entre le nombre de *reads* de l'échantillon et celui de la référence pour chaque fenêtre génomique et d'appliquer une correction basée sur le pourcentage en GC. La correction repose sur l'utilisation d'une régression locale ou LOESS pour *LOcal regrESSion* qui calcule de multiples fonctions locales sur des sous-ensembles locaux de données (Cleveland, 1979).
2. La seconde étape consiste à effectuer un lissage des données afin de diminuer la variabilité aléatoire de la profondeur du séquençage qui est d'autant plus importante lorsque le séquençage est effectué à très faible couverture. Ce lissage est effectué ici selon la méthode de Huang *et al.* (2007) où l'effet aléatoire est modélisé par une distribution de Cauchy.
3. La troisième étape consiste à normaliser la distribution des ratios. Ceci est effectué en assignant aux ratios les différents niveaux de ploïdie présents dans l'échantillon via un modèle de mélanges gaussiens. En effet, les ratios présentent une densité

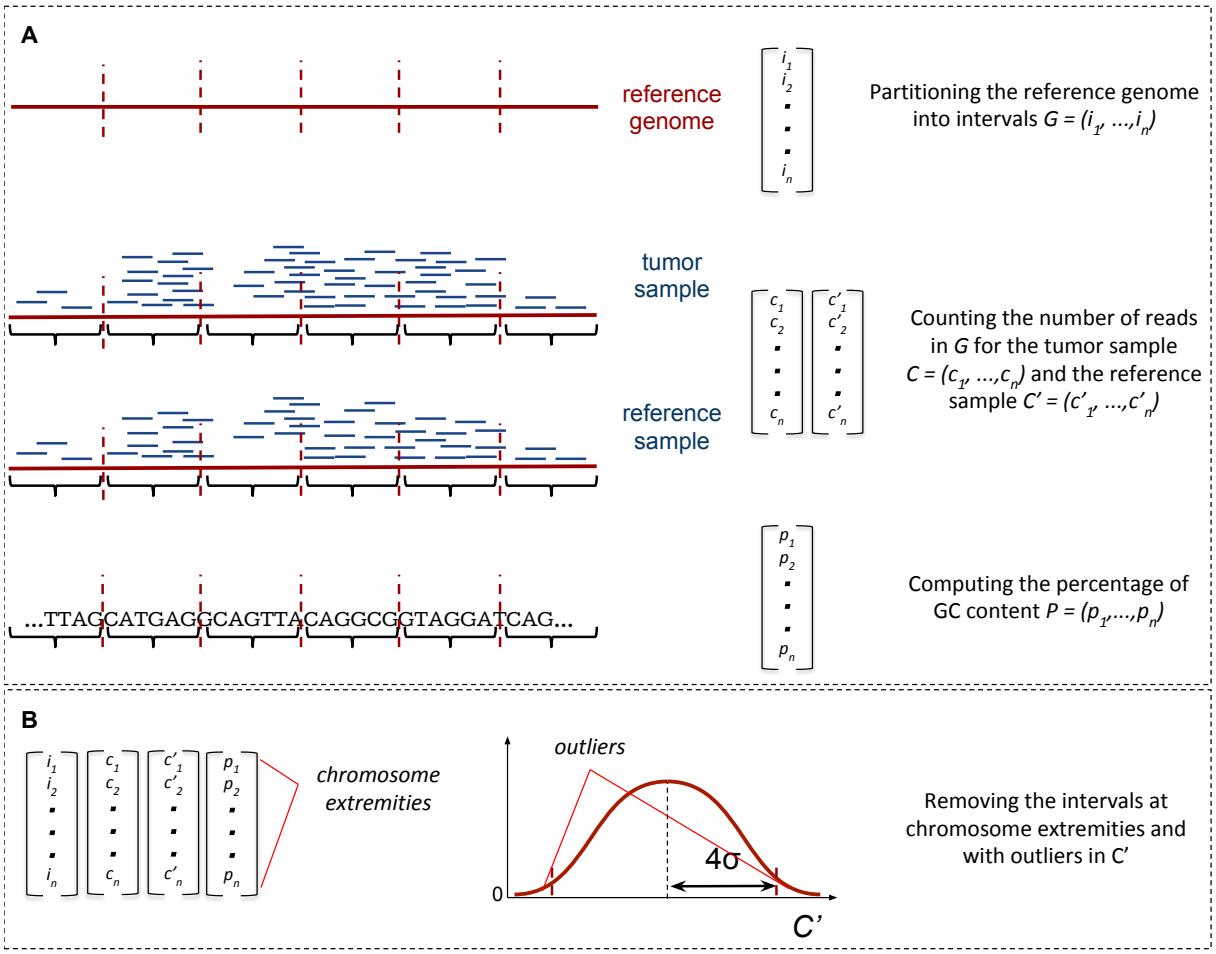


FIGURE 3.2 – *Description des étapes de pré-traitement des données.* A. Le génome de référence, ici représenté par un seul bras chromosomique, est d'abord partitionné en intervalles. Le nombre de *reads* est comptabilisé pour l'échantillon de référence et tumoral et le pourcentage en GC est calculé pour chacune des intervalles. B. Les intervalles situées aux extrémités des bras chromosomiques ou présentant des valeurs aberrantes dans l'échantillon normal sont supprimées.

normale multi-modale et les modes de la distribution correspondent aux ratios théoriques $\{0.5, 1, 1.5, 2, ...\}$ et ainsi aux différents niveaux de ploïdie $\{1, 2, 3, 4...\}$. La relation entre les modes et les niveaux de ploïdie est alors modélisée par un modèle linéaire simple dans lequel l'absence de certains niveaux est autorisée. Chaque mode observé est alors assigné à un mode théorique et la ploïdie modale, c'est à dire la ploïdie majoritaire dans l'échantillon, est assignée au pic maximisant la densité. Le coefficient de normalisation génomique δ correspondant au facteur multiplicatif entre les ratios théoriques et les ratios observés est alors calculé à partir de la ploïdie modale. Les ratios sont alors normalisés par δ. La figure 3.3 illustre la densité des ratios normalisés et ramenés à la ploïdie pour l'échantillon H11 de l'étude trans-Horgen.

4. La dernière étape consiste à corriger les ratios à partir du calcul de la contamination par le tissu sain. Il est admit que la contamination par le tissu sain induit la contraction du signal vers un ratio de 1 de manière linéaire pour l'ensemble des CNAs. Ainsi, la proportion de contamination par le tissu sain est calculé à partir des différences entre les ratios observés et les ratios attendus calculé à partir de la ploïdie correspondant au nombre de copie de 2. La proportion de la contamination par le tissu sain correspond alors à la moyenne des proportions calculées pour l'ensemble des niveaux de ploïdie identifiés dans l'étape précédente.

La segmentation des ratios normalisés est alors effectuée à l'aide de la méthode DNAcopy. DNAcopy approche le problème de la détection de CNAs comme un problème de détection de point de changement (*change-point detection problem*) qu'il résout avec une segmentation binaire circulaire (*Circular Binary Segmentation* ; CBS). L'algorithme CBS teste de manière récursive si la séquence des ratios présente un point où la fonction de distribution des données en amont est significativement différente des données en aval et permet de partitionner chaque chromosomes en segments présentant un *copy number* constant.

3.2.3 Application aux données

La préparation des données à été effectuée l'aide du script `bam2windows.pl` mis à disposition par les auteurs de CNAnorm sur le site web `www.precancer.leeds.ac.uk/software-and-datasets/cnanorm/`.

Pour chaque échantillon tumoral indépendamment, ce script prend en entrée le fichier *bam* correspondant, ainsi qu'un échantillon de référence associé. Dans notre cas, l'ADN constitutionnel n'ayant pas été séquencé, l'échantillon de référence est un échantillon construit à partir d'un pool d'échantillons de femmes saines mis à disposition par les auteurs de CNAnorm. De même, un fichier `gc1000Base.txt.gz`, mis à disposition sur le site web des auteurs, contient le contenu en GC tous les 1000 bp le long du génome humain (GRCh37/hg19) et permet de calculer le pourcentage en GC pour chaque fenêtre génomique. Nous avons alors effectuer le filtrage des fenêtres génomiques comme décrit dans la sous-section 3.2.1 pour l'ensemble des échantillons.

Après pré-traitement, l'ensemble des données est analysée à partir de deux jeux de paramètres d'entrée différents pour CNAnorm. Le premier jeu de paramètres est le même appliqués pour tout les échantillons avec une taille de fenêtre de 200 Kb et les paramètres par défaut. Le choix de la taille de fenêtres fixée permet d'obtenir un nombre de *reads* moyen par fenêtre compris entre 30 et 180 comme suggéré par les auteurs. Cette standardisation permet d'effectuer l'ensemble des analyses de la section 3.3. Le second jeux de

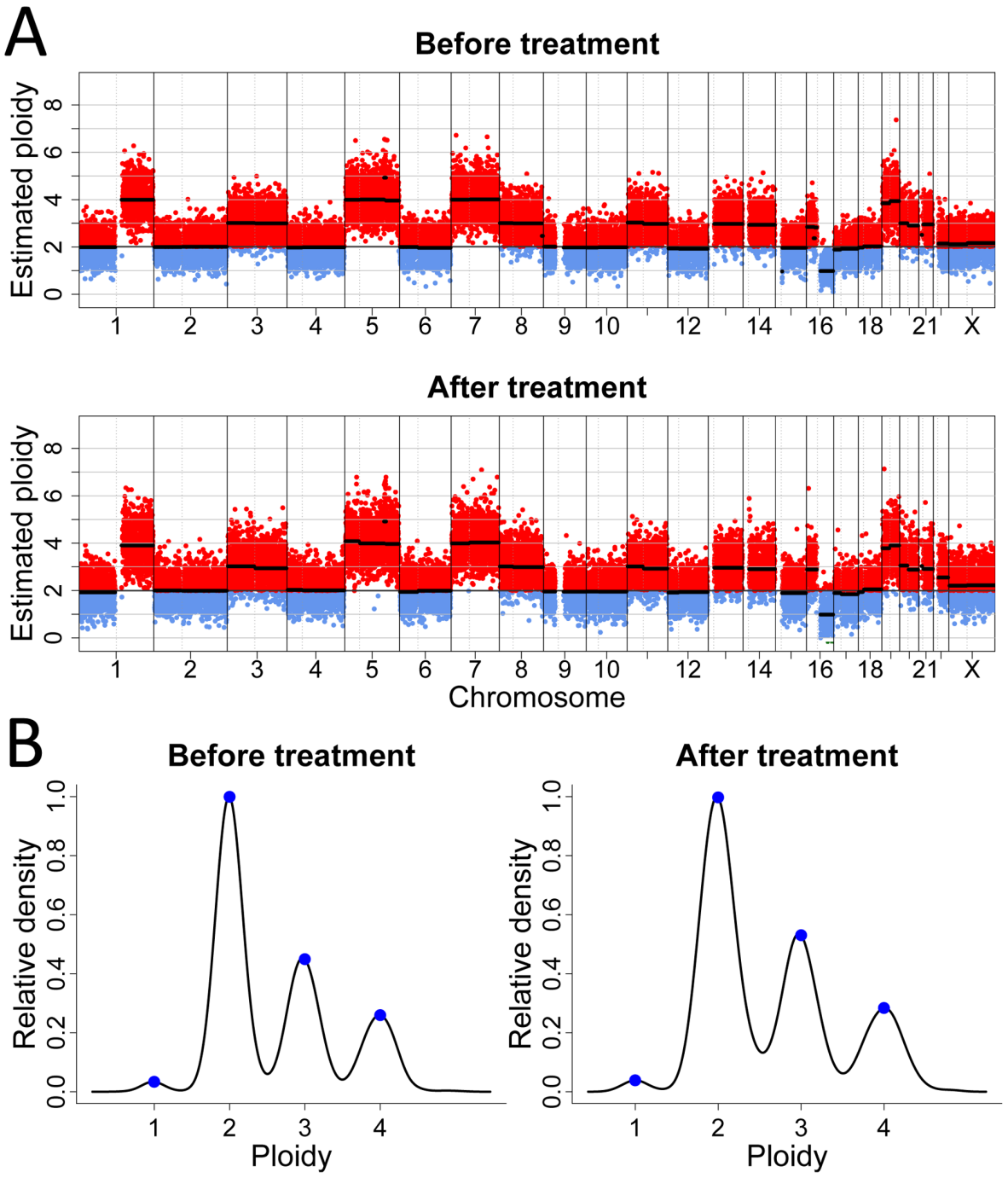


FIGURE 3.3 – *Profils de copy number avant et après traitement pour la tumeur H11.* Les profils de la tumeur H11 illustrent comment CNAnorm assigne la ploïdie. A : profils de *copy number*. Le profil H11 ne change pas après le traitement. B : Distribution de densité. La distribution de densité des ratios de l'ensemble des fenêtres génomiques est constitué de quatre pics correspondant à 1, 2, 3 et 4 copies correspondant aux niveaux de ploïdie.

paramètres est propre à chaque échantillon (tableau annexe 6) et permet de générer des profils de *copy number* observables et comparables visuellement (annexe 6.1).

3.3 Enrichissement statistiques et biologiques des résultats

3.3.1 Clustering des profils

Méthode

Le *clustering* des échantillons tumoraux permet de visualiser les différences entre les échantillons de notre étude. Afin de procéder au *clustering*, les ratios normalisés lissés par CNAnorm pour l'ensemble des fenêtres génomiques sont centrés sur la médiane puis clusterisés.

Le centrage sur la médiane permet de rendre les valeurs indépendantes de l'échantillon de référence utilisé lors du calcul des ratios en ajustant la valeur de chaque ratio de sorte qu'elles reflètent leur variation par rapport à l'ensemble des valeurs de l'échantillon. Le choix de l'utilisation de la médiane plutôt que la moyenne a été fait car l'opération de normalisation par la médiane est plus robuste vis à vis des valeurs extrêmes. La méthode de classification choisie est une classification ascendante hiérarchique. Cet méthode commence par un nombre de classes égal à celui du nombre d'individus n puis les rassemble en $n-1, ..., 1$ classes en fusionnant une à une les classes qui ont la plus grande similarité. La mesure de similarité choisie est la corrélation de Pearson et l'algorithme de *clustering* choisie est le *Centroid Linkage Clustering*. La corrélation de Pearson entre deux vecteurs de données $x = \{x_1, ..., x_n\}$ et $y = \{y_1, .., y_n\}$ est définie tel que :

$$r = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sigma_x} \frac{(y_i - \bar{y})}{\sigma_y}$$

Le *Centroid Linkage Clustering* va, pour chaque nouvelle classe, calculer un nouveau vecteur qui correspond à la moyenne des vecteurs de tous les éléments de la classe. L'algorithme s'arrête lorsqu'il reste une seule dernière classe.

Résultats

Le *clustering* des données est réalisé avec l'outil Cluster (de Hoon *et al.* (2004) ; version 3.0) qui implémente la méthode de classification présentée ci-dessus et visualisé via l'outil Treeview (Page (2002) ; version 1.1.6r4) (figure 3.4).

Sur la figure 3.4, nous observons qu'à l'exception des tumeurs H06 et H13, l'ensemble des échantillons issus de la même tumeur avant et après traitement sont regroupés.

L'absence de cellules tumorales dans l'échantillon H06Ch2 peut être une explication à son profil complètement plat engendrant un échec de clusterisation avec H06CH1. L'échantillon de H13Ch1 quant à lui n'est pas groupé avec H13Bx1 puisqu'il semble que le profil soit sensiblement plus plat après traitement.

Comme on pouvait s'y attendre pour le cancer du sein ERα+, de nombreuses tumeurs présentent des profils peu réarrangés, ou appelés «simplex» (échantillons de la branche principale du dendrogramme à gauche). Les gains de 1q / 16p et la perte de 16q sont les anomalies génétiques les plus fréquentes dans le cancer du sein. Les profils de l'ensemble

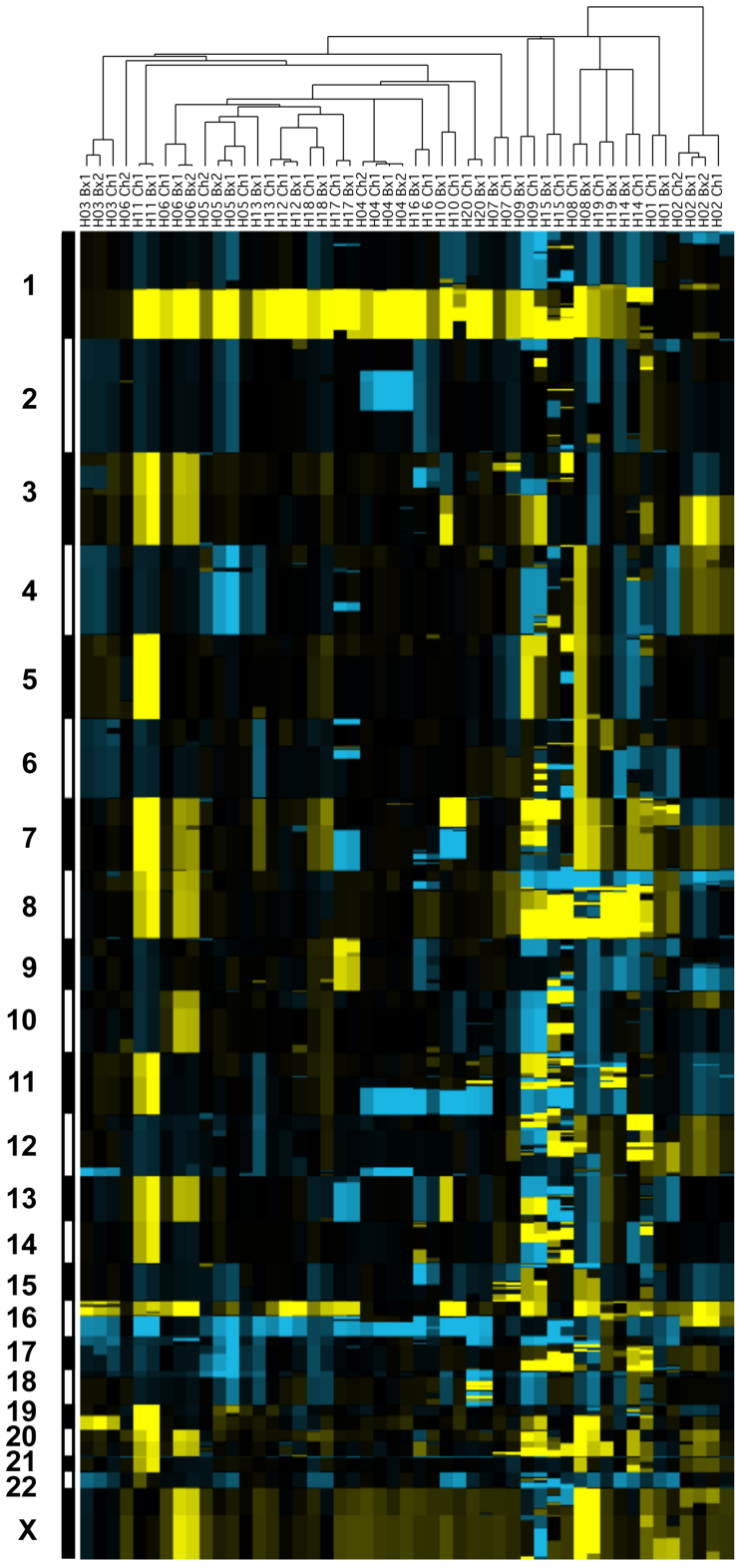


FIGURE 3.4 – *Clustering hiérarchique des 20 échantillons (H01-H20) avant (Bx) et après (Ch) traitement.* L'ensemble des tumeurs, à l'exception de H06CH2 présentent une altération sur le chromosome 1 et/ou 6. Le jaune représente un nombre de copies supérieur à 2 et le bleu inférieur à 2. Une branche a été modifiée afin de permettre la comparaison entre les échantillons H13Bx1 et H13Ch1.

des tumeurs présentent une ou deux de ces anomalies. Les chromosomes réarrangés proviennent typiquement de la recombinaison au niveau des centromères des chromosomes 1 et 16, première étape connue dans la carcinogenèse mammaire. La précocité de ce phénomène implique que le remaniement est partagé par l'ensemble des sous-clones tumoraux.

De plus, nous pouvons observer que les anomalies de ces chromosomes présents avant traitement le sont également après traitement dans l'ensemble des échantillons (sauf H06Ch2).

Par ailleurs, le *clustering* montre que les différences entre les tumeurs de différentes patientes sont en général beaucoup plus grandes que les différences avant et après traitement de la même patiente.

3.3.2 Différences locales avant/après traitement

Dans cette sous-section, nous allons nous intéresser à une analyse détaillée du cas H09 de l'étude trans-Horgen. Cette tumeur présente des changements locaux dans le profil de *copy number* après traitement.

Méthode

D'après l'observation des profils de nombre de copies, certaines tumeurs semblent contenir des amplicons, la plupart étant généralement présents avant et après traitement. Afin de mettre en avant le potentiel oncogène de ces amplicons et observer leur évolution au cours du traitement, nous leur avons associé les gènes qu'ils contiennent et regardé leur *copy number* avant et après traitement. Pour cela, un enrichissements des données segmentées est établi à partir du fichier de référence `refseq.txt` (NCBI). Chaque gène est ainsi assigné à l'ensemble des segments via l'utilisation de leurs coordonnées génomiques sur le génome de référence hg19.

Résultats

Nous nous sommes particulièrement intéressés à certains gènes d'intérêt dans le cas du cancer du sein présents dans ces amplicons : ESR1, ATG5, MSH2, PPM1D, PAK1 et NCOA3. La résistance à la thérapie hormonale peut résulter de mécanismes multiples. Le mécanisme sous-jacent le plus évident est l'amplification du gène du récepteur aux œstrogènes (ESR1) permettant de conférer une activation de la transcription indépendante de l'estradiol (Weis *et al.*, 1996). Ce phénomène a été identifié dans les cancers métastatiques du sein traités par la thérapie hormonale (Robinson *et al.*, 2013). Par ailleurs, de nombreux gènes *driver* classiques sont connus pour interagir avec ERα. Par exemple, l'augmentation de l'expression de coactivateurs du récepteur aux œstrogènes, tel que NCOA3 (Lavinsky *et al.*, 1998) et MSH2 (Wada-Hiraike *et al.*, 2005), est un mécanisme supplémentaire de résistance à la thérapie hormonale. De même, le gène PAK1 joue un rôle dans l'activation transcriptionnelle du gène ESR1 (Oladimeji *et al.*, 2016) tandis que le gène PPM1D, en plus d'être un régulateur négatif de la voie p53, stimule l'activité des récepteurs hormonaux dont ERα (Proia *et al.*, 2006). Enfin, une augmentation de l'activité autophagique, notamment par l'augmentation de l'expression du gène ATG5, semble être un événement précurseur dans la formation de cancers du sein et de la prostate (Kim *et al.*, 2011).

