Subclonal evolution in neuroblastoma

Paul Deveau

NNT : 2017SACLS140

Thèse de doctorat de l'Université Paris-Saclay préparée à l'Institut Curie

Ecole doctorale n°582
École doctorale de Cancérologie : biologie - médecine - santé
Spécialité de doctorat: Recherche clinique, innovation technologique, santé publique

par

M. Paul Deveau

Évolution sous-clonale dans le neuroblastome

Thèse présentée et soutenue à l'Institut Curie, le 27 Juin 2017.

Composition du Jury :

Mme.	Valentina Boeva	Chargée de recherche Institut Cochin	(Co-Directrice de thèse)
M.	Olivier Delattre	DRCE Institut Curie	(Co-Directeur de thèse)
Mme.	Gudrun Schleiermacher	Practicien Hospitalier Institut Curie	(Co-Directrice de thèse)
M.	Christian Auclair	Professeur des universités Gustave Roussy	(Président du jury)
M.	Fabien Calvo	Professeur des universités - practicien hospitalier Gustave Roussy	(Examinateur)
M.	Jan Koster	Postdoctoral fellow Universiteit van Amsterdam	(Rapporteur)
M.	Hugues Roest Crollius	Directeur de recherche École Normale Supérieure	(Rapporteur)

GUSTAVE ROUSSY
CANCER CAMPUS GRAND PARIS

UNIVERSITÉ PARIS SUD
FACULTÉ DE MÉDECINE

ÉCOLE DOCTORALE DE CANCÉROLOGIE

Titre : Évolution sous-clonale dans le neuroblastome

Mots clefs : Génomique, évolution clonale, neuroblastome

Résumé : Le neuroblastome est le cancer solide extra-cranial le plus fréquent chez l'enfant. Il est caractérisé par une très grande hétérogénéité tant au niveau clinique que moléculaire.
Alors que certains patients rentrent spontanément en rémission, on peut se demander quels facteurs permettent la réémergence du cancer chez d'autres malgré traitement.
Pour répondre à cette question, il convient d'identifier chez les patients ayant rechuté, les différentes populations clonales coexistant au diagnostic et/ou à la rechute. Cela permet, entre autre, d'étudier les voies différemment altérées entre ces deux temps.
Dans cette optique, nous présentons ici QuantumClone, un algorithme de reconstruction clonal à partir de données de séquençage, ainsi que son application à une cohorte de patients souffrant d'un neuroblastome. Sur ces données, l'application de notre méthode a permis d'identifier des différences dans le ratio de variants prédits fonctionnels par rapport à ceux prédits passagers entre les populations ancestrales, enrichies à la rechute ou appauvries à la rechute.

Title : Subclonal evolution in neuroblastoma

Keywords : WGS, clonal evolution, neuroblastoma

Abstract : Neuroblastoma is the most frequent solid extra-cranial cancer of childhood. This cancer displays a high heterogeneity both at clinical and molecular levels.

Even though in some patients spontaneous remission can be observed, some others relapse despite treatment and surgical resection. It may be wondered which are the factors that distinguish these two cases.
In order to answer this question, identification of populations coexisting at diagnosis and/or relapse in the patients which have relapsed is a prerequisite. This would allow, between other things, to study the pathways differently altered in clones that are specific to each time point.
With this in mind, we hereby present QuantumClone, a clonal reconstruction algorithm from sequencing data. In addition, we applied this method to a cohort of patients suffering from neuroblastoma.

On these data, our method identified differences in the functional mutation rate, i.e. the number of putative functional variants by total number of variants, between the ancestral clones, clones expanding at relapse, and clones shrinking at relapse.

Dedicated to all my friends and family, who have born with me during this adventure.

Acknowledgments

Black Knight: 'Tis but a scratch!

King Arthur: A scratch? Your arm's off!

— Monty Python*, Monty Python and the Holy Grail, 1975*

I would like to thank my supervisors Valentina, Olivier and Gudrun for allowing me to join this scientific adventure, providing feedback and keeping me on tracks all this time. I would also like to thank the labs that have welcomed me be it in Curie buildings or Cochin. I could have been a stranger in four groups, but I have always received comments, help, scripts, and much more whenever I needed. For that I am grateful to the Sysbio, RTOP, Génétique et biologie des cancers and Computational Epi-Genetics of Cancer teams! I would also like to express gratitude towards the jury for taking the time to assess this work. I would also like to express my gratitude towards Emmanuel Barillot who has offered me the opportunity to work at the Institut Curie.

This is also the place to thank the Ministère Supérieur de l'Enseignement Supérieur et de la Recherche for funding me and the Université Paris Sud for giving me the opportunity to teach throughout the three years of my PhD.

A special thank to everyone I have interacted with during this project: Angela, Léo, Mathieu, Simon, Isabelle, both Olivier, Sandrine, Benjamin, Cécile, Judith, Marine, Élodie (teaching me bash was not easy every day), Pierre (what would we be without Hadley Wickham?), Victorin, Isabelle, Jennifer, Pauline, Urszula, Gaelle, Laura, Loredana, Stéphane and Arnau. I am sure that I am forgetting people, I'll make sure to make it up to you during the defense if so.

All journeys are made of ups and downs, and this one did not escape. Laurence I owe you for allowing me to rant for hours on the reason(s) why I should quit.

I have already dedicated this work to my friends and family, but it is not enough,

so I'll go further here: I would not have achieved that much without you. Thank you for questioning my choices (a.k.a 'Why would you do research after engineering school? Don't you want to start working already?'), and remind me that life does not stop at the lab's door. Thank you Mathieu, Bertrand, Edmée, Romain, Stéphanie, Marie and Julien.

I will finish by acknowledging my parents, grand-parents and Macha. I don't think I would have made it this far without your support and comments, so this is as much yours as it is mine.

P.S.: All other quotes are much more serious than this one, but I felt compelled to break the emotional burden of the acknowledgments with humor. I will leave the interpretation of this quote to the reader.

Contents

List of Figures

List of Tables

L'ADN se cache depuis des millions d'années dans nos cellules. Nous sommes en train de le dérouler.

— Frank Thilliez*, GATACA, 2011*

L'homme n'est qu'un enfant dans l'échelle de l'évolution. Une bête sauvage qui se croit évoluée.

— Maxime Chattam*, Prédateurs, 2007*

Les lapins courent plus vite que les renards simplement parce qu'ils courent pour survivre.

— Frank Thilliez*, GATACA, 2011*

Abbreviations

A: **A**denine
AIC: **A**kaike **I**nformation **C**riterion
BAF: **B**-**A**llele **F**requency
bp: **b**ase **p**air
BGI: **B**eijing **G**enomic **I**nstitute
BIC: **B**ayesian **I**nformation **C**riterion
ddNTP: **d**i**d**eoxyribo**N**ucleotide **T**ri**P**hoshpate
C: **C**ytosine
CG: **C**omplete **G**enomics
CNA: **C**opy **N**umber **A**berration
CNG: **C**entre **N**ational de **G**énotypage
DNA: **D**eoxyribo**n**ucleic **A**cid
EM: **E**xpectation **M**aximization
GO: **G**ene **O**ntology
HTS: **H**igh **T**hroughput **S**equencing
HMM: **H**idden **M**arkov **M**odel
G: **G**uanine
LoH: **L**oss **o**f **H**eterozygosity
MAD:**M**edian **A**bsolute **D**eviation
MCMC: **M**arkov **C**hain **M**onte **C**arlo
NGS: **N**ext **G**eneration **S**equencing
NMI: **N**ormalized **M**utual **I**nformation
RF: **R**andom **F**orest
T: **T**hymine
TFBS: **T**ranscription **F**actor **B**inding **S**ite
SNV: **S**ingle **N**ucleotide **V**ariant
SV:**S**tructural **V**ariant
VAF: **V**ariant **A**llele **F**requency
VBMM: **V**ariational **B**ayesion **M**ixture **M**odel
WES: **W**hole **E**xome **S**equencing
WGS: **W**hole **G**enome **S**equencing

Mathematical notations

In all this thesis, we will use the following notations:
$\mathrm{B}_{(n,p)}$ for the binomial distribution of size n and probability p
c for cellularity
$\mathcal{L}$ for the likelihood
ℓ for the log-likelihood
ℓ^2 for the norm derived from the scalar product
ω for weights
$\mathbb{R}_+^*$ for the group of real numbers strictly higher than 0

Synthèse

Que l'on me donne six heures pour couper un arbre, j'en passerai quatre à préparer ma hache.

—Abraham Lincoln

Le neuroblastome est le cancer extra-cranial solide le plus fréquent chez l'enfant. Il est caractérisé par une très grande hétérogénéité tant au niveau clinique que moléculaire. En effet, on observe une rémission spontanée chez certains patients alors que la maladie peut progresser pour d'autres malgré une intervention thérapeutique chirurgicale et médicamenteuse. Dans ces conditions, on peut se demander quels facteurs différencient les premiers de ceux dont la maladie survient à nouveau aprés traitement.

Pour répondre à cette question, nous nous sommes attachés aux différentes populations cellulaires constituant la tumeur au diagnostic et à la rechute. Pour cela nous disposons de vingt-deux patients pour lesquels l'ADN constitutif, la tumeur au diagnostic et celle à la rechute ont été séquencés par séquençage à haut débit de génome complet.

Afin de détecter au mieux les différentes populations coexistantes, il a été nécessaire de développer un algorithme de reconstruction adapté à la problématique du neuroblastome. C'est-à-dire un algorithme prenant en compte les possibles altérations chromosomiques (gains et pertes), tout en se satisfaisant d'un faible nombre de variations de nucléotides uniques. Dans ce cadre, nous présentons QuantumClone, ainsi que les différentes techniques mathématiques permettant la résolution efficace du problème de reconstruction clonale. Les améliorations par rappot à l'existant, tant au niveau de la qualité de la prédiction du clustering que de la vitesse de calcul, ont été validés par comparaison sur des simulations numériques avec deux méthodes déjà publiées, nommément sciClone et pyClone.

Cependant, au travers d'une compétition — le DREAM Meta challenge — nous montrons d'une part que le nombre de variants appelés par les outils de

variant calling contiennent un nombre important de faux positifs, et d'autre part qu'il est difficile de recréer les erreurs recontrées dans les données biologiques. En effet, nous mettons en exergue la grande disparité existante dans cette compétition entre les échantillons simulés et les échantillons issus de patients. Pour cette raison, nous proposons une série de filtres permettant de retirer les faux positifs et reposant sur des raisonnements biologiques pour pallier le manque de résultats des algorithmes d'apprentissage supervisés.

Une fois les différentes populations clonales extraites des données, il est important de pouvoir caractériser leurs particularités biologiques. Pour cette raison, nous proposons de diviser les variations génomiques en deux groupes non exclusifs : celles à faible variance et haute qualité qui seront utilisés pour le clustering d'une part, et celles ayant un impact biologique connu ou prédit d'autre part. En effet, par attribution a posteriori des variations ayant un intérêt biologique, nous pensons pouvoir mettre en avant les mécanismes biologiques expliquant l'apparition ou la disparition de populations clonales entre le diagnostic et la rechute — l'une des forces étant la sélection négative par le traitement.

Cependant, très peu de gènes comportent des mutations récurrentes dans le neuroblastome. Nous pouvons citer comme exemple les plus fréquents ALK (avec une fréquence d'occurrence de 6 à 12% en fonction des cohortes), ou ATRX (inférieur à 10%). Il apparaît alors comme raisonnable de s'intéresser non pas à des gènes uniques mais à des ensembles de gènes, qui eux peuvent être touchés de façon récurrente par la maladie. Pour ce faire, il a fallu dans un premier déterminer les gènes candidats. Nous avons donc utilisé plusieurs outils de prédiction d'impact de variations dans les régions codantes (SIFT, Polyphen-2) et non codantes (Funseq2). En utilisant la liste des gènes contenant au moins une variation prédite délétère, nous avons pu comparer le nombre de gènes mutés dans un processus biologique et le nombre attendu par hasard. Cette comparaison a été réalisée en utilisant ACSNMineR, développé et publié pendant cette thése, et qui repose sur les connaissances biologiques agrégées dans l'Atlas of Cancer Signalling Network (ACSN). Nous montrons que différents processus biologiques déjà connus sont à l'œuvre dans le neuroblastome touchant aussi bien le cycle cellulaire, l'apoptose, la transition épithélio-mésenchymateuse, la réparation de l'ADN, la survie cellulaire ou la neurogenése. Plus spécifiquement, nous pouvons citer les voies de signalisation WNT (canonique et non-canonique), AKT/mTOR, ou enfin les MAPKinases.

Finalement, nous avons pu comparer les différentes voies affectées dans

les clones occupant une plus grande fraction de la tumeur au diagnostic ou à la rechute. Alors que le nombre de variations double entre le diagnostic et la rechute, le nombre de variations prédites délétères dans des processus biologiques enrichis en variations reste stable. De la même manière, les mêmes processus sont ciblés au diagnostic et à la rechute.

Afin d'expliquer ces résultats, nous formulons l'hypothése suivante : sachant que la capacité d'adaptation des cellules diminue avec l'accumulation de variations fonctionnelles, les populations ayant un avantage sélectif au diagnostic — du fait d'un grand nombre de mutations dans les processus cancérogènes — sont sélectionnées négativement par le traitement. Après traitement, seules des populations ayant un faible de taux de variations fonctionnelles ont survécu. La pression de sélection dûe au traitement étant relâchée, une nouvelle compétition intratumorale peut prendre place. Les populations accumulent alors à nouveau des variations fonctionnelles dans les mécanismes liés à la tumorigenèse, expliquant le nombre comparable de variations fonctionnelles au diagnostic et à la rechute.Le fait que les cellules accumulent de manière régulière des variations avec chaque division permet quant à lui d'expliquer que le taux de mutation fonctionnel est réduit à la rechute par rapport au diagnostic.

General overview

We detail below the structure of the thesis and the interdependence of chapters to one another.

- **Chapter 1** is an introduction to the relevant concepts of biology that will be used and developed in the following work.
- **Chapter 2** introduces the machine learning concepts that will be used throughout the manuscript, such as **clustering** and **classification**. We will also detail the clonal reconstruction problem and the existing literature on that subject.
- **Chapter 3** details the mathematical aspect of QuantumClone, the algorithm developed to solve the clonal reconstruction task. This chapter assumes that all variants provided are true positives. In the next chapter we will see how to remove the noise from the sequencing output.
- **Chapter 4** illustrates the fact that data from biological sequencing and shows that High Throughput Sequencing contains a high proportion of false positives. The DREAM meta challenge was used to illustrate the difficulty of finding correct features to discriminate true and false positives. Knowledge derived from this experience is then applied to the results of whole genome sequencing from 23 patients.
- **Chapter 5** focuses on the extraction of biological meaning from sequencing data. For that reason, we detail functional annotation tools and pathway enrichment analyses. This lead to the development of ACSNMineR, an R package to compute enrichment of variants in a biological module.
- **Chapter 6** is the specific application of all previous chapters to the neuroblastoma cohort presented in chapter 4, and the biological conclusions and hypothesis resulting from this application.
- **Chapter 7** concludes this work and gives possible tracks to continue and expand it.

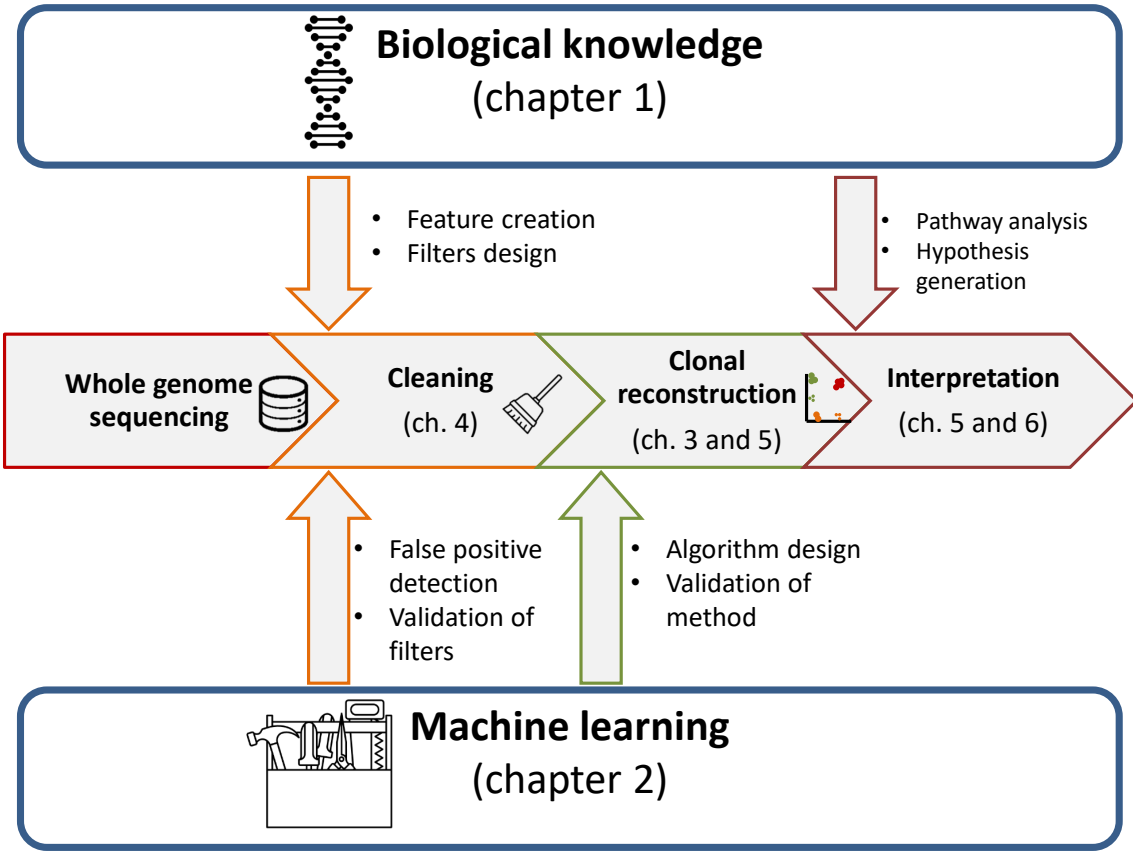


Figure 1: **From data to biological results**. We represent the interaction between the different chapters of the manuscript. Biological knowledge and machine learning are prerequisites to this work and intervene at different stages of the manuscripts, highlighted by specific points. The workflow in the middle represents the general chronology of a data analysis project, which may not reflect the structure of the manuscript presented here or the chronology of the thesis itself.

Chapter 1

Biological preamble

> *I have called this principle, by which each slight variation, if useful, is preserved, by the term of Natural Selection, in order to mark its relation to man's power of selection.*
>
> — Charles Darwin, *The Origin of Species, 1859*

In this work, many different aspects of cancer biology and computational biology are described. Although this introduction may not be as detailed as a biologist or a data scientist would like, it is provided so that both can exchange with the same vocabulary. We will first introduce the topic of cancer and detail the characteristics of neuroblastoma. Then we will describe the sequencing techniques called "High Throughput Sequencing" (HTS), and we will finally conclude with a brief discussion of statistics applied to genomic data.

1.1 A small history of cancer

The first historical description of cancer date as far back as 3000 b.c. in ancient Egypt, with treatments of breast cancer by cauterization[1]. Hippocrates is believed to be the first user of the word "carcinoma" to describe uncontrolled proliferative swellings [1]. It is only in the 18th century that John Hunter developed the idea that cancer could be cured through surgery, only when the tumor had not invaded nearby tissues.

1.1.1 Hallmarks of cancer

It is difficult to talk about cancer and not cancers, as many different diseases are classified under this name: from liquid tumors such as leukemia, pediatric can-

[1] American Cancer Society. 2009

cers such as Ewing Sarcoma, adult cancers such as the prostate cancer, sporadic or hereditary breast cancers. First classifications were based on the localization of the tumor, and are still used for some specific cancers: we are all too familiar with the terminology of colon, breast or prostate cancer. However a dissection of these tumors leads to a molecular classification of tumors based on cellular (think of Non-Small Cell lung cancer) or molecular (Estrogen Receptor positive breast cancer) markers. We can then wonder what all these diseases have in common.

In January 2000, one of the most renown reviews in oncology (attracting more than 15,000 citations) was published by Douglas Hanahan and Robert Weinberg [2]. This review defines eight hallmarks of cancer and was followed by an update 11 years later [3], adding four new hallmarks, bringing the figure to twelve key mechanisms for cancer development: sustaining proliferative signaling, evading growth suppressors, activating invasion and metastasis, enabling replicative immortality, inducing angiogenesis, resisting cell death, avoiding immune destruction, tumor promoting inflammation, deregulating cellular energetics and genome instability and mutation.

In this thesis, we will focus on the analysis of a pediatric cancer called neuroblastoma.

1.1.2 The oncological context: neuroblastoma

Pediatric tumors are less frequent than in adults, and represent only 1% of diagnosed cancers[2]. Neuroblastoma is the most common extra-cranial solid cancer of childhood, representing 7.6% of pediatric cancers in Europe [4]. This cancer stems from neural crest cells, and is characterized as sympaticoadrenal lineage neural-crest derived tumors [5, 6, 7]. This leads to a wide range of tumor localization such as the adrenal gland, neck or pelvis (Fig. 1.1).

In the next paragraphs we address the specificity of this disease both as heterogeneity between patients and within a patient.

Heterogeneity between patients

Neuroblastoma landscape is characterized by a small number of recurrent alterations, whether copy number alterations or mutations. In fact, recurrent alterations are *MYCN* amplifications (16%), 17q gain (48%), 11q loss (21%) or 1p loss

[2] www.e-sante.fr

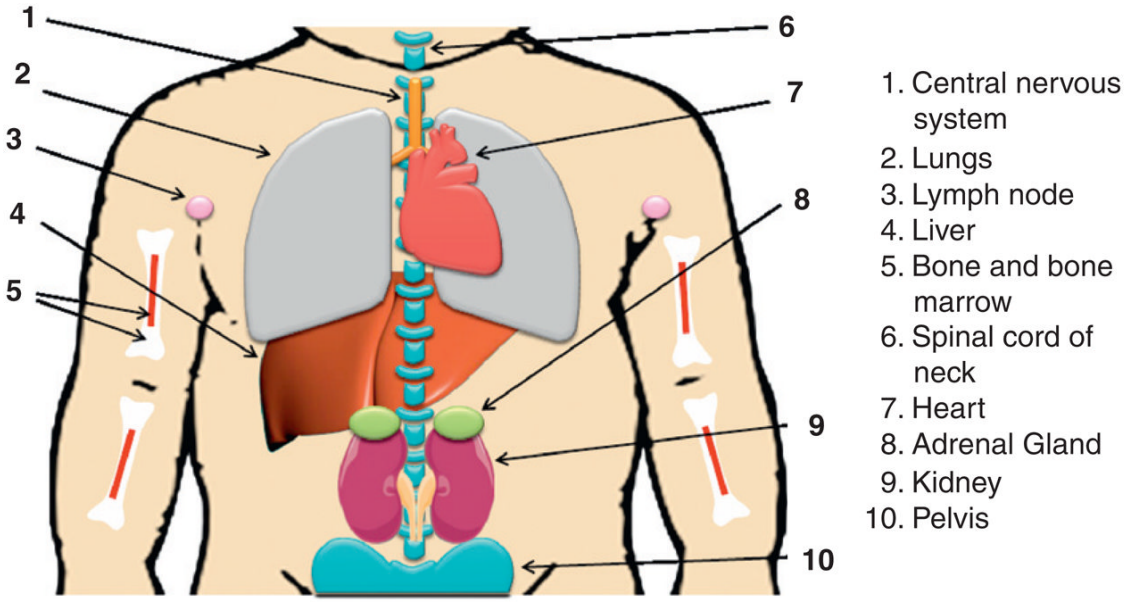


Figure 1.1: **Neuroblastoma localization.** Neuroblastoma primary tumors derive from precursor cells of the peripheral (sympathetic) nervous system and can arise anywhere along the sympathetic chain, most frequently in the adrenal gland (position 8 as shown). Neuroblastoma may also develop from spinal cord of neck (position 6) and pelvis (position 10). Neuroblastomas mainly metastasize to lymph nodes (position 3), liver (position 4), bone and bone marrow (position 5), and also spread to central nervous system (position 1) and lungs (position 2) in infants.[5]

(23%) [7], while mutations occur in *ALK* (6 - 12%) [8] or *ATRX* ($< 10\%$) genes [9].

This diversity is also represented in the prognosis of patients where spontaneous remission can be observed in younger patients, while older children (> 2 years old) with chromosomal imbalance have a poor outcome prognosis despite chemotherapy and ablative treatments. In addition, only an estimated 1 to 5% of neuroblastoma cases appear as hereditary [10, 5], with mutations in *PHOX2B* or *ALK* genes[11, 12] and predisposition loci in chromosomes 16p12–13 and 4p16 [5]. This has to be put in perspective with other cancers with a high hereditary burden such as retinoblastoma. Indeed, in retinoblastoma, *RB1* gene mutations are often dominant with near complete penetrance ($> 99\%$)[13, 14]

Intratumoral heterogeneity

Cancer is one of the very few times in a lifetime where one can be confronted to the principle of *Natural Selection*. Evolution is usually set on a time scale of multiple generations, and it is difficult to experience this in a single lifetime except with microorganisms or smaller organisms such as *Drosophila Melanogaster* or *C. Elegans*. An example of these reported phenomena is the change of color of the Peppered Moth, *Biston Betularia*, where the melanic phenotype of the moth

was associated to a lower predation by birds in regions with high atmospheric pollution, as first hypothesized in 1896 by J.W. Tutt [15]. Even in this occurence, it took decades for the natural selection to shift the most frequent allele from the typical to the melanic.

We can assume that our body is an ecosystem of its own, with bacteria and cells of different types coexisting, competing, and collaborating for the sake of the organism. Mutations are one of the driving forces of evolution in species, as new traits with potential benefits are more likely to be passed on to the next generations. It is also one of the forces underlying cancer development, and will push the cell toward an uncontrolled proliferative state - similar to what can be observed with invasive species.

It is then important to understand what are the signals that deregulated the cell, this means to be able to figure out the phylogeny of the cells and find the common ancestor to all these (Fig. 1.2). While it may not be useful to reconstruct the whole tree, some essential nodes may be of interest, for example the genotype of the cell that gave rise to the relapse could give clues and insights in the mechanisms of resistance to treatment.

1.2 High throughput sequencing analysis

Next generation sequencing (NGS) is a set of technologies that allowed faster and cheaper sequencing of the DNA molecule. We will focus here on the Illumina technology that is used for the whole exome (WES) and whole genome (WGS) analyses.

1.2.1 Principle

The modern sequencing methods rely on the reaction set up by Frederick Sanger in 1977, for which he received the Nobel Prize in 1980. For this reason, we will first explain the historical sequencing reaction, now dubbed "Sanger sequencing", before moving on to the Illumina protocol.

Sanger sequencing

Sanger sequencing mainly relies on the Polymerase Chain Reaction (PCR), which is used to amplify a DNA sequence (Fig. 1.3). This reaction can be terminated

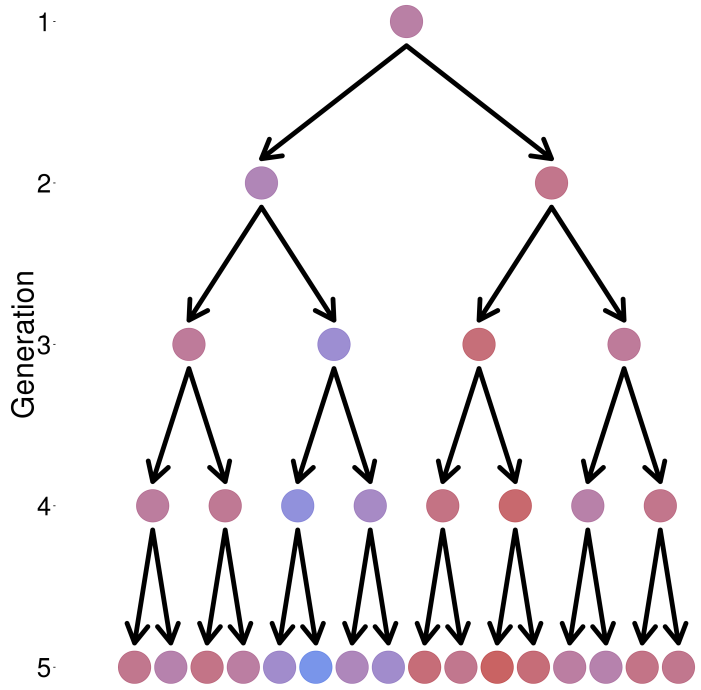


Figure 1.2: **Phylogenetic tree of a cancer.** Colors represent the different alterations characterizing each cell. With each division alterations accumulate so that each cell is distinct from all others be it from the same generation, from its ancestry or progeny.

during the elongation by the use of dideoxynucleotides triphosphate (ddNTP) which cannot form the phosphodiester bond required to link a nucleotide with the following. As a result, the reaction will create chains of various sizes, depending on where the ddNTP was incorporated. Having a different radioactively labeled ddNTP in each reaction mix allows to reveal the order in which adenine (A), thymine (T), guanine (G), and cytosine (C) are incorporated by electophoresis on a polyacrylamide gel. This has now been simplified with each ddNTP labeled with a fluorophore re-emitting at a different wavelength. The sequence can then be automatically read by monitoring light re-emission of the migrating molecules.

Illumina

In order to reduce costs and increase speed of sequencing alternative methods had to be designed to sequence the genome. Indeed, the Human Genome Project, which aimed at sequencing the first human genome, cost approximately 300 million USD for the first draft and an additional 150 million USD to refine this draft, using cloning and Sanger sequencing. With refined techniques, the same draft would have cost an estimated 14 million dollars in 2006 and can now be

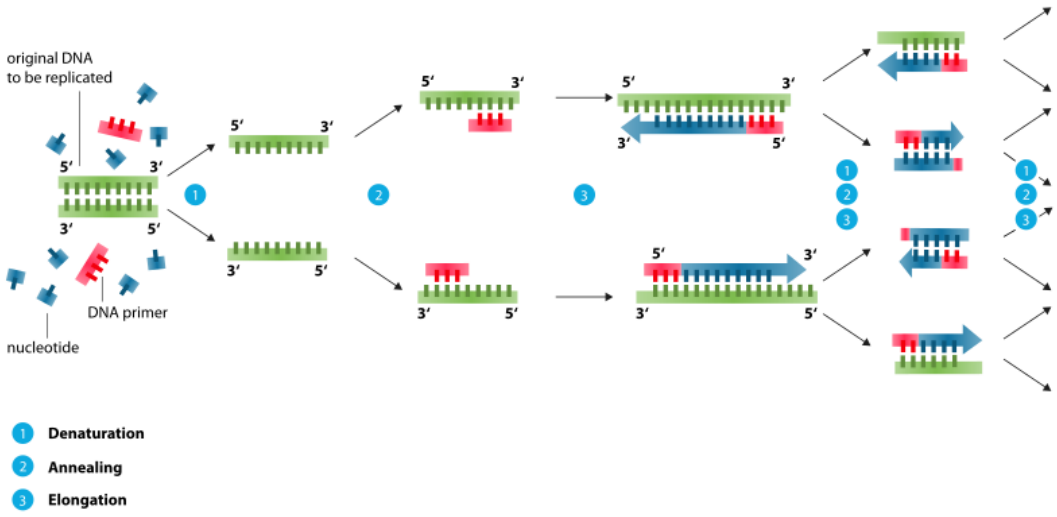


Figure 1.3: **Polymerase Chain Reaction**. The PCR mainly consists of three steps forming one cycle. First the denaturation at high temperature ensures that the DNA fragments are single stranded. Then the annealing phase at lower temperature allows the PCR primers to bind the DNA fragment to replicate. Finally the elongation allows the polymerase to synthesize the complementary strand of DNA with the available deoxynucleotides. This cycle is repeated several times - usually 10 to 30. Adapted from Enzoklop - Own work, CC BY-SA 3. https://commons.wikimedia.org/w/index.php?curid=32003643

achieved for 4000USD[3] in 2016, as shown in figure 1.4.

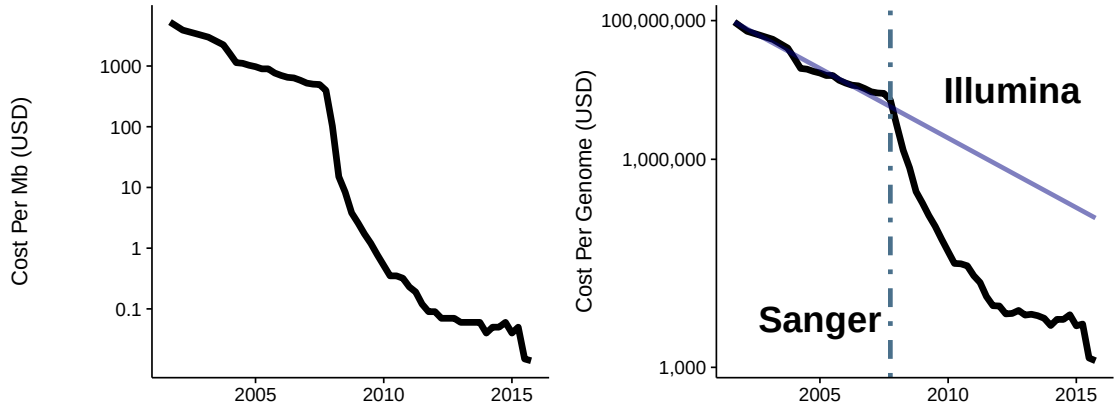


Figure 1.4: **Cost of sequencing** Evolution of the cost of sequencing of 1 million bases (Mb, left) or a full human genome (right) in USD, from September 2001 to October 2015, based on data from genome.gov. In addition, an equivalent of the Moore law (blue) is added on the right chart, to show a price divided by two every 18 months, and periods annotated with the technology used.

[3] https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome

Illumina sequencing relies on the reversible terminated chemistry concept described in Canard et al [16]. First the DNA fragments are digested to small fragments, and adapters are attached. These adapters will then be used to anchor the fragment on one end of the flow cell. Using a primer, the sequence is then amplified, with all copies of the sequence being localized in the neighborhood of the first fragment, due to the anchoring. This creates a cluster of identical sequences - forward and reverse strands. To sequence these clusters, a nucleotide engineered with reversible termination and a chromophore is added to the plate. Unused reactants are washed away, and a picture of the chip is taken. The deblocking step allows the incorporated nucleotide to bind to another nucleotide and the cycle is repeated until the full DNA molecule is sequenced.

This sequencing technique can be used to read the sequence from a single end of the molecule (single end sequencing), or from both ends (paired end sequencing, see Figure 1.5). Another technique, called mate pair sequencing circularizes the DNA fragment before fragmentation. Because of circularization, two sequences separated by more than 1kb can be brought together, which can be useful to detect complex genomic rearrangements.

In the case of paired end sequencing, the expected distance between the two pairs is known, and can be used by the aligner to map the sequence on the reference genome, as we will see in the next subsection.

Complete Genomics

Complete Genomics (CG) uses a proprietary technology optimized for human sequencing. In the cohort of neuroblastoma samples provided by John Marris from the Children's Hospital of Philadelphia, paired end sequencing was used. It is to be noted that the reads coming from this technology bear a deletion of a few base pairs (see Figure 1.6), limiting possible operations (such as indel realignment) on the file.

1.2.2 Read alignment

The sequences obtained from Illumina or Sanger are small fragments of the donor genome, with a length of 36 to 600 base pairs (bp). This is to be compared to the 3×10^9 bp of the human reference genome. It is thus required to align the donor sequences on the reference genome.

Figure 1.5: **Illustration of paired end sequencing read mapping.** Reads aligned on the genome from left to right are shown in red, those from right to left in blue. A grey line links the pair of read. Depth of coverage (number of reads overlapping the position) is shown on top in grey. Visualization of the reads and sequences made with the Integrative Genomics Viewer (IGV) [17]. Image courtesy of Léo Colmet-Daage.

Principle of read mapping

In theory, for a combination of 4 letters and length 100, a total of $4^{100} = 1.61 \times 10^{60}$ unique sequences exist. This number is far superior to the size of the genome and should guarantee that each sequence of 100 bp can be accurately attributed to its position.

However, our genome may have emerged from two whole genome duplications [18] at vertebrate stage, meaning that many genes have paralogs with very similar sequences[4]. Approximately $811,737,329$ bp are identified as a part of segmental duplication and $319,296,434$ bp as simple repeated element in the human genome assembly hg19.[5] We will detail further repeated elements and duplications in chapter 3.

[4] This is termed the 2R hypothesis, for **two r**ounds of duplications, or Ohno's hypothesis.

[5] Numbers are based on the UCSC genome browser tracks Segmental Duplication and Simple Repeats.

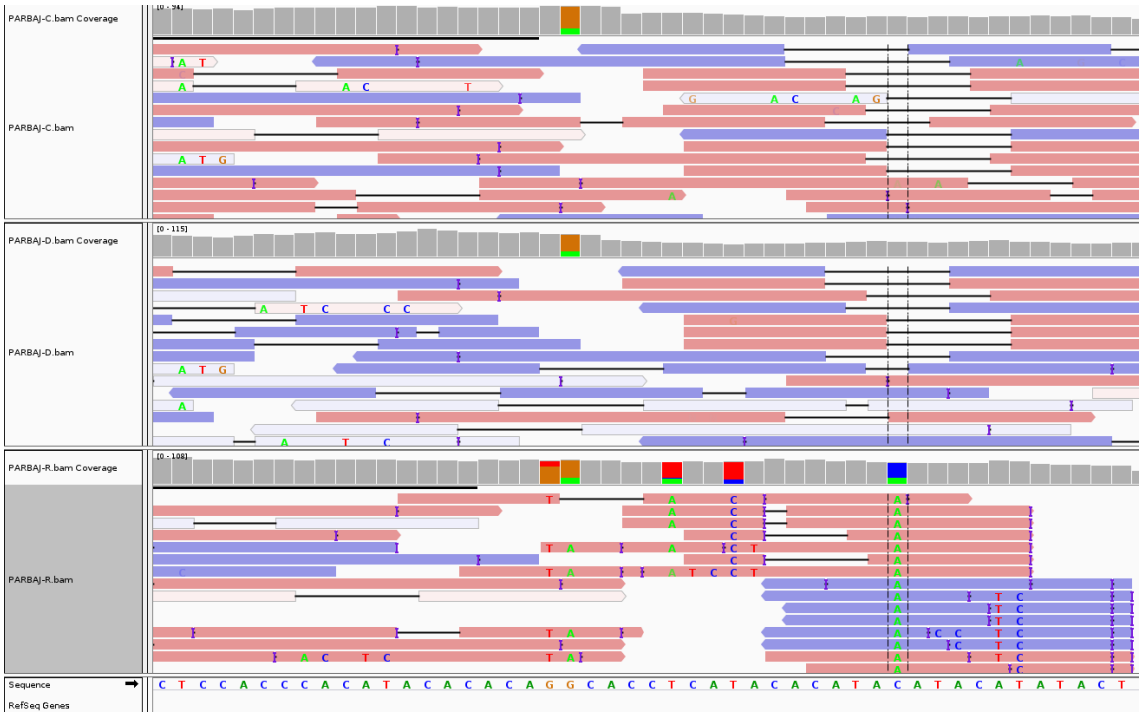


Figure 1.6: **Illustration of Complete Genomics sequencing output.** Contrary to the previous figures, lines here show a deletion inside a read. Purple brackets denote an insertion compared to the reference genome. **From top to bottom:** germline sequence, tumor at diagnosis, and tumor at relapse from patient PARBAJ. Highlighted is a variant (C to A) specific to the relapse sample.

Moreover, we align the sequences on a reference, and, on average 3.3 million differences [19] exist between any individual human genome and the reference. They correspond to Single Nucleotide Polymorphisms (SNPs) or private variations. This does not include sequencing errors. As a result, a number of allowed mismatches between the reference and the read sequence has to be set. This is especially of interest in the case of insertions or deletions, which will introduce many changes between the sequence read and the reference, as shown in table 1.1, if a gapless alignment is used or if the correct alignment was missed by the aligner.

Reference	ATACGACGAAGCTAC
1 mismatch	ATACGAGGAAGCTAC
2 mismatches	ACACGAGGAAGCTAC
1bp deletion	ATACGAC-AAGCTACA
2bp insertion	ATACGACGGCAAGCT

Table 1.1: **Example of variation impact on a sequence.** All sequences have the same length, the mutation is represented in red and subsequent changes in the gapless alignment are shown in orange. We can see that a deletion or insertion inserts many mismatches in the local alignment, such errors can be resolved through various techniques.

Variant calling

Once that reads have been mapped to the closest sequence of the reference, we can look for positions where it differs from the reference. This is done thanks to the variant calling.

In tumoral context, it is important to distinguish variations that would come from the tumor, called somatic, and the variants that are also present in the healthy tissue, termed germline. The former may give indications on the course of development of the disease while the latter may be linked to predisposition. A third type of variants exist, but will not be discussed in this manuscript: germinal mutations that appear in gonadal tissues, and that can potentially be passed to the descendants.

As a result, somatic variant callers first look for differences between the reads aligned at a given position and the reference, and if there is a difference, it will dig into the matching germline sample to see if the same variation can be found in its genome.

Many different algorithms have been developed to do so, among which we can cite Varscan2 [20], Mutect [21] or Strelka [22].

Varscan2, that will be used for the whole genome analysis, uses Fisher's exact test [23] on the number of reads supporting the variant in the germline and tumoral samples to distinguish somatic and germline variants.

1.3 Available and relevant data

1.3.1 Is it big data?

In the recent years, an unprecedented stream of information poured in different domains of computer science, including bioinformatics. We detailed in previous sections the principles of High Throughput Sequencing technologies. It can be noted that one sequenced read has to store the information about the sequence of each nucleotide as well as the confidence in the result of the nucleotide read being correctly guessed — called base quality. After mapping, we also have to encode the position of the genome where the sequence has been mapped, and the confidence in the mapping of this specific read on the reference genome —- called mapping quality. As a result, a whole genome sequenced with an average depth of $100\times$ and stored in a compressed format (called BAM), represents

$\sim 500GB$ of data.

However, we have said that roughly a million positions are supposed to be different from the reference genome, so less than one in a thousand. With reads of only 100bp, with the hypothesis of variants being heterozygous and a depth of sequencing of $100\times$, this would lead to as little as a read for 20 being informative. The mutated positions not being known *a priori*, this cost to efficacy can hardly be compressed. In practice, whole exome sequencing (WES) (where only the exonic regions are sequenced, based on a capture technique) or targeted sequencing (where only predetermined regions are sequenced) can be used. However, these techniques introduce a bias in the data and the analysis. In whole exomes, a bias in read distribution is often observed, meaning that the probability to observe reads in G/C poor or G/C rich regions is lowered. In addition, the regions captured will depend on the capture technology used. This will lead to a different coverage of genes and exons. Targeted sequencing is biased towards genes that have already been shown to have an interest in the disease, and will be less likely to be used in a prospective study.

At a time where "big data" has become a buzzword and marketing strategy, the disparity between useful and total information has been dubbed "Fat data". Indeed, data quantity is often understood as an increase in quality. The relevant information in our data for this study can be downsized to a few gigabytes worth, with the copy number alterations on the one hand and variant calling results on the other hand.

Another layer of information could be retrieved by looking at structural variants. Structural variants can be detected using dedicated algorithms such as SVDetect [24] or BreakDancer [25]. This aspect will not be further detailed in this thesis.

In chapter 2 we will see how to analyze genomic data with respect to the question of clonal evolution in cancer. We have seen here that the genomic data could be described as "fat data", reaching large sizes with little information contained. We also describe here another limitation of biological data: the scarcity of samples often outweighs the number of parameters to test.

1.3.2 The small p large n curse

Machine learning problems often require that the number of observations largely exceed the number of parameters to tune. This is required to avoid spurious

correlations.

In genomic, this assumption cannot be held true, and this is referred to as the "Small p large n" issue. This can be easily illustrated in Genome Wide Association Studies (GWAS), where a million polymorphisms (SNPs) are tested for association with a disease, and to a lower extent in cancer, where the association of the 19,033 protein coding genes with the disease is tested, to which we can add an extra 6,732 non-coding RNA genes (number correspond to unique HUGO symbols)[6]. These SNPs can be associated with (and sometimes responsible of) an increased risk of a few percents. To uncover association of a SNP that has a 10% increase of the risk, and a cohort of 1,000,000 participants equally affected by the disease or healthy, the p-value associated with the SNP will be ~ 0.1 (see table 1.2 for details).

	Minor allele	Major allele	total
Healthy	$4,762$	$95,238$	10^5
Disease	$5,238$	$94,762$	10^5
Total	1×10^4	1.9×10^5	2×10^5

Table 1.2: **Allelic imbalance**: for a cohort of 2×10^5 individuals, with an allele in 1% of the population and 10% more frequent in the disease population. This SNP would have a p-value computed by fisher test of 1.092×10^{-6}. After Bonferroni correction, and considering that SNP arrays currently have $\sim 10^6$ positions covered, we can estimate the corrected p-value to be: 0.1, higher than the 0.05 threshold, even though the cohort has 200,000 individuals.

The evolution of the cohort needed to estimate a linkage between a SNP and a disease depending on the cohort size, is shown in figure 1.7. It is important to note that this kind of studies cannot uncover low risk factors even in studies with large cohorts. For example, a SNP associated with a risk factor of 1% in the example cannot be found significant, even for a cohort of 10^8 persons.

As a result, it may be difficult to extract relevant features, i.e. explicative variables, from the data, especially in the case of low risk factors. In our case, with a cohort of ~ 20 patients, the relevance of a variant in the disease can hardly be detected, and we will limit the scope of our study to descriptive results.

[6] Detection of pathways enriched in somatic variants will be further detailed in chapter 5

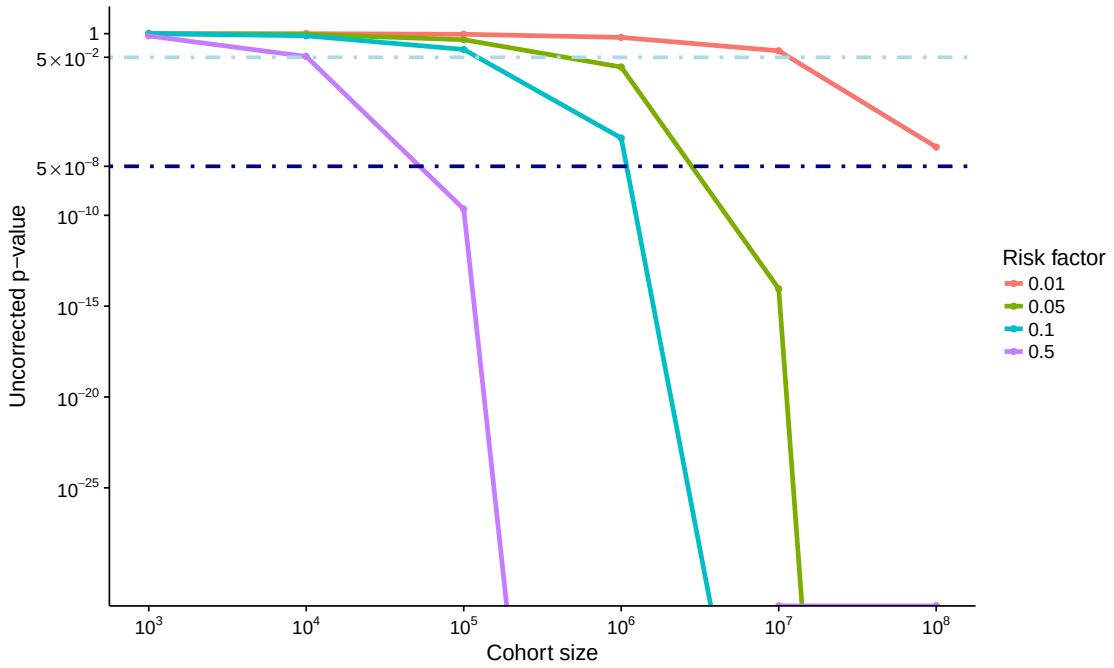


Figure 1.7: **p-value and cohort size** Evolution of significance of a SNP given the risk factor and the cohort size. An equal partition of healthy and disease patients is assumed. For readability p-values below 10^{-30} are not shown. The light blue line corresponds to the traditional 0.05 threshold, and the dark blue line corresponds to the threshold taking into account one million test (Bonferroni correction)

Chapter 2

Mathematical perspective of the clonal reconstruction

> *All models are wrong but some are useful.*
> — George Box*, Robustness in the strategy of scientific model building, 1979*

Machine learning is translated in French as *"apprentissage statistique"*, or "statistical learning". This terminology accurately represents the idea that computers do not learn by themselves, but that we fit a statistical model to the data that is gathered.

2.1 Introduction to machine learning

> *Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed*
> — *attributed to* Arthur l. Samuel*, circa 1959*

A more recent definition of machine learning has been given by Mitchell in 1997: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." [26]

In these 40 years, machine learning had been applied to various tasks, such as classification (when groups are known beforehand), regression, or clustering (where groups are unknown beforehand).

The task of finding clones using a group of variants is a **clustering** issue: the number and characteristics in terms of mutations and cellular prevalence are

unknown before the clustering (see chapter 3). Looking for relevant features explaining if a variant called is a true or false positive is a **classification** task (see chapter 4). In the following sections we will provide background for classification task first then for clustering.

2.1.1 Classification: supervised learning

"The term 'supervised learning' is rooted in statistical learning/machine learning parlance, where it describes the analysis of data via a focused structure. "[27]

In the data used for training, the groups or classes are known beforehand and usually provided by a human expert — hence the supervised denomination. A classification algorithm will then learn on a set of characteristics (or features) that are given for each entry, and will combine them in a way to try to predict the outcome. The algorithm will try to minimize the error given a metric which can be accuracy or recall for example.

Many classes of algorithms can be used to classify data, among which we can cite logistic regressions, Support Vector Machines (SVM), partition trees (and their extension random forests), and neural networks.

Recent trends in machine learning: neural networks

Neural networks have received a lot of attention from the media in the last years due to their progress in image recognition[1], transformation (generating dreams[2]), natural language processing[3], or beating human champions at Go games[4].

However, neural networks do not represent the entirety of classification algorithms. In fact, Kaggle — a data science competition platform — shows that the gradient boosted random forest package XGboost was dominating discussions in 2016, in front of Keras, a neural network library. Both these methods have reached this status because of their ease of use and efficiency in solving tasks.

[1] The Revolutionary Technique That Quietly Changed Machine Vision Forever, MIT Technology review, September 9th, 2014

[2] On a testé pour vous... Deep Dream, la machine à « rêves » psychédéliques de Google, Le Monde, July 9th, 2015

[3] Computer Wins on 'Jeopardy!': Trivial, It's Not, The New York Times, February 6th, 2015

[4] Google's AI Wins Fifth And Final Game Against Go Genius Lee Sedol, Wired, March 15th, 2016

We will see that simpler model can provide insights on the data, insights that can then be used as guidelines for patient handling for example. In more complex models, the structure of the classification is difficult to apprehend. This can be illustrated by the difficulty to understand what a specific neuron in the network sees from the image, except by extracting image parts that activate such neuron [28].

Focus on Random forests

We describe in the rest of this paragraph a standard procedure for classification tasks, that will illustrate the use of random forests, that will be subsequently used in chapter 4.
We will use the Wisconsin Breast Cancer data set [29, 30, 31, 32] as a toy example. In this data set, the classification task should discriminate between benign and malignant tumors, using nine features, such as uniformity of cell shape and size or the clump thickness. All features are evaluated on a scale of 1 to 10. This data set has been curated, and contains 699 entries, of which 16 contain at least one unobserved feature. These 16 entries have been removed from our analysis.

Previous studies have shown that based on these features an accuracy of up to 95.9% could be achieved [29, 33], using $2/3$ of the data set as training and the remaining as validation.

The principle of a partition tree is to find a feature that best separates between the classes. In our case, we would like to find the relevant features that separate between malignant and benign tumor (see Figure 2.1). For each partition, the algorithm will try to find the best value to separate between malignant and benign, and repeat the procedure until either:

- Adding a partition does not increase prediction metric - here the metric is the accuracy;

- The complexity of the tree is higher than parameters given by user.

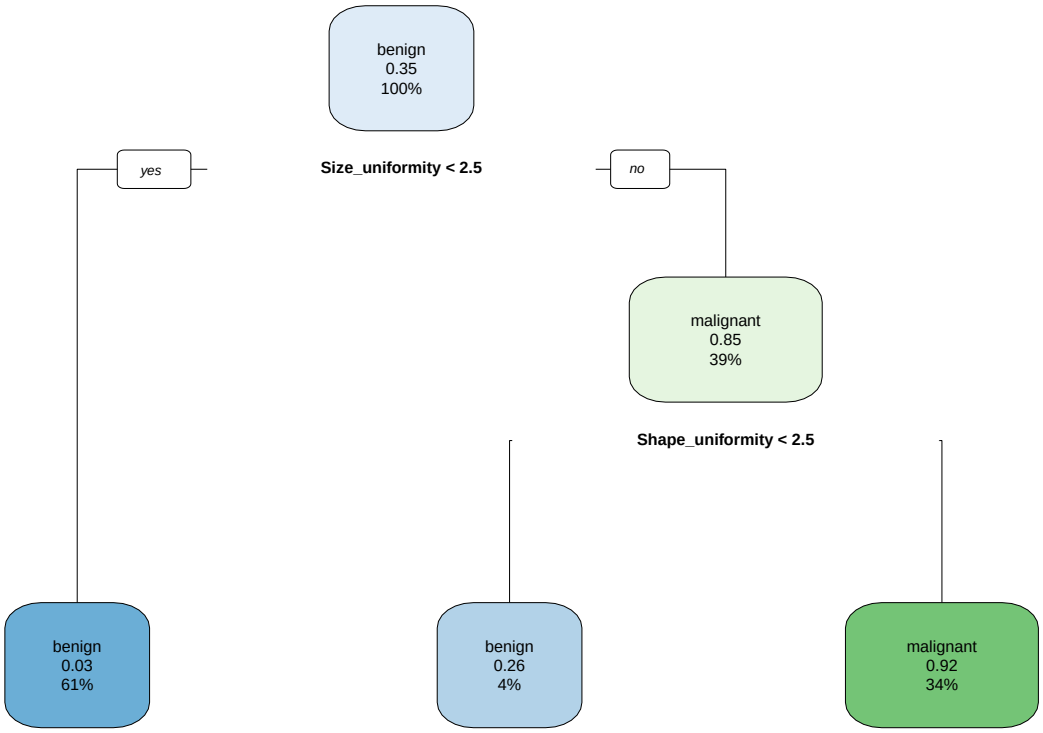


Figure 2.1: **Partition tree on Wisconsin breast cancer data.** This tree was trained on two thirds of the the Wisconsin data set. The label of the nodes shows the prediction, with the proportion of the malignant entries in the partition. The percentage of entries in the partition is shown on the third row. The representation of the partition tree was made using `rpart.plot`

In order to find the optimal complexity of the tree, a five-fold cross-validation is used. This means that for each value of the complexity, five random samplings of the training set will be done and accuracy will tested. Finally only the most accurate model will be kept, corresponding to the model presented in Figure 2.1. In our case, the best model achieved an accuracy of 94.8%, relatively close to the state of the art in 1992.