La figure 3.5.A représente les profils de *copy number* pour la tumeur H09. Une nette apparition d'amplicons sur les chromosomes 1, 2 et 6 est observable après traitement. Les gènes potentiellement responsables de l'initiation et/ou de la progression tumorale sont

mis en évidence par des flèches dans six des amplicons de la tumeur H09 et les rapports entre les *copy number* avant et après traitement est montré (figure 3.5.B points bleus pour H09, et noirs les autres tumeurs). Ces rapports entre les *copy number* avant et après traitement confirment l'observation à partir des profils : de nouveaux amplicons contenant ESR1, ATG5 et MSH2 sont apparus après traitement (figure 3.5.A). La droite représente la régression linéaire de l'ensemble des valeurs après traitement en fonction des valeurs avant traitement pour l'ensemble des tumeurs. La régression est réalisée indépendamment pour chaque segment génomique observé. La technologie *Fluorescent in situ hybridisation* (FISH) est utilisée pour confirmer la présence de l'amplification du gène ESR1 (figures 3.5 C et D). D'après les images, ce gène est fortement amplifié après traitement, mais, de façon inattendue, il est également amplifié avant le traitement. Ces résultats indiquent que l'apparition de l'amplicon ESR1 dans le profil de nombre de copies après traitement reflète une variation clonale plutôt qu'une amplification *de novo*. Cette variation peut être due à une sélection clonale par le traitement ou encore à l'échantillonnage de différentes régions au sein d'une tumeur hétérogène.

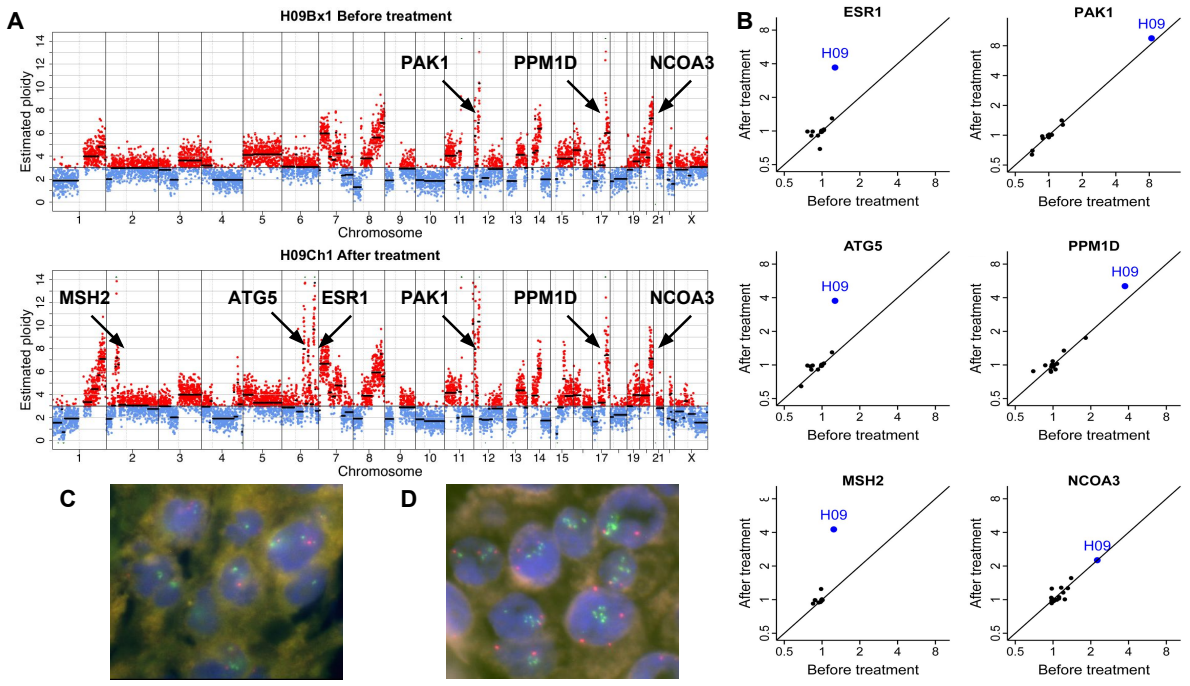


FIGURE 3.5 – *Analyse génomique avant et après traitement de la tumeur H09.* A. Profils génomiques montrant les gènes en A marqués par des flèches. B. Ratios des nombre de copies avant vs après traitement des gènes présents dans les régions amplifiées (H09, point bleu, autres tumeurs, points noirs). Le nombre de copies d'ESR1, ATG5 et MSH2 augmente après traitement, tandis que celui de PPM1D, PAK1 et NCOA3 reste inchangé. C et D. FISH pour ESR1 avant traitement (C) et après traitement (D). La sonde rouge est spécifique au centromère du chromosome 6, la sonde verte au gène ESR1.

3.3.3 Différences globales avant/après traitement

Dans cette sous-section, nous allons nous intéresser aux changements globaux de *copy number* intervenus durant le traitement pour l'ensemble des tumeurs de l'étude trans-

Horgen.

Méthode

Si l'on prend l'exemple de la tumeur H11 (figure 3.3), il n'y a pas de différence entre les profils avant et après traitement. CNAnorm trouve les pics correspondants aux principales valeurs de ploïdie (points bleus dans la représentation graphique des densités). Cependant, l'espacement des pics est différent dans les deux courbes de densité. En effet, certains paramètres intrinsèques à l'échantillon, tels que le pourcentage de contamination par le tissus sain, peuvent entraîner un phénomène de compaction du signal et induire des différences de niveaux de ploïdies plus ou moins importantes. Il est ainsi nécessaire de ramener les données à la même échelle afin de s'assurer que l'espacement des pics soit le même avant et après traitement. Cet ajustement permet alors d'estimer la corrélation entre les différents pics des échantillons avant et après traitement et ainsi la comparaison des deux profils.

Nous avons mis en place une approche, résumée dans la figure 3.6 permettant de projeter les valeurs segmentées pour l'ensemble des fenêtres génomiques n d'un échantillon i à l'échelle de mesure d'un échantillon j et ainsi ramener les échantillons dans deux gammes de valeurs comparables.

Pour cela, nous avons construit un modèle linéaire simple afin d'exprimer les valeurs segmentées X_j d'un échantillon j en fonction des valeurs segmentées X_i d'un échantillon j tel que :

$$X_j = \beta_0 + \beta_1 X_i + \epsilon_i$$

où ϵ est le terme d'erreur aléatoire du modèle et β_0 et β_1 sont deux paramètres à estimer. Ici l'estimation des paramètres est effectuée par la méthode des moindres carrés. Ainsi, les paramètres β_0 et β_1 choisis minimisent la somme des carrés des résidus tels que :

$$\text{Argmin}_{\beta_0,\beta_1} \sum_{i=1}^{n} (x_j - \beta_1 x_i - \beta_0)^2$$

où n est le nombre de fenêtres génomiques.

Les valeurs attendues X_e sont alors obtenues par l'application du modèle aux données X_i. Les valeurs attendues X_e sont ensuite soustraites aux valeurs observées X_j afin de générer des différences brutes en nombre de copies. Le score de différences $S_{i,j}$, entre l'échantillon i et l'échantillon j, est l'écart-type (sd) de ces différences tel que :

$$S_{i,j} = sd(X_j - X_e).$$

Afin d'éviter la distorsion des résultats par les valeurs extrêmes, les segments de taille inférieure à 20 Mb sont supprimés avant le calcul du score de différences. La variabilité intra-échantillon $V = \{V_i, V_j\}$ où V_i est la variabilité observée dans l'échantillon i et V_j dans l'échantillon j, est mesurée par la médiane des écarts entre les ratios normalisés $R = \{R_i, R_j\}$ et la valeur des segments associés $X = \{X_i, X_j\}$ tel que :

$$V = median(R - X).$$

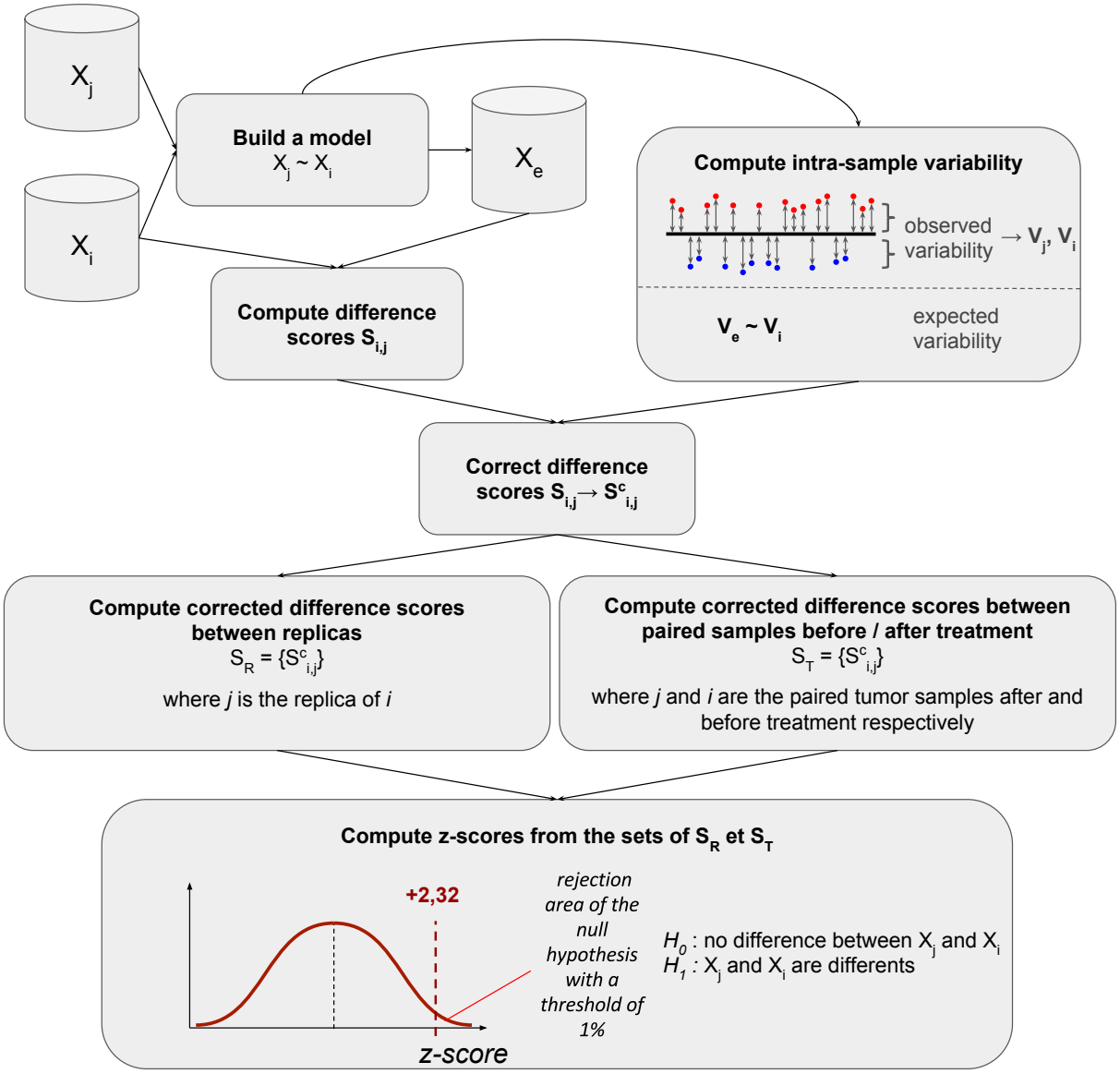


FIGURE 3.6 – *Résumé de l'approche pour la détection de changements de copy number avant / après traitements.* X_i : ratios segmentés observés de l'échantillon i ; X_j : ratios segmentés observés de l'échantillon j ; X_e : ratios segmentés attendus de l'échantillon i ; V_i :variabilité observée de l'échantillon i ; V_j : variabilité observée de l'échantillon j ; V_e : variabilité attendue de l'échantillon i ; Si, j : score de différences ; $S^c_{i,j}$: score de différences corrigés ; S_R score de différences corrigés entre réplicas ; S_T score de différences corrigés entre paire d'échantillons avant / après traitement

La variabilité intra-échantillon des données attendues après traitement V_e est obtenu en multipliant le bruit intra-échantillon de l'échantillon i par le coefficient de régression du modèle linéaire β_1 tel que :

$$V_e = V_i * \beta_1.$$

Plus le bruit est important, plus le score de différence peut-être considéré comme le résultat de la variabilité intra-échantillon plutôt qu'une réelle différence. Le maximum du bruit intra-échantillon observé V_j et attendu V_e sont alors utilisés afin d'obtenir un score de différences corrigé $S^c_{i,j}$ tel que :

$$S^c_{i,j} = \frac{S_{i,j}}{max(V_j, V_e)}$$

Deux ensembles de scores corrigés sont alors calculés :

1. $S_R = \{S^c_{i,j}\}$ avec i, j tel que j est un réplica de i
2. $S_T = \{S^c_{i,j}\}$ avec i, j tel que j correspond à l'échantillon tumoral après traitement de l'échantillon avant traitement i

Soit m_R la moyenne et σ_R l'écart-type des scores de l'ensemble S_R, un z-score z est alors calculé pour chaque paire d'échantillon de l'ensemble S_T tel que :

$$z = \frac{S^c_{i,j} - m_R}{\sigma_R}, S^c_{i,j} \in S^T.$$

Ce z-score correspond au multiple de l'écart-type des différences entre les réplicas et permet d'estimer la variabilité des différences sous l'hypothèse nulle selon laquelle deux tumeurs ont le même profil.

Le seuil de significativité p fixé à 1% et le test est unilatéral. Ainsi, d'après la table des z-scores, les z-scores supérieurs à 2.32 sont ainsi considérés comme significatifs.

Résultats

Le tableau 3.1 montre les résultats obtenus pour un modèle linéaire construit à partir de l'ensemble des chromosomes (colonne *z (all chr)*) et pour un modèle utilisant seulement les chromosomes 1 et 16 comme chromosomes de référence invariants (colonne *z (1/16)*) pour chaque paire d'échantillons avant / après traitement comparés (colonne *Samples compared*). L'utilisation des chromosomes 1 et 16 comme chromosomes de référence est justifiée par le fait que les changements dans ces chromosomes sont susceptibles d'être présents dans tous les sous-clones. En effet, la translocation der (1 ; 16) est généralement le premier événement oncogène à se produire dans la transformation des cellules tumorales des tumeurs ERα+.

Les tumeurs significativement différentes le sont avec les deux modèles. Sept paires de tumeurs se sont révélées significativement différentes (H08, H09, H10, H13, H14, H15 et H19 ; tableau 3.1). Le changement le plus courant après traitement correspond à une disparition d'anomalies. En effet, dans quatre des sept tumeurs, le profil s'est simplifié par la disparition des gains et des pertes de chromosomes entiers ou de bras chromosomiques (figure 3.7) tandis que les altérations sur les autres chromosomes restent. Nous pouvons exclure que ces changements résultent de la dilution du signal par le tissu normal car le

gain de 1q est encore clairement observable dans toutes les tumeurs affectées. La tumeur H13 présente la disparition d'un gain du chromosome 7 et des pertes des chromosomes 4, 6, 11, 12, 17 et 18, ne conservant que les altérations sur les chromosomes 1 et 16. La tumeur H08 présente une simplification des profils pour les chromosomes 2, 7 et 15 tandis que l'on observe la persistance d'autres réarrangements sur le profil global. La tumeur H10, les altérations sur les chromosomes 3, 13 et 18 disparaissent et la perte des chromosomes 10 et 15 apparaît. Enfin, la tumeur H19 présente une disparition des altérations sur les chromosomes 3 et 10 ainsi qu'un gain du bras 20q.

En résumé, les profils génomiques montrent des différences après traitement dans un tiers des cas. La simplification des profils observée dans 4 des cas peut être une conséquence de l'échantillonnage de différents sites au sein d'une tumeur composée de plusieurs populations clonales. Une analyse d'échantillons prélevés à partir de plusieurs sites dans l'échantillon chirurgical serait un moyen de répondre à cette interrogation. Une autre explication pourrait être que les différences reflètent la sélection clonale induite par le traitement avec élimination des clones réarrangés.

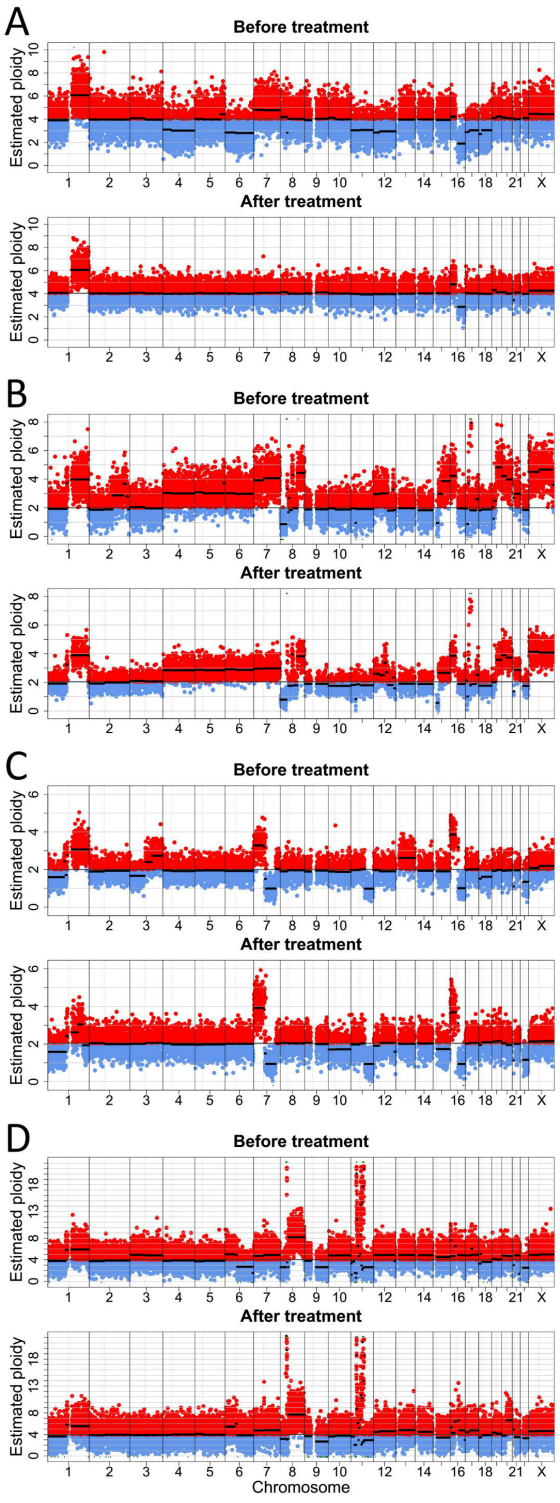


FIGURE 3.7 – *Profils copy number avant et après traitement.* A : Tumeur H13. B : Tumeur H08. C : Tumeur H10. D : Tumeur H19. La disparition de gains et de pertes de chromosomes entiers ou de bras chromosomiques et/ou l'apparition de nouvelles altérations sont observables après le traitement.

Chapitre 4

Hétérogénéité intra-tumorale

Dans ce chapitre, nous allons nous intéresser au développement d'une méthode d'analyse de l'hétérogénéité intra-tumorale à partir d'altérations somatiques de type SNAs (*Single Nucleotide Alteration*) et CNAs (*Copy Number Alteration*) identifiées via le séquençage haut débit d'échantillons tumoraux.

L'hétérogénéité intra-tumorale correspond à la complexité liée au mélange de différentes cellules composant l'échantillon tumoral. En effet, lorsque l'on séquence une tumeur, le matériel biologique utilisé est composé de cellules d'origines diverses : le tissu normal provenant du micro-environnement, les cellules immunitaires infiltrantes et les cellules tumorales. Cependant, la contamination de l'échantillon tumoral par le tissu sain n'est pas la seule source d'hétérogénéité. En effet, les cellules tumorales ne sont pas toutes identiques, elles forment plusieurs sous-populations. Il est admis que ces différentes sous-populations sont toutes issues d'une seule et même sous-population ancestrale qui peut être présente ou absente de l'échantillon tumoral (Nowell, 1976). En effet, lors de l'évolution tumorale, les cellules se divisent et acquièrent de nouvelles altérations et caractéristiques sous la pression de sélection de leur environnement comparable à un processus Darwinien.

L'intérêt de caractériser les différentes sous-populations cellulaires est motivé par le fait que ces dernières possèdent des caractéristiques individuelles propres influençant la réponse au traitement ou encore la formation de métastases. Ainsi se pose le *problème de reconstruction clonale* qui consiste en l'identification et la caractérisation des différentes populations cellulaires tumorales, aussi appelés *clones*, composant l'échantillon tumoral à partir des *reads* issus du séquençage de ce dernier. La distributions des *reads* et les altérations qu'ils comportent permet en effet d'identifier des altérations spécifiques aux clones, restreintes dans ce travail aux CNAs et SNAs somatiques, et ainsi reconstruire la composition de l'échantillon tumoral. De nombreuses méthodes ont été développées ces dernières années afin de répondre à ce problème de reconstruction bien que ces dernières restent encore insatisfaisantes à ce jour (voir la sous-section 1.2.6 de l'introduction).

C'est dans ce contexte que l'*ICGC-TCGA DREAM Somatic Mutation Calling - Tumour Heterogeneity Challenge* (SMC-Het) a été mis en place. L'objectif de ce challenge est d'améliorer les méthodes standards de reconstruction clonale. Idéalement, une méthode de reconstruction clonale devrait permettre de quantifier et assigner chaque altération aux différentes populations tumorales présentes dans un échantillon tumoral.

Dans la première partie de ce chapitre nous allons définir les différentes règles de calcul régies par les altérations de *copy number*, l'hétérogénéité intra-tumorale et la contamination par le tissu sain, lors du séquençage d'un échantillon tumoral. La seconde partie est consacrée à la méthode que nous avons développé afin de résoudre le problème de

reconstruction clonale et ainsi reconstituer la composition de l'échantillon tumoral.

4.1 Définitions

L'analyse de séquences d'un échantillon tumoral a pour objectif sa caractérisation par les altérations génétiques qu'il comporte. En effet, comme présenté dans l'introduction, le processus de tumorigenisation résulte de l'accumulation d'altérations génétiques (sous-section 1.2.6 de l'introduction). Dans ce travail, nous allons seulement considérer les variations somatiques et aborder le problème de reconstruction clonale à partir de deux types d'altération, les SNAs et les CNAs.

En effet, les variations identifiées dans un échantillon tumoral peuvent être de nature somatique ou germinale. Afin de prendre en compte seulement les altérations somatiques, l'une des solutions est d'effectuer le séquençage du tissu sain en parallèle de celui de l'échantillon tumoral.

Comme décrit dans l'introduction (sous-section 1.2.4), pour un génome humain normal diploïde, chaque locus génomique présente deux allèles parentaux différents, l'allèle correspondant au génome de référence est dénoté A et sa forme alternative B. Ainsi, pour un génome diploïde, les génotypes G possibles à un locus donné sont homozygote AA, hétérozygote AB ou homozygote BB, soit $G = \{AA, AB, BB\}$.

Dans ce travail, nous prenons en compte seulement les locus homozygotes AA dans l'échantillon sain devenus hétérozygotes AB dans l'échantillon tumoral sous l'hypothèse qu'une altération n'arrive pas deux fois à un même locus.

Dans cette sous-section, nous allons poser les définitions régissant le calcul de la fréquence allélique d'un variant (VAF) qui sera par la suite la donnée primordiale pour l'analyse de l'hétérogénéité clonale.