An extension of partition trees is random forest. Instead of training a single tree, multiple trees are trained on the data set. To prevent generation of identical trees, only a fraction of the features is used for training. The number of features simultaneously used is a parameter chosen by the user, and is tuned in our case using three folds of cross-validation. For prediction, each generated tree will give its prediction, weighted by its accuracy. With an "out-of-the-box" method, a 98.28% accuracy was reached on the same set of training and validation as previously. This in particular shows the quick progresses that have been made in the recent years. Such pace has also been highlighted in Jeremy Howard's course on neural

networks[5], with few lines of code achieving better predictions than state-of-the-art publication few years before.

This increase in accuracy is detrimental to the intelligibility of the model. While it is easy to grasp the model underlying a partition tree, which makes it easily implementable in diagnostic or every-day life, random forests are much more complex and require a prediction from the machine to classify new data.

2.1.2 Clustering

Clustering is the way to create groups based on the intrinsic architecture of the data. These groups are not necessarily known beforehand - but knowledge can be used to test accuracy of the method. The number of clusters k can be a parameter of the model, or can be selected using an information criterion. Information criteria balance the complexity of the clustering and the power of the model to explain accurately the data.

We describe in the two next parts two common clustering methods that will be used in chapter 3 and chapter 4 for clonal reconstruction and detection of similar pipelines.

Hierarchical clustering

Hierarchical clustering is a technique that aims at reconstructing a tree, which gives information on the whole structure of the dataset. Once the tree is completed, k clusters can be created by cutting the tree at a point where it contains exactly k branches, each branch defining a cluster.

The hierarchical clustering relies on a simple algorithm, described in algorithm 1. We can separate *agglomerative* and *divisive* clustering. Agglomerative clustering will start with as many classes as there are observations and groups classes by merging observations (bottom-up), whereas divisive clustering starts from a single class and removes observations one at a time (top-down).

Either two groups are merged together because they have minimal distance or dissimilarity, or they are split because they have maximal distance. The way the distance between clusters is computed can vary. In *single-link* clustering, the distance between two classes is the distance between two observations (one from each cluster) that are the closest from one another. In contrast, with *complete-link*

[5] http://course.fast.ai/

the distance used is the one between two observations of the clusters that have maximal distance[27] (see Figure 2.2). These two methods have shown to lead to extreme cases due to the *chaining problem* with single-link, and tightly contained clusters with complete link.

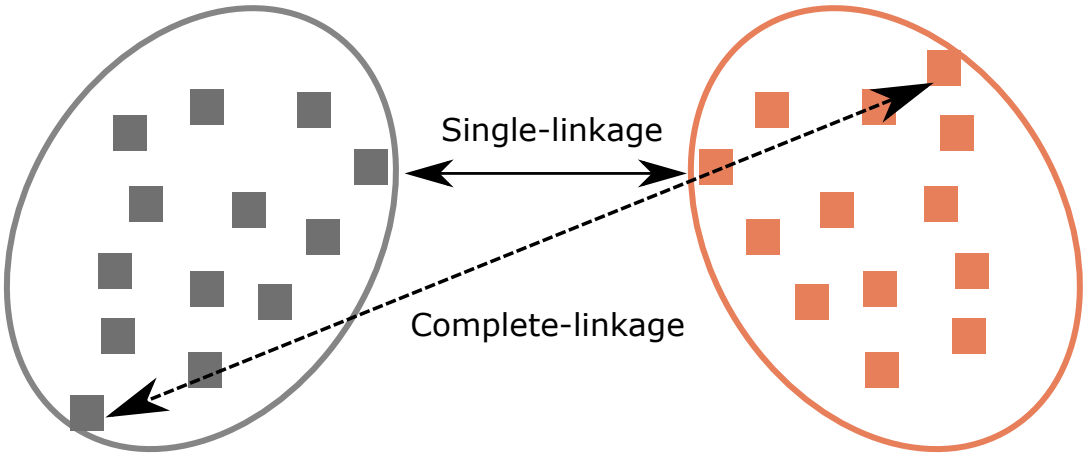


Figure 2.2: **Example of single linkage and complete linkage inter-cluster distance.** Two clusters (grey and orange), containing multiple observations are shown. For single linkage, the distance between the two clusters corresponds to the distance between the closest points, whereas the distance for complete linkage corresponds to the distance between the two farthest points.

A compromise of the two methods for inter-cluster distance computation is the *average-link* that approximates the average dissimilarity of the two clusters. The same intuition is used in the Ward method [34], where two clusters are merged if they minimize intra-cluster dissimilarity.

A combination of direction (bottom-up or top-down) and inter-cluster distance computation methods creates the algorithm, an example of which described in algorithm 1.

To compute the heterogeneity between two classes, a distance has to be chosen. Usual choices are ℓ^2 for numeric values, or Jaccard for binary values. We can write the Jaccard distance for two binary vectors of length n, with the number of events where both vectors equal 1 (respectively 0) $M_{1,1}$ (respectively $M_{0,0}$):

$$J = 1 - \frac{M_{1,1}}{n - M_{0,0}}$$

In chapter 3 we will also show a distance derived from the probability of two observations to belong to the same cluster).

Algorithm 1: Agglomerative hierarchical clustering pseudo-algorithm

```
Each observation is in its own class;
A distance is defined: Jaccard, ℓ², ...;
A method to compute inter-cluster distance is defined: Ward,
 single-link, ...;
while Number of classes > 1; do
    for i,j ∈ classes; do
        Compute distance between i and j;
    end
    Merge classes i and j that have the minimal distance to one
     another;
end
```

We refer the reader to chapter 11 of *Statistical data analysis*[27] or chapter 9 of the book *Data mining et statistique décisionnel*[35] for more details.

k-means and k-medoid

k-means is a widespread method used for clustering. A point will be attributed to a cluster if it is closest to the cluster center.

In the k-means clustering[36], the number of clusters is a user input. It generally comes from knowledge or previous analyses. The pseudo algorithm for the k-means is shown in algorithm 2.

Algorithm 2: k-means pseudo-algorithm, as described in Piegorsh (2015)[27]

```
Initialization: k points designated as centroids;
while Cluster assignment changes; do
    for i observations; do
        Compute distance between observation i and centroids;
        Attribute observation to cluster with closest centroid;
    end
    Update centroids as mean of observations in the cluster;
end
```

k-means can be highly sensitive to outliers, and an alternative has been proposed by using medoids[6] instead of means, resulting in the k-medoid algorithm[37].

[6] A medoid is a point of the cluster that minimizes intra-cluster dissimilarity.

Evaluation of number of clusters and information criteria

The number of clusters is usually unknown before clustering. In order to select a correct number of clusters based on the data, several approaches have been proposed, including silhouette analysis [38] and information criterion.

From the latest, we present the two most widely used criteria that are Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Information criterion rely on the idea that with more parameters it is easier to create a model that will explain the data. However, this can result in overfitting the data, which means that the statistical model has learned features that are specific to the dataset and do not reflect the general behavior (see Figure 2.3).

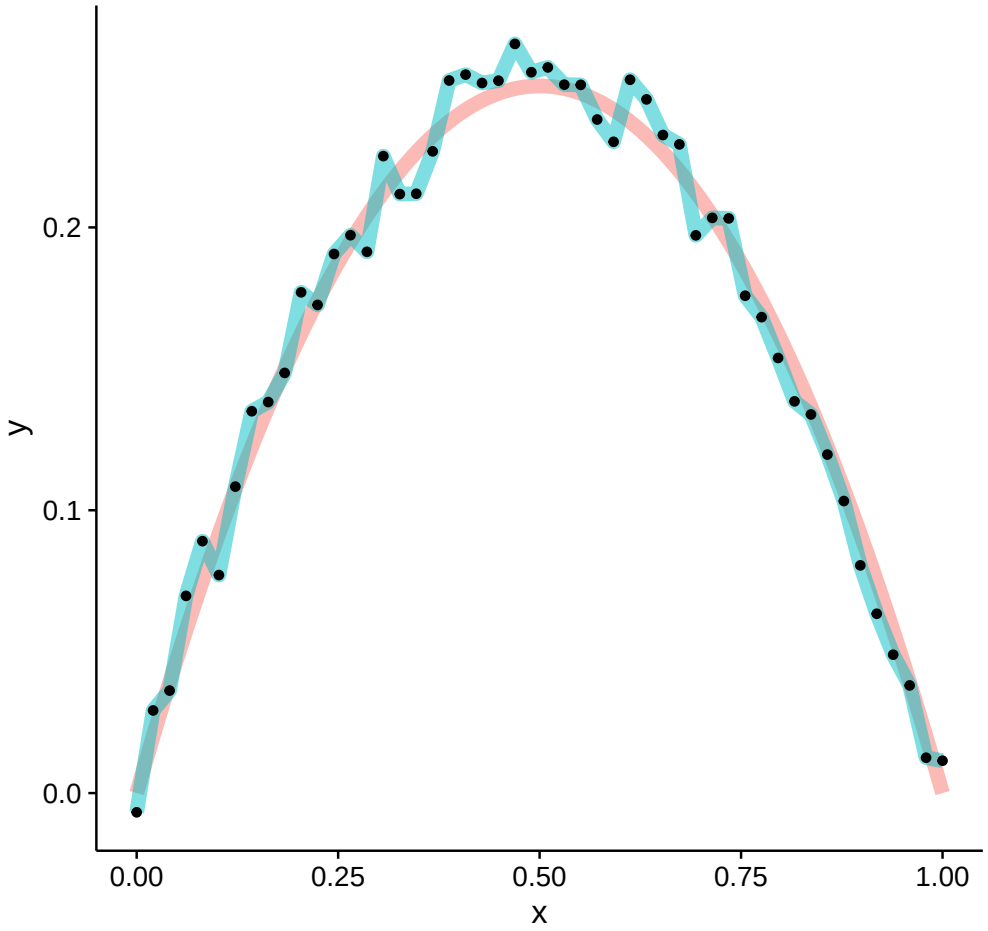


Figure 2.3: **Example of overfitting the data** The black dots represent observations, they have been generated by adding a small amount of noise to the function $f_{(x)} = x \times (1 - x)$, shown in pink. The blue line represents overfitting, as the function goes through each point of the training dataset. The algorithm thus has learned the model and the noise, which may not be relevant for future applications.

In order to limit overfitting by addition of new parameters, the information crite-

rion introduces a balance between the increase in parameters and the accuracy of the model. Mathematically, the BIC is written:

$$BIC = k \times \log(n) - 2\log(\mathcal{L}) \tag{2.1}$$

With k the number of parameters, n the number of observations, and ($\mathcal{L}$ the likelihood of the model.

Similarly, the AIC is written:

$$AIC = 2 \times k - 2\log(\mathcal{L}) \tag{2.2}$$

We can see that the BIC uses information from the number of observations of the model, which is not true for the AIC. In both cases, the model that will be chosen is the one that minimizes the information criterion.

Usage of clustering in biology

Clustering in biology has often been used for patient stratification [39], or extracting signature of variants[40] or differentially expressed pathways [41].

In our case, clustering will be use to differentiate between clonal populations that coexist in the tumor. We detail our model and published algorithms in the next section.

2.2 The clonal reconstruction task

2.2.1 Phylogeny

Phylogenetics is the study of the evolutionary structure underlying evolution of organisms. We can represent a set of cancer cells as different organisms evolving under a pressure of selection, be it selection by a strive for nutrients, escaping the immune system, or simply accumulation of deleterious mutations.

The phylogeny of a cancer holds information on the mutations that arise first, or those that confer an advantage to the population, expanding the population that carries them.

Imagine an experiment where one throws a handful of bouncing balls. At first, most of them would have the same trajectory, behaving in the same way. But as time passes, they would differentiate and move in different directions. With the

same idea, cells that have the same genotype would behave in the same fashion, yet they would accumulate mutations with each cells division. This leads to a differential evolution between the populations given enough generations (Figure 2.4).

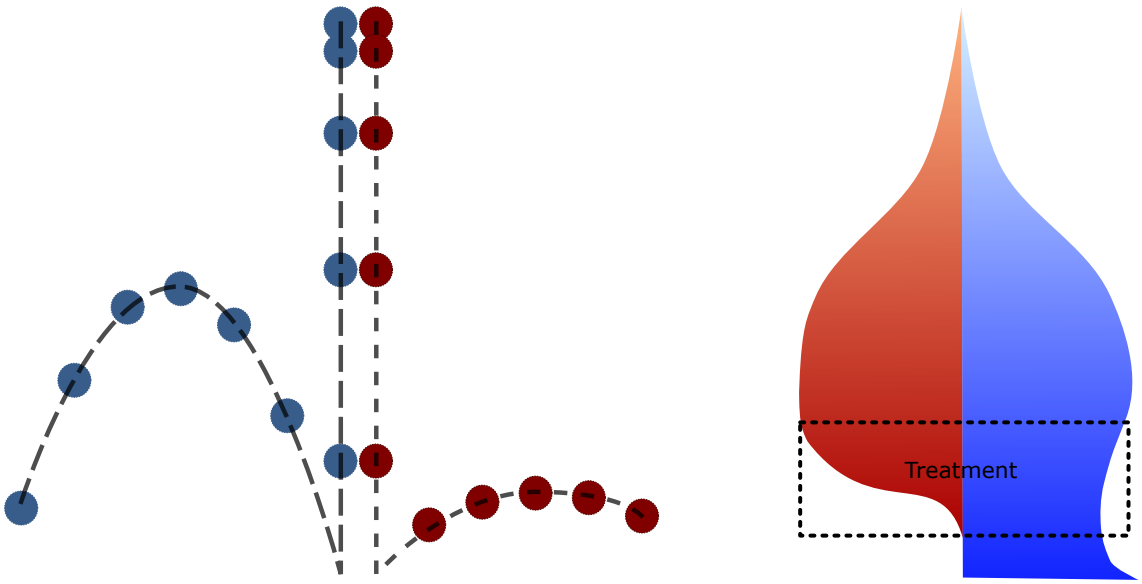


Figure 2.4: **Left: Bouncing balls**: two ball are simultaneously let loose. During the free fall stage, the trajectories are identical. Due to internal differences, when they meet the ground the trajectories differ. **Right: evolution of the clonal population.** Both clones expand when there is no pressure of selection. However, the red clone disappears after treatment, while the blue one resists treatment and expands afterwards.

2.2.2 Assumptions used for the clonal reconstruction

In order to identify different cell populations we have to rely on a set of axioms, or assumptions, that rely on biological insights of the tumor mechanistic:

- **Infinite loci**: The probability that a mutation appears twice is null. This is justified by the randomness of mutations, and the very large size of the human genome [42, 43].
- **Diploid contamination**: the cells infiltrating the tumor exclusively have a diploid genome.
- **Cellular prevalence and observed allele frequency** are proportional to one another, given correction by the number of alleles mutated and the number of copies of the locus in the tumor

2.2.3 Mathematical model of clonal reconstruction

Read sampling during the sequencing process

Modeling the link between the proportion of reads carrying an alternative allele (and *a fortiori* the number of chromosomes) and the observed number of reads is a sampling issue. The easiest way to model this phenomenon is to use a binomial law, where the probability to draw an alternative allele is the proportion of alternative alleles in the population (figure 2.5).

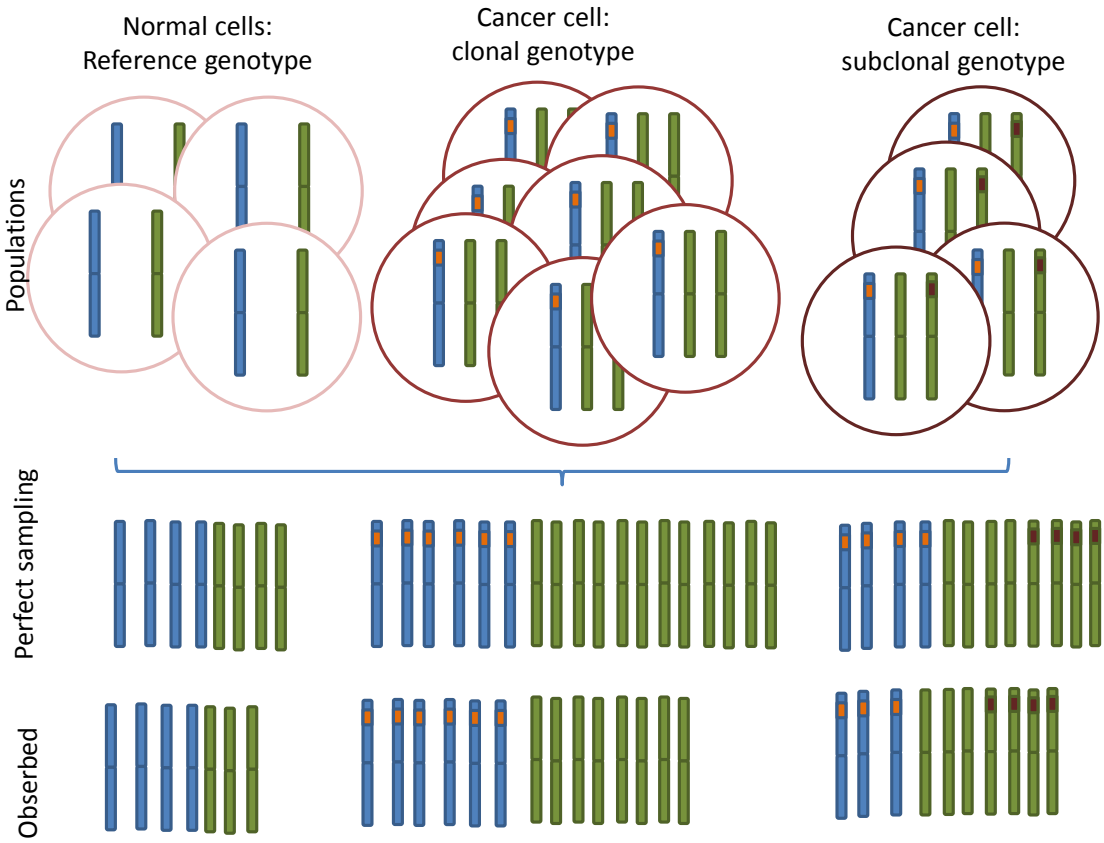


Figure 2.5: **Sampling model.** Alleles from paternal and maternal alleles are represented in blue and green. Mutations are represented with orange and red boxes. We here show what would be the result of a perfect sampling of alleles (i.e. if the proportions of the different alleles was exactly preserved), and an example of realistic sampling.

The sampling of reads can be affected by experimental conditions, for example depth of coverage in WGS data is highly linked to the G/C content of the reads. For a given depth of coverage, we model the sampling issue by a binomial law:

$$Alt \sim \mathrm{B}\left(n, p\right) \tag{2.3}$$

Where:

- Alt is the number of reads supporting alternative allele;
- n is the depth of coverage at the positions;
- p is the proportion of alternative alleles in the sampling populations. The observed p, often noted $\hat{p}$ in statistics, is the Variant Allele Frequency (VAF) in our problem.

In the case of a mixture of diploid populations without loss of heterozygosity (LoH), the proportion of alternative alleles is directly linked to the cellular prevalence of a variant, i.e. the fraction of cells bearing the variant (ç_{alt}):

$$p = \frac{\text{ç}_{alt}}{2} \tag{2.4}$$

However in the case of multiple populations with different number of copies of the locus, the probability of drawing the alternative allele has to be re-written:

$$p = \frac{\sum_i N_i \times \text{ç}_i}{\sum_i G_i \times \text{ç}_i} \tag{2.5}$$

With:

- N_i the number of copies of the variant allele in the cell population i;
- ç_i the fraction of cells with genotype i;
- G_i the number of copies of the locus in the population i.

In the hypothesis that the tumor has only one genotype in the sampled population, and that infiltrating cells all are diploid, we can rewrite the equation as:

$$p = \frac{\text{ç} \times N}{(1 - \text{ç}_{conta}) \times G + \text{ç}_{conta} \times 2} \tag{2.6}$$

Where ç_{conta} is the fraction of normal cells infiltrating the tumor, ç is the faction of cells bearing the alternative variant, N the number of copies of the variant, and G the number of copies of the locus in the tumoral cells. We can verify that in a triploid tumor AAB, without contamination ($\text{ç}_{conta} \to 0$, the variant in the ancestral clones will have probability $1/3$ if $N = 1$ or $2/3$ if $N = 2$, as expected.

Given that the number of copies of the locus can be inferred through statistical analysis of the data [44, 45, 46, 47], the problem to solve is to detect the different populations in the data and the number of copies of the variant in hyperdiploid loci.

Overview of the literature

Many different mathematical solutions have been proposed to solve the clonal reconstruction task [48, 49, 50, 51]. A summary and comparison of these methods

is the subject of a DREAM challenge[7] as well as a review[52].

We reproduced the list from Beerenwinkel et al. [52] summarizing the different algorithms and their principles (Table 2.1).

Software	Data	Model/Inference
PhyloSub[53]	SNV	Tree-stick-breaking process, binomial / MCMC
PyClone[49]	SNV	Dirichlet Process, beta-binomial / MCMC
SciClone[48]	SNV	Beta mixture model
Clomial[54]	SNV	Binomial / EM
Trap[55]	SNV	Exhaustive search under constraints
CloneHD[56]	SNV + CNA	HMM, EM, Variational Bayes
ThetA[45]	CNA	Maximum likelihood
cancerTiming[57]	CNA	Maximum likelihood
GRAFT[58]	CNA	Partial maximum likelihood
MEDICC[59]	CNA	Finite state transducer, Minimum-event distance
TuMult[60]	CNA	Breakpoint distance
TITAN[46]	CNA	HMM / EM

Table 2.1: Existing algorithms: we here reproduce the list of table from Beerenwinkel et al[52].SNV: Single Nucleotide Variant; CNA: Copy Number Aberration; MCMC: Markov-Chain Monte Carlo; EM: expectation maximization: HMM: Hidden Markov Model.

CNA based algorithms

In this section we will discuss the principles of the class of algorithms dealing with CNA that use two possible inputs: coverage and B-allele frequency (BAF).

This class of algorithms heavily relies on segmentation algorithms: the goal is to detect breakpoints (i.e. changes in the signal).
The depth of coverage of a portion of the genome heavily depends on intrinsic factors of the region (GC content, mappability), as well as experimental factors (sequencing kit). As a result, the data has to be normalized by a control, usually the constitutive DNA of the patient sequenced with the same protocol. After normalization and segmentation, changes in the depth of coverage indicate changes in the number of copies of the locus compared to a baseline. Tumors, however, can be hyperdiploid, meaning that the average number of copies of the tumor is higher than two.

[7] ICGC-TCGA-DREAM Somatic Mutation Calling Challenge – Tumor Heterogeneity and Evolution

Moreover, the median B-allele frequency can help distinguishing regions with allelic imbalance, as the B-allele frequencies will be shifted from the 50% position, as explained by Equation 2.6. The algorithm thus has to estimate the following parameters that best fit the model:

- $\dot{\varsigma}_{conta}$: the proportion of normal cells;
- A_iB_i: the number of copies A and B alleles of segment i that explain both BAF and depth of coverage.

Usually, numbers of A and B alleles are coerced to integer values, and the output then is the profile of the major clone (see Figure 2.6). For example, an heterozygous SNV in a triploid (AAB) locus without contamination should be observed either at 33% (B-allele) or 66% (A-allele). If we add contamination, we have to consider that normal cells contribute to the BAF by bringing one copy of the B-allele but also one copy of the A-allele. We can then write:

$$BAF_i = \frac{\dot{\varsigma}_{conta} + (1 - \dot{\varsigma}_{conta})B_i}{2 \times \dot{\varsigma}_{conta} + (1 - \dot{\varsigma}_{conta})(A_i + B_i)} \tag{2.7}$$

This formula can also be used to estimate the contamination on loci with odd number of copies. However, detection of sub-optimal solutions on segments can reveal subclonal copy number changes. The algorithm then has to estimate:

- $\dot{\varsigma}_{conta}$: the proportion of normal cells;
- n: the number of populations;
- $\dot{\varsigma}_j$: the fraction of cells with genotype j;
- $A_{i,j}B_{i,j}$: the number of copies A and B alleles of segment i in population j so that $\sum_j \dot{\varsigma}_j A_{i,j}B_{i,j}$ explains both BAF and depth of coverage.

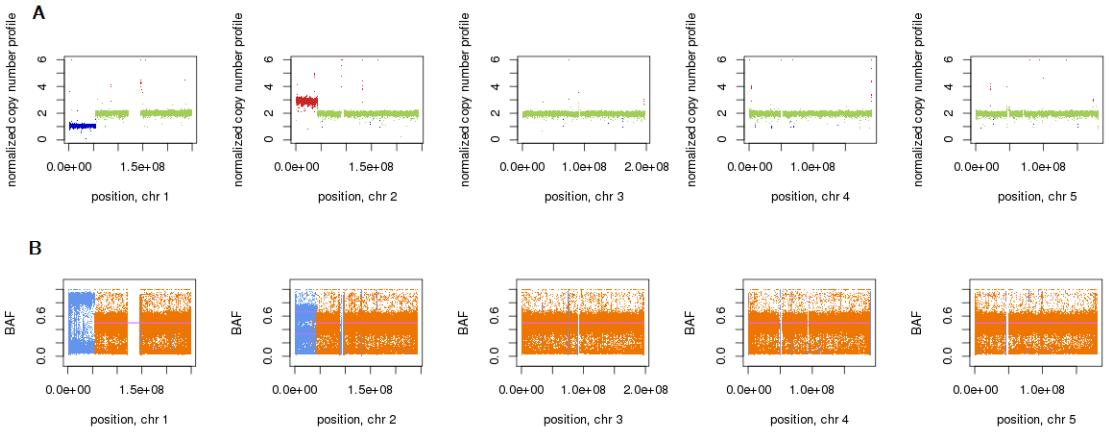


Figure 2.6: **Illustration of the copy number reconstruction for sample NB1361-D (diagnosis) of chromosomes 1 to 5. (A)** Copy number status of the different loci. The normal, diploid, state is shown in green, losses in blue and gains in red. The 1p loss is associated in this case with the 17q gain (not shown). **(B)** B-allele frequency: for each SNP, the VAF of the minor allele (in the population) is shown. For an heterozygous position in a diploid locus a VAF of 50% is observed. For odd number of copies, two states are shown, corresponding to major and minor alleles (see Equation 2.7). Due to contamination, the median BAF for single copy locus is higher than 0, and higher than 33% in triploid locus.

SNV based algorithms

While CNA class can function on their own, SNV based algorithms require that variants are called first to estimate the number of alternative and reference allele of the variant in all related samples (i.e. samples coming from the same patient). In addition, some algorithms require information from copy number and contamination as input, as we have seen in Equation 2.6 that the probability to draw a variant is related to its copy number status and contamination by normal cells.