4.1.1 Calcul de la fréquence allélique d'un variant

Définition 4.1. La *fréquence allélique* d'un variant (VAF) pour un locus l est égale au nombre de copies c^B supportant l'allèle alternatif B relatif au nombre total de copies $c^T = c^A + c^B$, où c^A est le nombre de copies supportant l'allèle de référence A, tel que :

$$\mathrm{VAF}_l = \frac{c^B}{c^A + c^B} = \frac{c^B}{c^T} \tag{4.1}$$

Ainsi, pour un génome diploïde, l'ensemble des valeurs théoriques de VAF sont $\mathrm{VAF} = \{0, 0.5, 1\}$ pour chaque génotype dans $G = \{AA, AB, BB\}$ respectivement et sans CNA.

Nous allons maintenant voir comment l'hétérogénéité de l'échantillon tumoral affecte le calcul de la VAF. En effet, le nombre de copies du locus considéré, la proportion de cellules tumorales portant l'altération et la contamination par le tissus sain vont influencer la valeur de la VAF. Nous allons présenter ce phénomène par une série d'exemples de complexité croissante.

Exemple 1 : échantillon tumoral homogène - sans CNA

Soit un échantillon tumoral homogène composé d'une seule population de cellules tumorales $S = \{\mathcal{C}\}$ (figure 4.1.A). Soit le chromosome Chr_1 ne présentant pas de CNA. Soit un locus l_1 de génotype AB, présent sur le chromosome Chr_1, B étant l'allèle altéré. Dans ce cas, la valeur de la VAF théorique de l'altération au locus l_1 est calculée selon l'équation 4.1 :

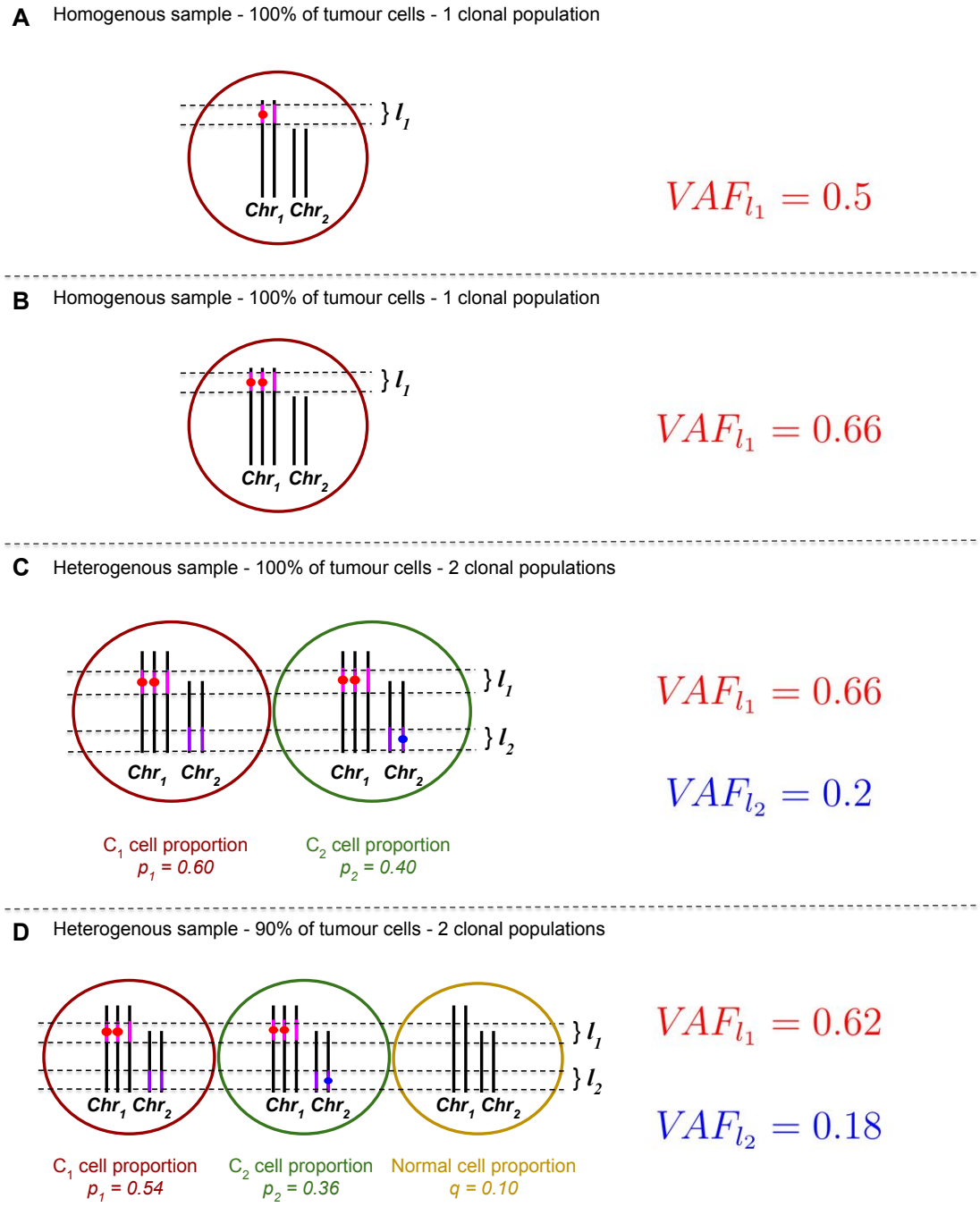


FIGURE 4.1 – *Influence de la composition de l'échantillon tumoral sur le calcul de la VAF.* L'influence des altérations de type CNAs, de l'hétérogénéité intra-tumorale et de la contamination par le tissus sain sur le calcul de la VAF est présenté par des exemples de complexité croissante (A-D) pour le locus l_1 présent sur le chromosome 1 (Chr_1) et le locus l_2 présent sur le chromosome 2 (Chr_2).

$$\text{VAF}_{l_1} = \frac{1}{1+1} = \frac{1}{2} = 0.5$$

avec $c^A = 1$, $c^B = 1$ et $c^T = 2$.

Exemple 2 : échantillon tumoral homogène - avec CNA

Soit un échantillon tumoral homogène composé d'une seule population de cellules tumorales $S = \{\mathcal{C}\}$ (figure 4.1.B). Soit le chromosome Chr_1 présentant une altération de type CNA correspondant à une duplication d'un des deux chromosomes parentaux. Soit un locus l_1 de génotype ABB, présent sur le chromosome Chr_1.

Dans ce cas, la valeur de la VAF théorique de l'altération au locus l_1 est calculée selon l'équation 4.1 :

$$\text{VAF}_{l_1} = \frac{2}{1+2} = \frac{2}{3} = 0.66$$

avec $c^A = 1$, $c^B = 2$ et $c^T = 3$.

Exemple 3 : échantillon tumoral hétérogène - avec CNA

Si l'on considère un échantillon tumoral hétérogène, la VAF va alors dépendre des proportions des différentes populations de cellules tumorales portant l'altération considérée et de leur CNA associée.

Soit un échantillon tumoral hétérogène composé de plusieurs populations de cellules tumorales $S = \{\mathcal{C}_1, \mathcal{C}_2\}$ de proportion $P = \{p_1, p_2\}$ respectivement (figure 4.1.C). Soit le chromosome Chr_1 présentant une altération de type CNA correspondant à une duplication d'un des deux chromosomes parentaux. Soit le chromosome Chr_2 ne présentant pas de CNA.

Soit un locus l_1 de génotype ABB, présent sur le chromosome Chr_1, dans l'ensemble des populations parmi $S = \{\mathcal{C}_1, \mathcal{C}_2\}$ (figure 4.1.C). Pour ce locus, l'altération étant commune à l'ensemble des populations, la valeur de la VAF théorique de l'altération au locus l_1 est calculée selon l'équation 4.1 :

$$\text{VAF}_{l_1} = \frac{2}{1+2} = \frac{2}{3} = 0.66$$

avec $c^A = 1$, $c^B = 2$ et $c^T = 3$.

Soit un locus l_2 présent sur le chromosome Chr_2, de génotype AA dans la population $\mathcal{C}_1$ et de génotype AB dans la population $\mathcal{C}_2$. Pour ce locus, les nombres de copies altérées sont différents dans les populations $\mathcal{C}_1$ et $\mathcal{C}_2$. Le calcul de la VAF doit alors prendre en compte cette différence en pondérant le nombre de copies portant l'allèle altérée par la proportion de la population correspondante.

De manière générale, un échantillon tumoral hétérogène est composé de n populations de cellules tumorales $S = \{\mathcal{C}_1, ..., \mathcal{C}_n\}$ de proportion $P = \{p_1, ..., p_n\}$ respectivement avec $\sum_{i=1}^{n} p_i = 1$.

Définition 4.2. Étant donné un échantillon tumoral hétérogène, le *nombre de copies* c^B *supportant l'allèle alternatif* B au locus l est égal à :

$$\begin{aligned} c^B &= c_1^B \times p_1 + ... + c_n^B \times p_n \\ &= \sum_{i=1}^{n} (c_i^B \times p_i) \end{aligned} \tag{4.2}$$

où c_i^B est le nombre de copies supportant l'allèle alternatif B dans la population cellulaire $\mathcal{C}_i$ de proportion cellulaire p_i.

De même, le nombre total de copies au locus considéré est susceptible de varier entre les différentes populations composant l'échantillon. Le calcul de ce dernier doit alors être pondéré.

Définition 4.3. Étant donné un échantillon tumoral hétérogène, le *nombre total de copies* c^T au locus l est égal à :

$$\begin{aligned} c^T &= (c_1^A + c_1^B) \times p_1 + ... + (c_n^A + c_n^B) \times p_n \\ &= c_1^T \times p_1 + ... + c_n^T \times p_n \\ &= \sum_{i=1}^{n} (c_i^T \times p_i) \end{aligned} \tag{4.3}$$

où $c_i^T = c_i^A + c_i^B$ est le nombre total de copies présent dans la population cellulaire $\mathcal{C}_i$ de proportion cellulaire p_i.

Nous allons maintenant définir le calcul de la VAF dans le cas général correspondant. Ainsi, par combinaison des équations 4.1, 4.2 et 4.3, le calcul de la VAF est modifié lorsque l'on considère un échantillon tumoral hétérogène.

Définition 4.4. Étant donné un échantillon tumoral hétérogène, la *fréquence allélique* d'un variant (VAF) pour un locus l est égale à la somme du nombre de copies c^B, défini dans l'équation 4.2, relatif au nombre total de copies c^T, défini dans l'équation 4.3, tel que :

$$\mathrm{VAF}_l = \frac{\sum_{i=1}^{n} (c_i^B \times p_i)}{\sum_{i=1}^{n} ((c_i^A + c_i^B) \times p_i)} = \frac{\sum_{i=1}^{n} (c_i^B \times p_i)}{\sum_{i=1}^{n} (c_i^T \times p_i)}. \tag{4.4}$$

Ainsi, la VAF théorique du locus l_1 reste inchangée (figure 4.1.C) selon l'équation 4.4 :

$$\mathrm{VAF}_1 = \frac{2 \times 0.6 + 2 \times 0.4}{(1+2) \times 0.6 + (1+2) \times 0.4} = \frac{2}{3} = 0.66$$

avec $c_1^A = 1$, $c_1^B = 2$, $c_1^T = 3$, $c_2^A = 1$, $c_2^B = 2$, $c_2^T = 3$, $p_1 = 0.6$ et $p_2 = 0.4$.

La VAF théorique au locus l_2, selon l'équation 4.4, est quant à elle de :

$$\mathrm{VAF}_2 = \frac{0 \times 0.6 + 1 \times 0.4}{2 \times 0.6 + 2 \times 0.4} = \frac{0.4}{2} = 0.2$$

avec $c_1^A = 2$, $c_1^B = 0$, $c_1^T = 2$, $c_2^A = 1$, $c_2^B = 1$, $c_2^T = 2$, $p_1 = 0.6$ et $p_2 = 0.4$.

Exemple 4 : échantillon tumoral hétérogène contaminé par du tissu sain - avec CNA

Enfin, la contamination par le tissus sain va aussi influencer la valeur théorique de la VAF. Dans notre cas, nous nous intéressons seulement aux locus homozygotes AA dans le tissus sain. Il en découle que pour le calcul de la VAF des altérations somatiques, le tissus sain n'est autre qu'une population de cellules pour laquelle, à un locus l, le nombre de copies c^B est systématiquement égal à 0 et le nombre de copies c^A est systématiquement égal à c^T soit 2.

Soit un échantillon tumoral hétérogène composé de plusieurs populations de cellules tumorales $S = \{\mathcal{C}_1, \mathcal{C}_2\}$ de proportion $P = \{p_1, p_2\}$ respectivement (figure 4.1.D).

Soit q la proportion de cellules normales, appelée aussi proportion de contamination par le tissus sain, présente dans l'échantillon avec $q = 1 - \sum_{i=1}^{n} p_i$. Par convention, la population de cellules normales n'est pas inclue dans S.

Soit le chromosome Chr_1 présentant une altération de type CNA correspondant à une duplication d'un des deux chromosomes parentaux. Soit le chromosome Chr_2 ne présentant pas de CNA. Soit un locus l_1 présent sur le chromosome Chr_1, de génotype ABB dans l'ensemble des populations parmi $S = \{\mathcal{C}_1, \mathcal{C}_2\}$. Soit un locus l_2 présent sur le chromosome Chr_2, de génotype AA dans la population $\mathcal{C}_1$ et de génotype AB dans la population $\mathcal{C}_2$.

Le calcul du nombre de copies total c^T présent dans l'échantillon doit alors prendre en compte la proportion de cellules saines q.

Définition 4.5. Étant donné un échantillon tumoral hétérogène contaminé par une proportion q de tissus sain, le *nombre total de copies* c^T au locus l est égal à :

$$\begin{aligned} c^T &= (c_1^A + c_1^B) \times p_1 + ... + (c_n^A + c_n^B) \times p_n + 2 \times q \\ &= c_1^T \times p_1 + ... + c_n^T \times p_n + 2 \times q \\ &= \sum_{i=1}^{n} (c_i^T \times p_i) + 2 \times q \end{aligned} \tag{4.5}$$

où $c_i^T = c_i^A + c_i^B$ est le nombre total de copies présent dans la population cellulaire $\mathcal{C}_i$ de proportion cellulaire p_i et q la proportion de cellules saines.

Le calcul de la VAF théorique est alors modifié.

Définition 4.6. Étant donné un échantillon tumoral hétérogène contaminé par une proportion q de tissus sain, la *fréquence allélique* d'un variant (VAF) pour un locus l est égale à la somme du nombre de copies c^B, défini dans l'équation 4.2, relatif au nombre total de copies c^T, défini dans l'équation 4.5, tel que :

$$\text{VAF}_l = \frac{\sum_{i=1}^{n} (c_i^B \times p_i)}{\sum_{i=1}^{n} ((c_i^A + c_i^B) \times p_i) + 2 \times q} = \frac{\sum_{i=1}^{n} (c_i^B \times p_i)}{\sum_{i=1}^{n} (c_i^T \times p_i) + 2 \times q} \tag{4.6}$$

Les VAFs théoriques aux loci l_1 et l_2 selon l'équation 4.6 sont alors de :

$$\begin{aligned} \text{VAF}_{l_1} &= \frac{2 \times 0.54 + 2 \times 0.36}{3 \times 0.54 + 3 \times 0.36 + 2 \times 0.1} \\ &= \frac{2 \times 0.9}{3 \times 0.9 + 2 \times 0.1} = 0.62 \end{aligned}$$

avec $c_1^A = 1$, $c_1^B = 2$, $c_1^T = 3$, $c_2^A = 1$, $c_2^B = 2$, $c_2^T = 3$, $p_1 = 0.54$, $p_2 = 0.36$ et $q = 0.1$.

$$\begin{aligned}\text{VAF}_{l_2} =& \frac{0 \times 0.54 + 1 \times 0.36}{2 \times 0.54 + 2 \times 0.36 + 2 \times 0.1} \\ =& \frac{0.36}{2} = 0.18\end{aligned}$$

avec $c_2^A = 2$, $c_2^B = 0$, $c_2^T = 2$, $c_2^A = 1$, $c_2^B = 1$, $c_2^T = 2$, $p_1 = 0.54$, $p_2 = 0.36$ et $q = 0.1$.

4.1.2 Observation des altérations

Nous venons de voir comment le calcul de la VAF est régit par la composition en cellules tumorales de l'échantillon et des altérations de *copy number*. Ce sont ces différentes règles qui vont nous permettre de résoudre le problème de reconstruction clonale. Dans le cas pratique, nous allons nous servir des données de séquençage afin d'identifier et caractériser les différentes populations de cellules composant un échantillon tumoral.

Ainsi, à partir des données de séquençage, nous avons accès aux comptes de *reads* qui supportent l'allèle de référence r^A et l'allèle alternative r^B permettant le calcul de la VAF pour un locus l, tel que :

$$\text{VAF}_l = \frac{r^B}{r^A + r^B} = \frac{r^B}{r^T} \tag{4.7}$$

où $r^T = r^A + r^B$.

Dans notre cas, nous allons prendre en compte seulement les loci de génotype AB et leurs dérivés altérés par duplication ou délétion, tel que AAB, ABB ou BB *etc.*

Par ailleurs, lorsque l'on observe les valeurs de VAFs, nous n'avons pas d'information directe sur les différentes proportions de populations tumorales composant l'échantillon. Cependant, les valeurs de VAFs nous renseignent sur la fraction cellulaire comportant l'altération ou fraction cellulaire altérée. Nous définissons ci dessous, le calcul de la fraction cellulaire altérée pour un échantillon tumoral hétérogène.

Soit un échantillon tumoral hétérogène composé de n populations de cellules tumorales $S = \{\mathcal{C}_1, ..., \mathcal{C}_n\}$ de proportion $P = \{p_1, ..., p_n\}$ respectivement avec $\sum_{i=1}^{n} p_i = 1$.

Définition 4.7. Soit un sous-ensemble de k populations de cellules tumorales inclues dans S avec $k \leq n$ partageant le même génotype au locus l, alors la *fraction cellulaire altérée* ACF au locus l est égale à la somme des k proportions cellulaires tel que :

$$\text{ACF}_l = \sum_{j=1}^{k} p_j \tag{4.8}$$

Enfin, en pratique, nous ne connaissons pas le génotype des différents loci mais nous disposons de l'état de *copy number* sous la forme de l'*Allele Specific Copy Number* (ASCN). L'ASCN correspond au nombre de copies de chaque chromosome parental quantifié par deux entiers $\{n_{\text{min}}, n_{\text{maj}}\}$ où n_{min} est le nombre de copies d'allèle parental minoritaire et n_{maj} est le nombre de copies d'allèle parental majoritaire (figure 4.2.A). Cependant, nous ne savons pas où se trouve l'altération B. Le nombre de copies altérées c^B peut alors prendre différentes valeurs :

— n_{min} dans le cas où l'altération est portée par l'ensemble des chromosomes dont le nombre d'allèles est égal au nombre d'allèles parentaux minoritaires n_{min} ; dans ce cas l'altération de SNA a eu lieu avant l'événement de CNA.

— n_{maj} dans le cas où l'altération est porté par l'ensemble des chromosomes dont le nombre d'allèles est égal au nombre d'allèles parentaux majoritaires n_{maj} ; dans ce cas l'altération de SNA a eu lieu avant l'événement de CNA.

— 1 dans le cas où l'altération est portée par un seul chromosome ; dans ce cas l'altération de SNA a eu lieu après l'événement de CNA. Dans ce travail, nous considérons que les événements de type CNA arrivent une seule fois par chromosome parental, nous ne considérons donc pas les autres entiers inférieur à n_{min} ou n_{maj}.

Par ailleurs, les populations tumorales composant l'échantillon peuvent présenter des états de *copy number* différents : chaque segment peut avoir plusieurs états de *copy number* dans l'échantillon (figure 4.2.B). En pratique, les algorithmes de calcul d'ASCN limitent leur nombre à deux états : $ASCN_1 = \{n_{\text{min}_1}, n_{\text{maj}_1}\}$ et $ASCN_2 = \{n_{\text{min}_2}, n_{\text{maj}_2}\}$. Ainsi, pour chaque locus, le nombre de copies altérées c^B peut alors prendre la valeur de 1, n_{min_1} ou n_{maj_1}, n_{min_2} ou n_{maj_2} représentant l'ensemble de combinaisons de solutions associés à chaque locus.

Remarque. Dans le cas où les sommes $(n_{\text{min}_1} + n_{\text{maj}_1})$ et $(n_{\text{min}_2} + n_{\text{maj}_2})$ sont différentes, le calcul de c^T doit être quant à lui pondéré par les proportions de cellules correspondantes.

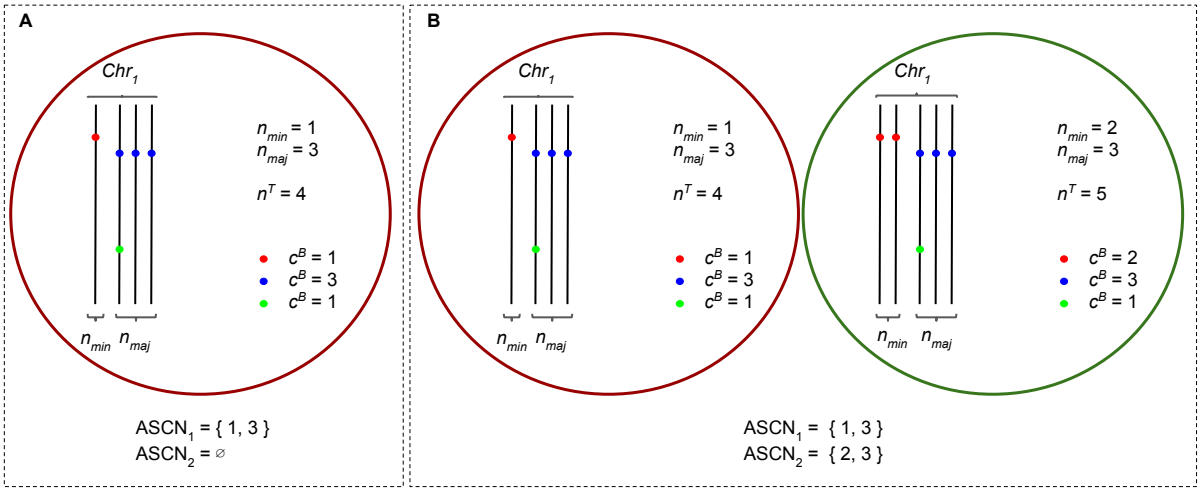


FIGURE 4.2 – *Illustration de l'état de copy number.* A. Exemple d'un état de *copy number* homogène d'$ASCN$ $\{1,3\}$ pour le chromosome 1 (Chr_1) : c^B peut ainsi prendre les valeurs parmi $\{1,3\}$. B. Exemple d'un état hétérogène composé des $ASCN_1$ $\{1,3\}$ et $ASCN_2$ $\{2,3\}$ pour le chromosome 1 (Chr_1) : c^B peut ainsi prendre les valeurs parmi $\{1,2,3\}$ ..

4.2 Méthode

4.2.1 Approche

L'objectif de la méthode que nous avons développée est de résoudre le problème de reconstruction clonale à partir du séquençage d'un échantillon tumoral. Le problème de

reconstruction clonale consiste en l'identification et la caractérisation des populations cellulaires (saine et tumorales) composant un échantillon tumoral. Dans ce travail, nous abordons ce problème à partir des altérations somatiques de type SNAs et CNAs identifiées pour un ensemble de loci.