We detail here two methods that will be used for comparison to our clonal reconstruction algorithm: sciClone[48] and pyClone [49]. These algorithms have been selected because of their widespread use [61, 62, 63, 64, 65].

These two methods rely on approximate inference. Markov Chain Monte Carlo (MCMC), used in pyClone, is a stochastic technique. ‘Given infinite computational resource, they [stochastic techniques] can generate exact results’[66]. Variational techniques, on the contrary, are based on a deterministic approximation of the posterior distribution.

Markov Chain Monte Carlo and pyClone

Before detailing complex Bayesian model, we first describe the basic concept of a graphical model as it can be easier to represent complex Bayesian models

using a graphical model. For the joint distribution $p_{(a,b)}$ of the two variables a and b, we can write:

$$p_{(a,b)} = p_{(b|a)}p_{(a)}$$

This would be represented by Figure 2.7A. If we extend to three variables, adapting the example from Bishop et al [66], we can write (using Bayes Theorem):

$$p_{(a,b,c)} = p_{(c|a,b)}p_{(b|a)}p_{(a)}$$

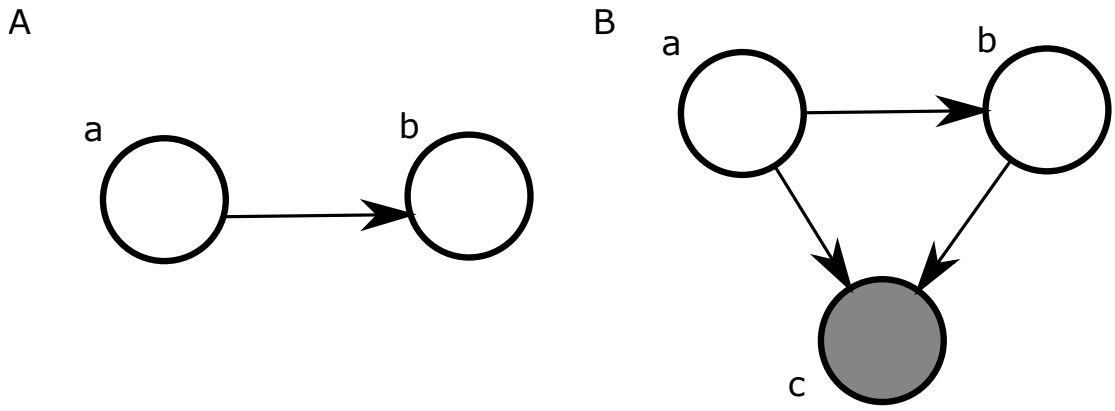


Figure 2.7: **Graphical model representation (A)** Representation of $p_{(b|a)}p_{(a)}$ **(B)** Representation of $p_{(c|a,b)}p_{(b|a)}p_{(a)}$. The c variable is colored to show that it is observed.

Using this representation, the model for pyClone can be written as in Figure 2.8.

This model shows (on the left hand side), that a proposed cellular prevalence for variant m will first be sampled from a Dirichlet process with uniform distribution (in $[0; 1]$), and accepted or rejected based on the user defined parameters a_α and b_α. Due to the Dirichlet process, even though an infinite number of states exists, those states are discrete. As a result, a variant can either take a cellular prevalence Φ^m that is new or used. If two variants use the same cellular prevalence, they are considered to belong to the same cluster. More details about the implementation of pyClone can be found in the supplentary Figure 3 and supplementary note of Roth et al[49].

We can note that this sampling method gives access to the whole distribution of cellular prevalence in the sample. However, the first pass of the algorithm requires $O(sm^2)$ operations, and each subsequent pass will require $O(skm)$ operations, where k is the number of remaining clusters, m the number of variants, and s the number of samples. This high complexity of the model illustrate the note

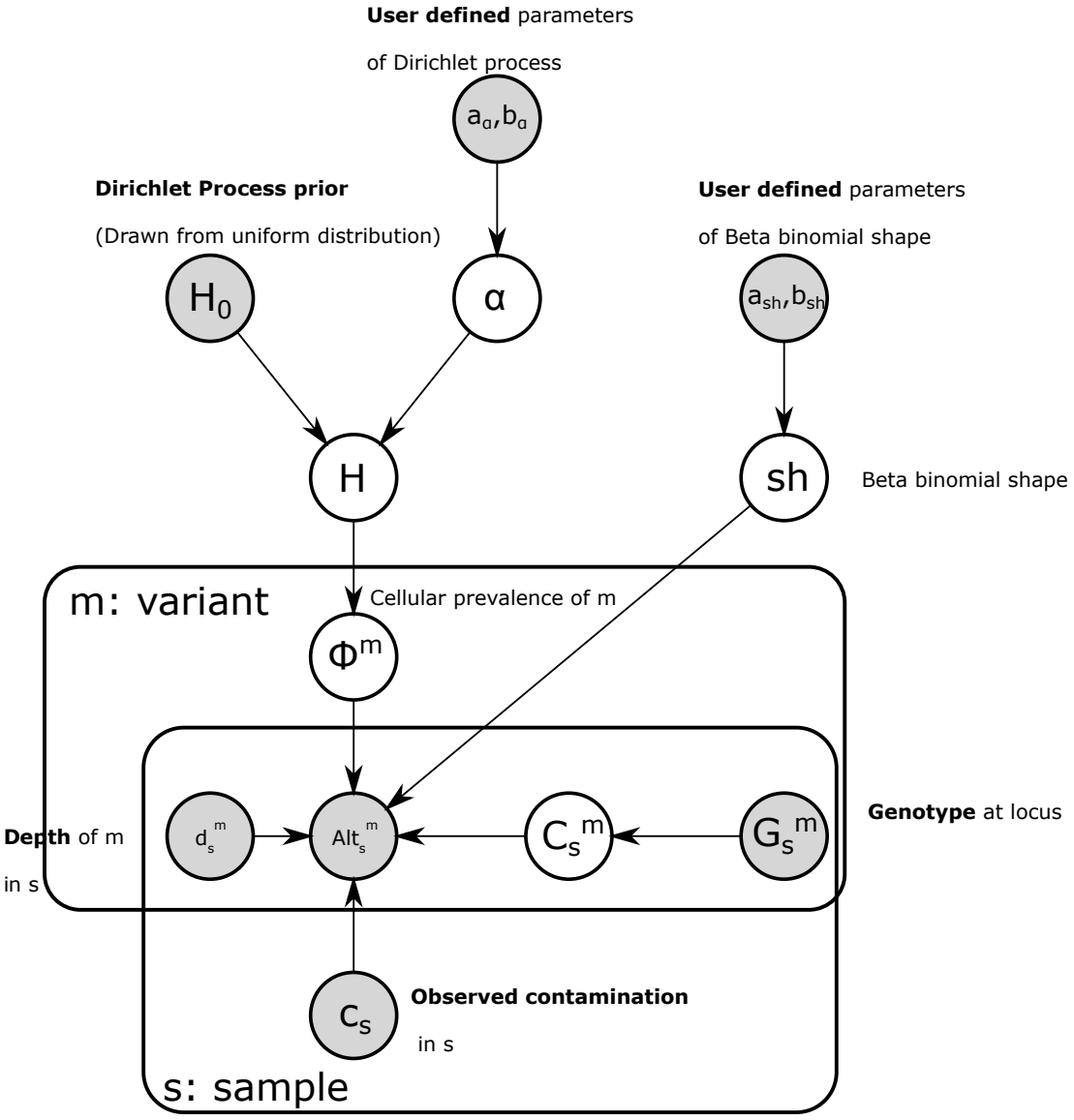


Figure 2.8: **Graphical representation of the pyClone model**. This representation shows how genotype information, user defined parameters and observed depth of a variant in a sample are used to estimate the cellular prevalence of a clone. This figure has been adapted from Roth et al[49] for consistency with notations and clarity. In their model Φ^m and sh are combined using a beta binomial distribution.

from Bishop et al[66]: 'In practice, sampling methods can be computationally demanding, often limiting their use to small-scale problems.'

Variational Bayesian Mixture Model and sciClone

sciClone relies on a variational bayesian mixture model (VBMM)[66]. Variational inference is a deterministic method, that can find approximate solutions. The log marginal probability can be decomposed as following:

$$\log p_{(X)} = \mathcal{L}_{(q)} + KL_{(q||p)} \tag{2.8}$$

Where KL is the Kullback-Leibler divergence:

$$KL_{(q||p)} = -\int q_{(Z)} \log\left(\frac{p(Z|X)}{q(Z)}\right) dZ \tag{2.9}$$

and

$$\mathcal{L}_{(q)} = \int q_{(Z)} \log\left(\frac{p(X,Z)}{q(Z)}\right)$$

Here X is the random variable and Z is a set of latent variables.

Variational inference will then minimize the Kullback-Leibler divergence for a given class of functions q. sciClone makes the assumption that the data follows a beta-binomial mixture model to take into account possibly higher variability in the observed VAF. Miller et al make the classic assumption that the q distributions can be factorized (Supplementary Text S1 section C[48]), but add the condition that latent variables must be independent and non-overlapping.

In the section C of supplementary text S1, Miller et al. [48] also define the expectation with respect to the distributions q_j ($j \neq i$) of the approximate posterior distribution:

$$E_{j\neq i}[\log p(X,Z)] \equiv \int \log p(X,Z) \prod_{j\neq i} q_j \partial Z_j$$

From this, they conclude that at each pass, the value of q_i that minimizes the KL divergence is:

$$\log q_i \propto E_{j\neq i}[\log p(X,Z)]$$

The algorithm reaches convergence when the maximal difference between two passes is lower than 10^{-4}.

Contrary to pyClone, sciClone voluntarily restricts clonal reconstruction to variants in copy neutral and LoH-free regions. In the next chapter we will present a new solution to the clonal reconstruction task that relies on an expectation maximization procedure, called QuantumClone.

Chapter 3

Efficient solving of the clonal reconstruction task

> *Can a genetic accident of unpredictable biological properties be taken into account in the Seldon plan?*
>
> — Isaac Asimov, *Second Foundation, 1953*

In this chapter we will detail our implementation of a clonal reconstruction algorithm: QuantumClone.

3.1 Implementation

In this section we will extensively explain the implementation of QuantumClone.

3.1.1 Expectation Maximization

In order to uncover the parameters of the clonal populations in the data, we chose to use an EM method.

We use the algorithm described in algorithm 3 in diploid cases.

We define a stopping threshold $\eta \in \mathbf{R}_+^*$, and we say that the convergence criterion is reached,if between two rounds of optimization we have:

$$\max_k \left(|\dot{\varphi}_{k,s,n} - \dot{\varphi}_{k,s,n+1}|, |\omega_{k,n}\omega_{k,n+1}| \right) < \eta$$

Where $\dot{\varphi}_{k,s,n}$ is cellular prevalence of a cluster k in sample s for iteration n, and $\omega_{k,n}$ is the weight of cluster k at iteration n. This principle is illustrated in figure 3.1, which shows updates of probabilities on E-steps and changes in clone centers and weights during M-steps. The exact values after each step are given in Table 3.1.

If we look at the clone centers and weights after each maximization step we see:

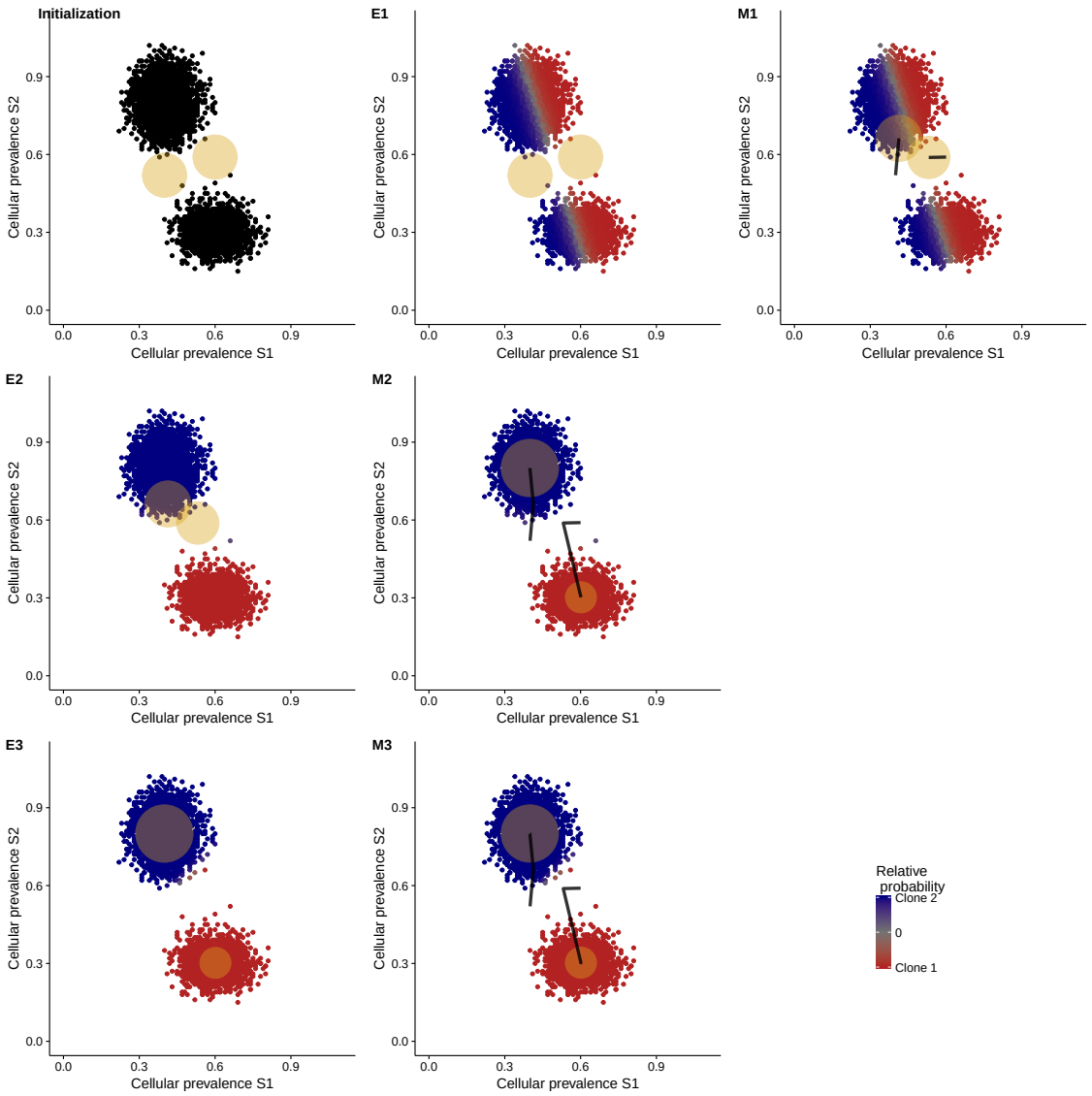


Figure 3.1: **Example of EM convergence**. Two clones are generated with cellular prevalence 0.6 and 0.3 for clone 1 and 0.4 and 0.8 for clone 2. Clone 1 bears 65% of the 5000 generated variants, all with a depth of 200$\times$.
In yellow we show centers of the clusters, with the size of the point proportional to the weight of the cluster. The color of the variant shows the relative probability of a variant to belong to cluster 1 or 2, given by the following law:
$p = \frac{p2-p1}{p2+p1}$, where $p2$ (respectively $p1$) is the probability to belong to cluster 2 (resp. 1)
The trajectory of each clone center is shown after each update (M steps).

Algorithm 3: Expectation Maximization algorithm underlying QuantumClone

```
for c in number of clones; do
    for I in initializations; do
        initialize distribution parameter: ç_{k,0}, ω_{k,0} the cellularity
         and weight of clone k;
        while not converged; do
            E-step: update f_{i,k} probability of variant i to belong
             to k;
            M-step: find ç_{k,n}, ω_{k,n} to maximize log-likelihood ℓ,
             knowing f_{i,k};
        end
    end
end
```

Iteration	Clone1 S1	C1 S2	Clone2 S1	C2 S2	Weight C1	Weight C2
Ground truth	0.600	0.300	0.400	0.800	0.650	0.350
Start	0.600	0.590	0.400	0.520	0.500	0.500
M1	0.531	0.588	0.413	0.662	0.521	0.479
M2	0.601	0.302	0.399	0.800	0.652	0.348
M3	0.601	0.303	0.399	0.801	0.651	0.349

Table 3.1: **Cluster values for EM example** The algorithm converges toward the real position of the clone centers. Simultaneously, the proportion of cells belonging to a clone converges to the weights used to generate the data. Finally, the algorithms stops when the difference between two observations is smaller than 1%. **Abbreviations: C**lone, **S**ample

Solving the unknown number of copies of a variant

Equation 2.6 shows that the number of copies of the variant is required to compute the probability of a variant to belong to a clone. The status of each variant is unknown before the clustering unfortunately. As a consequence, the algorithm has to find the correct number of copies of the variant among all possible statuses, as shown in figure 3.2.

To do so, all possible states of a variant are described and will be used in a first round of EM. Then, only the values that are the most likely — i.e. the ones with the highest probability to belong to a clone — are kept, as shown in figure 3.3. In order to avoid that contribution increases with the number copy status possibilities, the contribution of each variant to the model is normalized to 1. Then another round of EM is started with only the selected value for the copy number of the variant.

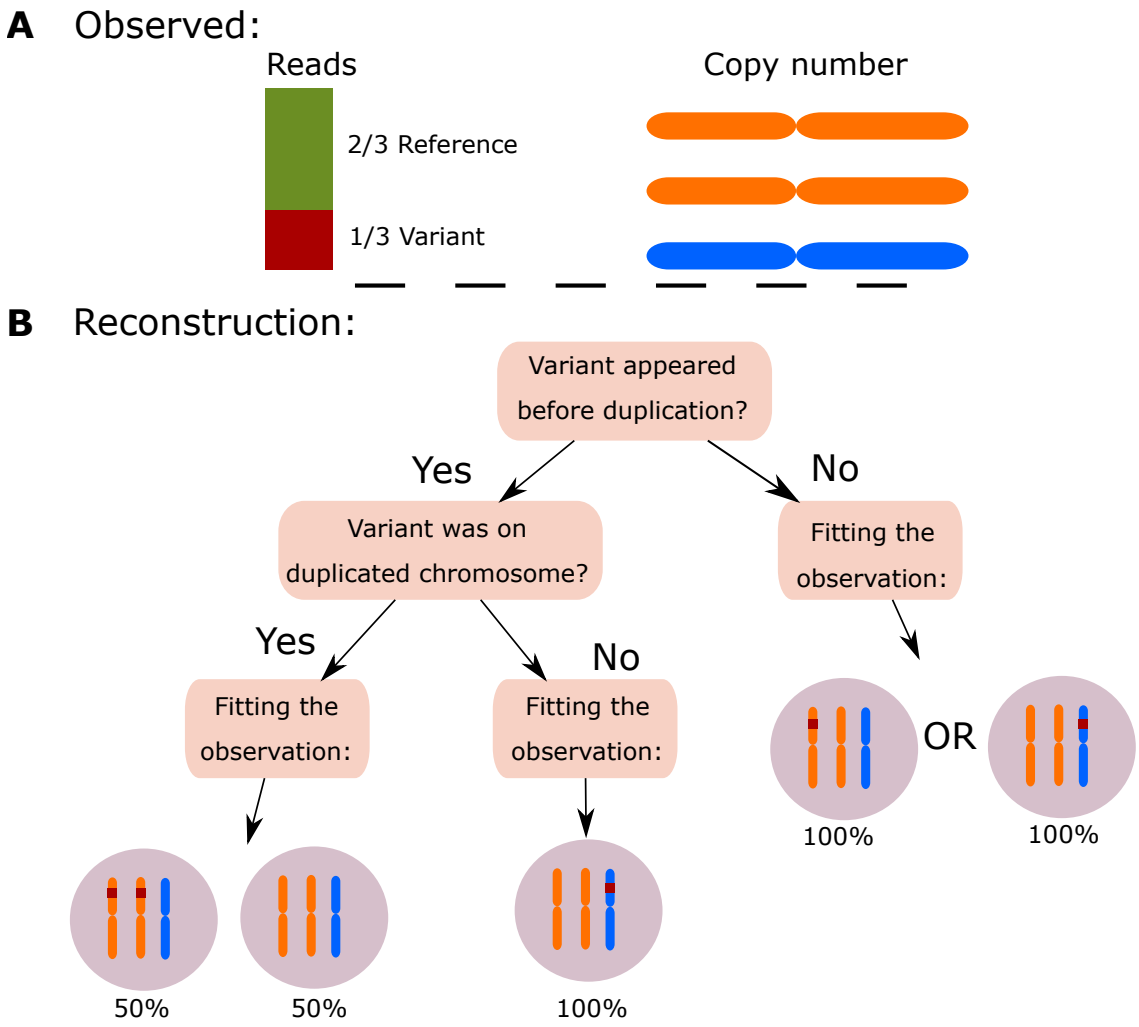


Figure 3.2: **Copy number status of a variant. (A)** In this example, a variant is detected on $1/3$ of the reads, and the genotype of the locus is AAB. **(B)** With these information, several cases fitting the data are shown.

In the case of polyploid tumors, the algorithm uses two rounds of EM, as described in algorithm 4.

Finally the best model from the $c \times I$ computed is chosen thanks to an information criterion, either Akaike (AIC), Bayesian (BIC), or modified Bayesian.

3.1.2 Incremental upgrades to the maximization step

In order to converge, we needed to maximize the log-likelihood function that can be written in our case:

$$\ell = \sum_{i \in variant} \sum_{k \in clones} \sum_{s \in samples} \sum_{p \in possibilities(i)} \beta_{(i,p)} f_{(i,k)} \log \left(P_{i,s,p} \left(\text{¢}_{i,s} | \text{¢}_{k,s} \right) \right) \quad (3.1)$$

With $\beta_{(i,p)}$ a coefficient so that for a given variant i, $\sum_p \beta_{(i,p)} = 1$. This has been

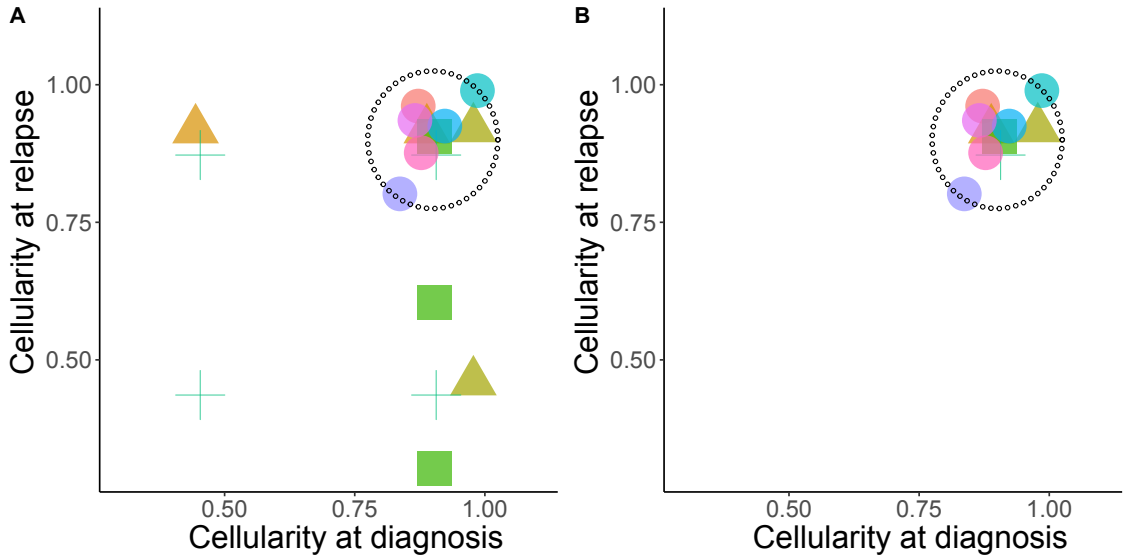


Figure 3.3: **Principle of selection of variant cellularity for variants in polyploid regions.** Mutations located in regions of copy number aberration can be present on several chromosomal copies; they can thus be assigned to several cellular prevalence values (panel A). After the Expectation Maximization (EM) step each mutation is attributed the most likely cellular prevalence value (panel B). Each mutation is represented by a specific color. Mutations located in AB regions (circles); mutations located at relapse in regions of gain (squares), mutations located in regions of gain both at diagnosis and relapse (triangles).

added to prevent higher contribution to the model from variants with higher copy number status. Indeed, as shown in Figure 3.2, a variant in a triploid region has two possible copy status whereas a variant in a diploid region only has one. Without this correction, a variant in a triploid region would contribute twice as much to the model, but only one of the two possibilities would be true.

At first, the maximization step used the `optim` function in R, using Broyden-Fletcher-Goldfarb-Shanno ('BFGS ') algorithm. This method relied on a numeric differentiation of the log-likelihood function. This box constrained optimization on the $[0;1]^{n\times s}$ space, where n is the number of clones and s the number of samples, requires many calls to the computation function.

In order to reduce the computational time, the exact gradient was provided (see annex 8.1 for computation). This can effectively decrease computational time when variants are in hyperdiploid loci. However, looking at the exact formula of the gradient (see 8.1), we can see that when all variants have the same adjusting factor to go from VAF to cellular prevalence, then it is easy to find the exact 0 of the function 3.2. This is only possible when all variants are in either haploid or diploid regions. Then, the adjusting factor simply is a transition from observed

Algorithm 4: Expectation Maximization algorithm underlying QuantumClone, when at least one variant lies in hyperdiploid or LoH region.

for *c in searched clone value;* **do**
  for *I in initializations;* **do**
    initialize distribution parameter: $\dot{\varsigma}_{k,0}, \omega_{k,0}$ the cellularity and weight of clone k;
    while *not converged;* **do**
      E-step: update $f_{i,k}$ probability of possibility i to belong to k;
      M-step: find $\dot{\varsigma}_{k,n}, \omega_{k,n}$ to maximize log-likelihood ℓ, knowing f_i, k;
    end
    select most likely position
    while *not converged;* **do**
      E-step: update $f_{i,k}$ probability of variant i to belong to k;
      M-step: find $\dot{\varsigma}_{k,n}, \omega_{k,n}$ to maximize log-likelihood ℓ, knowing f_i, k;
    end
  end
end

VAF to cellular prevalence and is the same for all variants.

$$\frac{\partial \ell}{\partial \dot{\varsigma}_{k,s}} = 0 \Leftrightarrow \dot{\varsigma}_{k,s} = \frac{\sum_{i \in variants} t_{i,k,s} \times Alt_{i,s}}{\alpha \times \sum_{i \in variants} t_{i,k,s} \times Depth_{i,s}} \tag{3.2}$$

With $t_{i,k,s}$ the contribution of possibility p to cluster k and α the adjusting factor, as explained in section 8.1.

Effectively computing the zero of the function has allowed a decrease in computational time of orders of magnitude, and explains the gains in performance compared to other published methods as we will see in the following part.

3.1.3 Improvements in the initialization procedure

It is a common practice to use the result of a k-mean clustering algorithm to initialize EM algorithms[1]. We first used a k-medoid algorithm from `fpc` R package.

In order to improve clustering results, we came with a new initialization procedure. We use a hierarchical clustering of the variants using as dissimilarity the p-value obtained from the z-score of two variants being from the same population.