L'hétérogénéité intra-tumorale et les multiples altérations de *copy number* présents dans l'échantillon tumoral entraînent une combinatoire de solutions possibles de composition de l'échantillon. Malgré de nombreuses méthodes développées ces dernières années, ce problème reste un domaine émergent (section 1.2.6).

Afin de résoudre ce problème, nous avons développé une méthode qui permet le regroupement des SNAs et ainsi retrouver les différentes fractions cellulaires altérées (ACFs) composant l'échantillon. Ce regroupement est effectué à partir de leurs fréquences alléliques, ou VAFs, qui reflètent leurs niveaux de présence dans l'échantillon c'est à dire le nombre de cellules et de chromosomes qui les comportent. En effet, comme décrit dans la section précédente, le calcul de la VAF d'un locus est régit par la proportion de cellules comportant l'altération, le nombre d'allèles portant l'altération et le nombre d'allèle total. Le regroupement des loci permet alors de considérer des sous-ensembles partageant les mêmes propriétés en matière de *copy number* et proportion cellulaire et ainsi à identifier l'ensemble des populations cellulaires tumorales et leurs caractéristiques.

4.2.2 Modèle gaussien

Comme nous l'avons vu dans l'introduction (voir la sous-section 1.2.2 de l'introduction), la profondeur de séquençage n'est pas constante dans un échantillon séquencé. En effet, les biais biologiques, technologiques et bioinformatiques entraînent une variation du nombre de *reads* couvrant chaque base nucléotidique.

Soit un échantillon tumoral homogène, sans CNA. Soit un ensemble de loci de génotype AB, les valeurs des VAFs observées seront distribuées autour de 0.5 (figure 4.3).

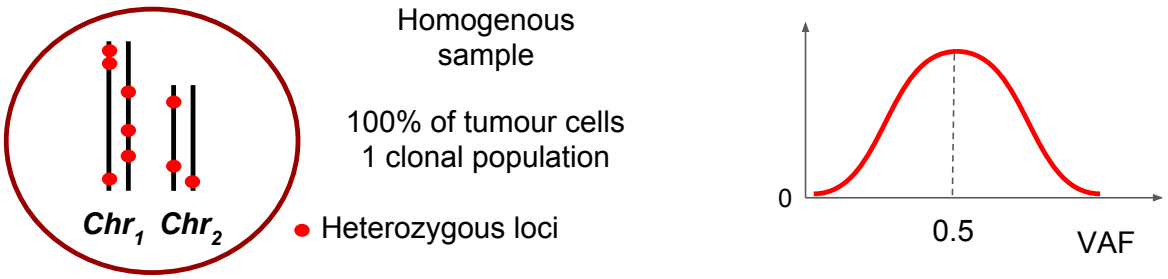


FIGURE 4.3 – *Modèle Gaussien de la distribution des VAFs.* Illustration de la distribution des VAF d'un ensemble de loci hétérozygotes sur les chromosomes 1 et 2 (Chr_1 et Chr_2) pour un échantillon tumoral homogène sans contamination de tissu sain.

Ainsi, lorsque l'on s'intéresse à la mesure des VAFs de multiples loci d'un échantillon nous pouvons modéliser leur observation par une loi normale de moyenne $\overline{\text{VAF}}$ correspondant à la VAF moyenne.

De plus, dans un échantillon, nous pouvons considérer l'ensemble des loci altérés $\mathcal{L}$ ou des sous-ensembles de ces loci $\mathcal{L} = \{\mathcal{L}_i\}$ tel que $\cup\mathcal{L}_i = \mathcal{L}$ et $\cap\mathcal{L}_i = \emptyset$.

Définition 4.8. Soit $\mathcal{L}$ un ensemble de loci, alors la *fraction allélique moyenne* $\overline{\text{VAF}}$ correspond à la moyenne de l'ensemble des VAFs des différents loci de $\mathcal{L}$, tel que :

$$\overline{\text{VAF}} = \frac{\sum_{l \in \mathcal{L}} (\text{VAF}_{l_j})}{|\mathcal{L}|} \tag{4.9}$$

Comme décrit dans la section 4.1, les valeurs de VAFs observées vont être influencées par l'état de *copy number* et l'hétérogénéité intra-tumorale. Ainsi, la distribution des VAFs n'est pas composée d'une seule distribution gaussienne mais de multiples gaussiennes. Ces multiples gaussiennes correspondent à des sous-ensembles de loci partageant à la fois le même génotype et le même ACF. Pour chaque sous-ensemble $\mathcal{L}$ de loci, l'objectif est alors de calculer la fraction cellulaire altérée à partir de $\overline{\text{VAF}}$ et de l'ASCN propre à ce sous-ensemble $\mathcal{L}$.

4.2.3 Calcul de la fraction cellulaire altérée

Comme dans la sous-section 4.1, de manière générale, un échantillon tumoral hétérogène est composé de n populations de cellules tumorales $S = \{\mathcal{C}_1, ..., \mathcal{C}_n\}$ de proportion $P = \{p_1, ..., p_n\}$ et de proportion de cellules normales q, avec $q = 1 - \sum_{i=1}^{n} p_i$.

De manière générale, la fraction cellulaire altérée peut être calculée en combinant les équations 4.6 et 4.9.

Soit un sous ensemble de loci $\mathcal{L}$ présents dans k populations tumorales ($k \leq n$). Chaque locus $l \in \mathcal{L}$ a un nombre de copies altérées égal à c^B et un ASCN de $\{n_{\text{min}}, n_{\text{maj}}\}$. Le nombre total de copies est donc tel que $c^T = n_{\text{min}} + n_{\text{maj}}$ et le nombre de copies altérées tel que $c^B \in \{1, n_{\text{min}}, n_{\text{maj}}\}$. Alors, pour tout $l \in \mathcal{L}$, le calcul de l'ACF$_l$ peut être décomposé comme suit :

$$\begin{aligned} \text{VAF}_l &= \frac{c^B \times \text{ACF}_l}{c^T \times \text{ACF}_l + 2 \times q} \\ &= \frac{c^B \times \text{ACF}_l}{(c^T - 2) \times \text{ACF}_l + 2} \end{aligned}$$

Suite à la mise en évidence de la variable ACF, ceci nous donne :

$$\text{ACF}_l = \frac{2 \times \text{VAF}_l}{c^B + \text{VAF}_l \times (2 - c^T)} \tag{4.10}$$

Nous allons voir, par une suite d'exemples de complexité croissante (figure 4.4), comment calculer les différentes fractions cellulaires altérées composant un échantillon tumoral à partir d'un ensemble d'ensembles de loci $\mathcal{L}$ obtenus par la séparation de la distribution de l'ensemble des VAFs en gaussiennes. Ainsi, chaque sous-ensemble L_i est caractérisé par une valeur $\overline{\text{VAF}}$ et un état de *copy number* sous la forme de l'ASCN.

Exemple 1 : échantillon tumoral homogène - avec CNA

Soit un échantillon tumoral homogène composé d'une seule population de cellules tumorales (figure 4.4.A). Soit un ensemble d'ensembles de loci $\mathcal{L} = \{L_1, L_2, L_3\}$. Soit

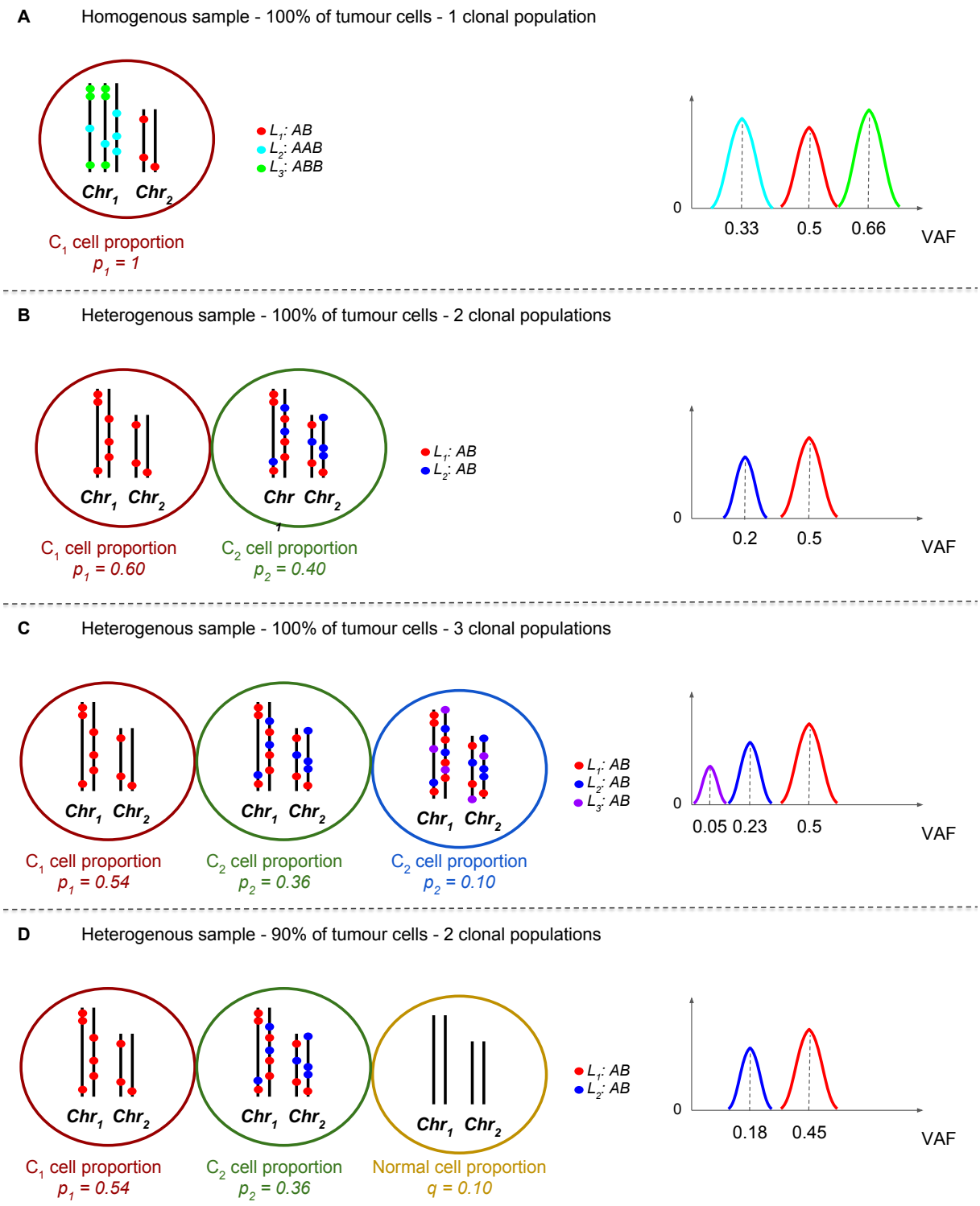


FIGURE 4.4 – *Observation des VAFs selon la composition de l'échantillon tumoral.* Illustration de l'observation des VAFs, par des exemples de complexité croissante (A-D), pour des sous-ensembles de loci (L_1, L_2 et L_3) présent sur les chromosomes 1 et 2 (Chr_1 et Chr_2).

l'ensemble de loci L_1 de génotype AB, d'ASCN $\{1,1\}$, et de VAF moyenne observée $\overline{\text{VAF}}_{L_1}$ de 0.5. La fraction cellulaire altérée ACF_{L_1} est de :

$$\text{ACF}_{L_1} = \frac{2 \times 0.5}{1 + 0.5 \times (2-2)} = \frac{1}{1} = 1$$

avec $c^T = 2$ et $c^B = 1$.

Soit l'ensemble de loci L_2 de génotype AAB, d'ASCN de $\{1,2\}$ et de VAF moyenne observée $\overline{\text{VAF}}_{L_2}$ de 0.33. La fraction cellulaire ACF_{L_2} est de :

$$\text{ACF}_{L_2} = \frac{2 \times 0.33}{1 + 0.33 \times (2-3)} = \frac{0.66}{1 - 0.33} = \frac{0.66}{0.67} \approx 1$$

avec $c^T = 3$ et $c^B = 1$.

Soit l'ensemble de loci L_3 de génotype ABB, d'ASCN de $(1,2)$ et de VAF moyenne observée $\overline{\text{VAF}}_{L_3}$ de 0.66. La fraction cellulaire ACF_{L_3} est de :

$$\text{ACF}_{L_3} = \frac{2 \times 0.66}{2 + 0.66 \times (2-3)} = \frac{1.32}{2 - 0.66} = \frac{1.32}{1.34} \approx 1$$

avec $c^T = 3$ et $c^B = 2$.

Exemple 2 : échantillon tumoral hétérogène - sans CNA

Soit un échantillon tumoral hétérogène composé de plusieurs populations de cellules tumorales $S = \{\mathcal{C}_1, \mathcal{C}_2\}$ de proportion $P = \{p_1, p_2\}$ respectivement. Soit un ensemble d'ensembles de loci $\mathcal{L} = \{L_1, L_2\}$. Soit l'ensemble de loci L_1 de génotype AB, d'ASCN $\{1,1\}$, et de VAF moyenne observée $\overline{\text{VAF}}_{L_1}$ de 0.5. La fraction cellulaire ACF_{L_1} est de :

$$\text{ACF}_{L_1} = \frac{2 \times 0.5}{1 + 0.5 \times (2-2)} = \frac{1}{1} = 1$$

avec $c^T = 2$ et $c^B = 1$.

Soit l'ensemble de loci L_2 de génotype AB, d'ASCN de $\{1,1\}$ et de VAF moyenne observée $\overline{\text{VAF}}_{L_2}$ de 0.2. La fraction cellulaire ACF_{L_2} est de :

$$\text{ACF}_{L_2} = \frac{2 \times 0.2}{1 + 0.2 \times (2-2)} = \frac{0.4}{1 - 0} = 0.4$$

avec $c^T = 2$ et $c^B = 1$.

Exemple 3 : échantillon tumoral hétérogène - sans CNA

Soit un échantillon tumoral hétérogène composé de plusieurs populations de cellules tumorales $S = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3\}$ de proportion $P = \{p_1, p_2, p_3\}$ respectivement. Soit un ensemble d'ensembles de loci $\mathcal{L} = \{L_1, L_2, L_3\}$. Soit l'ensemble de loci L_1 de génotype AB, d'ASCN $\{1, 1\}$, et de VAF moyenne observée $\overline{\text{VAF}}_{L_1}$ de 0.5. La fraction cellulaire ACF_{L_1} est de :

$$\text{ACF}_{L_1} = \frac{2 \times 0.5}{1 + 0.5 \times (2 - 2)} = \frac{1}{1} = 1$$

avec $c^T = 2$ et $c^B = 1$.

Soit l'ensemble de loci L_2 de génotype AB, d'ASCN de $\{1, 1\}$ et de VAF moyenne observée $\overline{\text{VAF}}_{L_2}$ de 0.23. La fraction cellulaire ACF_{L_2} est de :

$$\text{ACF}_{L_2} = \frac{2 \times 0.23}{1 + 0.23 \times (2 - 2)} = \frac{0.46}{1 - 0} = 0.46$$

avec $c^T = 2$ et $c^B = 1$.

Soit l'ensemble de loci L_3 de génotype AB, d'ASCN de $\{1, 1\}$ et de VAF moyenne observée $\overline{\text{VAF}}_{L_3}$ de 0.05. La fraction cellulaire ACF_{L_3} est de :

$$\text{ACF}_{L_3} = \frac{2 \times 0.05}{1 + 0.05 \times (2 - 2)} = \frac{0.1}{1 - 0} = 0.1$$

avec $c^T = 2$ et $c^B = 1$.

Exemple 4 : échantillon tumoral hétérogène contaminé par du tissus sain - sans CNA

Soit un échantillon tumoral hétérogène composé de plusieurs populations de cellules tumorales $S = \{\mathcal{C}_1, \mathcal{C}_2\}$ de proportion $P = \{p_1, p_2\}$ respectivement et contaminé par une population de cellules saines de proportion q avec $q = 1 - (p_1 + p_2)$. Soit un ensemble d'ensembles de loci $\mathcal{L} = \{L_1, L_2\}$. Soit l'ensemble de loci L_1 de génotype AB, d'ASCN $\{1, 1\}$, et de VAF moyenne observée $\overline{\text{VAF}}_{L_1}$ de 0.45. La fraction cellulaire ACF_{L_1} est de :

$$\text{ACF}_{L_1} = \frac{2 \times 0.45}{1 + 0.45 \times (2 - 2)} = \frac{1}{1} = 0.9$$

avec $c^T = 2$ et $c^B = 1$.

Soit l'ensemble de loci L_2 de génotype AB, d'ASCN de $\{1, 1\}$ et de VAF moyenne observée $\overline{\text{VAF}}_{L_2}$ de 0.18. La fraction cellulaire ACF_{L_2} est de :

$$\text{ACF}_{L_2} = \frac{2 \times 0.18}{1 + 0.18 \times (2 - 2)} = \frac{0.36}{1 - 0} = 0.36$$

avec $c^T = 2$ et $c^B = 1$.

4.2.4 Algorithme

Le problème de reconstruction clonale d'un échantillon tumoral peut être résumé par trois sous-problèmes :

— le calcul de la fraction de cellules tumorales,

— le calcul du nombre de populations clonales et leur proportion,

— l'assignation de l'ensemble des SNAs à chacune des populations.

La difficulté calculatoire est due à l'ensemble de valeurs possibles que peuvent prendre c^B à partir de l'ASCN. Il est alors nécessaire de résoudre la combinatoire sous-jacente afin de sélectionner parmi l'ensemble des solutions celle qui est la plus probable. Pour cela, nous avons développé une méthode regroupant les SNAs partageant la même ACF et le même ASCN à partir de leur VAF. Comme présenté dans la sous-séction précédente, plusieurs groupes d'altérations coexistent selon le nombre de copies de chromosomes qui la portent, le nombre de copies de chromosomes total pour ce locus mais aussi la proportion des cellules présentant cette altération dans l'échantillon tumoral. Ainsi, la distribution des VAFs des différents groupes d'altérations forme un mélange complexe. Notre méthode va alors décomposer ce mélange complexe en différents signaux simples via un algorithme de modèle de mélanges gaussiens (GMM).

Dans le cas de nos données, nous ne connaissons pas le nombre de clusters, ou sous populations tumorales. Afin de déterminer le nombre de clusters le plus probable, nous utilisons le critère d'information bayésien (BIC) qui maximise la vraisemblance tout en pénalisant par la taille de l'échantillon et le nombre de paramètre Le regroupement, ou *clustering*, des altérations va alors permettre de retrouver le nombre de populations cellulaires tumorales en considérant l'impact de l'hétérogénéité intra-tumorale différemment selon l'ASCN. Pour cela, la méthode est composée de 4 étapes (figure 4.5) :

1. *La décomposition de l'ensemble des VAFs en différents sous-ensembles.* Cette étape est effectuée séparément pour chaque chromosome par un premier GMM. Le nombre de composantes est sélectionné de manière à minimiser le critère BIC. Ainsi, pour chaque chromosome, nous obtenons un ensemble d'ensembles de loci partageant tous la même ACF et le même ASCN.
2. *Le calcul de l'ensemble des ACFs possibles.* Pour chaque sous-ensemble de loci, de chaque chromosome, pour chaque état de *copy number* l'algorithme calcule l'ensemble des ACFs à partir de la VAF moyenne en faisant varier le nombre de copies altérées dans l'ensemble des solutions $\{1, n_{\text{min}_1}, n_{\text{maj}_1}, n_{\text{min}_2}, n_{\text{maj}_2}\}$ et le nombre de copies total dans $\{(n_{\text{min}_1} + n_{\text{maj}_1}), (n_{\text{min}_2} + n_{\text{maj}_2})\}$. L'ensemble des ACFs est ensuite réduit à l'intervalle $]0 - 1[$.
3. *L'estimation du nombre de populations de cellules tumorales à partir de l'ensemble des ACFs possibles via un second GMM.* De la même manière, le critère BIC est utilisé afin de déterminer le nombre de composantes. Le nombre de composantes est alors réduit au nombre de clusters composés de sous-ensembles de loci présents dans une seule composante. Ceci est effectué dans le but de ne pas prendre en compte les clusters composés de groupes de loci pouvant être classés dans d'autres composantes. De plus, les clusters comportant moins de 10 loci sont filtrés.
4. *L'assignation de chaque SNA à une population clonale.* Cette étape est effectuée via la méthode de clusteuring *k-means* sur l'ensemble des ACFs possibles calculé dans la seconde étape à partir du nombre de clusters calculé dans la troisième. Chaque SNA est alors associé à la sous population pour laquelle la distance à la moyenne

du cluster est minimale. Pour chaque cluster, l'ACF moyen et le nombre de SNAs le composant peut-être alors calculé.

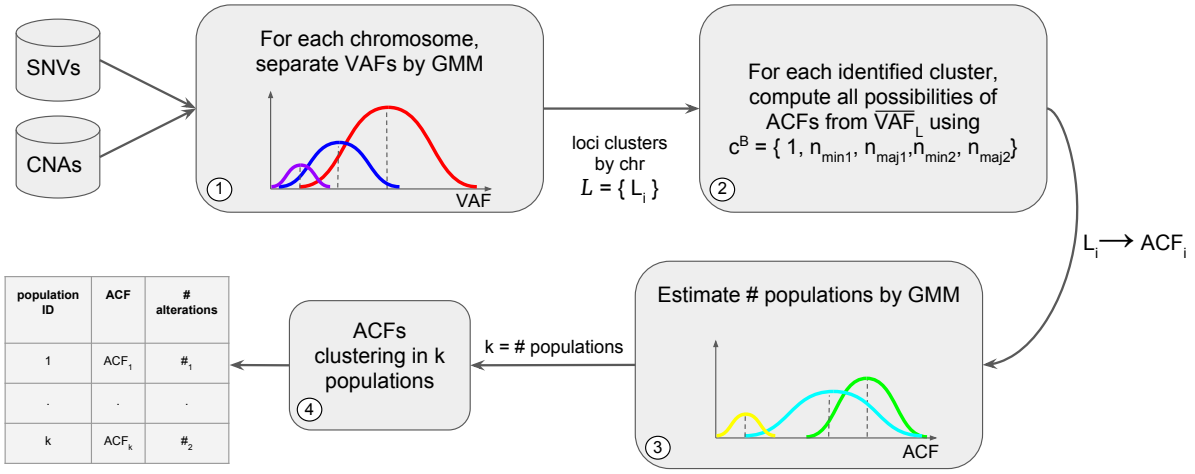


FIGURE 4.5 – *Méthode.* Notre approche est composée de 4 étapes : (1) la séparation des VAFs en différents clusters à partir d'un GMM, (2) le calcul de l'ensemble des ACFs possibles par combinaison des VAFs moyenne par cluster et des CNAs, (3) l'estimation du nombre de populations tumorales à partir d'un GMM sur les ACFs, (4) l'assignation de chaque SNAs à une population tumorale permettant ainsi le calcul de la proportion de chacune d'entre elles.

Dans ce travail, nous assumons que les mutations sont clonales et conservées durant la progression tumorale c'est à dire qu'une population tumorale issue d'une première population conserve les même altérations. Par conséquent, la fraction cellulaire altérée maximale correspond à la proportion de tissu tumoral présent dans l'échantillon, la proportion de cellules normales étant égale à 1 moins cette ACF maximale.

Cet algorithme est implémenté en langage Python et utilise, entre autres, la librairie de *Machine Learning* : *scikit learn*.

4.3 Application aux données du challenge DREAM SMC-Het

4.3.1 Matériel

Les données utilisées dans ce chapitre sont des données simulées mises à disposition par le DREAM challenge. Elles ont été construites à partir de profils tumoraux observés dans différents types de cancers présentant des altérations de *copy number* et des mutations somatiques partagées ou propres à plusieurs sous-populations tumorales. Un éventail de pureté, de ploïdie et de profils mutationnels est proposé. Les données sont fournies dans des formats de résultats standards des outils permettant la détection des altérations correspondantes :

— Les altérations somatiques, ou SNAs, dans le format *Variant Calling Format (VCF)*.
— Les altérations de *copy number* ou état de *copy number* de l'ensemble des chromosomes ou segments chromosomiques, dans le format *Battenberg*.

Pré-traitement des données

Dans notre cas, seules les altérations somatiques de type substitution de 1 paire de bases sont renseignées. Dans le fichier VCF nous avons un ensemble d'informations pour chaque altération. Plus précisément, nous disposons du nombre de *reads* couvrant l'allèle de référence r^A, alternative r^B et total r^T, la qualité de base moyenne des *reads* couvrant cette position et la VAF résultante. Ces informations permettent d'effectuer un filtrage des altérations sur leur qualité et leur couverture de séquençage.

Le pré-traitement des données consiste à fusionner les fichiers VCF et Battenberg afin de créer une grande table de données reliant chacun des SNA au segment chromosomique sur lequel il se trouve.

4.3.2 Description des échantillons

Nous avons à notre disposition 6 échantillons tumoraux simulés. La proportion de cellules tumorales, le nombre de sous populations, leur proportions et le nombre de SNAs qu'elles contiennent sont résumés dans la figure 4.6 pour l'ensemble des échantillons triés par complexité croissante.

4.3.3 Résultats

Les altérations dont la qualité est inférieure à 30 ou dont la couverture de l'allèle altérée est inférieure à 10 sont filtrées. De plus, les altérations appartenant à un ensemble de loci dont l'état de *copy number* est très peu représenté dans l'échantillon sont aussi filtrées. Ainsi, si une altération est présente dans un sous-ensemble de loci L_i regroupés par l'état de copy number dont la taille est inférieure à 10 tel que $|L_i| < 10$, cette dernière n'est pas prise en compte. Enfin, les altérations présentes sur les chromosomes X et Y ne sont pas prises en compte.

Afin d'illustrer les étapes de notre méthode présenté dans la sous-section 4.2.4, nous allons prendre l'exemple de la tumeur 4.

La première étape consiste à effectuer un GMM de manière indépendante pour chaque chromosome. L'hypothèse est que chaque cluster obtenu est alors composé de loci présentant une ACF et un état de *copy number*. C'est ce que l'on observe, par exemple, pour le chromosome 4 dans la figure 4.7 où l'on observe que l'algorithme a identifié deux clusters. Ces deux clusters ont une VAF moyenne de $\overline{\text{VAF}}_{L_0} = 0.204393$ pour le cluster 0 et de $\overline{\text{VAF}}_{L_1} = 0.839333$ pour le cluster 1.

La seconde étape calcule alors l'ensemble des ACFs possibles pour l'ensemble des clusters issus de la première étape à partir des différentes combinaisons d'ASCN selon l'équation 4.10. Les différentes valeurs associées aux loci du chromosome 4 sont les suivantes : $\{n_{\text{min}_1} = 0, n_{\text{maj}_1} = 2, n_{\text{min}_2} = \emptyset, n_{\text{maj}_2} = \emptyset\}$ auxquelles nous rajoutons la valeur de 1 dans le cas où la mutation serait après la duplication (voir la sous section 4.1.2). Ainsi, les différentes valeurs d'ACFs possibles sont de :

Cluster 0

— $c^B = 1$:

$$\frac{2 \times 0.204393}{1 \times +0.204393 \times (2-2)} = 0.408786$$

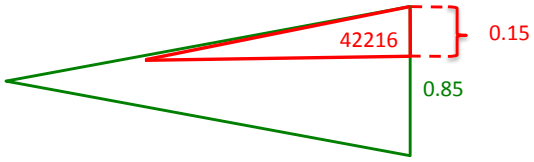
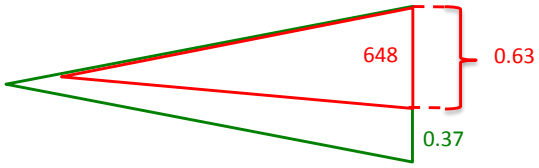
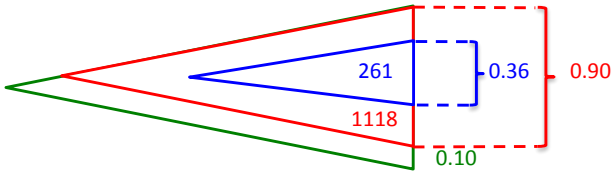
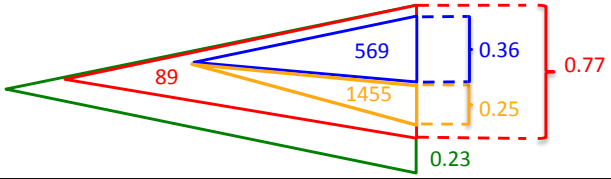
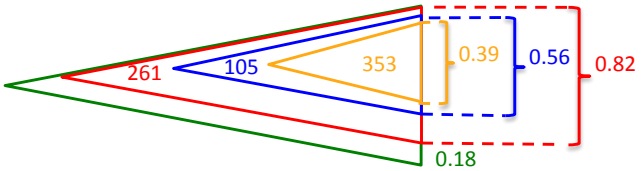
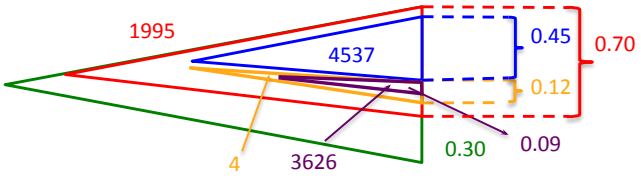
Tumor ID	Cancer cell proportion	# clones	Composition
T6	0,15	1	
T3	0,63	1	
T2	0,9	2	
T7	0,77	3	
T4	0,82	3	
T1	0,7	4	

FIGURE 4.6 – *Description des échantillons.* L'identifiant des échantillons, la proportion de cellules tumorales, le nombre de populations différentes de cellules tumorales et la composition sous forme graphique, en matière d'ACF et de comptes de SNAs, est décrit pour chaque échantillon.

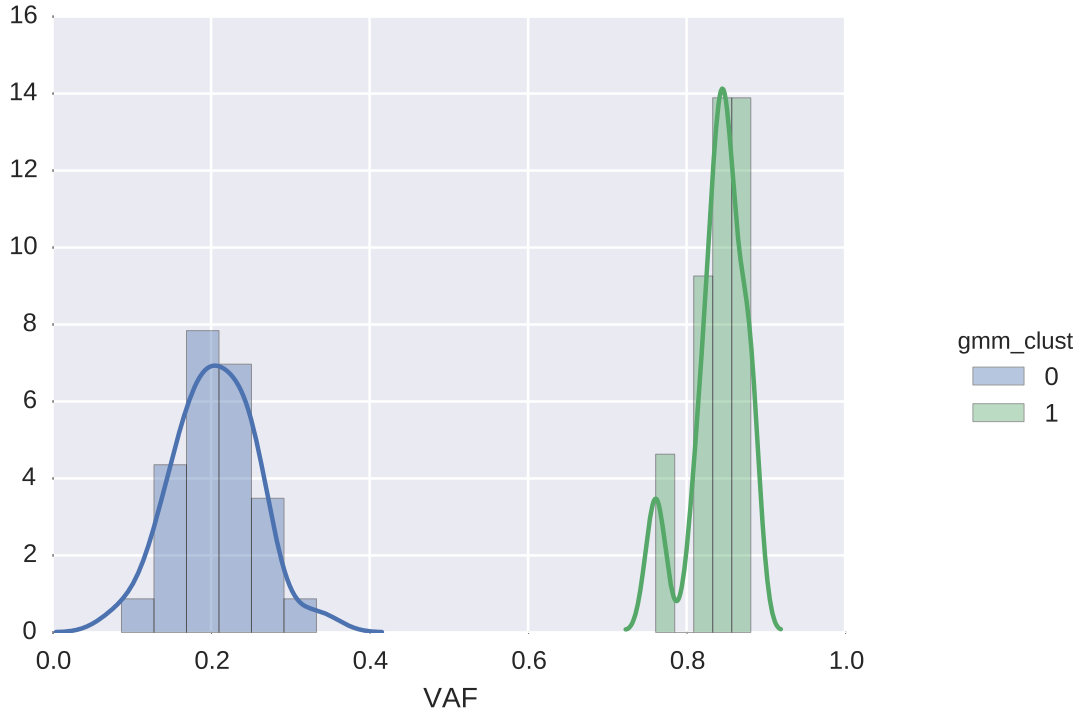


FIGURE 4.7 – *Distribution des VAFs du chromosome 4.* Les VAFs sont séparées en 2 clusters : le pic bleu correspond aux VAFs composant le cluster 0 et le pic vert aux VAFs composant le cluster 1.

— $c^B = n_{\text{min}_1} = 0$: Non définit (division par 0)
— $c^B = n_{\text{maj}_1} = 2$:

$$\frac{2 \times 0.204393}{2 + 0.204393 \times (2-2)} = 0.204393$$

Cluster 1

— $c^B = 1$:

$$\frac{2 \times 0.839333}{1 + 0.839333 \times (2-2)} = 1.678667$$

— $c^B = n_{\text{min}_1} = 0$: Non définit (division par 0)
— $c^B = n_{\text{maj}_1} = 2$:

$$\frac{2 \times 0.839333}{2 + 0.839333 \times (2-2)} = 0.839333$$

La valeur 1.678667 est exclue des possibilités puisqu'elle est supérieure à 1.

Après le calcul de l'ensemble des ACFs pour l'ensemble des clusters de loci pour tous les chromosomes, la troisième étape consiste à effectuer un nouveau GMM. La réduction du nombre de composantes au nombre de clusters composés d'individus classés de manière unique et d'au moins 10 loci nous permet alors d'estimer le nombre de populations tumorales composant notre échantillon. Ainsi, dans le cas de la tumeur 4, 3 clusters, ou populations sont identifiées (figure 4.8).

Enfin, à partir de l'ensemble des valeurs d'ACFs calculés dans la seconde étape et un nombre de clusters $k = 3$, chaque SNA est affecté à un cluster via un *clustering k-means*. Cette assignation nous permet ainsi de connaître pour chaque SNA ses valeurs de c^B et c^T associées et alors de calculer les ACFs moyennes pour chacune des populations tumorales.

Les résultats pour l'ensemble des échantillons sont résumés dans le tableau 4.1.

Bien que le nombre de populations semble parfois être sur-évalué, les résultats obtenus sont concordant avec ceux attendus pour la majorité des tumeurs (voir la figure 4.6). Ces résultats ont été classés dans les 5 premières méthodes lors de la soumission de notre algorithme au DREAM challenge sur un nombre de 32 équipes internationales participantes. De nombreuses améliorations pourraient être apportées à ce travail exploratoire. Une des améliorations directe pourrait être de prendre en compte l'hétérogénéité du nombre de copies total, comme expliqué dans la remarque 4.1.2, différemment de l'algorithme actuel.

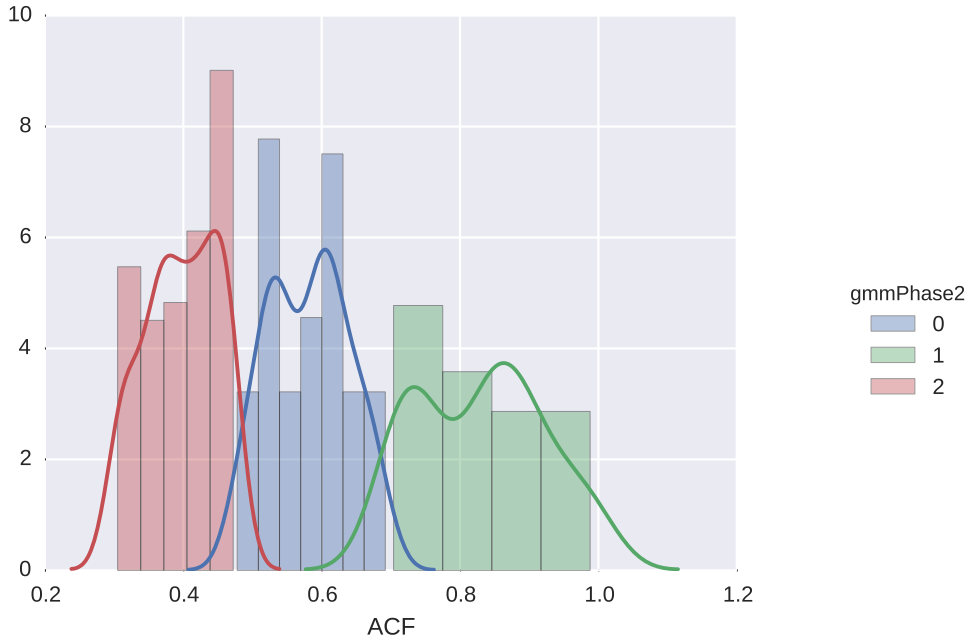


FIGURE 4.8 – *Distribution des ACFs pour l'ensemble des chromosomes.* Les ACFs sont séparés en 3 clusters correspondant aux différentes populations cellulaires : le pic bleu correspond aux ACFs composant le cluster 0, le pic vert aux VAFs composant le cluster 1 et le pic rouge aux VAFs composant le cluster 2.

Tumor ID	**# clusters**	**Cancer cell proportion**	**Population ID**	**ACF**	**# SNAs**
T6	1	0.20	1	0.20	33722
T3	2	0.59	1 2	0.17 0.59	121 621
T2	4	0.88	1 2 3 4	0.33 0.57 0.80 0.88	190 79 299 637
T7	3	0.71	1 2 3	0.11 0.26 0.71	1370 406 64
T4	3	0.83	1 2 3	0.37 0.57 0.83	164 437 60
T1	3	0.92	1 2 3	0.32 0.48 0.92	2162 5259 200

TABLE 4.1 – Résumé des résultats obtenus pour l'ensemble des tumeurs.

Chapitre 5

Discussion

Dans cette thèse, nous nous sommes intéressés au développement de nouvelles méthodes computationnelles d'analyse de données de séquençage d'échantillons tumoraux. Dans le second chapitre, l'objectif était de permettre la détection de SNVs dans un contexte de séquençage ciblé de transcrits pour une cohorte de patientes atteintes de cancers du sein (EORTC 10994). Nous avons vu que l'adaptation des approches classiques permet une détection efficace lorsque le séquençage est effectué avec la technologie NGS 454. En effet, l'établissement d'un score de correspondance entre chaque *read* et chaque transcrit alternatif nous a permis d'identifier des variations de manière transcrit-spécifique en assignant chaque *read* au transcrit maximisant ce score.

Cette approche s'est cependant révélée inefficace lorsque les *reads* provenaient de la technologie PacBio. Nous avons alors développé une nouvelle méthode, MICADo, permettant la distinction entre les SNVs et les autres altérations présentes dans les *reads*. L'originalité de cette méthode est qu'elle ne nécessite pas au préalable d'effectuer un alignement des *reads* sur une séquence de référence. En effet, l'utilisation des graphes de de Bruijn permet de s'affranchir de cette étape connue pour être particulièrement propice aux erreurs en présence d'indels qui sont les erreurs de séquençage prédominantes de la technologie PacBio. De plus, MICADo tire profit du bruit de fond commun de la cohorte et utilise cette information pour effectuer un test statistique distinguant les mutations réelles des altérations systématiques et récurrentes causées par différents biais biologiques et technologiques. La comparaison de notre nouvelle méthode avec deux approches classiques, VarScan et GATK, ainsi que sa validation sur des données PacBio (autres que EORTC 10994) de séquençage ciblé pour un ensemble d'échantillons tumoraux nous conforte dans la pertinence de notre méthode ainsi que dans son utilisation future. Afin d'assurer la généralisation de notre méthode, il est nécessaire de permettre une utilisation facile et intuitive aux biologistes et cliniciens grâce à une interface graphique. Ainsi, le développement d'un *wrapper* Galaxy pourrait répondre à ce besoin.

Dans le troisième chapitre, nous avons vu comment mettre en évidence des changements locaux et globaux en matière de *copy number* intervenus durant une thérapie néo-adjuvante à partir de données de séquençage *Whole Genome* de très faible couverture. De part son faible coût, cette approche pourrait être d'une aide efficace pour l'identification de CNAs avant traitement en vu d'une prédiction de l'évolution tumorale suite au traitement. Nous avons ainsi mis en place une méthodologie d'analyses statistiques robustes post-identification de segments présentant des CNAs et permettant de comparer ces altérations pour des échantillons pairés avant et après traitement. Cette méthode a été validée sur la cohorte HORGEN où nous avons pu identifier les différences entre

les profils de *copy number* avant et après traitement. La question reste ouverte quant à l'origine de ces différences. En effet, ces différences peuvent être d'origine biologique et ainsi être dans ce cas le reflet de l'adaptation de la tumeur au traitement ou être le reflet de l'hétérogénéité intra-tumorale impliquant une sous-sélection de populations cellulaires lors de la biopsie.

Dans le quatrième chapitre, nous avons proposé une formulation du problème de reconstruction clonale. Ce formalisme nous a permis de développer une nouvelle méthode permettant l'identification et la caractérisation des différentes populations cellulaires composant un échantillon tumoral à partir des altérations de type CNA et SNA. Notre approche est basée sur les modèles de mélanges gaussiens et permet de décomposer le signal mutationnel le long des chromosomes en clusters de SNAs mis en correspondance avec un profil de CNA cohérent. Plusieurs pistes d'améliorations sont envisageables. Par exemple, plusieurs états de *copy number*, peuvent être à l'origine des données observées, auquel cas, nous disposons d'un ensemble de solutions équivalentes. Notre méthode recherche les différentes proportions de cellules tumorales et retient une des solutions de façon déterministe. Il pourrait être opportun de développer un meilleur critère de choix de solution afin de retenir la solution maximisant la vraisemblance entre les données observées et la solution de reconstruction proposée. Aussi, il serait intéressant de voir si l'annotation des loci permettrait une meilleure évaluation du nombre de populations. De plus, différents types de *clustering* pourraient être testés aux différentes étapes de l'algorithme. Par ailleurs, afin d'être validée, cette méthode devra être testée sur d'autres données simulées ou réelles tout en faisant varier les différents algorithmes d'identification de SNAs et CNAs afin de s'assurer de l'indépendance des résultats par rapport aux algorithmes utilisés en amont.

Nous avons ainsi vu dans cette thèse des étapes prédominantes dans l'analyse de séquences issues d'échantillons tumoraux : l'identification de variations ponctuelles et structurales (spécifiquement, celles de *copy number*) ainsi que leur mise en commun pour la caractérisation des populations clonales composant ces échantillons. Ainsi notre apport méthodologique s'inscrit dans un domaine de recherche très actif et qui tend à devenir de plus en plus standardisé au grand bénéfice d'applications cliniques.

Néanmoins, le développement de nouvelles technologies permet de mettre en évidence l'impact d'autre types d'altérations impliqué dans le développement tumorale, telles que par exemple les phénomènes épigénétiques. Le développement de méthodes d'analyse de ces données est actuellement en plein essor et nous pouvons envisager que dans quelques années elles aussi contribueront au choix de thérapies adaptées. Afin d'amener l'utilisation de ce type de données en clinique, il est nécessaire de rendre leur analyse robuste et standardisée. L'importante complexité de ces données représente un challenge important pour la recherche en bioinformatique dans ce domaine.

Beaucoup d'autres types de données "omics" moins conventionnelles (proteomique, lipidomique etc.) ainsi que les données d'imagerie mais aussi l'ensemble des méta-données (cliniques, environnementales, liées aux habitudes de vie) sont extrêmement pertinentes pour obtenir une vision complète des pathologies complexes que sont les cancers. Le développement de nouvelles méthodes pour intégrer, analyser et extraire de la connaissance de l'ensemble de ces données hétérogènes est nécessaire pour une caractérisation toujours plus fine du cancer dans l'objectif de proposer des thérapies efficaces s'inscrivant dans la perspective de la médecine personnalisée.