[1] Francis Bach, *K-means, EM, Mélanges de Gaussiennes, Théorie des graphes*, http://www.di.ens.fr/~fbach/courses/fall2010/cours3.pdf

The number of clusters to look for is a parameter provided by the user — as a range of values. For each tested value n, we can initialize the algorithm with n centers found by cutting the hierarchical tree in n groups. The value of the centers is then taken as an average of the cellular prevalence of variants from the cluster weighted by the depth of sequencing of each variant (see Equation 8.1 in chapter 8). This second method provided more accurate starting points, reducing the number of steps required to converge — and doing so, the computational time. It also provides better results with a decreased chance of falling in a local minimum.

3.2 Experimental comparison of methods

In this section we will focus on the validation of our method, called QuantumClone, through simulated data, and we will compare it to sciClone and pyClone.

3.2.1 Comparison methodology

Simulating cancer samples

The simulations of genomic data from cancer samples have been made using `QuantumCat`, part of the `QuantumClone` R package. First, a phylogenetic tree is created with the following set of properties:

- In this tree, in a given sample the summed cellular prevalence of progeny cannot be higher than its ancestor;
- The number of clones — nodes and leaves of the tree bearing at least one observed variant — is a parameter of the simulation;
- As clones can be nodes of the tree, a variant can only belong to a single clone;
- The tumor stems from a single ancestral clone.

Then, for each variant in each sample, a depth of coverage of the position is simulated using a negative binomial distribution using the chosen mean depth of coverage and parameters fitted on experimental data: coverage of variants called from the patients published in Eleveld et al [67]. In hyperdiploid cases, the genotype and number of copies of each variant is also simulated.

Then the number of alternative reads for each variant is drawn with a binomial law, as explained in 2.3.

To validate this model, we hereby present the simulations of a $50\times$ average depth of sequencing, where all simulated variants are heterozygous, and the observed germline VAF distribution from patient NB0784 (see Figure 3.4).

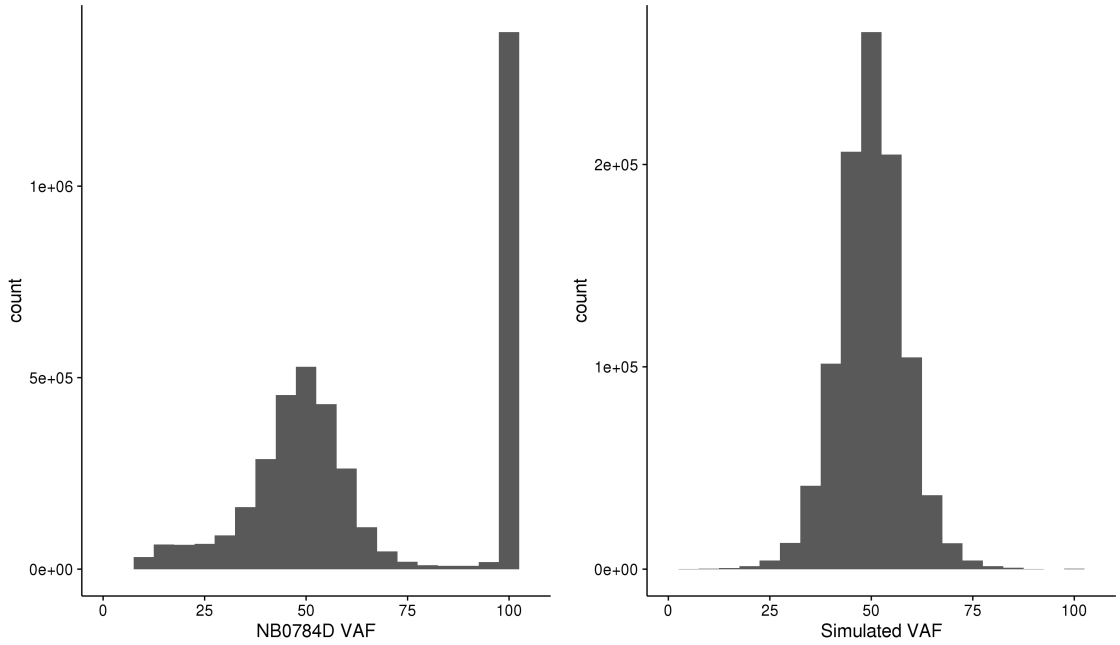


Figure 3.4: **Distribution of variant allele frequency (VAF). Left:** VAF distribution for variants called in the diagnosis sample from patient NB0784 and annotated as germline by varscan 2. **Right**: the simulated VAF distribution of heterozygous variants for a $50\times$ average coverage (10^6 simulated variants).

Evaluation of algorithms

The clonal reconstruction by the three algorithms is then assessed using normalized mutual information (NMI), euclidian distance error (ℓ^2 error), and computational time. NMI and ℓ^2 will be detailed below.

The NMI assesses the mutual information of two group partitions, one being the classes created by the simulation and the second being the reconstructed clusters. In this section we will note Ω the set of clusters ω_k and $\mathbb{C}$ the set of clones c_j . Note that for NMI, the two sets are interchangeable. Also, we will note $|X|$ the cardinal of set X , i.e. the number of elements in X. The NMI can then be written:

$$NMI_{(\Omega,\mathbb{C})} = -2 \times \frac{\sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} \log\left(\frac{N \times |\omega_k \cap c_j|}{|\omega_k||c_j|}\right)}{\sum_k \frac{|\omega_k|}{N} \log\left(\frac{|\omega_k|}{N}\right) + \sum_j \frac{|c_j|}{N} \log\left(\frac{|c_j|}{N}\right)}, \tag{3.3}$$

Where N is the number of observations (variants used for the clustering in our case). NMI is a positive function of the reconstruction quality and is bounded by 0 (no mutual information between reconstruction and ground truth) and 1 —- case of a perfect clustering.

The average ℓ^2 error measures the average distance in ℓ^2 norm between the cellular prevalence of the cluster center attributed to the variant and the real cellular prevalence of this variant.

For each parameter tested, 50 simulations were generated. Each simulation was stored and given as input to all three algorithms. In our simulation experiments, the following parameters varied within realistic ranges: depth of sequencing (100$\times$ to 1000$\times$), fraction of contamination by normal cells (from 0 to 70%), number of variants used for the clonal reconstruction (from 50 to 200), number of tumor samples used for each patient (from 1 to 5) and number of distinct clones per cancer (from 2 to 10).

3.2.2 Results from *in silico* experiments

This section has been adapted from Clonal assessment of functional mutations in cancer based on a genotype-aware method for clonal reconstruction, Deveau et al (see section 9.2).

Our analysis showed that QuantumClone is equivalent to or better than the best published algorithm in clustering quality (Figure 3.5A) for diploid genomes. Although in terms of NMI QuantumClone showed similar performances compared to pyClone, QuantumClone generally outcompeted sciClone for NMI ($p-value < 2.2 \times 10^{-16}$, Welch two sample t-test). In particular, in samples with 50% contamination by normal cells QuantumClone drastically outperformed sciClone ($p-value = 3.6 \times 10^{-10}$ Welch one-side two-sample t-test). On average, QuantumClone decreased the ℓ^2 mean error by 69% compared to sciClone and 22% compared to pyClone, significantly improving predictions compared to both methods ($p - value < 2.2 \times 10^{-16}$). At high values of sequencing depth, all methods accurately estimated prevalence of variants (Figure 3.5B, ℓ^2 mean error < 0.059 at 1000$\times$ for all methods). However, at depth of sequencing of 100$\times$, which is the depth of sequencing currently used for the majority of WES and WGS experiments, QuantumClone consistently gave better predictions than pyClone ($p - value = 1.5 \times 10^{-6}$, Welch one-sided two-sample t-test) and sciClone ($p - value = 4.9 \times 10^{-9}$). In addition, compared to the other methods, QuantumClone took the best advantage of data when multiple samples were provided for

the analysis ($p - value = 2.4 \times 10^{-10}$ and $< 2.2 \times 10^{-16}$ for pyClone and sciClone respectively, Welch one-sided two-sample t-test, for simulated tumors with five samples).

Also, the average computational time was significantly decreased using QuantumClone compared to sciClone (median 35 fold improvement), or pyClone (median 46 fold improvement, Figure 3.5C). In the case of highly heterogeneous tumors (e.g. tumors with 10 simulated clones), the gain in computational time was of 41 fold ($p - value < 2.2 \times 10^{-16}$) compared sciClone and 44 fold ($p - value < 2.2 \times 10^{-16}$) compared to pyClone. Similarly, when five samples were provided, we observed a 74.1 fold ($p - value < 2.2 \times 10^{-16}$) compared to pyClone and 74.2 fold ($p - value < 2.2 \times 10^{-16}$) compared to sciClone.

We expect that in addition to the parameters discussed above, the degree of genome rearrangement and chromosome duplication significantly affects the quality of the mutation clustering and consecutive clonal reconstruction. Indeed, values of cellular prevalence are linked to VAF values through the parameters representing the number of copies of the variant and the number of copies of the reference allele. Given an observed VAF value, a variant occurring in a high copy number locus has more possibilities for values of cellular prevalence: a variant with an observed allele frequency of 25% can only be linked to a cellular prevalence of 50% in a AB locus, while this variant can arise from cellular prevalence values of 33.3%, 50% or 100% if the genotype at this locus is AAAB.

In order to validate QuantumClone on diploid and hyper-diploid genomes, we simulated variants in loci of genotype AB, AAB, AABB, and in a nearly diploid genome, where all possible genotypes can be observed (Figure 3.6). We excluded sciClone from this experiment as it cannot use variants from non-diploid regions.

In all types of regions, QuantumClone and pyClone performed equally in terms of NMI (Figure 3.6A), but QuantumClone outperformed pyClone in terms of mean ℓ^2 error with an improvement of 31% (Figure 3.6B, $p - value = 5.7 \times 10^{-11}$). In addition, QuantumClone without parallelization was faster than pyClone in three out of four setting (from 6.3 fold slower to 61.5 faster; 15.6 times faster on average), while the distributed algorithm outcompeted pyClone in all settings (average computational time decreased by a 43 fold compared to pyClone, Figure 3.6C).

In addition, in the majority of cases QuantumClone correctly assumed the ex-

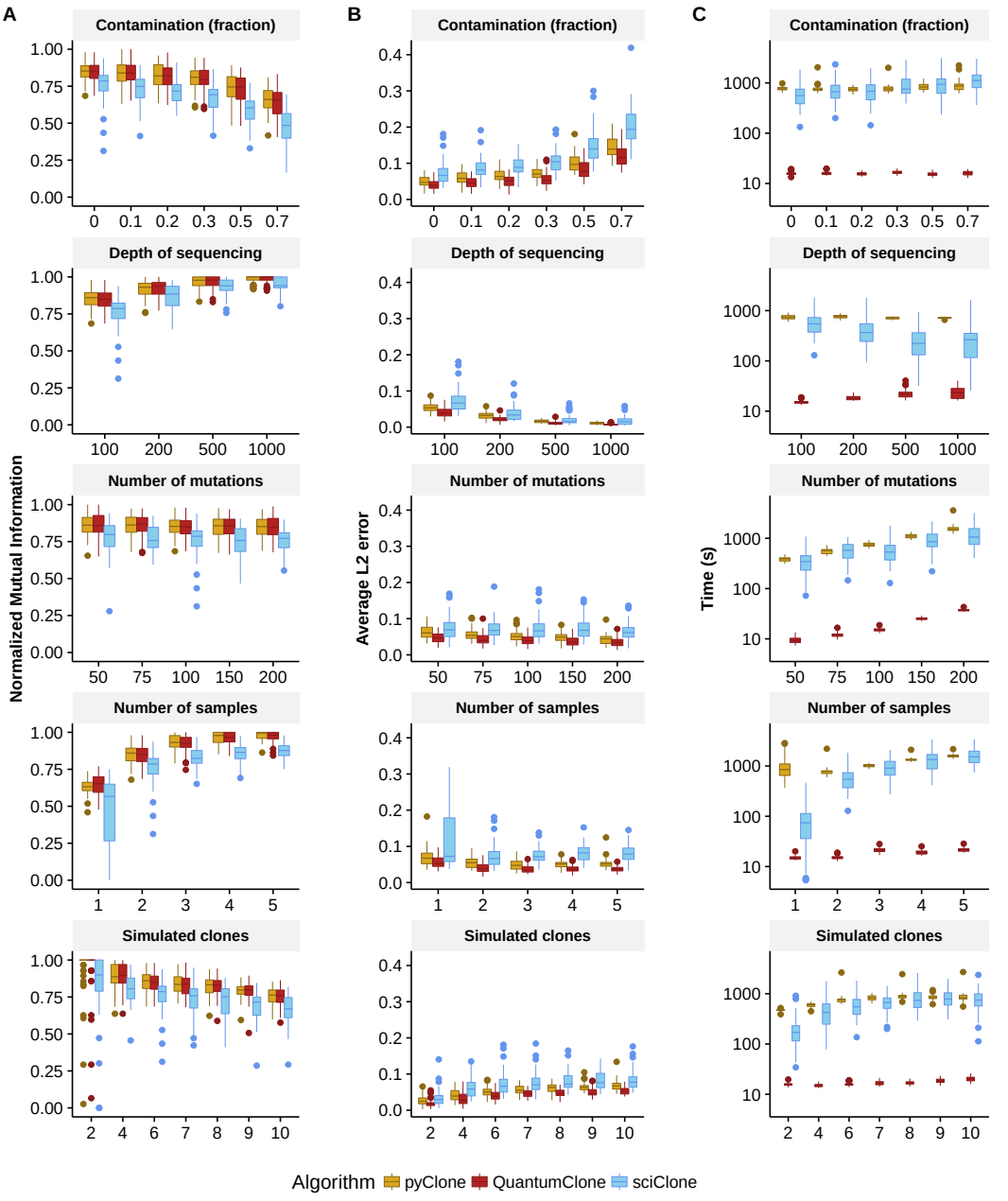


Figure 3.5: **Comparison of QuantumClone to existing methods. (A)Normalized Mutual Information (NMI)** is used to assess the quality of variant clustering on simulated data, with a single parameter varying in each test. This measure evaluates correct assignment of two variants to the same cluster. QuantumClone (red) shows equivalent performance to the best tool in each settings. **(B) L2 average error** is used to assess the error for each clustered variants between its simulated position and its reconstructed position. **(C) Computational time** necessary to complete the clustering with each algorithm. Default parameters: two tumor samples without contamination sequenced at 100×; 6 clones; 100 mutations used for clustering.

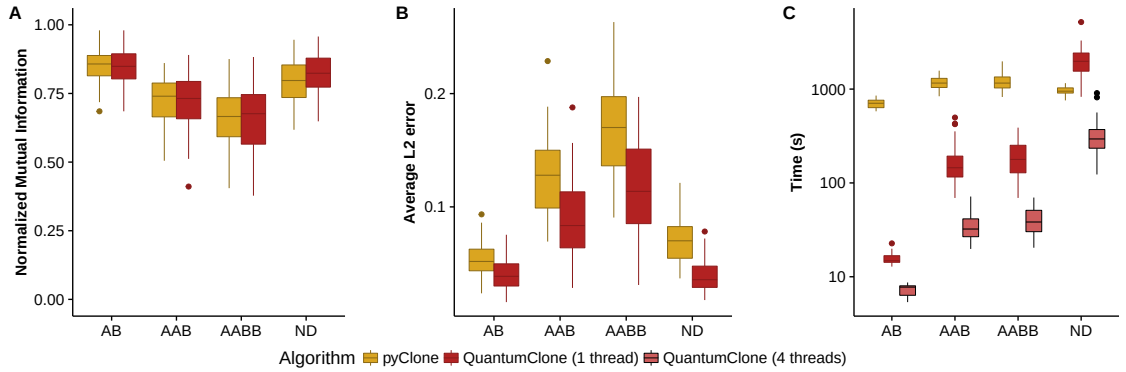


Figure 3.6: **Quality of clonal reconstruction for mutations located in regions of altered copy number. (A) Normalized Mutual Information** shows equivalent performances of pyClone and QuantumClone in diploid, triploid and tetraploid tumors, or nearly diploid (ND), whereas the **average L2 error (B)** shows significantly better results for QuantumClone.**(C)** QuantumClone can use parallelization to handle longer computations that can be due to visiting all possible variant copy states.

act number of copies of a variant in polyploid regions (average accuracy $= 68.9\%$, $p - value < 2.2 \times 10^{-16}$, Figure 3.7).

3.2.3 Validation of the algorithm for hyperdiploid genomes

In order to demonstrate the validity of our approach on hyperdiploid samples, we hereby show results from the comparison of QuantumClone to a version that is forced to predict all variants at the single copy level, extending the results from Figure 3.7.

As suggested by figure Figure 3.7, the correct selection of copy number status by QuantumClone greatly improves clustering quality with higher NMI, and decreases the average ℓ^2 error (see Figure 3.8).

3.2.4 Improvements in the QuantumClone algorithm

In this section, we illustrate the improvements made in QuantumClone through time on simulated data. Even though the reconstruction algorithm has been highly modified during the three years, the QuantumCat function for data generation has been mostly conserved. This allows comparison of the different versions of the tool.

As ℓ^2 average error was only considered as a possible metric between July 2016 and January 2017, it is not displayed in Figure 3.9.

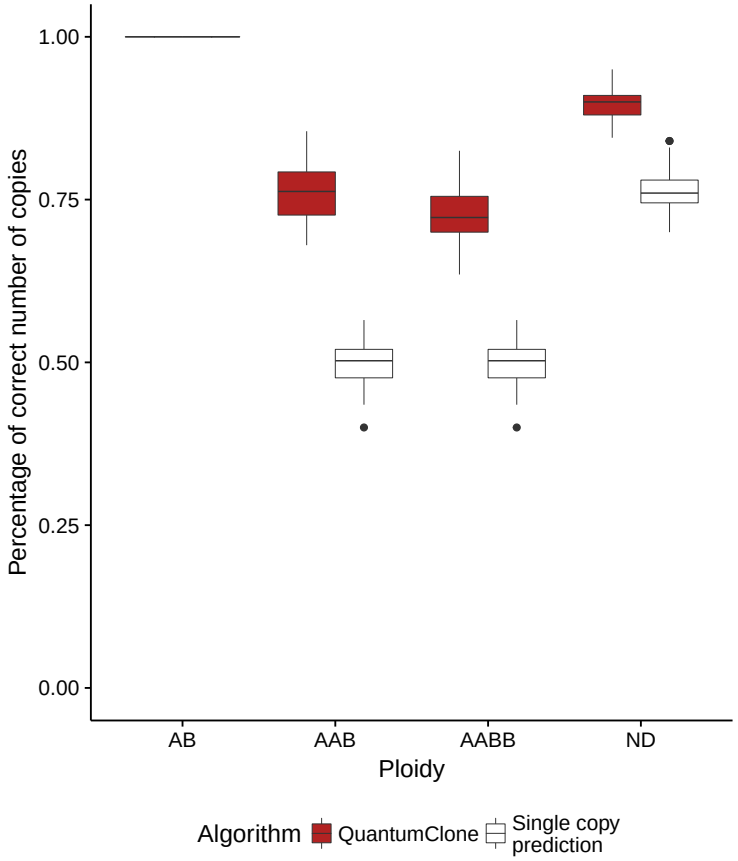


Figure 3.7: **Fraction of correct guesses**: QuantumClone is compared to a predictor that would assume that all variants are at the single copy state. As only a single state is available for AB regions, both predictors achieve 100% accuracy. In hyperdiploid regions, the maximal number of copies is determined by the number of copies of the A-allele.

The first thing to notice is that the latest version is the one that has been the most thoroughly evaluated.

Secondly, it may be surprising that the first version better dealt with single sample data. This was in fact due to an error in the phylogenetic tree generation for single samples that also prevented correct evaluation of QuantumClone on this parameter a year later. For all other parameters we can see the incremental gain in reconstruction quality. This is especially brought to light by the varying number of simulated clones. This is also due to the fact that the default number of simulated clones has changed between 2016 and 2017, and went from four to six clones.

If the increase in quality between 2015 and 2016 was detrimental to computational time — as the gradient descent step was becoming more complex — by

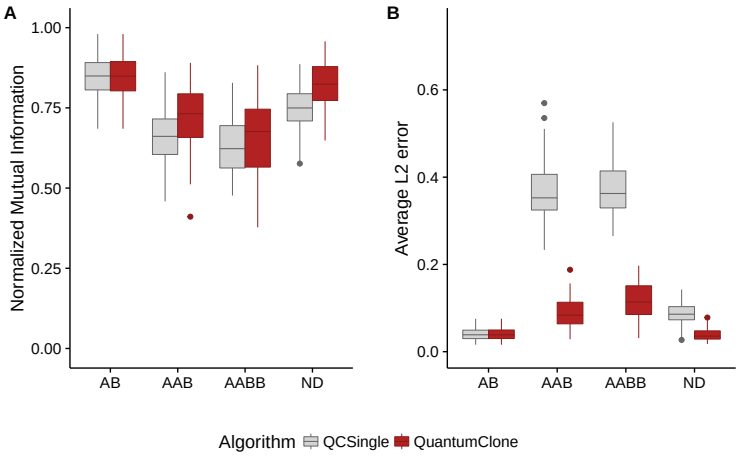


Figure 3.8: **QuantumClone vs QuantumCloneSingle**: QuantumClone is compared to a its derivative where all variants are predicted at the single copy state. **(A)** Comparison on normalized mutual information shows poorer performance of the Single algorithm, with decreased NMI, especially in the case of nearly diploid tumors **(B)** The ℓ^2 error shows drastically increased error when the algorithm is forced to select the single copy level, especially in strictly triploid and stricly tetraploid tumors.

switching to an accurate computation of gradient zeros we were able to drastically decrease the computational time. Another factor that can explain the changes between the penultimate and final versions of QuantumClone is the initialization procedure, as described in subsection 3.1.3.

Last but not least, the total computational time for the simulated tests depicted here (i.e. QuantumClone on diploid simulations only) amounts to 362,134.8 s or $\sim 100h$.

3.3 QuantumClone guidelines harnessed from simulations

In this section, we will distinguish intrinsic factors of the tumor, that cannot be known *a priori*, such as the fraction of normal cells in the sample, the number of mutations or the number of clones in the sample, and the extrinsic factors: depth of sequencing or number of samples sequenced. We provide guidelines to chose appropriate values for factors extrinsic to the tumor.

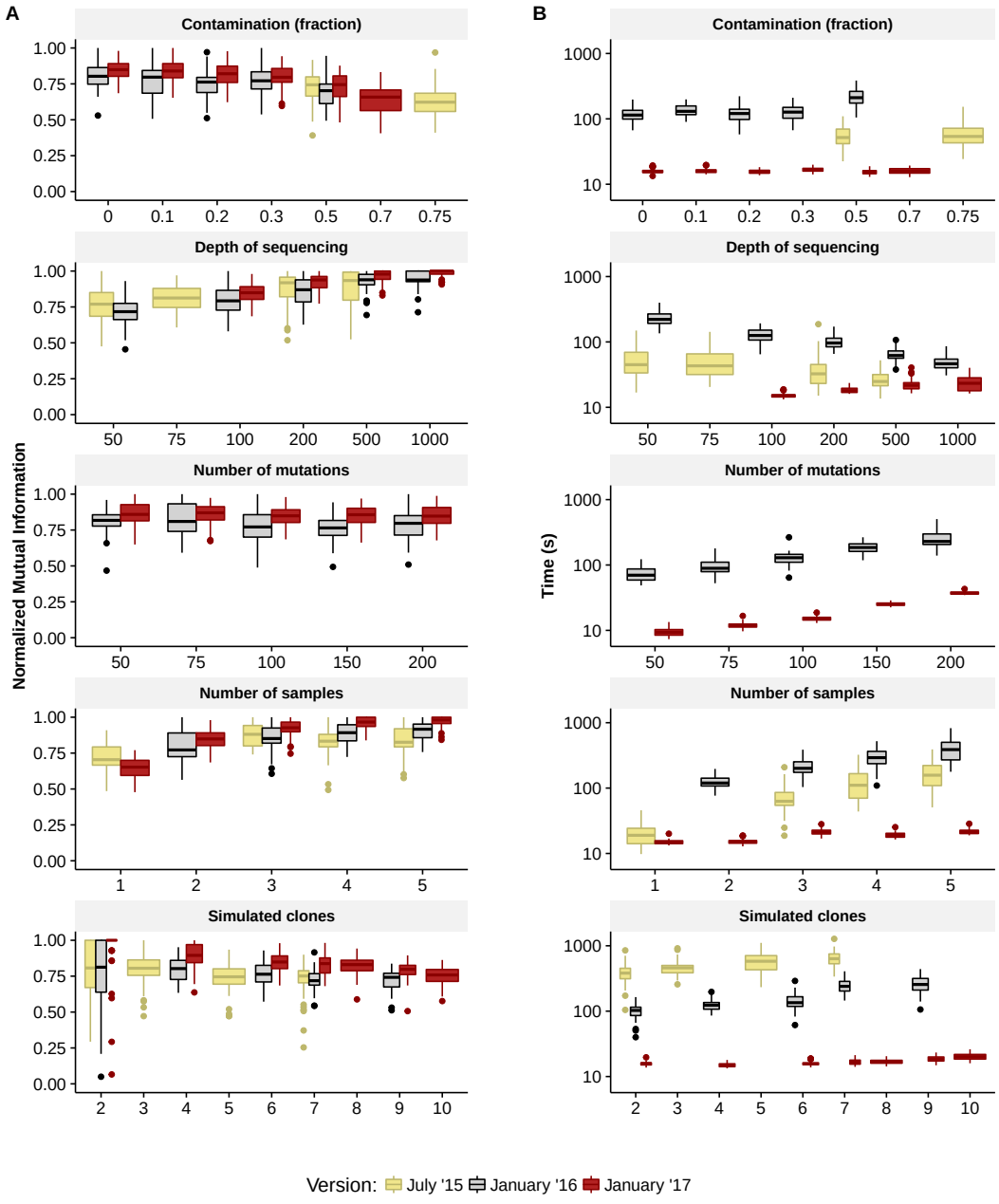


Figure 3.9: **Comparison of QuantumClone versions. (A) Normalized Mutual Information (NMI)** is used to assess the quality of variant clustering on simulated data, with a single parameter varying in each test. This measure evaluates correct assignment of two variants to the same cluster. **(B) Computational time** necessary to complete the clustering with each algorithm. Default parameters: two tumor samples without contamination sequenced at 100×; 100 mutations used for clustering.

3.3.1 Impact of extrinsic factors on the reconstruction

From the simulations, we can extract a few rules of thumb. First, doubling the sequencing depth — going from 100$\times$ to 200$\times$, only increases the NMI by an average of 8.7% (from 0.85 to 0.92) whereas doubling the number of samples (going from two to four samples) results in an increase of 13.1% (from 0.85 to 0.96). In addition, using several samples from the same patient can ease uncovering the phylogenetic tree of the tumor. Note that the samples may not necessarily come from different time points but can also come from different localization of the same tumor, or a combination of both space and time changes. Note that the observed ℓ^2 error decreases of 46% (from 0.041 to 0.022) for a doubling of sequencing depth compared to 7% (from 0.041 to 0.038) if the number of samples is doubled.

It should be stressed that the decrease in ℓ^2, is affected by the number of dimensions (d) of the working space. If a model makes an ϵ error in each direction, the ℓ^2 error will be:

$$\ell^2 = \sqrt{d \times \epsilon^2} = |\epsilon|\sqrt{d}$$

As a result, if the error in each direction stays the same, doubling the number of dimensions would multiply the ℓ^2 error by $\sqrt{2} \approx 1.41$. To accurately compare the error between two dimensions, we define ϵ the average error per dimension:

$$\epsilon = \frac{\ell^2}{\sqrt{d}}$$

We recapitulate the results in Table 3.2. There, we can see that the average ϵ decrease for a doubling of the number of samples (34%) is only slightly smaller than the average decrease for a doubling of sequencing depth (45%) .