Chapitre 6

Annexes

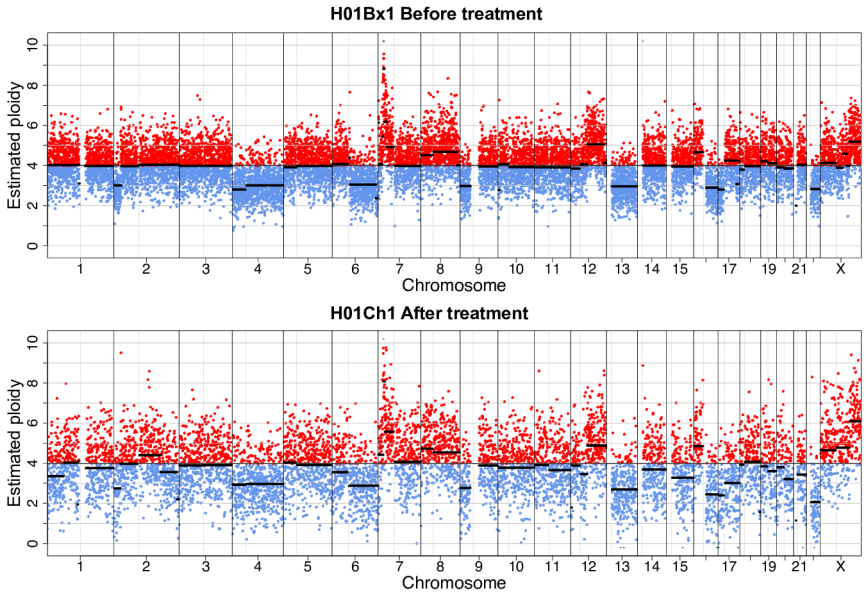
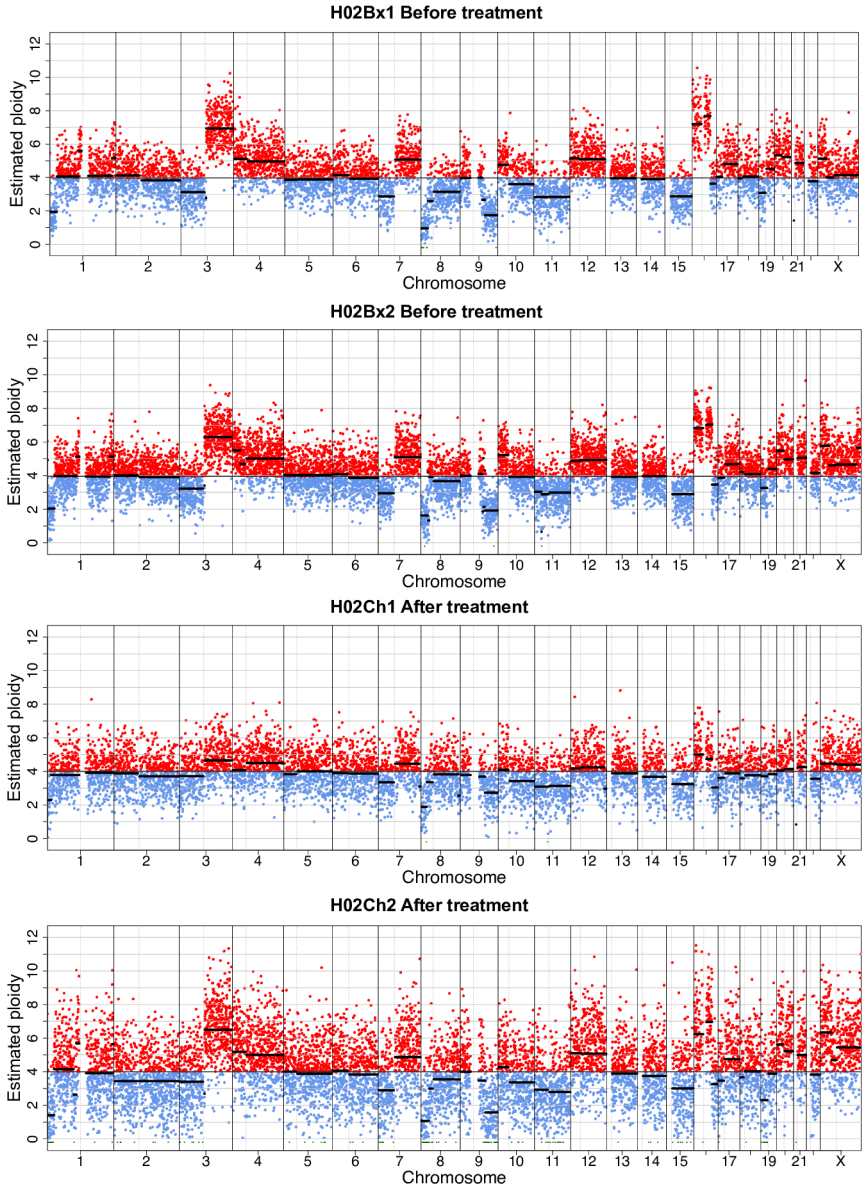
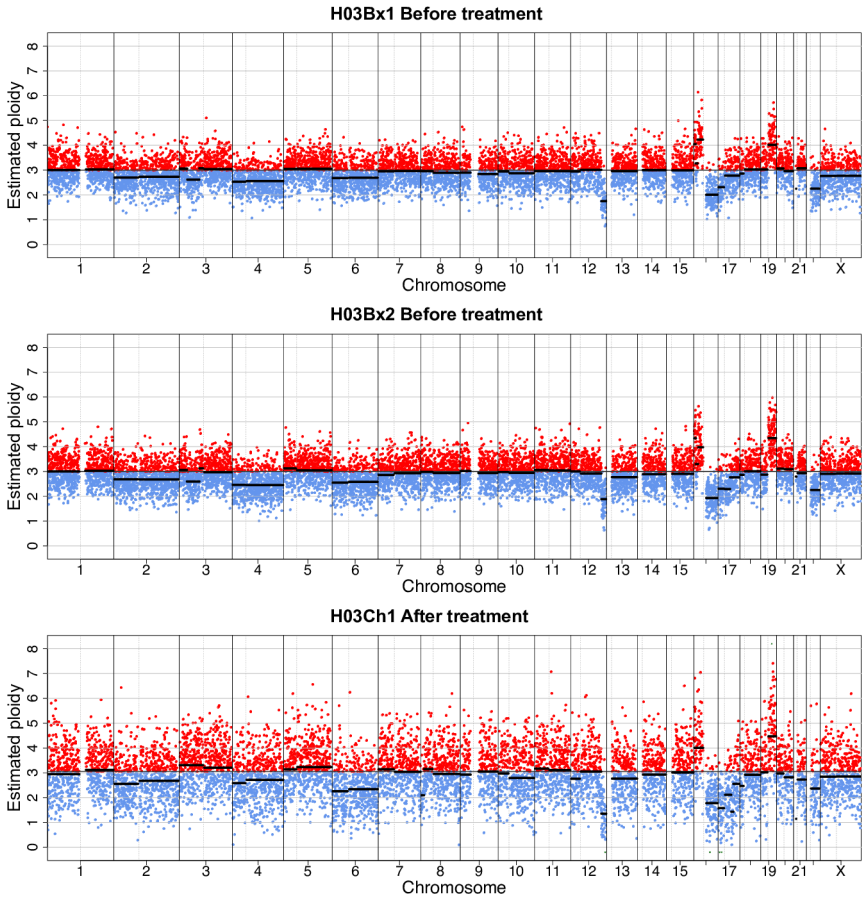
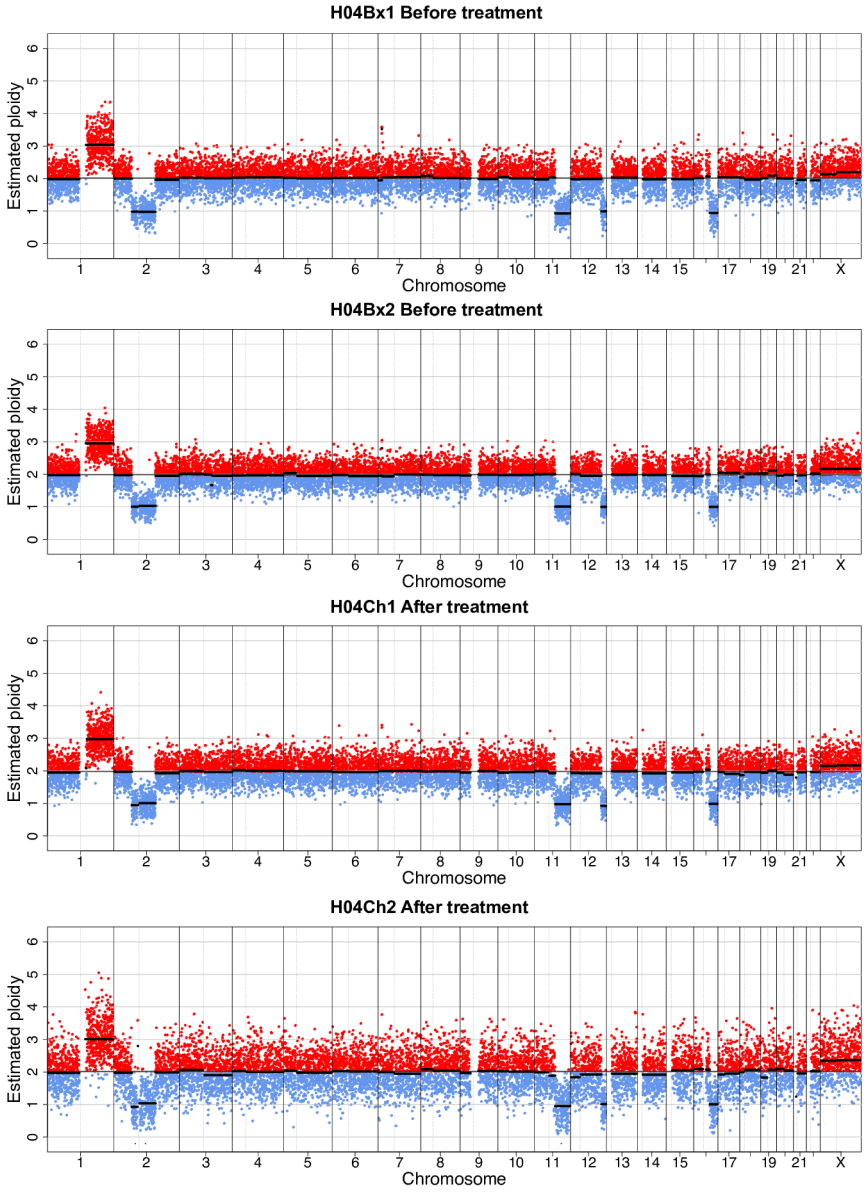
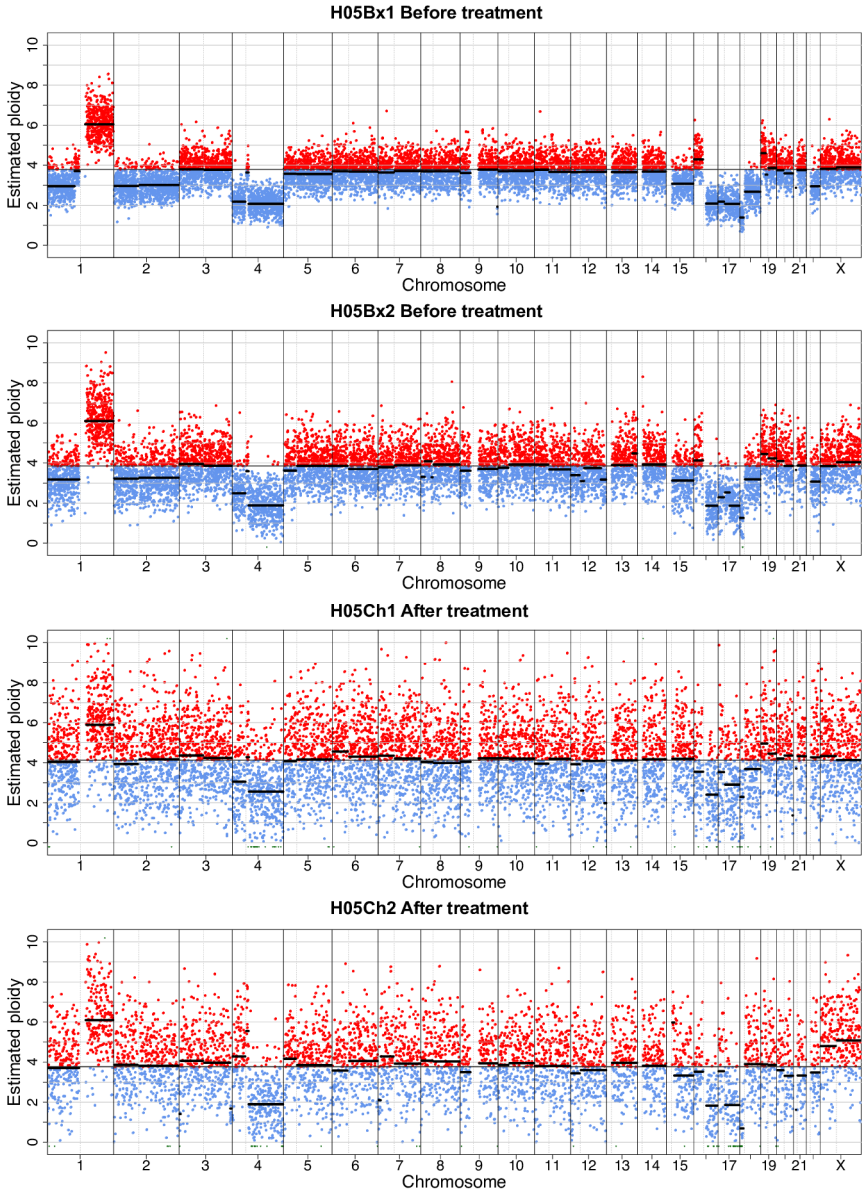


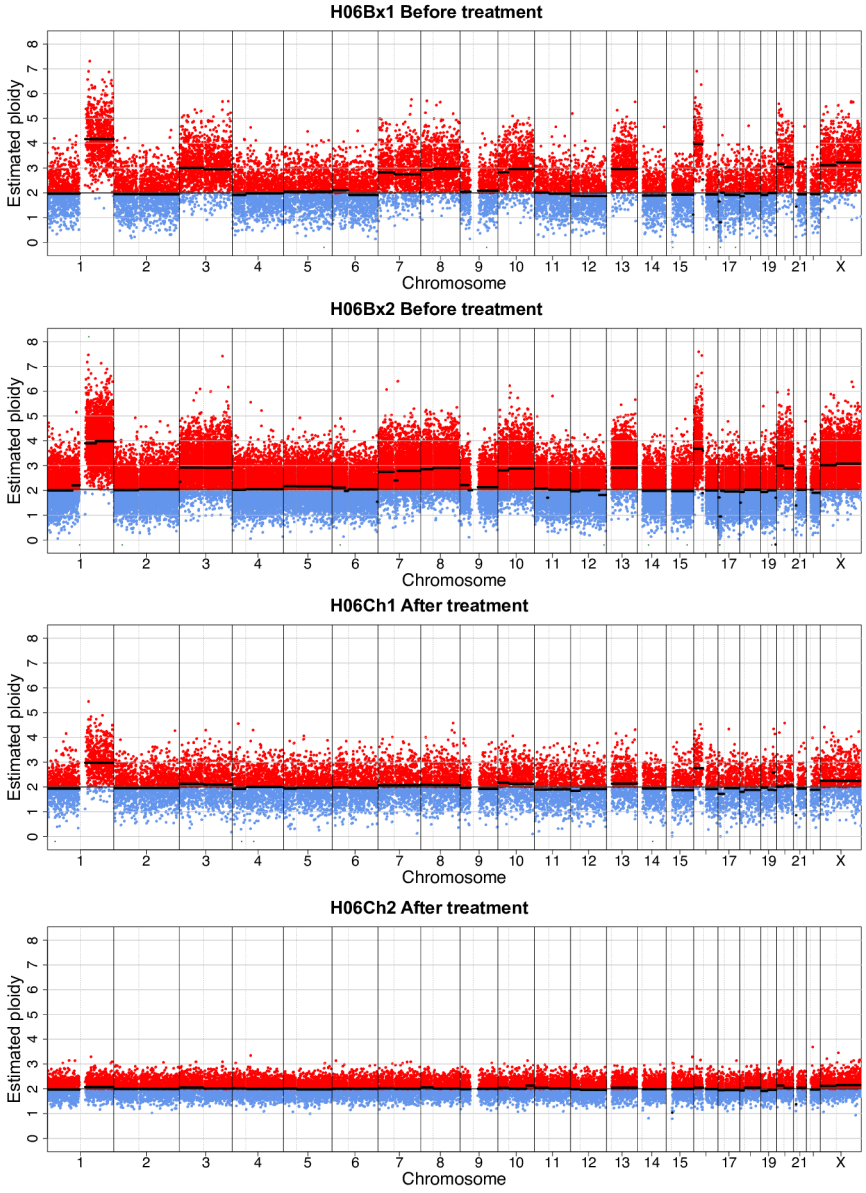
FIGURE 6.1 – *Profils génomiques de l'ensemble des tumeurs de l'étude.* Pour faciliter l'identification des changements, le nombre de copies calculé avec CNAnorm a été ajusté pour donner la même ploïdie modale avant et après traitement.

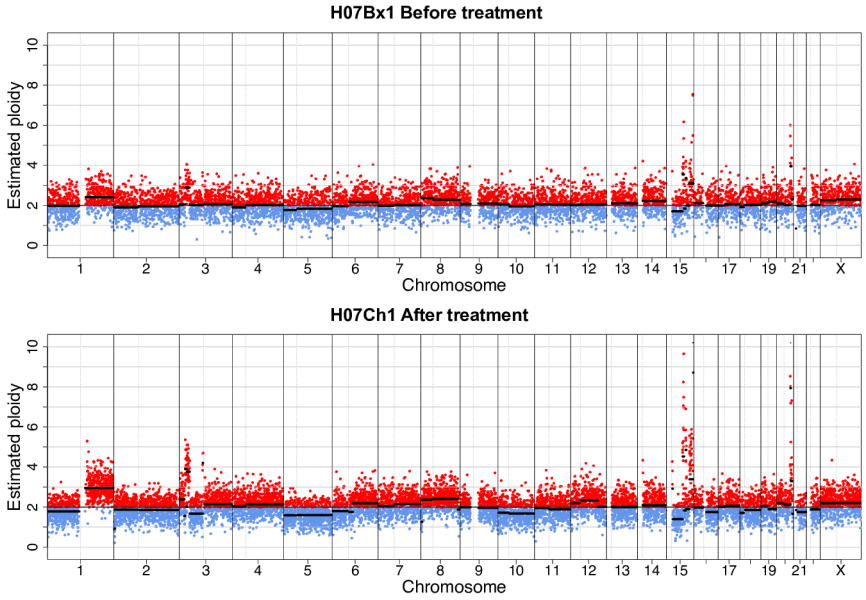

H02Bx1 Before treatment
Estimated ploidy
0
2
4
6
8
10
12
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome
H02Bx2 Before treatment
Estimated ploidy
Chromosome
H02Ch1 After treatment
Estimated ploidy
Chromosome
H02Ch2 After treatment
Estimated ploidy
Chromosome

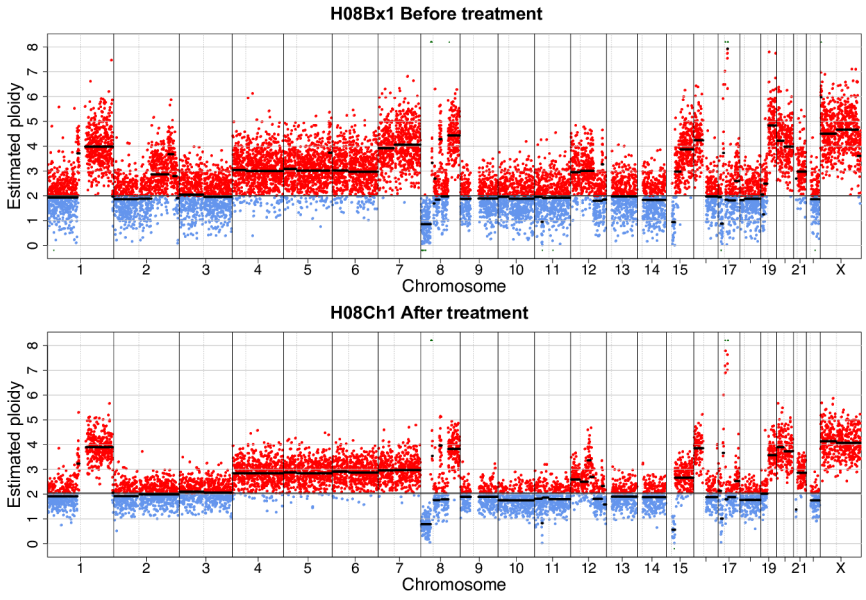

H03Bx1 Before treatment
Estimated ploidy
0 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome
H03Bx2 Before treatment
Estimated ploidy
0 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome
H03Ch1 After treatment
Estimated ploidy
0 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome

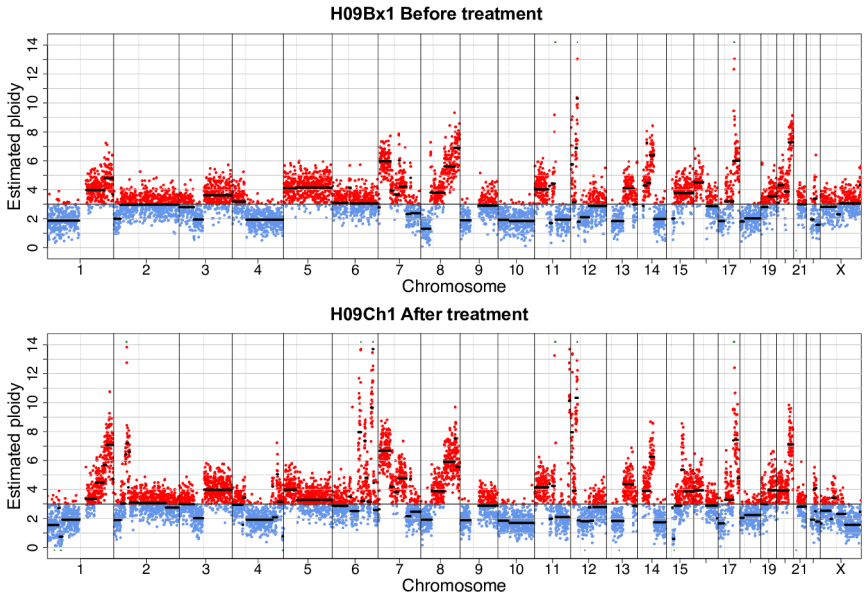

H04Bx1 Before treatment
Estimated ploidy
0 1 2 3 4 5 6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome
H04Bx2 Before treatment
Estimated ploidy
0 1 2 3 4 5 6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome
H04Ch1 After treatment
Estimated ploidy
0 1 2 3 4 5 6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome
H04Ch2 After treatment
Estimated ploidy
0 1 2 3 4 5 6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome

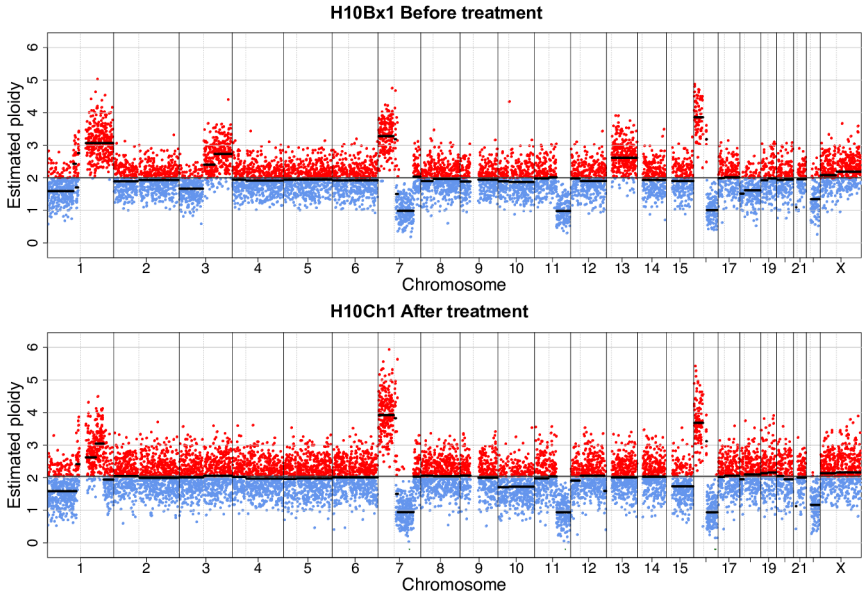

H05Bx1 Before treatment
Estimated ploidy
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome
H05Bx2 Before treatment
Estimated ploidy
Chromosome
H05Ch1 After treatment
Estimated ploidy
Chromosome
H05Ch2 After treatment
Estimated ploidy
Chromosome

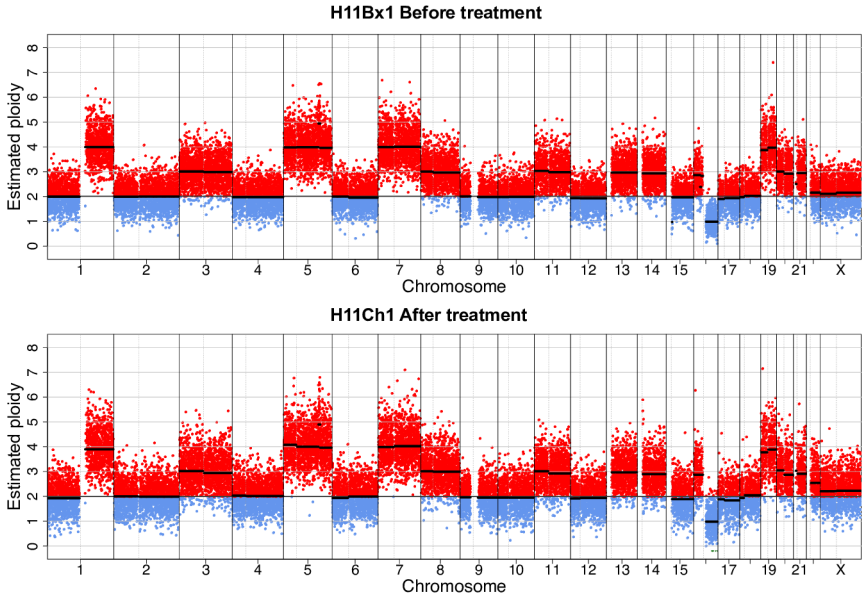

H06Bx1 Before treatment
Estimated ploidy
0 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome
H06Bx2 Before treatment
Estimated ploidy
0 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome
H06Ch1 After treatment
Estimated ploidy
0 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome
H06Ch2 After treatment
Estimated ploidy
0 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome

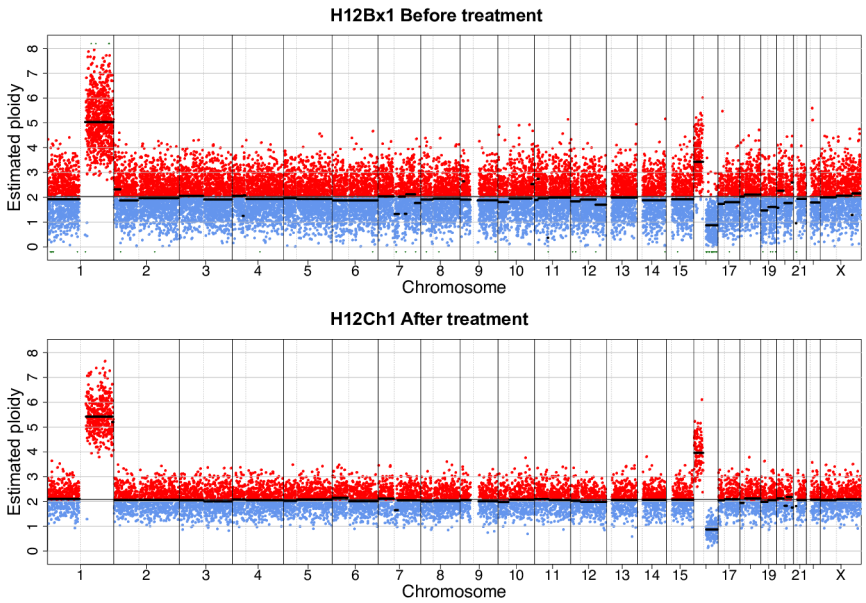

H07Bx1 Before treatment
Estimated ploidy
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome
H07Ch1 After treatment
Estimated ploidy
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome

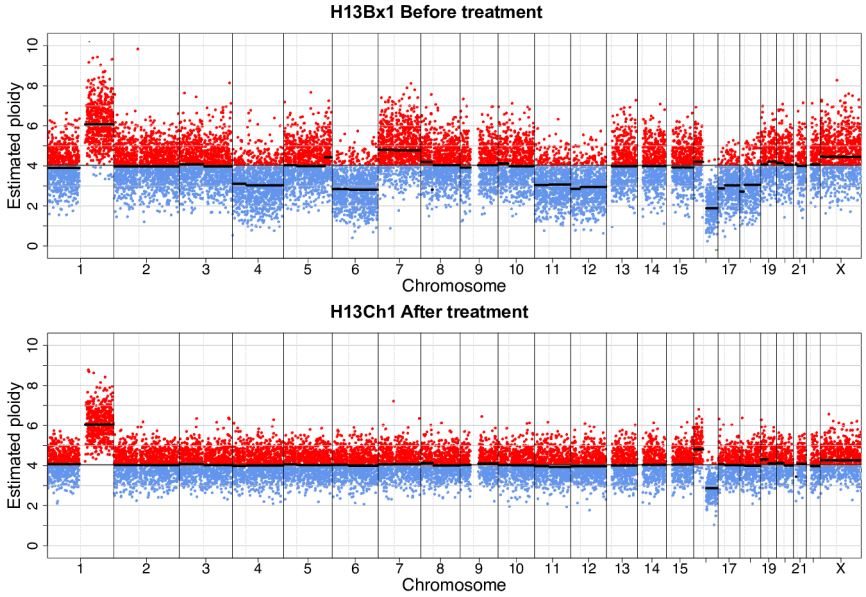

H08Bx1 Before treatment
Estimated ploidy
0
1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome
H08Ch1 After treatment
Estimated ploidy
0
1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome

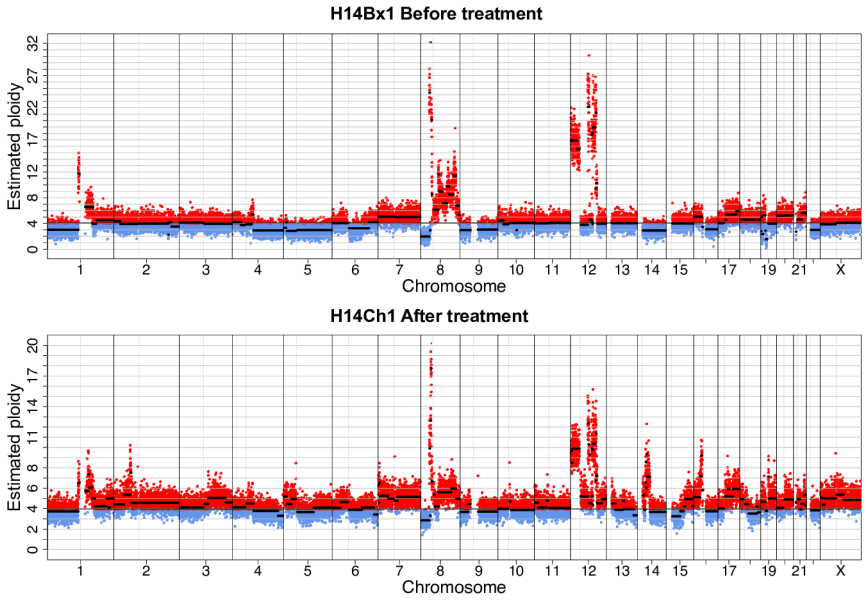

H09Bx1 Before treatment
Estimated ploidy
0 2 4 6 8 10 12 14
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome
H09Ch1 After treatment
Estimated ploidy
0 2 4 6 8 10 12 14
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome

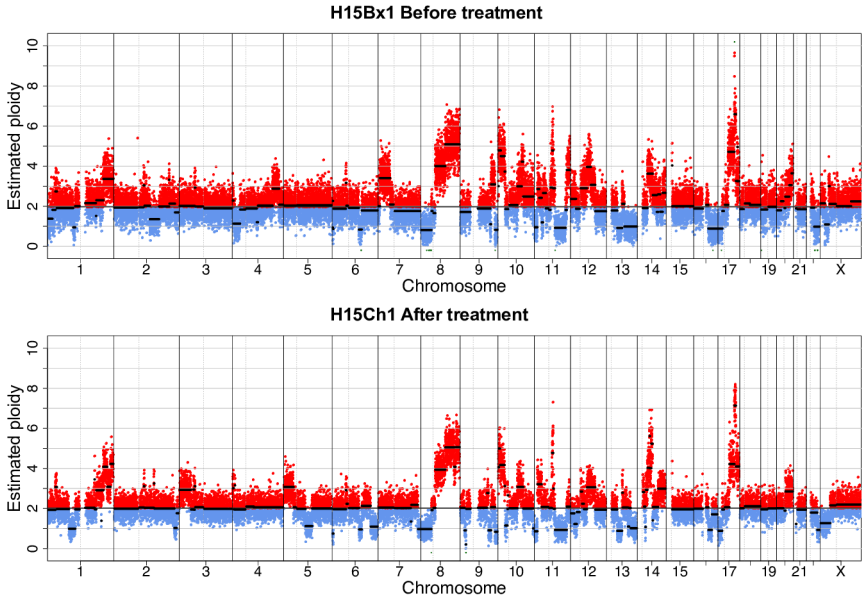

H10Bx1 Before treatment
Estimated ploidy
0 1 2 3 4 5 6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome
H10Ch1 After treatment
Estimated ploidy
0 1 2 3 4 5 6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome

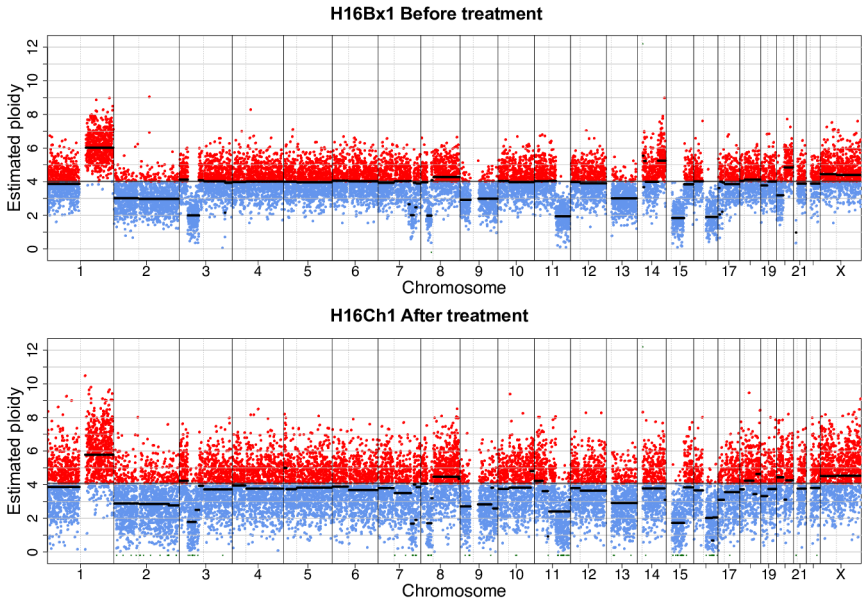

H11Bx1 Before treatment
Estimated ploidy
0 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome
H11Ch1 After treatment
Estimated ploidy
0 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome

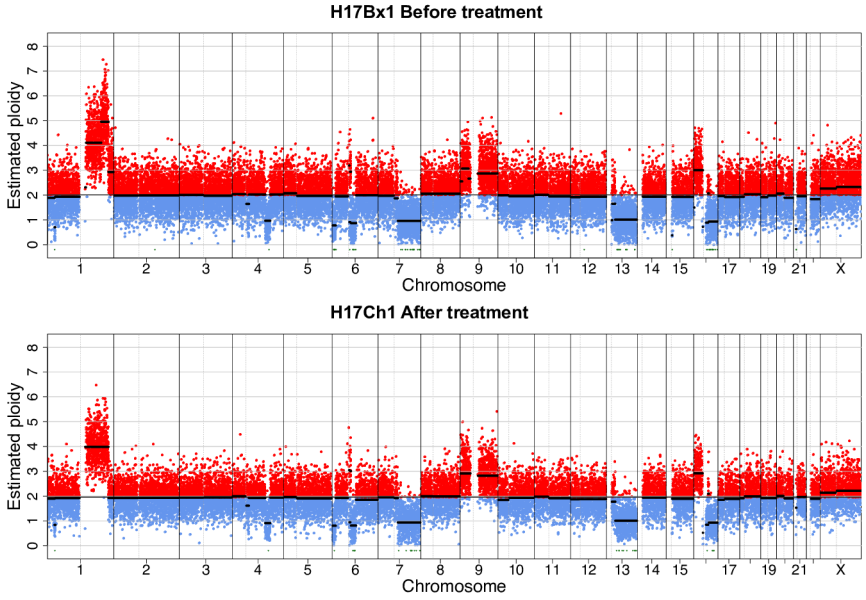

H12Bx1 Before treatment
Estimated ploidy
0 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome
H12Ch1 After treatment
Estimated ploidy
0 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome

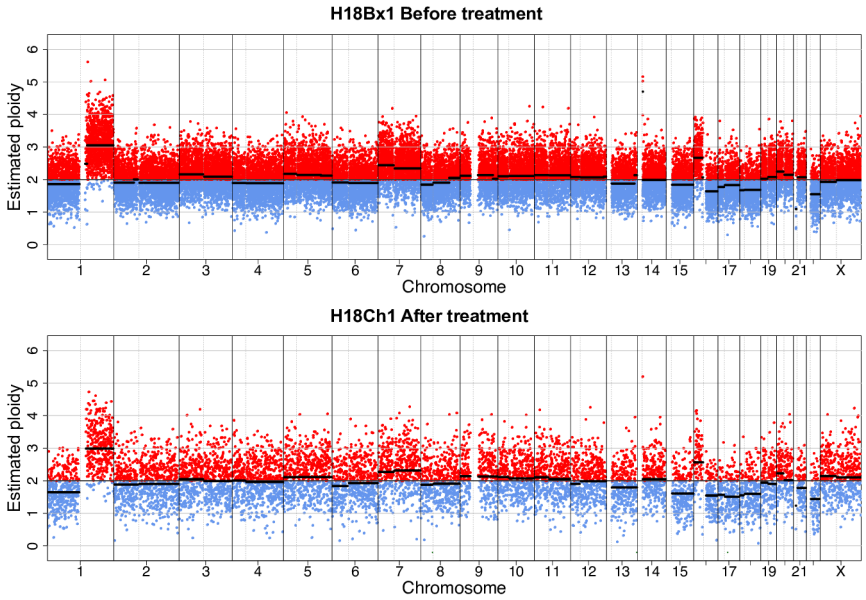

H13Bx1 Before treatment
Estimated ploidy
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome
H13Ch1 After treatment
Estimated ploidy
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome

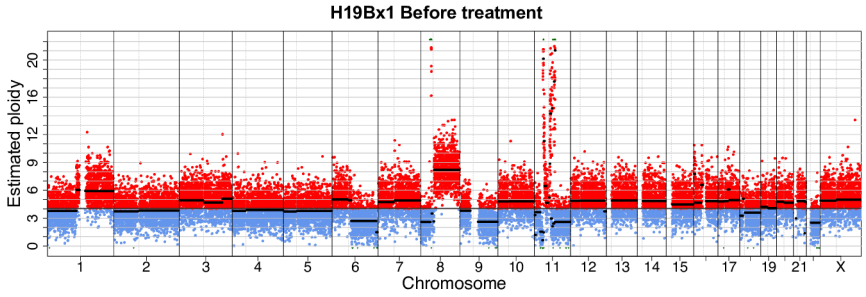

H14Bx1 Before treatment
Estimated ploidy
0
4
8
12
17
22
27
32
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome
H14Ch1 After treatment
Estimated ploidy
0
2
4
6
8
11
14
17
20
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome

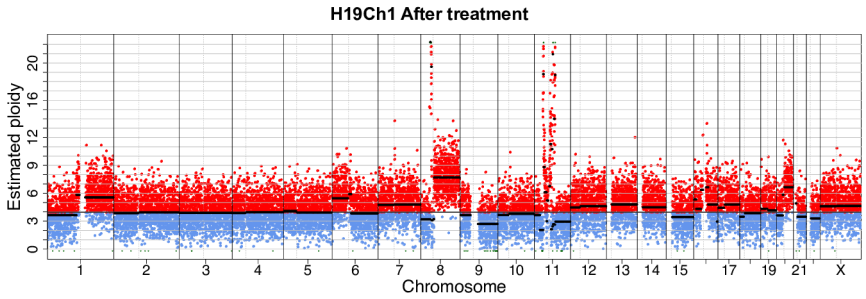

H15Bx1 Before treatment
Estimated ploidy
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome
H15Ch1 After treatment
Estimated ploidy
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome

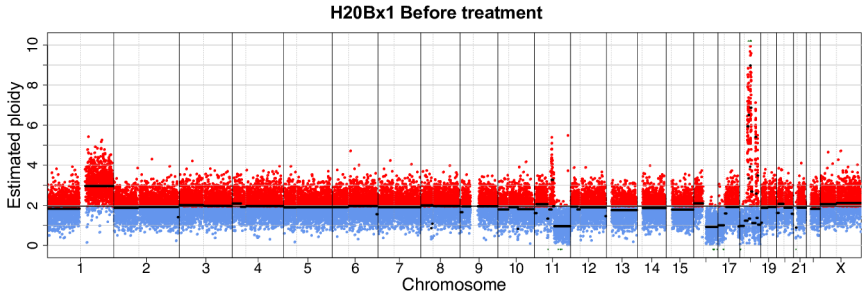

H16Bx1 Before treatment
Estimated ploidy
0
2
4
6
8
10
12
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome
H16Ch1 After treatment
Estimated ploidy
0
2
4
6
8
10
12
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome

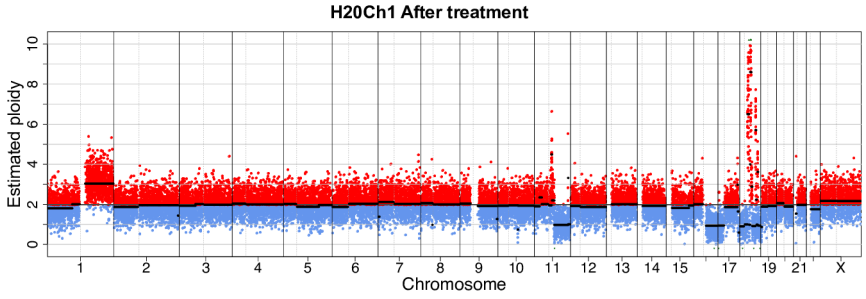

H17Bx1 Before treatment
Estimated ploidy
0 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome
H17Ch1 After treatment
Estimated ploidy
0 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome


H18Bx1 Before treatment
Estimated ploidy
0 1 2 3 4 5 6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome
H18Ch1 After treatment
Estimated ploidy
0 1 2 3 4 5 6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21 X
Chromosome


H19Bx1 Before treatment
Estimated ploidy
0
3
6
9
12
16
20
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome


H19Ch1 After treatment
Estimated ploidy
0
3
6
9
12
16
20
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome


H20Bx1 Before treatment
Estimated ploidy
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome


H20Ch1 After treatment
Estimated ploidy
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X
Chromosome

sample	PT	#read	λ	method	adjust	addP	maxR	exclude
H01Bx1	80	50	3	density	1.5	+2	5	2,5-7,9-12,14,15,17-22,M,Y
H01Ch1	50	100	3	density	1.5	+2	5	2,5-7,9-12,14,15,17-22,M,Y
H02Bx1	60	100	5	density	1	+2	6	1-3,7-10,14-22,M,Y
H02Ch1	25	100	2	density	1.25	+2	6	1-3,7,8,10-14,16-22,M,Y
H02Bx2	60	150	5	density	1	+2	6	1-3,7-10,14-22,M,Y
H02Ch2	50	150	15	density	1.25	+2	6	1-3,7-10,14-22,M,Y
H03Bx1	70	75	5	density	2	+1	4	2,6,7-11,13-15,17-22,M,Y
H03Ch1	70	75	5	density	2	+1	4	1,2,6-11,13-15,17-22,M,Y
H03Bx2	50	75	3	density	2	+1	4	2,6-11,13-15,17-22,M,Y
H04Bx1	90	75	3	density	2	+0	3	M,Y
H04Ch1	90	75	3	density	2	+0	3	M,Y
H04Bx2	80	75	3	density	3	+0	3	M,Y
H04Ch2	50	75	3	density	4	+0	3	M,Y
H05Bx1	90	75	3	density	1.5	+2	5	2,5-11,13-22,M,Y
H05Ch1	80	150	9	density	1.5	+2	5	2,5,7-11,13-15,20-22,M,Y
H05Bx2	60	75	7	density	1	+2	5	2,5-11,13-22,M,Y
H05Ch2	70	150	10	density	1	+2	5	2,5-11,13-22,M,Y
H06Bx1	70	50	1	density	1.25	+0	4	M,Y
H06Ch1	70	50	1	density	1.25	+0	4	M,Y
H06Bx2	60	50	1	density	1.25	+0	4	M,Y
H06Ch2	40	50	/	mode	/	+0	4	M,Y
H07Bx1	50	75	2	density	2	+0	5	4-22,M,Y
H07Ch1	50	75	1	density	1.5	+0	5	M,Y
H08Bx1	70	75	2	density	1	+1	4	2,5,8-22,M,Y
H08Ch1	50	75	2	density	1	+1	4	2,5,8-22,M,Y
H09Bx1	90	75	3	density	1.25	+1	7	M,Y
H09Ch1	90	75	3	density	1.25	+0	7	M,Y
H10Bx1	80	50	3	density	3	+0	3	1,3-10,12,14-22,M,Y
H10Ch1	60	50	1	density	3	+0	3	M,Y
H11Bx1	70	50	3	density	1.25	+0	4	M,Y
H11Ch1	70	50	3	density	1.25	+0	4	M,Y
H12Bx1	50	50	3	density	1	+0	4	2-14,17-22,M,Y
H12Ch1	80	50	2	density	3	+0	4	1-14,17,18-22,M,Y
H13Bx1	70	75	3	density	1	+2	5	M,Y
H13Ch1	60	75	1	density	2	+2	5	M,Y
H14Bx1	80	75	3	density	1.25	+2	16	1-4,6,8,9,12,13,17-22,M,Y
H14Ch1	70	150	3	density	1.25	+2	10	1-4,6,8,9,12,13,17-22,M,Y
H15Bx1	70	75	3	density	3	+0	5	M,Y
H15Ch1	60	75	3	density	3	+0	5	M,Y
H16Bx1	90	75	3	density	1.25	+2	6	M,Y
H16Ch1	40	150	7	density	1.5	+1	6	3-8,11,12,14-22,M,Y
H17Bx1	60	50	1	density	3	+0	4	M,Y
H17Ch1	90	50	3	density	10	+0	4	M,Y
H18Bx1	60	50	3	density	1.5	+0	3	M,Y
H18Ch1	40	50	3	density	2	+0	3	M,Y
H19Bx1	80	50	3	density	1.5	+3	11	M,Y
H19Ch1	50	50	4	density	1.5	+2	11	M,Y
H20Bx1	60	50	3	density	2	+0	5	M,Y
H20Ch1	50	50	3	density	2	+0	5	M,Y

TABLE 6.1 – *Résumé des paramètres utilisés pour l'obtention de profils génomiques.* sample : échantillon considéré ; PT : pourcentage de cellules tumorales ; #reads : nombre de *reads* moyen dans chaque fenêtre génomique ; λ : paramètre de lissage des ratios corrigés (voir 3.2.2), étape 2 ; method : méthode utilisée pour identifier les ratios provenant d'un locus diploïde et le pourcentage de tissus tumoral ; adjust : multiplicateur du paramètre *bandwith* (ou taille de fenêtre) durant l'estimation de la densité des pics lorsque la méthode utilisée est "density" ; addP : ajustement de la ploïdie majoritaire (si différente de 2) ; maxR : ratio maximum à afficher sur le profil génomique ; exclude : liste des chromosomes exclus pour estimé la ploïdie majoritaire.

Bibliographie

Abel, Haley J et Duncavage, Eric J, 2013. Detection of structural dna variation from next generation sequencing data : a review of informatic approaches. *Cancer genetics*, 206(12) :432–440.

Alkan, Can, Coe, Bradley P et Eichler, Evan E, 2011. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5) :363–376.

Baker, Monya, 2011. Sorting out sequencing data. *Nature methods*, 8(10) :799.

Barillot, Emmanuel, Calzone, Laurence, Hupe, Philippe, Vert, Jean-Philippe et Zinovyev, Andrei, 2012. *Computational systems biology of cancer*. CRC Press.

Beerenwinkel, Niko, Schwarz, Roland F, Gerstung, Moritz et Markowetz, Florian, 2015. Cancer evolution : mathematical models and computational inference. *Systematic biology*, 64(1) :e1–e25.