Sequencing depth	Number of samples	ℓ^2	ϵ	ℓ^2 decrease	ϵ decrease
$100\times$	2	0.041	0.029		
$200\times$	2	0.022	0.016	46%	45%
$100\times$	4	0.038	0.019	7%	34%

Table 3.2: **Sequencing depth and number of samples comparison** The ℓ^2 error can partially mask an improvement when the number of dimensions increases. To accurately compare the error decrease we show ϵ the decrease in ℓ^2 normalized by the number of dimensions.

The Pearson correlation between the sequencing depth and NMI was of $\rho = 0.611$, compared to $\rho = 0.816.$ for the increase in the number of samples and NMI (both p-values $< 2.2 \times 10^{-16}$). In terms of ℓ^2 error, the correlation between the

depth of sequencing and ℓ^2 was of -0.714 (p-value $< 2.2 \times 10^{-16}$), compared to -0.422 between number of samples and ℓ^2 (p-value$= 3.1 \times 10^{-12}$).

As a result, we can conclude that an increase in the number of samples should be favored when possible compared to an increase in the sequencing depth.

3.3.2 Impact of intrinsic factors on the reconstruction

The contamination by normal cells in the sequenced sample remains uncertain before sequencing, but can be estimated by pathologists. The number of variants in the tumor or the heterogeneity of the tumor are also unknown prior to the analysis.

We here show that the number of variants used to reconstruct the tumor barely affects the quality of the clustering, with a Spearman ρ equal to -0.263 (p-value $= 2.4 \times 10^{-5}$) for the ℓ^2 error and $\rho = -0.075$ (p-value $= 0.24$) for the NMI.

With an opposite behavior, the fraction of contaminating cells negatively impacted both NMI ($\rho = -0.629$, p-value $< 2.2 \times 10^{-16}$), and ℓ^2 ($\rho = 0.738$, p-value $< 2.2 \times 10^{-16}$). In the same way, the clonal heterogeneity negatively impacted NMI ($\rho = -0.629$, p-value $< 2.2 \times 10^{-16}$), and ℓ^2 error ($\rho = 0.738$, p-value $< 2.2 \times 10^{-16}$).

We will see in subsection 5.2.2 how to deal with higly remodeled tumors without loss of accuracy.

Chapter 4

Contributions to variant calling

> *The real risk with AI isn't malice but competence. A superintelligent AI will be extremely good at accomplishing its goals, and if those goals aren't aligned with ours, we're in trouble.*
>
> — Stephen Hawking, *Reddit Ask Me Anything, 2015*

Biological data have several issues in machine learning. We already mentioned the 'small p large n'issue, where the number of observations is largely inferior to the number of features. The second issue is the rather high variability and noise that exists in the data from biological experiments. One such example can be highlighted by the use of zero-inflated mixture models in the case of single cell experiments for example. In this part, we will focus on the noise associated to variant calling from DNA sequencing.

4.1 DREAM Challenge

In this section I will detail insights and models elaborated as a part of a team participating in a DREAM challenge. Only the models I personally implemented will be described, logistic regression and Negative Matrix Factorization models developed by Judith Abecassis will not be presented.

The ICGC-TCGA SMC-DNA Meta challenge[1] is an expansion of the ICGC-TCGA Dream Mutation calling challenge. It was organized by Paul C. Boutros (Ontario Institute for Cancer Research), Josh Stuart (University of California, Santa Cruz), Gustavo Stolovitzky (IBM, DREAM), Stephen Friend (Sage Bionetworks), and Thea Norman (Sage Bionetworks).

In the 'original' Mutation calling challenge, participants were asked to design

[1] https://www.synapse.org/

the best variant calling pipelines, meaning that they would be ranked both on precision and recall for the variants called, starting from BAM files.

The Mutation calling challenge was divided in subchallenges, either at different time points, or on different data (simulated *vs* patient, SNV *vs* SV). If we focus on the SNV challenges, the timeline allowed competitors to improve their pipelines between two subchallenges by using the results of the previous one. As a consequence, there is no evidence that the same pipeline would be used in all subchallenges, and it is reasonable to assume that no two submissions used the same pipelines between two subchallenges.

The organizers from the mutation calling challenge had found that, for each subchallenge, using a majority vote of the five best ranked submissions to predict true positives always outcompeted the best submissions from participants. In order to further improve results from variant calling, the Meta challenge aimed at finding the true positives among the calls made by participants of the Mutation calling challenge using machine learning techniques to aggregate predictions — by either using five or fifty pipelines. The ranking was made using the F1 score, which is defined by the harmonic mean of precision and recall. The true positives were given for the simulated samples to be able to train a supervised algorithm, and the predictions were assessed on the data from both simulation and real patients.

From the pipeline, only the number of the submission was provided, which could be linked to a team, but could not give information on the tools or parameters used.

It is to be noted that, as the pipelines used for the submissions in the Mutation calling challenge differed between two samples, it was impossible to learn features for a given pipeline — such as weighing prediction of the pipeline by its accuracy or recall.

4.1.1 Description

Data features provided by organizers

The data provided by organizers of the challenge contained 14 samples (four from simulations, five from colorectal cancer, and five from prostate cancer), and for each sample the organizers provided:

- All positions called by at least one pipeline;

- For each pipeline (referenced by its submission number), the status of each position;

- 13 genomic features at this position, such as base quality, number of reference and alternative reads, mapping quality, and strand bias.

In addition, ground truth was provided for the four simulated samples, and users could add any relevant biological feature. We chose to add information about the localization of the SNV inside a repeated region.

Synthetic samples 1 to 4 (noted IS1–4) are of increasing difficulty, with addition of contamination by normal cells, structural variants, and subclonal variants. In addition, the number of variants called in each sample varied by orders of magnitude (see Table 4.1).

Sample	Number of calls	Number of true positives	Number of pipelines
IS1	214541	3535	119
IS2	51108	4303	69
IS3	22884	7709	67
IS4	129091	15163	223

Table 4.1: Overview of DREAM training dataset. We here give the number of calls made at least by one pipeline, the number of these that are true positives, and the number of pipelines available.

We can see that progresses are made in terms of accuracy with time. This refinement of pipelines is also illustrated in Figure 4.1, for S1 to S3.

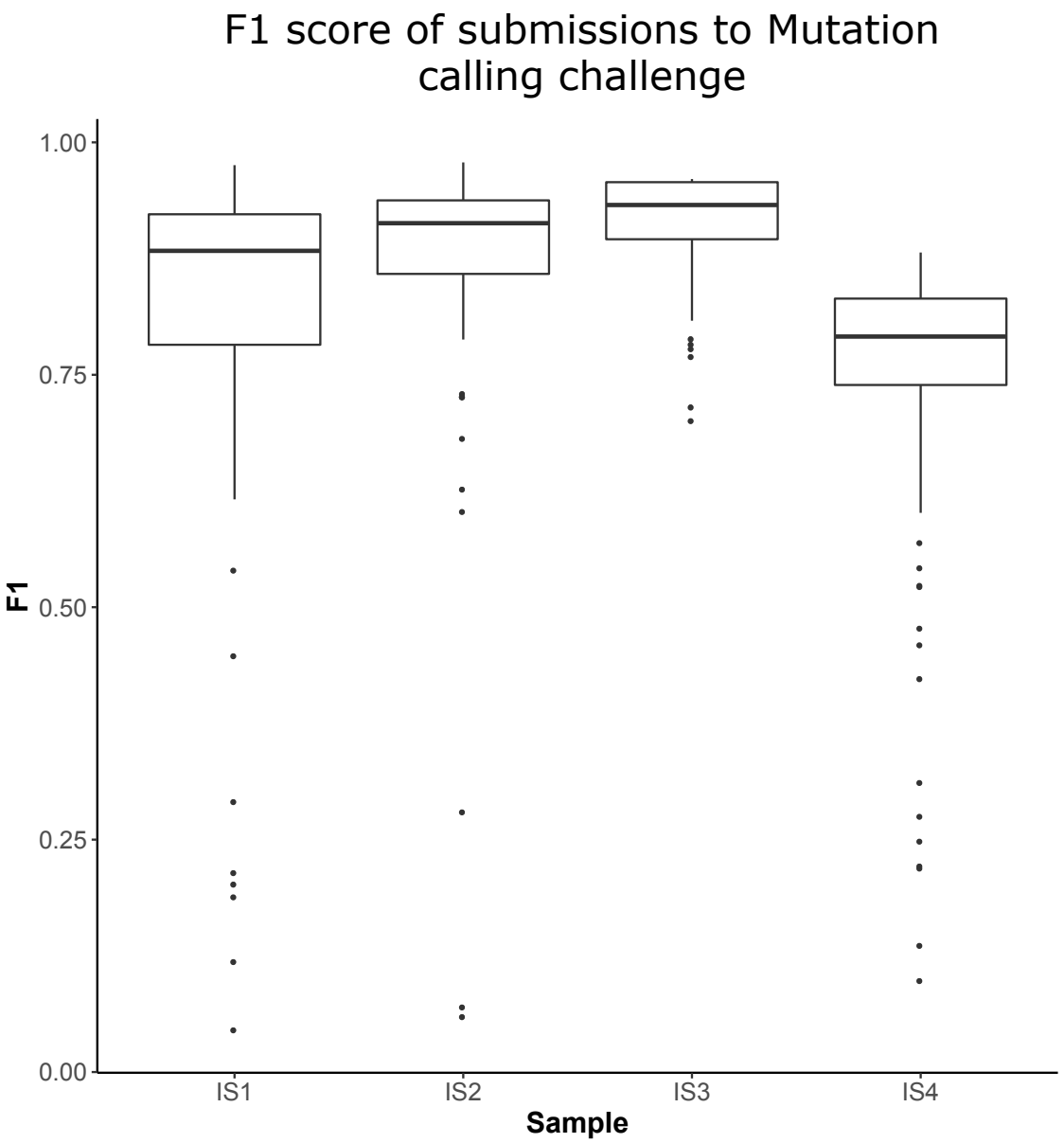


Figure 4.1: **F1 evolution in DREAM data set.** The F1 score on the simulated data set increased with each iteration — IS1 through IS3 — except for IS4. In that case the increased complexity between IS3 and IS4 drastically reduced F1 score.

A first data exploration also shows that the predictions of pipelines were highly conserved within each teams (Figure 4.2). This is visualized using a hierarchical tree clustering on all variants of a sample. We can also note that the consensus obtained by majority vote of all pipelines are close to one another, despite the poor performance of multiple pipelines.

In Figure 4.2 we can also see submissions from a team tends to cluster together, showing incremental changes in the pipeline.

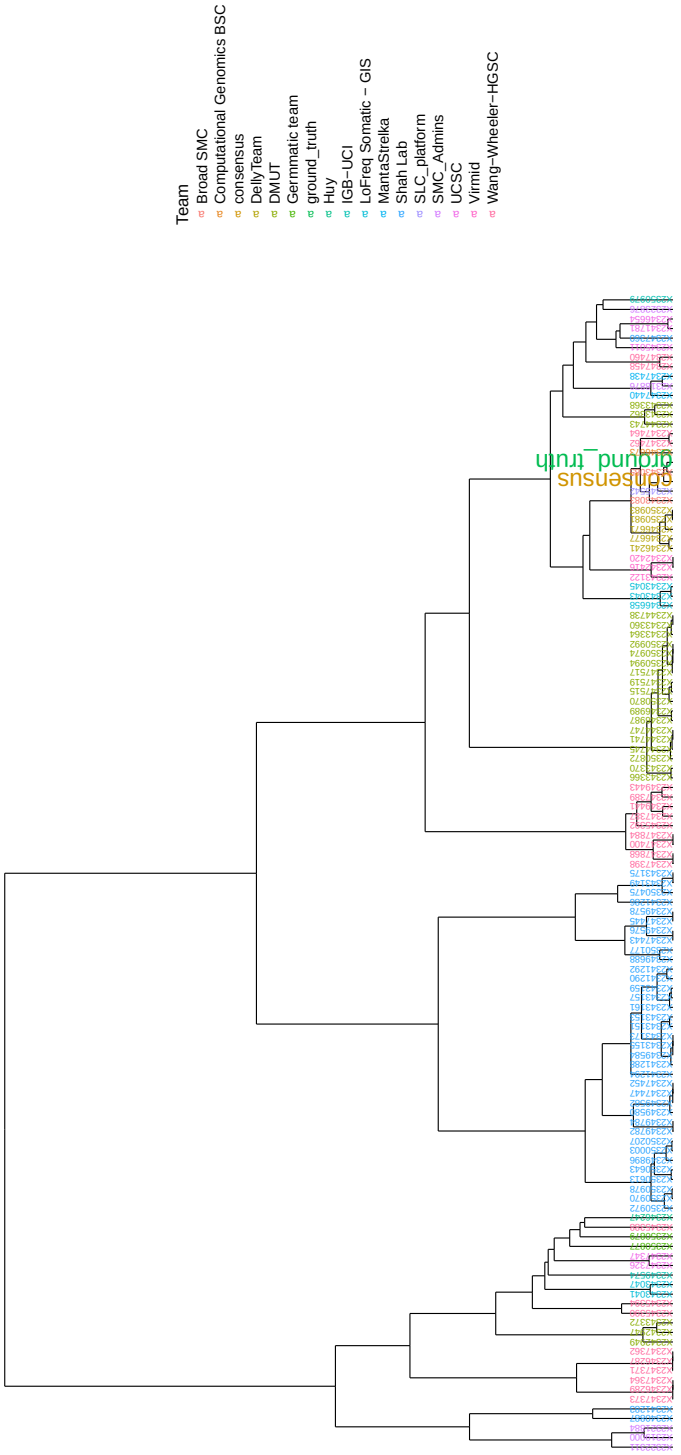


Figure 4.2: **Hierarchical clustering of pipelines based on predicted positions.** This clustering is based on a Jaccard distance between the pipelines. A ward method is then used to combine classes. The consensus obtained by majority vote of all the pipelines and the ground truth are also shown. The different colors of submission names reveal the team which has made the submission.

Feature augmentation and selection

Biological data can be error prone. The human genome contains repeated sequences that only have few nucleotides of difference. The mapping to such parts of the genome can be highly difficult, and wrongfully attributed sequences will result in predicted mismatches.

In order to avoid such bias, we added information from the UCSC repeated regions track. We also included the GC content (percentage of G and C nucleotides in a 50bp region around the variant), the variant allele frequency, and the homopolymer rate, defined by the sum of squared homopolymer lengths normalized by the length of the sequence. For example, the AAATTGAGG would have a homopolymer rate of $\frac{3^2+2^2+1^2+1^2+2^2}{9} = \frac{19}{9} \approx 2.11$. These features are there to indicate the potential sequencing error in sequences that integrate a high degree of repetitions.

In addition, we integrate the read base quality. The base quality per position was provided by challenge organizers.

Finally, we added the consensus ratio, as the number of selected pipelines that predicted a variant at this position. This feature assumes that the selection of pipelines will collectively behave in a similar trend in all samples. Nonetheless, the fact that pipelines changed between samples prevented learning directly on the pipelines.

4.1.2 Proposed model and cross-validation

Two challenges were created, one using five or less pipelines. In the second challenge participants could use a maximum of 50 pipelines. The selection of pipelines was left to the competitors.

In order to maximize the potential recall, we chose to select the pipelines maximizing the number of variants called.

In order to test models, we trained a model on three samples and tested on the fourth. This procedure was applied to all samples consecutively, and mean F1 as well as the median absolute divergence (MAD) was used to evaluate the model.

Pipeline selection

In the two sub-challenges, the maximal number of pipelines that could be used for predictions was lower than the number of submissions provided. This has for consequence that the first step of the analysis will be a selection of a given number of pipelines.

Two different strategies had to be balanced: one could provide pipelines with a very high accuracy - further filtering refining the selection of variants - or a very high recall. The variants that were not called by at least one of the selected pipelines could not be used for predictions.

We define a fictional consensus pipeline, as the hypothesized pipeline that for each tumor sample independently, only the variants predicted by a majority of pipelines are considered as called.

In order to maximize the accuracy, we selected pipelines that were the closest to the consensus - in terms of Manhattan distance. The strategy to maximize the recall was to select the pipeline with the highest number of variants, remove all these positions, then repeat until the desired number of pipelines is reached.

First models only learned on pipelines close to consensus, limiting recall. Maximizing recall however lead to numerous false positives. Balance was achieved by training three models: one on a few very stringent pipelines, the second on extensive pipelines, and the third on the aggregation of both stringent and extensive models (see Figure 4.3).

4.1.3 Results from the different pipelines

In this section, we will discuss results as provided by DREAM challenge organizers.

In silico data

The simplest idea tested was the majority vote of pipelines close to the consensus (see Table 4.2).

We here show that the majority vote of the five pipelines closest to consensus achieved a high F1 score, that was only lightly decreased by a higher threshold, mainly due to a decrease in sensitivity not balanced by the increased precision. Using 50 pipelines achieved an even higher F1 score, even if this procedure

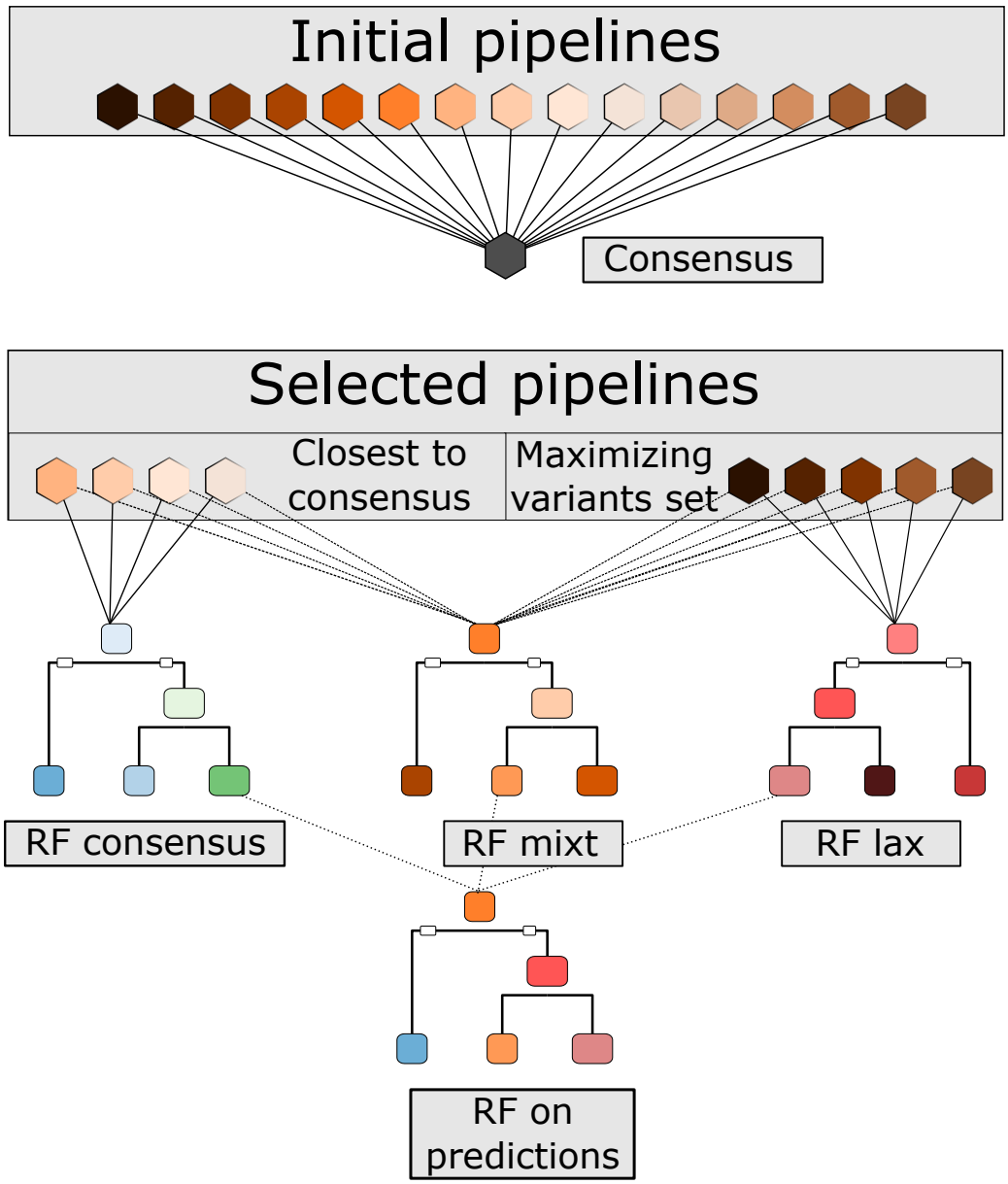


Figure 4.3: **Random forest (RF) model for DREAM.** The consensus pipeline is generated by majority vote of all pipelines. From the initial pipelines only a fraction of the closest to the consensus will be taken. To reach 5 or 50 pipelines, the set of pipelines is completed by addition of pipelines maximizing the set of variants called. Three different models are trained using the fraction of pipelines that have called a variant as well as other features - such as mappability or variant allele frequency. The prediction of the three models is aggregated using another random forest.

Number of pipelines	Threshold for prediction	F1	MAD
5	0.5	0.9451	0.0269
5	0.8	0.9414	0.0294
50	0.5	0.9572	0.0124

Table 4.2: **Results from majority vote *in silico***.The threshold for prediction is the fraction of the pipelines agreeing on a true positive to predictive a variant at this position. MAD: Median Absolute Deviation.

seems highly unlikely to be used in diagnosis, as the use of 50 different pipelines would be too much time-consuming for a single sample.

The score could be increased by using features from genomic context and variant context, as shown in Table 4.3. For the first three models, the training was realized on the same samples as the test, this means that the measure only shows how close to the training is the fit, and can result in overfitting the data. The line marked with a * has been trained only on sample IS4. With this, we show that by training the sample of higher complexity, it was possible to learn the model behind the simulation.

Number of pipelines	Consensus/Recall ratio	Filter	F1	MAD	Feature set
5	0.4	0.4	0.9754	0.0250	A
50	0.3	0.4	0.9774	0.0177	A
50	0.3	0.3	0.9679	0.0091	B
50*	0.3	0.3	0.9571	0.0415	B

Table 4.3: **Results from random forest models *in silico***.Training for set A contains mean base quality, allele frequency, tumor coverage, variant inside or outside of a repeated region, mean mapping quality, homopolymer rate and GC content. Set B additionally contained distance to closest SNP, and if the variant was inside intergenic or intronic regions. * Training was only realized on tumor sample IS4.

4.1.4 Discussion: difference between *in silico* and real data

First, we can note that results provided for real sequencing data were assessed using specificity and not F1 score contrary to simulated data. Specificity (also

called true negative rate) is the fraction of true negatives correctly identified as such. The issue with specificity is that it can be "hacked" by predicting many false positives.

We can see the poor performance of the pipelines in this setting, with a specificity twice as low as the F1 score from the previous experiment (Table 4.4). The fact that we did not use a supervised method for these submissions shows that there is an important discrepancy between simulated and real data in this challenge. This discrepancy is highlighted in Table 4.5, with performances lower than the consensus alone.

Number of pipelines	Threshold for prediction	Specificity	MAD
5	0.5	0.4104	0.0620
5	0.8	0.4160	0.0560
50	0.5	0.4224	0.0649

Table 4.4: **Results from majority vote cancer samples.**The threshold for prediction is the fraction of the pipelines agreeing on a true positive to predictive a variant at this position. MAD: Median Absolute Deviation.

Number of pipelines	Consensus/Recall ratio	Filter	F1	MAD	Feature set
5	0.4	0.4	0.3796	0.1508	A
50	0.3	0.4	0.3723	0.0431	A
50	0.3	0.3	0.3693	0.0343	B
50*	0.3	0.3	0.3577	0.0179	B

Table 4.5: **Results from random forest models cancer samples.**Training for set A contains mean base quality, allele frequency, tumor coverage, variant inside or outside of a repeated region, mean mapping quality, homopolymer rate and GC content. Set B additionally contained distance to closest SNP, and if the variant was inside intergenic or intronic regions. * Training was only realized on tumor sample IS4.

From these we can conclude that the model used to generate the data did not fit reality closely enough to extract relevant features for filtering. This is also illustrated by the fact that training on three simulated samples and testing on the fourth resulted in an F1 score of 0.979 ($MAD = 0.005$) for our more complex model.

As a result, for the clonal reconstruction problem, filters had to be redesigned to select only high quality variants. Final results of the challenge are not available yet.

4.2 Filters for clonal reconstruction

In this section we will detail the filters used for to limit the number of false positives in the variant calling for clonal reconstruction without necessity for visual inspection.

4.2.1 Presentation of the neuroblastoma WGS cohort

The cohort used in *Clonal assessment of functional variants in cancer based on a genotype-aware method for clonal reconstruction* consists of 22 patients (see Table 4.6) for which the germline DNA and tumoral DNA both at diagnosis and relapse were sequenced. Patients whose ids start with 'NB'come from the French cohort. Samples were sequenced at 50$\times$ for the germline DNA, and $100\times$ for the tumoral DNA. Samples of patients NB308, NB3099, NB804, NB1224, NB1269 and NB1382 were sequenced at the Beijing Genomic Institute (BGI), the remaining NB patients were sequenced at the Centre National de Génotypage (CNG). Samples whose ids start with 'PA'come from the US cohort and were sequenced at $100\times$ both for germline DNA and tumoral DNA at Complete Genomics (CG).

For these patients, estimation of copy number status was realized using Control-FREEC (version 7.2, see Figure 4.4) which also gave an estimation of the contamination by normal cell of the samples. This estimation is in agreement with purity estimation given by pathologist (Table 4.7).

4.2.2 Raw output of variant calling

Variant calling was performed using Varscan2 version 2.3.6[20]. Due to the very large size of the data, performing multiple variant callings and then aggregating them by majority vote was out of line because of both time and size constraints. For example, Strelka[22] requires an estimated 50 cpu-hours to complete a variant calling for an exome sequenced at 40 to 60$\times$. With 44 whole genomes sequenced at 100$\times$ (roughly 200$\times$ the size of the reference provided by Strelka authors), and assuming that the time required is linear with the size of the data, we can estimate the time for a single variant calling of the full cohort to be $44 \times 50 \times 200 = 440,000$ hours, or the equivalent of 50 years of computation.