Boeva, Valentina, Popova, Tatiana, Bleakley, Kevin, Chiche, Pierre, Cappo, Julie, Schleiermacher, Gudrun, Janoueix-Lerosey, Isabelle, Delattre, Olivier et Barillot, Emmanuel, 2012. Control-freec : a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, 28(3) :423–425.

Bonnefoi, Hervé, Piccart, Martine, Bogaerts, Jan, Mauriac, Louis, Fumoleau, Pierre, Brain, Etienne, Petit, Thierry, Rouanet, Philippe, Jassem, Jacek, Blot, Emmanuel *et al.*, 2011. Tp53 status for prediction of sensitivity to taxane versus non-taxane neoadjuvant chemotherapy in breast cancer (eortc 10994/big 1-00) : a randomised phase 3 trial. *The lancet oncology*, 12(6) :527–539.

Carter, Scott L, Cibulskis, Kristian, Helman, Elena, McKenna, Aaron, Shen, Hui, Zack, Travis, Laird, Peter W, Onofrio, Robert C, Winckler, Wendy, Weir, Barbara A *et al.*, 2012. Absolute quantification of somatic dna alterations in human cancer. *Nature biotechnology*, 30(5) :413–421.

Chelala, Claude, Khan, Arshad et Lemoine, Nicholas R, 2009. SNPnexus : a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, 25(5) :655–661.

Chin, Cheen Fei et Yeong, Foong May, 2010. Safeguarding entry into mitosis : the antephase checkpoint. *Molecular and cellular biology*, 30(1) :22–32.

Chin, Chen-Shan, Sorenson, Jon, Harris, Jason B, Robins, William P, Charles, Richelle C, Jean-Charles, Roger R, Bullard, James, Webster, Dale R, Kasarskis, Andrew, Peluso, Paul *et al.*, 2011. The origin of the Haitian cholera outbreak strain. *New England Journal of Medicine*, 364(1) :33–42.

Cibulskis, Kristian, Lawrence, Michael S, Carter, Scott L, Sivachenko, Andrey, Jaffe, David, Sougnez, Carrie, Gabriel, Stacey, Meyerson, Matthew, Lander, Eric S et Getz, Gad, 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3) :213–219.

Cleveland, William S, 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368) :829–836.

Compeau, Phillip EC, Pevzner, Pavel A et Tesler, Glenn, 2011. How to apply de Bruijn graphs to genome assembly. *Nature biotechnology*, 29(11) :987–991.

de Hoon, Michiel JL, Imoto, Seiya, Nolan, John et Miyano, Satoru, 2004. Open source clustering software. *Bioinformatics*, 20(9) :1453–1454.

DePristo, Mark A, Banks, Eric, Poplin, Ryan, Garimella, Kiran V, Maguire, Jared R, Hartl, Christopher, Philippakis, Anthony A, Del Angel, Guillermo, Rivas, Manuel A, Hanna, Matt *et al.*, 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5) :491–498.

Ding, Li, Wendl, Michael C, McMichael, Joshua F et Raphael, Benjamin J, 2014. Expanding the computational toolbox for mining cancer genomes. *Nature Reviews Genetics*, 15(8) :556–570.

Easton, Douglas F et Eeles, Rosalind A, 2008. Genome-wide association studies in cancer. *Human Molecular Genetics*, 17(R2) :R109–R115.

Eid, John, Fehr, Adrian, Gray, Jeremy, Luong, Khai, Lyle, John, Otto, Geoff, Peluso, Paul, Rank, David, Baybayan, Primo, Bettman, Brad *et al.*, 2009. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910) :133–138.

Feuk, Lars, Carson, Andrew R et Scherer, Stephen W, 2006. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2) :85–97.

Fonseca, Nuno A, Rung, Johan, Brazma, Alvis et Marioni, John C, 2012. Tools for mapping high-throughput sequencing data. *Bioinformatics*, page bts605.

Gerstung, Moritz, Beisel, Christian, Rechsteiner, Markus, Wild, Peter, Schraml, Peter, Moch, Holger et Beerenwinkel, Niko, 2012. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature communications*, 3 :811.

Gilliland, D. Gary et Griffin, James D., 2002. The roles of FLT3 in hematopoiesis and leukemia. *Blood*, 100(5) :1532–1542. doi :10.1182/blood-2002-02-0492.

Gordon, A et Hannon, GJ, 2010. Fastx-toolkit. *FASTQ/A short-reads preprocessing tools (unpublished) http ://hannonlab. cshl. edu/fastx_ toolkit.*

Gordon, David, Abajian, Chris et Green, Phil, 1998. Consed : a graphical tool for sequence finishing. *Genome research*, 8(3) :195–202.

Gronbaek, Kirsten, Hother, Christoffer et Jones, Peter A, 2007. Epigenetic changes in cancer. *Apmis*, 115(10) :1039–1059.

GUSNANTO, Arief, WOOD, Henry M, PAWITAN, Yudi, RABBITTS, Pamela et BERRI, Stefano, 2012. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics*, 28(1) :40–47.

HANAHAN, Douglas et WEINBERG, Robert A, 2000. The hallmarks of cancer. *cell*, 100(1) :57–70.

HANAHAN, Douglas et WEINBERG, Robert A, 2011. Hallmarks of cancer : the next generation. *cell*, 144(5) :646–674.

HARISMENDY, Olivier, NG, Pauline C, STRAUSBERG, Robert L, WANG, Xiaoyun, STOCKWELL, Timothy B, BEESON, Karen Y, SCHORK, Nicholas J, MURRAY, Sarah S, TOPOL, Eric J, LEVY, Samuel *et al.*, 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome biology*, 10(3) :R32.

HOFMANN, Ariane L, BEHR, Jonas, SINGER, Jochen, KUIPERS, Jack, BEISEL, Christian, SCHRAML, Peter, MOCH, Holger et BEERENWINKEL, Niko, 2017. Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers. *BMC Bioinformatics*, 18(1) :8.

HUANG, Jian, GUSNANTO, Arief, O'SULLIVAN, Kathleen, STAAF, Johan, BORG, Åke et PAWITAN, Yudi, 2007. Robust smooth segmentation approach for array cgh data analysis. *Bioinformatics*, 23(18) :2463–2469.

HUDSON, Thomas J, ANDERSON, Warwick, ARETZ, Axel, BARKER, Anna D, BELL, Cindy, BERNABÉ, Rosa R, BHAN, MK, CALVO, Fabien, EEROLA, Iiro, GERHARD, Daniela S *et al.*, 2010. International network of cancer genome projects. *Nature*, 464(7291) :993–998.

IGGO, Richard, RUDEWICZ, Justine, MONCEAU, Elodie, SEVENET, Nicolas, BERGH, Jonas, SJOBLOM, Tobias et BONNEFOI, Hervé, 2013. Validation of a yeast functional assay for p53 mutations using clonal sequencing. *The Journal of Pathology*, 231(4) :441–448. doi :10.1002/path.4243.
URL `http://dx.doi.org/10.1002/path.4243`

IQBAL, Zamin, CACCAMO, Mario, TURNER, Isaac, FLICEK, Paul et MCVEAN, Gil, 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics*, 44(2) :226–232.

JIANG, Yuchao, QIU, Yu, MINN, Andy J et ZHANG, Nancy R, 2016. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proceedings of the National Academy of Sciences*, 113(37) :E5528–E5537.

KATO, Shunsuke, HAN, Shuang-Yin, LIU, Wen, OTSUKA, Kazunori, SHIBATA, Hiroyuki, KANAMARU, Ryunosuke et ISHIOKA, Chikashi, 2003. Understanding the function–structure and function–mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proceedings of the National Academy of Sciences*, 100(14) :8424–8429.

KIM, Min Sung, SONG, Sang Yong, LEE, Ji Youl, YOO, Nam Jin et LEE, Sug Hyung, 2011. Expressional and mutational analyses of atg5 gene in prostate cancers. *Apmis*, 119(11) :802–807.

KNUDSON, Alfred G, 1971. Mutation and cancer : statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4) :820–823.

KOBOLDT, Daniel C, CHEN, Ken, WYLIE, Todd, LARSON, David E, MCLELLAN, Michael D, MARDIS, Elaine R, WEINSTOCK, George M, WILSON, Richard K et DING, Li, 2009. Varscan : variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17) :2283–2285.

KOBOLDT, Daniel C, ZHANG, Qunyuan, LARSON, David E, SHEN, Dong, MCLELLAN, Michael D, LIN, Ling, MILLER, Christopher A, MARDIS, Elaine R, DING, Li et WILSON, Richard K, 2012. Varscan 2 : somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3) :568–576.

LACAVE, Roger, LARSEN, Christian Jacques et ROBERT, Jacques, 2005. *Cancérologie fondamentale*. John Libbey Eurotext.

LANGMEAD, Ben, TRAPNELL, Cole, POP, Mihai et SALZBERG, Steven L, 2009. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3) :R25.

LAVINSKY, Robert M, JEPSEN, Kristen, HEINZEL, Thorsten, TORCHIA, Joseph, MULLEN, Tina-Marie, SCHIFF, Rachel, DEL-RIO, Alfonso Leon, RICOTE, Mercedes, NGO, Sally, GEMSCH, Joslin *et al.*, 1998. Diverse signaling pathways modulate nuclear receptor recruitment of n-cor and smrt complexes. *Proceedings of the National Academy of Sciences*, 95(6) :2920–2925.

LE, Si Quang et DURBIN, Richard, 2011. Snp detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome research*, 21(6) :952–960.

LEBLANC, Veronique G et MARRA, Marco A, 2015. Next-generation sequencing approaches in cancer : where have they brought us and where will they take us ? *Cancers*, 7(3) :1925–1958.

LEY, Timothy J, MARDIS, Elaine R, DING, Li, FULTON, Bob, MCLELLAN, Michael D, CHEN, Ken, DOOLING, David, DUNFORD-SHORE, Brian H, MCGRATH, Sean, HICKENBOTHAM, Matthew *et al.*, 2008. Dna sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, 456(7218) :66–72.

LI, Heng, 2011. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21) :2987–2993.

LI, Heng et DURBIN, Richard, 2009. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14) :1754–1760.

LI, Heng, HANDSAKER, Bob, WYSOKER, Alec, FENNELL, Tim, RUAN, Jue, HOMER, Nils, MARTH, Gabor, ABECASIS, Goncalo, DURBIN, Richard *et al.*, 2009. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16) :2078–2079.

Li, Heng et Homer, Nils, 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5) :473–483.

Liu, Biao, Morrison, Carl D, Johnson, Candace S, Trump, Donald L, Qin, Maochun, Conroy, Jeffrey C, Wang, Jianmin et Liu, Song, 2013. Computational methods for detecting copy number variations in cancer genome using next generation sequencing : principles and challenges. *Oncotarget*, 4(11) :1868–81.

Liu, Yewei, Yin, Ting, Feng, Yuanbo, Cona, Marlein Miranda, Huang, Gang, Liu, Jianjun, Song, Shaoli, Jiang, Yansheng, Xia, Qian, Swinnen, Johannes V *et al.*, 2015. Mammalian models of chemically induced primary malignancies exploitable for imaging-based preclinical theragnostic research. *Quantitative imaging in medicine and surgery*, 5(5) :708.

Medvedev, Paul, Stanciu, Monica et Brudno, Michael, 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nature methods*, 6 :S13–S20.

Metzker, Michael L, 2010. Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1) :31–46.

Mielczarek, M et Szyda, J, 2016. Review of alignment and snp calling algorithms for next-generation sequencing data. *Journal of applied genetics*, 57(1) :71–79.

Miller, Christopher A, Hampton, Oliver, Coarfa, Cristian et Milosavljevic, Aleksandar, 2011. Readdepth : a parallel r package for detecting copy number alterations from short sequencing reads. *PloS one*, 6(1) :e16327.

Miller, Christopher A, White, Brian S, Dees, Nathan D, Griffith, Malachi, Welch, John S, Griffith, Obi L, Vij, Ravi, Tomasson, Michael H, Graubert, Timothy A, Walter, Matthew J *et al.*, 2014. Sciclone : inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol*, 10(8) :e1003665.

Nielsen, Rasmus, Paul, Joshua S, Albrechtsen, Anders et Song, Yun S, 2011. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6) :443–451.

Nowell, Peter C, 1976. The clonal evolution of tumor cell populations. *Science*, 194(4260) :23–28.

Oesper, Layla, Mahmoody, Ahmad et Raphael, Benjamin J, 2013. Theta : inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome biology*, 14(7) :R80.

Oesper, Layla, Satas, Gryte et Raphael, Benjamin J, 2014. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics*, 30(24) :3532–3540.

Oladimeji, Peter, Skerl, Rebekah, Rusch, Courtney et Diakonova, Maria, 2016. Synergistic activation of erα by estrogen and prolactin in breast cancer cells requires tyrosyl phosphorylation of pak1. *Cancer research*, 76(9) :2600–2611.

Olshen, Adam B, Venkatraman, ES, Lucito, Robert et Wigler, Michael, 2004. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4) :557–572.

O'Rawe, Jason, Jiang, Tao, Sun, Guangqing, Wu, Yiyang, Wang, Wei, Hu, Jingchu, Bodily, Paul, Tian, Lifeng, Hakonarson, Hakon, Johnson, W Evan *et al.*, 2013. Low concordance of multiple variant-calling pipelines : practical implications for exome and genome sequencing. *Genome medicine*, 5(3) :28.

Pabinger, Stephan, Dander, Andreas, Fischer, Maria, Snajder, Rene, Sperk, Michael, Efremova, Mirjana, Krabichler, Birgit, Speicher, Michael R, Zschocke, Johannes et Trajanoski, Zlatko, 2014. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, 15(2) :256–278.

Page, Roderic DM, 2002. Visualizing phylogenetic trees using treeview. *Current Protocols in Bioinformatics*, pages 6–2.

Peng, Quan, Satya, Ravi Vijaya, Lewis, Marcus, Randad, Pranay et Wang, Yexun, 2015. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. *BMC genomics*, 16(1) :589.

Peng, Roger D, 2011. Reproducible research in computational science. *Science*, 334(6060) :1226–1227.

Pevzner, Pavel A, Tang, Haixu et Waterman, Michael S, 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17) :9748–9753.

Piccolo, Stephen R et Frampton, Michael B, 2016. Tools and techniques for computational reproducibility. *GigaScience*, 5(1) :30.

Pirooznia, Mehdi, Goes, Fernando S et Zandi, Peter P, 2015. Whole-genome cnv analysis : advances in computational approaches. *Frontiers in genetics*, 6.

Pirooznia, Mehdi, Kramer, Melissa, Parla, Jennifer, Goes, Fernando, Potash, James, McCombie, W Richard et Zandi, Peter, 2014. Validation and assessment of variant calling pipelines for next-generation sequencing. *Human genomics*, 8(1) :14.

Proia, David A, Nannenga, Bonnie W, Donehower, Lawrence A et Weigel, Nancy L, 2006. Dual roles for the phosphatase ppm1d in regulating progesterone receptor function. *Journal of Biological Chemistry*, 281(11) :7089–7101.

Quenel-Tueux, Nathalie, Debled, Marc, Rudewicz, Justine, MacGrogan, Gaetan, Pulido, Marina, Mauriac, Louis, Dalenc, Florence, Bachelot, Thomas, Lortal, Barbara, Breton-Callu, Christelle *et al.*, 2015. Clinical and genomic analysis of a randomised phase ii study evaluating anastrozole and fulvestrant in postmenopausal patients treated for large operable or locally advanced hormone-receptor-positive breast cancer. *British journal of cancer*, 113(4) :585–594.

Reddy, Karen L et Feinberg, Andrew P, 2013. Higher order chromatin organization in cancer. Dans *Seminars in cancer biology*, tome 23.2, pages 109–115. Elsevier.

ROBASKY, Kimberly, LEWIS, Nathan E et CHURCH, George M, 2014. The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics*, 15(1) :56–62.

ROBINSON, Dan R, WU, Yi-Mi, VATS, Pankaj, SU, Fengyun, LONIGRO, Robert J, CAO, Xuhong, KALYANA-SUNDARAM, Shanker, WANG, Rui, NING, Yu, HODGES, Lynda *et al.*, 2013. Activating esr1 mutations in hormone-resistant metastatic breast cancer. *Nature genetics*, 45(12) :1446–1451.

RUDEWICZ, Justine, SOUEIDAN, Hayssam, URICARU, Raluca, BONNEFOI, Hervé, IGGO, Richard, BERGH, Jonas et NIKOLSKI, Macha, 2016. Micado–looking for mutations in targeted pacbio cancer data : An alignment-free method. *Frontiers in Genetics*, 7.

SCHIRMER, Melanie, IJAZ, Umer Z, D'AMORE, Rosalinda, HALL, Neil, SLOAN, William T et QUINCE, Christopher, 2015. Insight into biases and sequencing errors for amplicon sequencing with the illumina miseq platform. *Nucleic acids research*, page gku1341.

SHUKLA, Sachet A, ROONEY, Michael S, RAJASAGI, Mohini, TIAO, Grace, DIXON, Philip M, LAWRENCE, Michael S, STEVENS, Jonathan, LANE, William J, DELLAGATTA, Jamie L, STEELMAN, Scott *et al.*, 2015. Comprehensive analysis of cancer-associated somatic mutations in class i hla genes. *Nature biotechnology*, 33(11) :1152–1158.

SIMS, David, SUDBERY, Ian, ILOTT, Nicholas E, HEGER, Andreas et PONTING, Chris P, 2014. Sequencing depth and coverage : key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2) :121–132.

SMITH, Catherine C, WANG, Qi, CHIN, Chen-Shan, SALERNO, Sara, DAMON, Lauren E, LEVIS, Mark J, PERL, Alexander E, TRAVERS, Kevin J, WANG, Susana, HUNT, Jeremy P *et al.*, 2012. Validation of ITD mutations in FLT3 as a therapeutic target in human acute myeloid leukaemia. *Nature*, 485(7397) :260–263.

STRATTON, Michael R, CAMPBELL, Peter J et FUTREAL, P Andrew, 2009. The cancer genome. *Nature*, 458(7239) :719–724.

TEO, Shu Mei, PAWITAN, Yudi, KU, Chee Seng, CHIA, Kee Seng et SALIM, Agus, 2012. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, 28(21) :2711–2718.

URICARU, Raluca, RIZK, Guillaume, LACROIX, Vincent, QUILLERY, Elsa, PLANTARD, Olivier, CHIKHI, Rayan, LEMAITRE, Claire et PETERLONGO, Pierre, 2014. Reference-free detection of isolated SNPs. *Nucleic Acids Research*. doi :10.1093/nar/gku1187.

VOGELSTEIN, Bert et KINZLER, Kenneth W, 2004. Cancer genes and the pathways they control. *Nature medicine*, 10(8) :789–799.

WADA-HIRAIKE, O, YANO, T, NEI, T, MATSUMOTO, Y, NAGASAKA, K, TAKIZAWA, S, OISHI, H, ARIMOTO, T, NAKAGAWA, S, YASUGI, T *et al.*, 2005. The dna mismatch repair gene hmsh2 is a potent coactivator of oestrogen receptor α. *British journal of cancer*, 92(12) :2286–2291.

WALERYCH, Dawid, NAPOLI, Marco, COLLAVIN, Licio et DEL SAL, Giannino, 2012. The rebel angel : mutant p53 as the driving oncogene in breast cancer. *Carcinogenesis*, 33(11) :2007–2017.

Wang, Qingguo, Jia, Peilin, Li, Fei, Chen, Haiquan, Ji, Hongbin, Hucks, Donald, Dahlman, Kimberly Brown, Pao, William, Zhao, Zhongming *et al.*, 2013. Detecting somatic point mutations in cancer genome sequencing data : a comparison of mutation callers. *Genome Med*, 5(10) :91.

Wang, Xuefeng, Chen, Mengjie, Yu, Xiaoqing, Pornputtapong, Natapol, Chen, Hao, Zhang, Nancy, Powers, R Scott et Krauthammer, Michael, 2015. Global copy number profiling of cancer genomes. *Bioinformatics*, page btv676.

Warden, Charles D, Adamson, Aaron W, Neuhausen, Susan L et Wu, Xiwei, 2014. Detailed comparison of two popular variant calling packages for exome and targeted exon studies. *PeerJ*, 2 :e600.

Weinberg, Robert, 2013. *The biology of cancer*. Garland science.

Weinstein, John N, Collisson, Eric A, Mills, Gordon B, Shaw, Kenna R Mills, Ozenberger, Brad A, Ellrott, Kyle, Shmulevich, Ilya, Sander, Chris, Stuart, Joshua M, Network, Cancer Genome Atlas Research *et al.*, 2013. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10) :1113–1120.

Weis, Karen E, Ekena, Kirk, Thomas, James A, Lazennec, Gwendal et Katzenellenbogen, Benita S, 1996. Constitutively active human estrogen receptors containing amino acid substitutions for tyrosine 537 in the receptor protein. *Molecular endocrinology*, 10(11) :1388–1398.

Wheeler, David A, Srinivasan, Maithreyan, Egholm, Michael, Shen, Yufeng, Chen, Lei, McGuire, Amy, He, Wen, Chen, Yi-Ju, Makhijani, Vinod, Roth, G Thomas *et al.*, 2008. The complete genome of an individual by massively parallel dna sequencing. *nature*, 452(7189) :872–876.

Wineinger, Nathan E, Kennedy, Richard E, Erickson, Stephen W, Wojczynski, Mary K, Bruder, Carl et Tiwari, Hemant, 2008. Statistical issues in the analysis of dna copy number variations. *International journal of computational biology and drug design*, 1(4) :368–395.

Wodarz, Dominik et Komarova, Natalia, 2005. *Computational biology of cancer : lecture notes and mathematical modeling*. World Scientific.

Wood, Henry M, Belvedere, Ornella, Conway, Caroline, Daly, Catherine, Chalkley, Rebecca, Bickerdike, Melissa, McKinley, Claire, Egan, Phil, Ross, Lisa, Hayward, Bruce *et al.*, 2010. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of dna from formalin-fixed paraffin-embedded specimens. *Nucleic acids research*, 38(14) :e151–e151.

Wu, Thomas D et Watanabe, Colin K, 2005. Gmap : a genomic mapping and alignment program for mrna and est sequences. *Bioinformatics*, 21(9) :1859–1875.

Yadav, Vinod Kumar et De, Subhajyoti, 2014. An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Briefings in bioinformatics*, page bbu002.

Yu, Zhenhua, Liu, Yuanning, Shen, Yi, Wang, Minghui et Li, Ao, 2014. Climat : accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data. *Bioinformatics*, 30(18) :2576–2583.

Zare, Habil, Wang, Junfeng, Hu, Alex, Weber, Kris, Smith, Josh, Nickerson, Debbie, Song, ChaoZhong, Witten, Daniela, Blau, C Anthony et Noble, William Stafford, 2014. Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput Biol*, 10(7) :e1003703.

Zhao, Min, Wang, Qingguo, Wang, Quan, Jia, Peilin et Zhao, Zhongming, 2013. Computational tools for copy number variation (cnv) detection using next-generation sequencing data : features and perspectives. *BMC bioinformatics*, 14(11) :S1.

Zook, Justin M, Chapman, Brad, Wang, Jason, Mittelman, David, Hofmann, Oliver, Hide, Winston et Salit, Marc, 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature biotechnology*, 32(3) :246–251.