Patient	Risk stratification	Stage	MYCN status	Gender	Age at diagnosis	Time to relapse	Time to last report	Status	Diagnosis
NB1178	H	4	N/Amp	M	30	21	24	Dead	Retroperitoneum
NB1224	L	2	N/Amp	M	15	8	18	Alive	Mediastinum
NB1269	H	4	Amp	M	14	9	11	Dead	Retroperitoneum
NB1382	H	4	Amp	M	4	50	64	Dead	Abdomen
NB308	L	2	N/Amp	F	2	21	91	Alive	Abdomen
NB399	L	4s	N/Amp	M	0	7	134	Dead	Subcutaneous nodule
NB804	I	4	N/Amp	F	2	26	56	Alive	Subcutaneous nodule
PAPVEB	L	2	N/Amp	M	57	9	40	Dead	Adrenal gland
PARBAJ	I	3	N/Amp	M	1	10	88	Alive	Retroperitoneum
PARHAM	I	4	N/Amp	F	11	1	81	Dead	Pelvis
PASHFA	H	3	Amp	F	13	7	11	Dead	Adrenal gland
PASNPG	I	3	N/Amp	F	10	10	63	Alive	Retroperitoneum
PATNKP	H	4	N/Amp	M	113	20	40	Alive	Retroperitoneum
PATYIL	I	4	N/Amp	F	11	8	16	Dead	Abdomen
PAUDDK	I	3	N/Amp	M	12	11	38	Alive	Pelvis
NB0784		4	N/Amp	F	26	12	94	Alive	Pelvis
NB1177		4s	N/Amp	M	13	10	49	Alive	Subcutaneous - para vertebral left
NB1361	L	4s	N/Amp	M	27	16	27	Dead	Surrenal mass
NB1385	H	4	Amp	M	147	8	14	Dead	MD
NB1434	H	4	Amp	F	26	10	27	Dead	Left mandibula
NB1457	L	L2	N/Amp	M	12	6	22	Alive	Retro-pharyngal
NB1471		1	N/Amp	M	64	7	20	Alive	Thorax

Table 4.6: **Neuroblastoma cohort**: Characteristics of neuroblastoma samples used for data analysis. Ages at diagnosis is given in months. **H**: High; **I**: Intermediate; **L**: Low; **N/Amp**; Non amplified; **Amp**: Amplified; **M**: Male; **F**: Female.

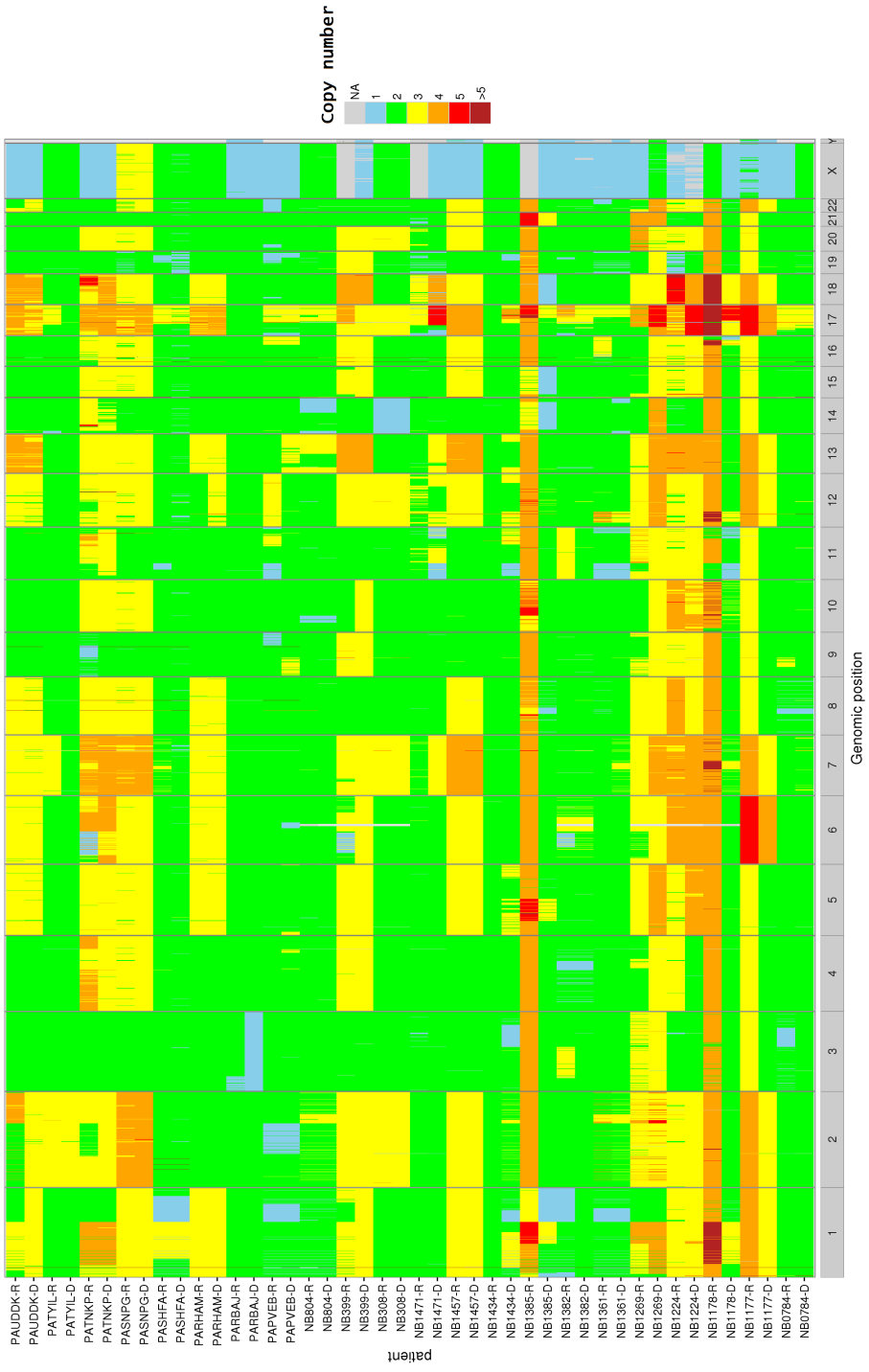


Figure 4.4: **Copy number summary.** Summary of the copy number profiles for all samples determined by Control-FREEC. The copy number is represented by a color, with the genomic position in abscissa and the sample on the y-axis

Patient	Contamination estimation (diagnosis)	Contamination estimation (relapse)	Purity diagnosis (Pathologist)	Purity relapse (Pathologist)
NB1178	0.060	0.028	90%	Proliferating tumoral cells on 20% of slice
NB1224	0.188	0.204	60%	90%
NB1269	0.208	0.150	60%? High tumoral burden	High tumoral burden
NB1382	0.131	0.079	60-70%	90%
NB308	0.600	0.600	50%	60%
NB399	0.166	0.412	90%	$>$60%
NB804	0.213	0.140	90%	Made of 60% viable tumoral cells
PAPVEB	0.065	0.024		
PARBAJ	0.083	0.067		
PARHAM	0.182	0.144		
PASHFA	0.550	0.024		
PASNPG	0.060	0.410		
PATNKP	0.072	0.122		
PATYIL	0.008	0.090		
PAUDDK	0.048	0.006		
NB0784	$>$0.70	0.014		
NB1177	0.034	0.018		
NB1361	0.038	0.012		
NB1385	0.150	0.700		
NB1434	0.145	$>$0.9		
NB1457	0.000	0.085		
NB1471	0.044	$>$0.75		

Table 4.7: **Purity estimation of samples**. We here present the contamination estimation evaluated by Control-FREEC and purity estimation by pathologists (when available). For clonal reconstruction, patients with at least one sample with contamination higher than 70% were withdrawn from the analysis. As a reminder, the link between contamination (c) and purity (c) is $p = 1 - c$

The raw output from Varscan shows that a bias exists between sequencing platforms (CNG vs BGI) and sequencing technology (Illumina vs Complete Genomics) (see Figure 4.5). Complete genomics reads have the particularity to bear a deletion in the middle of the read, making impossible to remap the reads or do a realignment around indels except with proprietary tools.

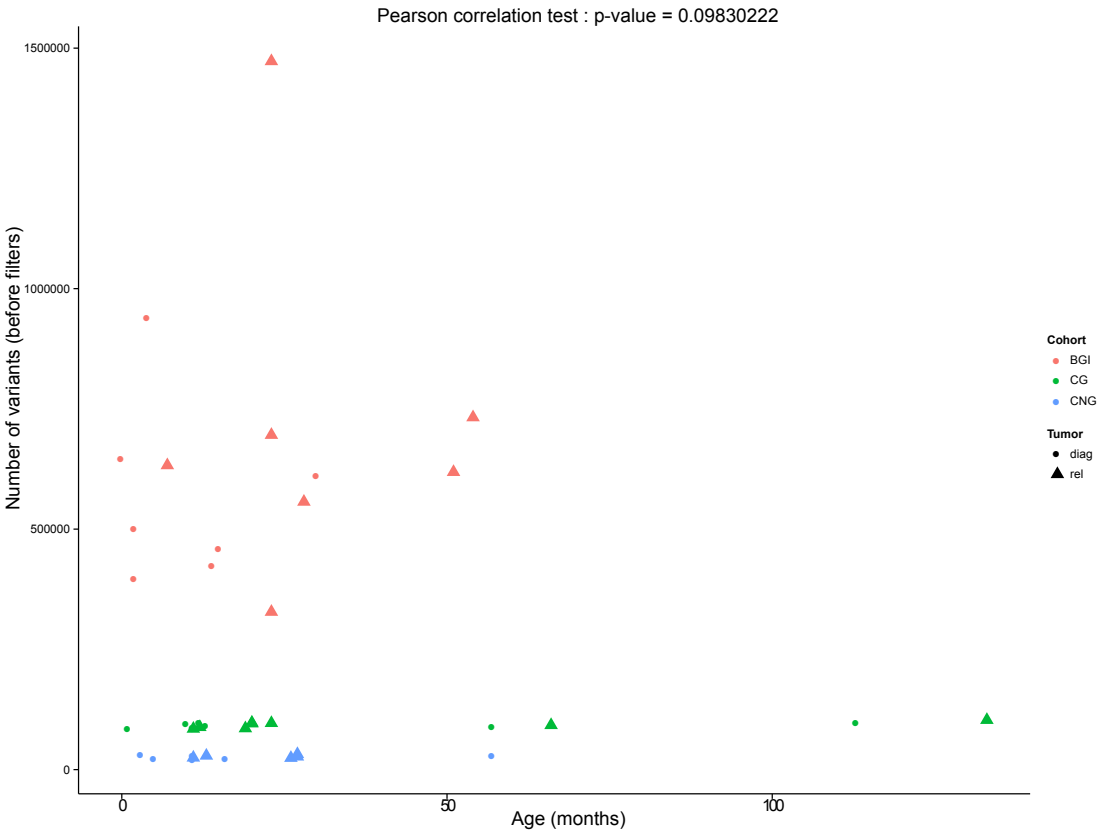


Figure 4.5: **Number of calls from Varscan2.** The number of variants has been previously correlated with the age of the patient in pediactric tumors [9]. Here we demonstrate a sequencing platform effect between samples sequenced at the BGI, CG, or CNG.

The goal of filters is then to:

1. Filter out false positives;
2. Reach a comparable number of variants called independently of the platform and technology.

4.2.3 Retrieving high fidelity variants

We define high fidelity variants as variants with low variance on the observed VAF, and lowest probability to be false positive as possible (see section 9.2).

To do so, we selected variants with at least 50× coverage at the position and required that *a minima* 10% of reads support the variant allele.
In addition, variants were required to be located in regions of maximal mappability, assessed by the UCSC 100bp mappability track.
Local mappability does not exclude regions that can be highly similar to other regions of the genomes. To exclude a bias coming from mismapping, we only included variants outside of repeat and duplicated genomic regions (assessed by the UCSC repeat masker, simple repeat, and segmental duplication region tracks).
We further filtered mutations that created a stretch of four or more identical nucleotides. By this we mean an A>C transition in a CCACC sequence for example, or T>G in GGGT (see Figure 4.6).

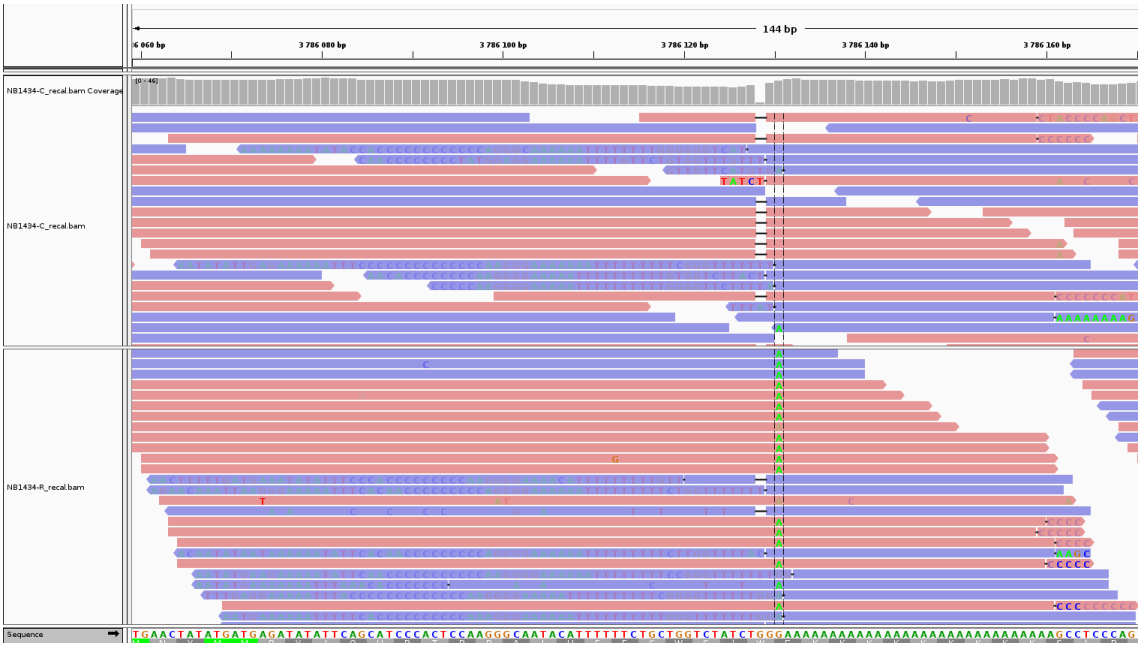


Figure 4.6: **Example of a likely false positive call extending a stretch of polynucleotides.** In the relapse sample we can observe a G>A substitution before a stretch of As. Reads with the substitution also present mappability issues, as shown by the high number of variants on those reads. The fact that the variation is specific to the relapse shows the variability of the mapping step.

We filtered out variants corresponding to polymorphisms present in more than 1% of the population (present in snp138, 1000Genomes, esp6500) except if it was a known cancer related variant (COSMIC database for coding and non-coding mutations). An usual way of filtering out sequencing and mapping artifacts is to remove all variants called in at least one sample of germline origin. However, it should be noted that this way of doing does not scale with cohort size (the bigger the cohort the more positions will be filtered out), and may introduce biases.

Indeed, *ALK* mutations have already been observed as germline variants, especially in hereditary neuroblastomas[11, 12].

Finally, we only kept mutations located in regions where the genotype evaluated by Control-FREEC was available.

4.2.4 Assessment of applied filters

It should be noted that no re-sequencing of the variants predicted as high fidelity has been realized. This can be legitimately explained by the cost and time of such procedure, and the scarcity of biological material. As a result, validation of the filters have been made by visual inspection of the sequences by Integrative Genome Viewer (IGV) [68, 17].

In addition, we used the correlation between age of patient and number of variants[9] to evaluate the quality of the filtering as reducing platform biases. After all filters, a significant correlation between age and variants called is found Figure 4.8.

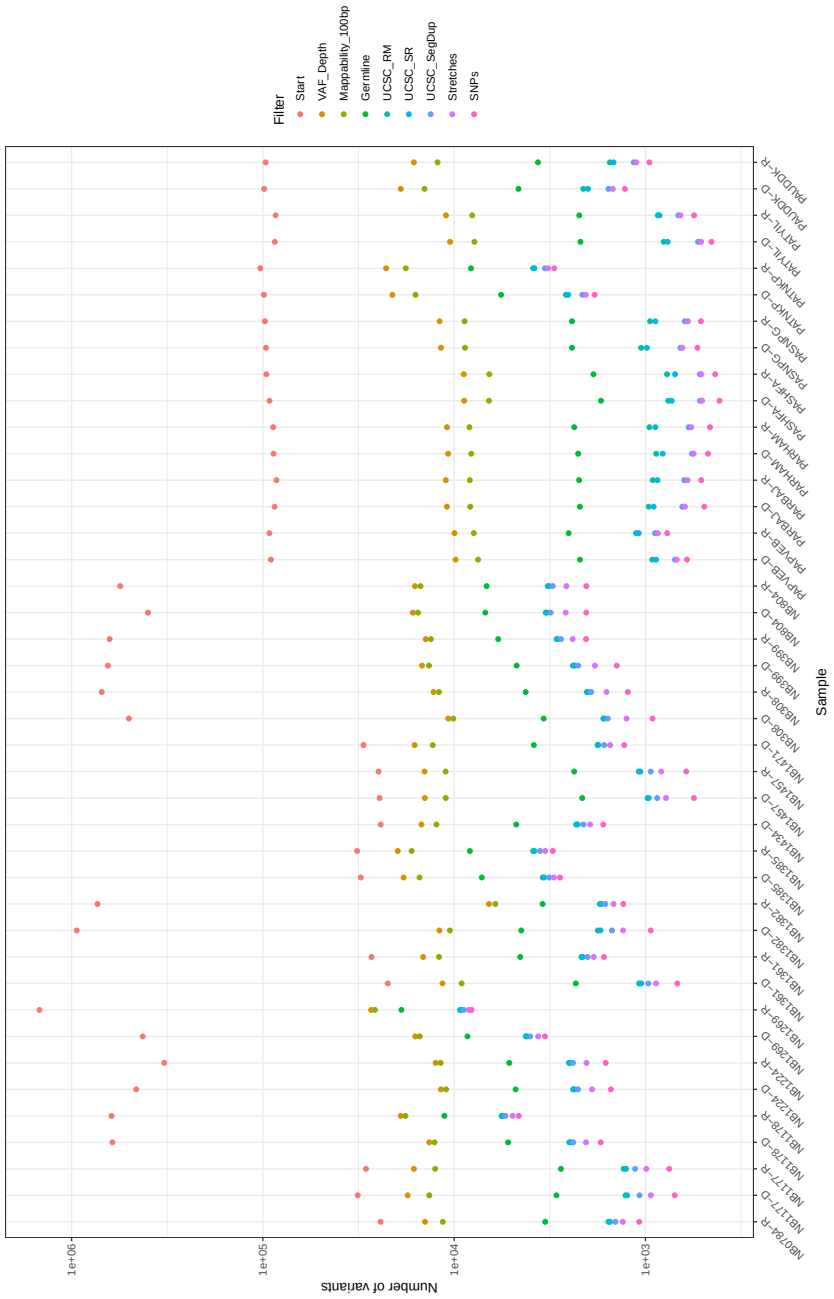


Figure 4.7: **Effect of filtering on the number of somatic variants called.** The initial number of variants corresponds to the raw output from Varscan2; the “VAF_Depth” filter selects variants with at least 10% of the reads supporting each variant and at least 50 reads mapping at the position of each variant; the “Mappability_100bp” filter requires the mappability of a 100 bp DNA sequence to be 1 at the position of the variant (the UCSC genome browser 100 bp mappability track); the "Germline" filter removes variants found at germline level in other patients of the cohort; the “Stretches” filter removes variants that would create a homopolymer of three or more identical bases; the “SNPs” filter removes variants that are present in the SNP databases at a frequency higher than 1% except when the variant is listed in the COSMIC database; finally, the “UCSC_RM”,“UCSC_SM”,“UCSC_SegDup” filter removes variants that are present in respectively repeat masker, simple repeat and segmental duplication regions from the corresponding UCSC genome browser track.

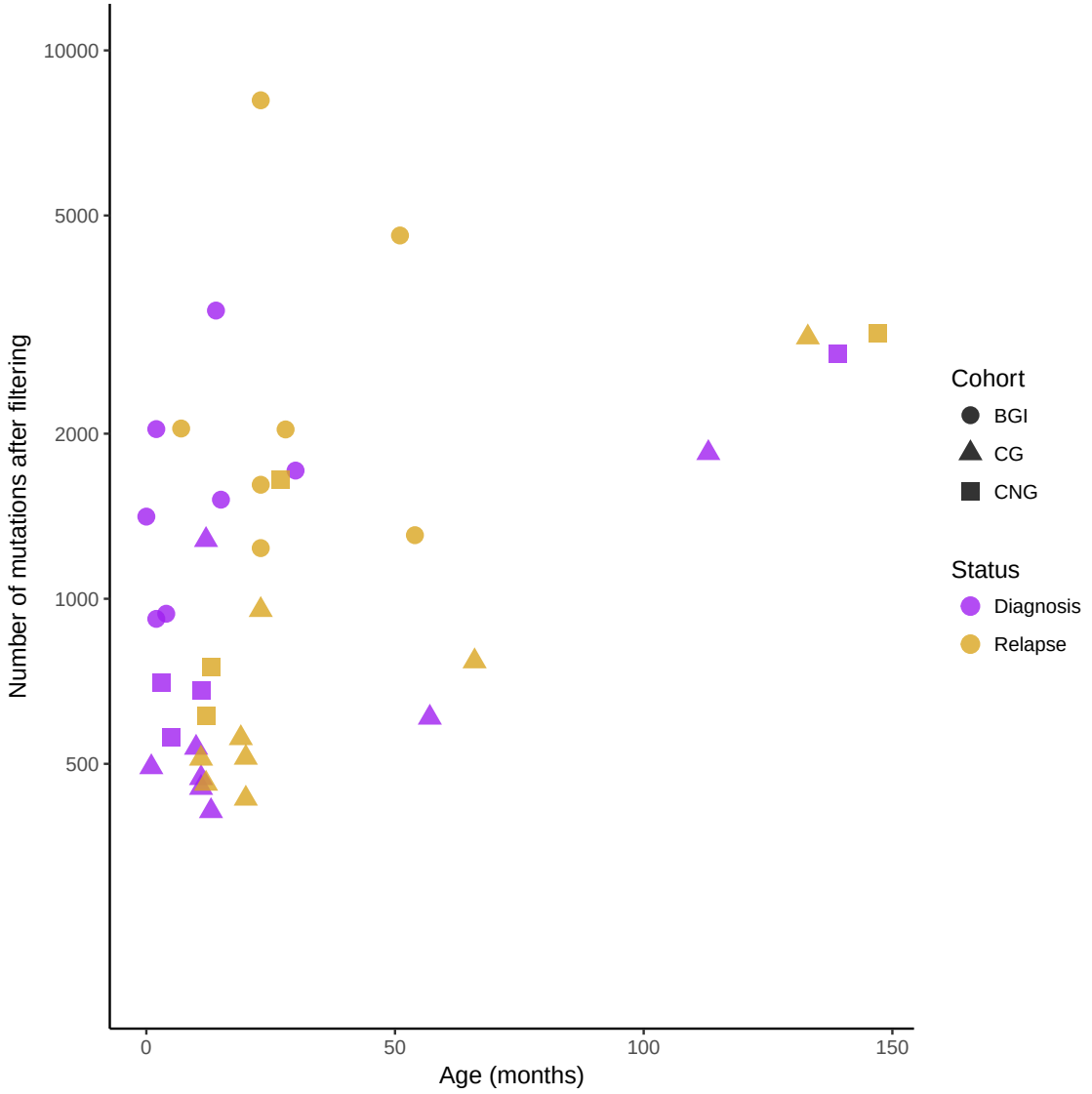


Figure 4.8: **After filtering, number of variants correlates with age.** As previously published[9], this effect is also found after filtering in our cohort: Spearman's $\rho = 0.44$, $p - value = 6.3 \times 10^{-3}$. Note that the number of variants at relapse is highly correlated to the number of variants at diagnosis (Spearman's $\rho = 0.93$, $p - value = 3.4 \times 10^{-6}$).

Chapter 5

Combining enrichment and clonal reconstruction results

> *EVERYTHING IS DEEPLY INTERTWINGLED. In an important sense there are no "subjects" at all; there is only all knowledge, since the cross-connections among the myriad topics of this world simply cannot be divided up neatly.*
>
> — Theodor Holm Nelson, *Computer Lib/Machine Dreams, 1974*

As exposed in the introduction, neuroblastoma has only very few recurrent genomic alterations, especially at the single nucleotide level. This raises the possibility that it is not specific genes that are altered but pathways. This implies that mutational signal should not be observed at the gene level but at a larger level.

5.1 How to find pathways enriched in mutations

Enrichment of a pathway will depend on the method of description of the pathway. In the case of networks (directed or undirected), diffusion strategies have been developed [69, 70, 71]. However, when the pathway is defined as a list of genes, and only discrete events are observed - such as mutations - the simplest model is to compare the number of mutations observed in genes from the pathway compared to the expected number of genes expected to be mutated if all pathways had equal distribution [72]. Finally, the last method consists in grouping genes by ontology, which is equivalent to the previous method except for the fact that it does not work directly with gene networks but with the ontology of genes[73, 74, 75, 76].

Nonetheless, all these methods require that variants with a *functional* impact are given as input, as passenger variants are not considered relevant for the analysis of disrupted pathways.

5.1.1 Finding variants with biological impact

In this section we will differentiate substitutions that are in protein coding regions (exons), or in non-coding regions (intronic, intergenic regions). In protein coding regions, substitutions can be silent (the amino acid is not changed), missense (the amino acid is changed for another one), or non-sense (also called stop-gain as the trinucleotide is replaced by a stop codon, truncating the protein). While, arguably, silent mutations are thought to be benign, in reality these mutations can be deleterious due to changes in the RNA and protein structure [77]. Stop-gain substitutions are often thought as highly deleterious for the protein. However, translation and transcription mechanisms can lead to a functionally active protein even in the case of a stop codon. For example, exon skipping (when the exon bearing the mutation is removed from the RNA by splicing events), stop readthrough (when the stop is ignored), or reinitiation (when the starts from a new ATG trinucleotide) [78, 79] events can potentially reduce the impact of a stop gain.

This shows that even in the simplest cases, it is difficult to assess the impact of a variation on the protein function. To solve this issue, we relied on prediction algorithms presented in the next paragraphs.

Predicting impact on protein structure

SIFT and Polyphen are two widely used prediction tools that help prioritizing candidate genes with putative deleterious variants. Kumar et al. [80] summarizes the differences between SIFT [81] and Polyphen [82] by the fact that SIFT solely uses sequence homology whereas Polyphen uses both sequence homology and protein structure from SWISS-Prot.

In more details, SIFT uses sequence homology, as described in Figure 5.1. It does not directly try to reconstruct the 3D structure of the protein, but looks at local conservation of a protein sequence to establish if the substituted amino acid has the same characteristics as a given proportion of homologous and paralogous sequences. Amino acid can be electrically charged, which often leads to an hydrophilic behavior, and would preferentially located at the surface of the protein, whereas neutral amino acids tend to have a more lipophilic and hydrophobic behavior, which would be more often found on the inside of proteins. As a result, change in those characteristics can lead to different structures which in turn could change the activity of the protein.

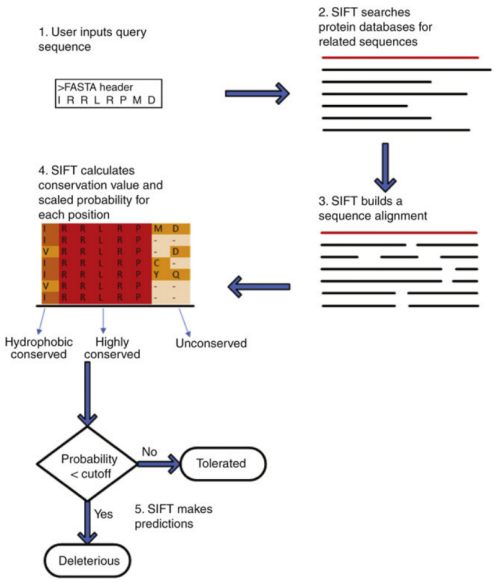


Figure 5.1: **SIFT workflow**, from [80]. For the protein sequence IRRLRPMD, first SIFT looks for sequences with high homology in homologous and paralogous genes. After alignment of the sequences, a conservation score for each position is computed, telling if the position requires a highly conserved amino acid, an hydrophobic / hydrophilic amino acid, or if the position is not conserved. The probability to observe the amino acid coming from the substitution is then compared to a threshold to separate benign, possibly damaging and damaging mutations.

Polyphen2 uses sequence conservation, protein structure and protein function annotation. These pieces of information are combined by a classifier to predict the probability of a variant being deleterious (Figure 5.2).

Funseq2

We detailed in the previous section, the prediction of impact of variants in the protein coding regions. In whole genome sequencing, we also have access to non-coding sequences, that could be altered by a genetic event and impact the behavior of the tumor. One such possibility is the disruption of transcription factor binding sites (TFBS), which could lead to a change in the gene expression level in the tumor (Figure 5.3).

Testing all possible motifs of all transcription factors would be extremely tedious, and would necessitate important statistical corrections. In order to avoid these issues, we relied on Funseq2, which estimates the impact of variations both in coding and non-coding regions.

The first step of Funseq2 [84, 85] relies on aggregating data from databases

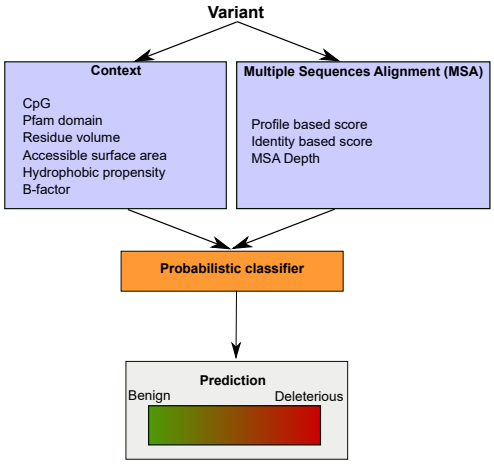


Figure 5.2: **Polyphen-2 workflow from [82].** Polyphen-2 uses as features for the probabilistic classifier 11 features, some of which are extracted from the multiple sequences alignment, whereas some others are physical (fluctuation of molecule around its position, also called B-factor) or biological properties (Pfam is a database of protein function domains). In the end, a prediction ranging from benign to deleterious and a confidence in the prediction are provided.

such as 1000Genomes, COSMIC, or GERP, to establish a list of genes that are conserved between species or targeted in cancer. After this step, the scoring (see Figure 5.4) uses information such as breaking or gaining motifs (consensus sequence of a TFBS), or centrality of the gene in a gene network.

All these information can be aggregated to predict variants of potential interest. However, the total number of variants predicted to be deleterious by at least one of the three algorithms is rather high, and not all deleterious variants can be of interest for the disease studied. In order to focus on highly relevant genes, and due to the very low frequency of recurrently altered genes in neuroblastoma, we extracted genes in pathways recurrently altered at the cohort level.

5.1.2 Diffusion networks and network based stratification

Introduction to networks

We define a biological network as a network of species represented as nodes and biological interactions represented by edges. The nodes in the graphs can be the DNA sequence of a gene, the transcribed RNA, or, when relevant, the protein associated with the gene. A biological interaction can be a direct interaction, for example two proteins that bind together, or can be through activation or repression of a gene.

Networks that use information from activation or repression tend to be di-

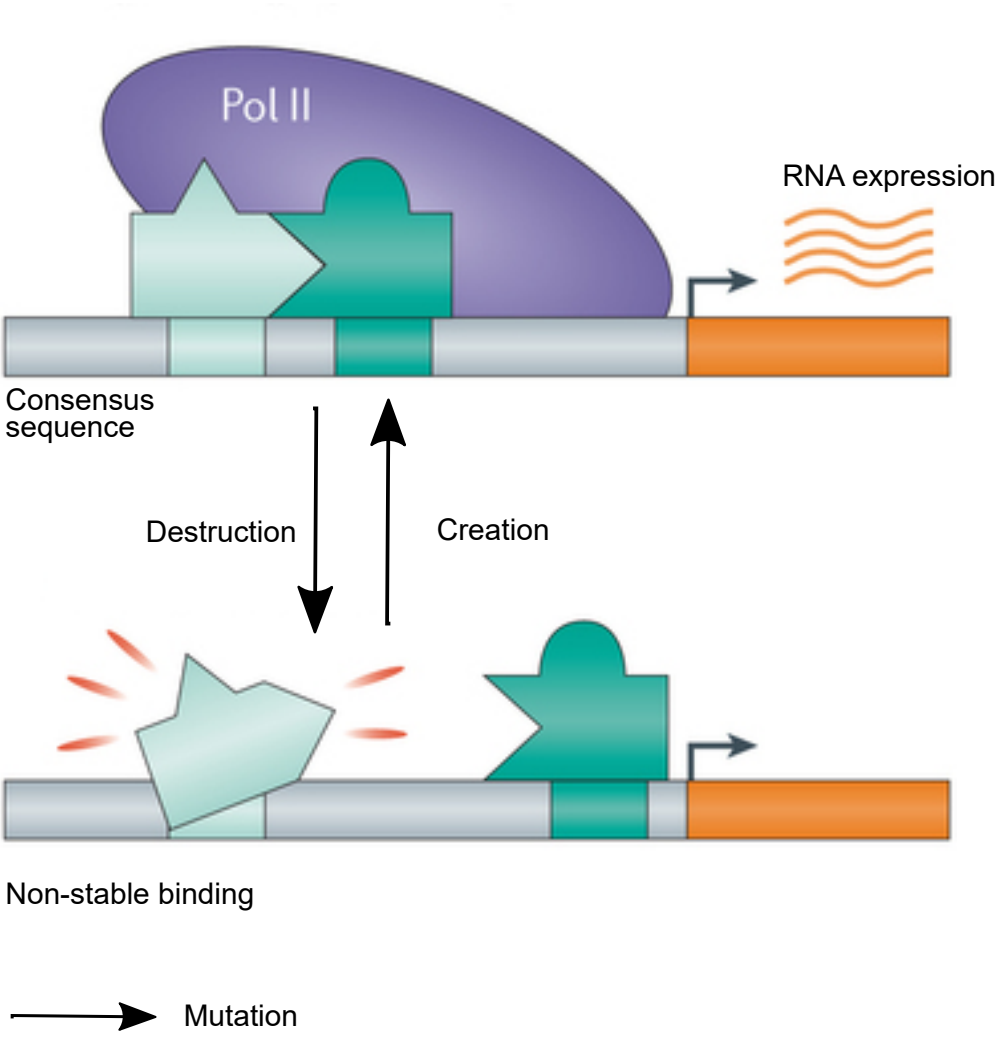


Figure 5.3: **Transcription factor binding site conversion**. A mutation in a sequence can either create a TFBS, leading to expression of the RNA, or disrupt the sequence, leading to loss of expression of the RNA. Adapted from *In pursuit of design principles of regulatory sequences*, [83]

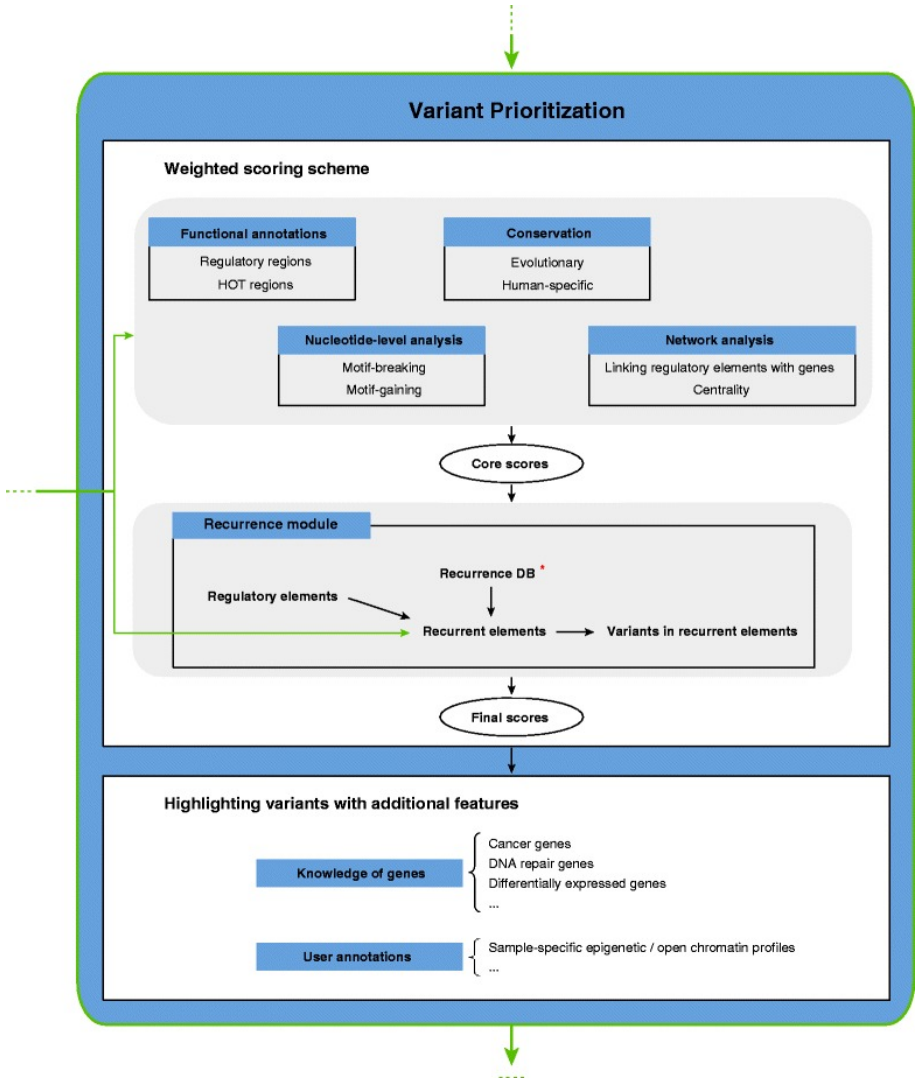


Figure 5.4: **Funseq2 prioritization**. The variant prioritization step will annotate input variants and then score them using the weighted scoring scheme. Features used in the weighted scoring scheme can be classified into 'functional annotations', 'conservation', 'nucleotide-level analysis', 'network analysis', and 'recurrence'. 'Recurrence' feature could be detected from user-input cancer samples and also from 'Recurrence DB' (* means optional. User can choose to use the 'Recurrence DB' or not). Different from other features, 'recurrence' depends on the user-input (for example, if user only uploads one sample and chooses not to use the 'Recurrence DB', then 'recurrence' feature will not be observed for any variant). Each feature is assigned a weighted score (Material and methods). Scores obtained from the top grey panel are called 'core scores', which is independent of the user's choice (see above for 'recurrence' feature). Variants with the 'recurrence' feature are assigned an additional score in the final output. In addition to features used in the scoring scheme, other features are used to highlight potentially interesting variants, such as variants associated with known cancer genes. Figure and legend from *Fu et al* [85]

rected, as a gene can effect its targets, but this effect is often unidirectional. The action of A on B will be written as $A \rightarrow B$. Protein-protein interaction are undirected as they only depict the fact that two proteins can be found spatially interacting. While it is always possible to convert a directed network to an undirected one, the opposite is false.

It can be difficult to assess the role of a perturbation in a network and general idea relies on diffusion equations. In the same idea that heat sources can diffuse energy through physical links, the perturbation of the network can be passed to neighboring chemical species through interaction. For example, if A is responsible for the expression of B and we disrupt A, we expect B not to be expressed anymore.

Diffusion networks [86, 87] rely on the idea that a variant affecting the functional properties of a protein or non-coding RNA (ncRNA) will not only affect the protein itself but also neighboring genes of the network (see Figure 5.5).

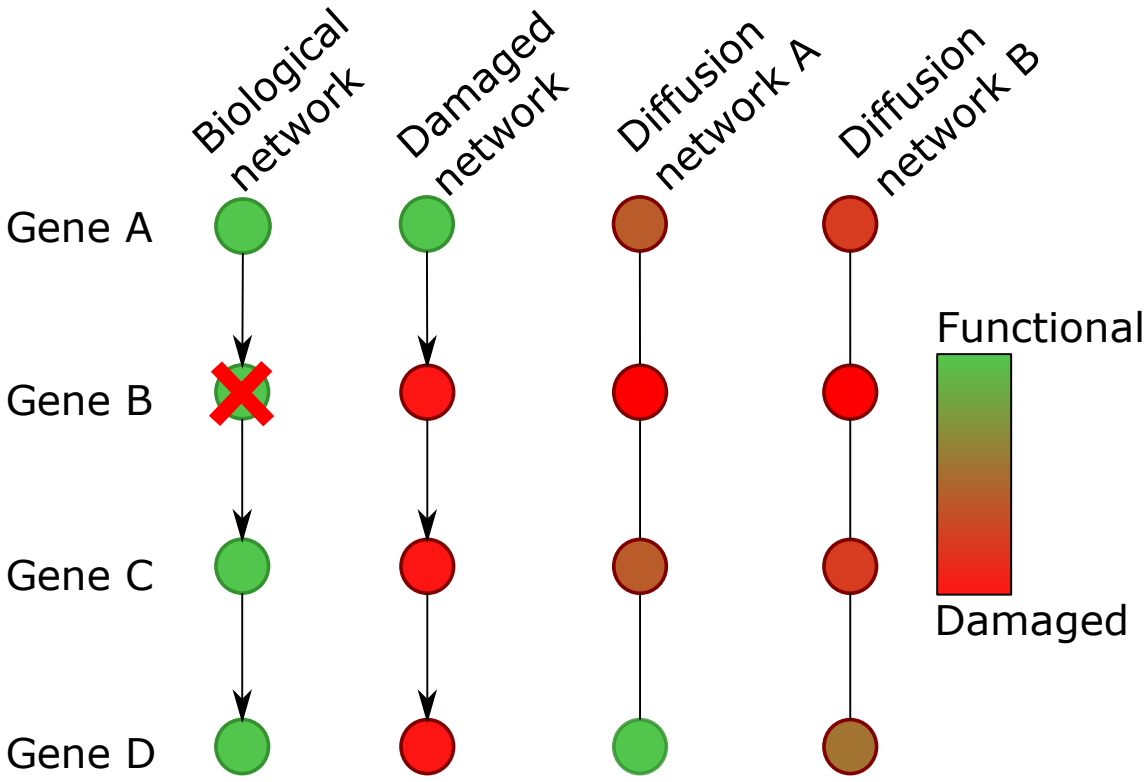


Figure 5.5: **Disruption of a linear pathway.** The biological network shows the healthy pathway. With gene B deleted, all downstream genes are also affected, as shown in the damaged network. When using a diffusion model with low diffusion rate, only genes A and C will also appear affected by deletion of gene B (network A), while a higher diffusion rate will lead to gene D also being affected (network B).

This class of algorithms requires as minimal knowledge an undirected network, meaning that interactions occur both in direction $A \rightarrow B$ and $B \rightarrow A$. However, testing all possible sub-networks of k genes for a network of N genes requires testing $\binom{k}{N} = \frac{N!}{k!(N-k)!}$ possibilities. Consequently, testing all possible sub-networks of sizes 1 to N would ask for $\sum_{k=1}^{N} \binom{k}{N} = 2^N - 1$ evaluations[1].

Diffusion model in linear (signaling) pathway

If we take the example of a linear pathway Figure 5.5, which can model a signaling pathway, the deletion of a gene in the middle of the pathway should lead to a disruption of the pathway integrity due to the lack of redundancy (Figure 5.5). A diffusion network will not predict that all genes downstream of the deletion are affected, but only neighbors, with an effect decreasing with the distance to the gene, the rate of the decrease being a parameter of the algorithm.

However, most of the time, pathways have more complex architectures, and the deletion of a gene may not be sufficient to disrupt the pathway. Some genes in the network play a more fundamental role than others, as they are used in the cross-talk between different elements of the network. These genes, termed hub genes, should be considered as good candidates of genes being targeted by cancer, if the pathway has to be shut down. TP53 is one such example of gene recurrently altered in cancers and is also central to many pathways [88, 89, 90] (see Figure 5.6).

Other times, however proxy genes (genes that have many neighboring genes somatically mutated, but non altered themselves) can be more relevant to understand the disease or be used as predictors of the disease [71]. This case is depicted in Figure 5.7, where the gene with the highest impact is not mutated itself, but has mutated neighbors.

Use of diffusion network assumes a high reliability in the edges of the graph, as the diffusion process can only occur between two entities that are linked to one another. In addition, this process can be computationally intensive in large networks as previously described. In order to bypass these restrictions, we detail enrichment of molecular mechanisms in the next subsection.

[1] Demonstration of this comes from the development of $(X+1)^N = \sum_{k=0}^{N} \binom{k}{N} \times 1^k \times 1^{(N-k)}$. For $X = 1$, we find $2^N = \sum_{k=0}^{N} \binom{k}{N} \times 1^k \times 1^{(N-k)}$

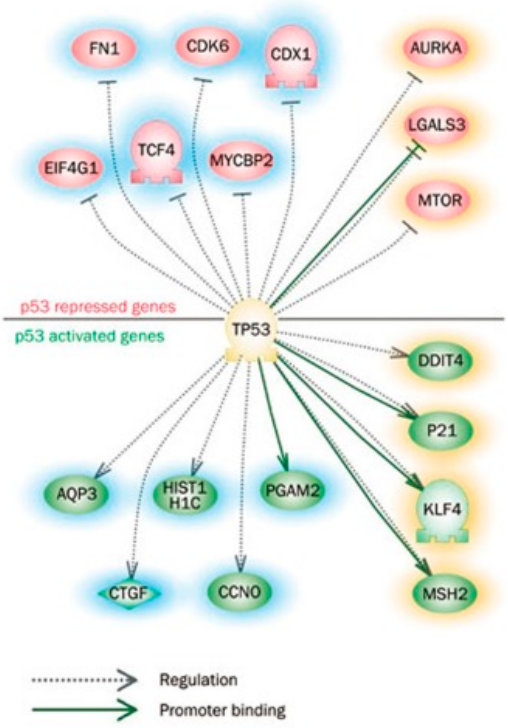


Figure 5.6: **The p53 network, adapted from Liu et al [91].**

5.1.3 Gene ontology

Two widely used algorithms to detect over represented pathways in a gene set are Panther [73, 74] and DAVID[75, 76]. Both rely on gene ontology (GO), which is defined as 'the framework for the model of biology. The GO defines concepts/classes used to describe gene function, and relationships between these concepts. It classifies functions along three aspects: molecular function [...], cellular component [...], [and] biological process.'[2]

Panther uses a binomial test to assess the over-representation or under-representation of a class in the input list [92]. An example is given in box 3 from Mi et al. [92], 440 genes map to the term 'induction of apoptosis'out of the roughly 20,000 genes of the human genome. This means that about $P_{apoptosis}2.2\%$ of the genes are considered as linked to this process. This means that for a gene list of $K = 500$ items, we would expect about $k_{apoptosis} = 11$ of them to be related to apoptosis. If more (respectively less) than that are found, we can estimate if this list is statistically enriched (respectively depleted). The corresponding p-value is computed with the formulas:

$$P_{enriched} = \sum_{k=k_{(C)}}^{K} \frac{K}{k} P_{(C)}^{K} (1 - P_{(C)})^{K-k} \tag{5.1}$$

[2] http://www.geneontology.org/

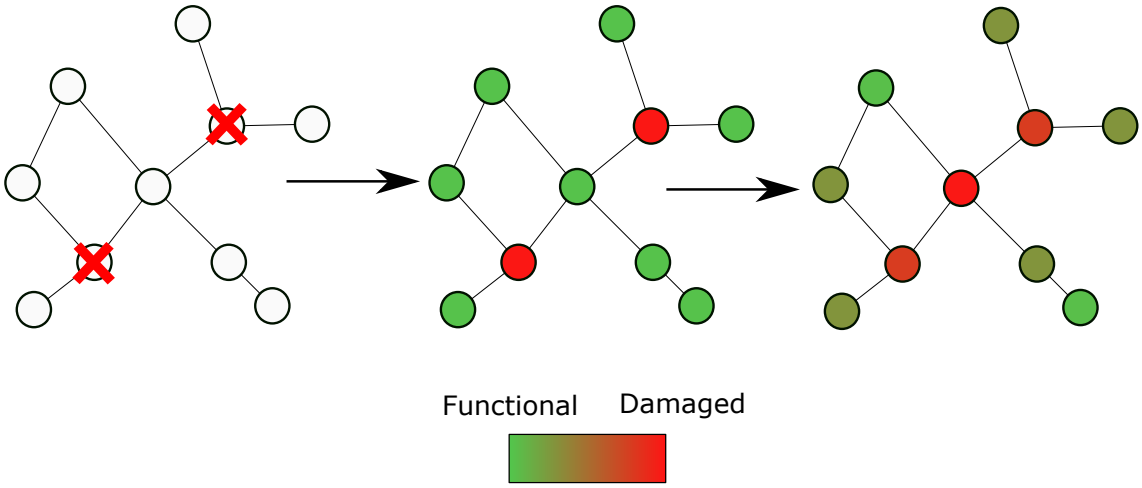


Figure 5.7: **Explanation of diffusion models: case of hub genes.** In this network, two genes were altered. By diffusion, the highest score will be found in a gene that is unaltered but central to the network. This gene can then be further studied, and linked to different aspect of the disease such as survival.

$$P_{depleted} = \sum_{k=0}^{k_{(C)}} \frac{K}{k} P_{(C)}^{K} (1 - P_{(C)})^{K-k} \tag{5.2}$$

DAVID uses a Fisher exact test to compute enrichment or depletion of a class in the list. It also incorporates annotations from databases outside of GO, such as KEGG [93] and BioCarta [94]. The results of the different databases are then aggregated using a κ measure, which evaluates agreement between the two sets. Both DAVID and Panther enrichment analysis have been used to assess the enrichment results in neuroblastoma from the ACSNMineR which will be presented in the next section.

5.1.4 Enrichment analysis of gene sets

N.B.: Gene Set Enrichment Analysis (GSEA) is a computational method maintained by the BROAD Institute. Here we describe enrichment analysis of gene sets, without capitals to avoid confusion.

Enrichment analysis of gene sets use ass input list(s) of genes or variants, and looks for statistical enrichment in this list. Contrary to diffusion networks, as the input is a list of genes, there are no information coming from interaction between two genes - which can be seen as more robust, but also as losing knowledge from the network.

Statistical model of gene enrichment

The goal of statistical is to compare a gene set, or **module**, to a control group. Usually, a module is compared to the statistical universe, which is the list of genes that can be observed in the experiment. For example, the universe from a chip analysis will depend on the probes of the chip, and will be limited to the genes that are quantified by the probe set.

The assumption behind gene set enrichment analysis is that all genes have the same probability to be mutated. As a result the observed proportion of genes mutated within a module can be compared to the proportion of genes outside the module.

Further refinement of the model can include the size of the genes to remove a possible bias towards modules with longer genes. Indeed, if we consider a model where the mutations are strictly random with uniform distribution along the genome - this corresponds to a random process of mutation without biological selection - then the probability to mutate a gene will be directly linked to its size. The direct consequence of this is that the probability of mutating a module will not only depend on the size (as the number of genes) of the module but on the total length of the genes making the module (this can be for example the sum of all gene transcript lengths).

A possibility to correct for that would be to compare the results to simulated sampling: given N variants, N being the observed number of variants in the biological data, and the probability to draw a gene being proportional to its size, it is possible to assess if the sampling is different from the observation.

This refinement however can be rendered difficult if not impossible when we take into account genetic events such as motif gain or disruption. Then the size of the gene should not only include its transcript, but also promoter regions. For transcription factors, a mutation in its targets could also be evaluated. This would mean that counting TFBS would artificially increase the length of the gene. In addition, regulatory regions could also be taken into account in the size of the 'gene'. Super enhancers are regions that can cover a few megabases, and are much longer than regular enhancers. Taking into account super enhancers would thus result in a drastic increase of the size of the gene. Taking into account all these possibilities would lead to a high variability in the 'gene'size, for that reason we limited the model to a uniform probability to mutate a gene independently of

its transcript size.

ACSN and ACSNMineR

This section has been adapted from Calculating Biological Module Enrichment or Depletion and Visualizing Data on Large-scale Molecular Maps with `ACSNMineR` *and* `RNaviCell` *Packages, Deveau et al [72] (see section 9.1).*

The Atlas of Cancer Signaling Network (ACSN) is a web-based database which describes signaling and regulatory molecular processes that occur in a healthy mammalian cell but that are frequently deregulated during cancerogenesis [95]. The ACSN atlas aims to be a comprehensive description of cancer-related mechanisms retrieved from the most recent literature.

Currently, ACSN maps cover signaling pathways involved in DNA repair, cell cycle, cell survival, cell death, epithelial-to-mesenchymal transition (EMT) and cell motility. Each of these large-scale molecular maps is decomposed in a number of functional modules. The maps themselves are merged into a global ACSN map. Thus the information included in ACSN is organized in three hierarchical levels: a global map, five individual maps, and several functional modules. Each ACSN map covers hundreds of molecular players, biochemical reactions and causal relationships between the molecular players and cellular phenotypes. ACSN represents a large-scale biochemical reaction network of 4,826 reactions involving 2,371 proteins (as of today), and is continuously updated and expanded.

We have included the three hierarchical levels in the `ACSNMineR` package, in order to be able to calculate enrichments at all three levels. The calculations are made by counting the number of occurences of gene symbols (HUGO gene names) from a given list of genes of interest in all ACSN maps and modules. Table 5.1 is detailling the number of gene symbols contained in all the ACSN maps.

The statistical significance of the counts in the modules is assessed by using either the Fisher exact test [23, 96] or the hypergeometric test, which are equivalent for this purpose [97].

The current ACSN maps are included in the `ACSNMineR` package, as a list of character matrices.

For each matrix, rows represent a module, with the name of the module in the first column, followed by a description of the module (optional), and then followed by all the gene symbols of the module. The maps will be updated according to every ACSN major release.

The main function of the `ACSNMineR` package is the `enrichment` function, which

Map	Total	Nb mod.	Min	Max	Mean
ACSN global	2239	67	2	629	79
Survival	1053	5	208	431	328
Apoptosis	667	7	19	382	136
EMT & Cell motility	634	9	18	629	137
DNA repair	345	21	3	171	45
Cell cycle	250	25	2	130	20

Table 5.1: **ACSN maps included in the `ACSNMineR` package**. Map: map name, Total: total number of gene symbols (HUGO) used to construct the map, Nb mod.: number of modules, Min: minimum number of gene symbols in the modules, Max: maximum number of gene symbols in the modules, Mean: average number of gene sybols per module. N.B.: one gene symbol may be present in several modules of the map.

is calculating over-representation or depletion of genes in the ACSN maps and modules. We have included a small list of 12 Cell Cycle related genes in the package, named `genes_test` that can be used to test the main enrichment function and to get familiar with its different options.

```
genes_test
 [1] "ATM"    "ATR"    "CHEK2"  "CREBBP" "TFDP1"  "E2F1"   "EP300"
 [8] "HDAC1"  "KAT2B"  "GTF2H1" "GTF2H2" "GTF2H2B"
```

The example shown below is the simplest command that can be done to test a gene list for over-representation on the six included ACSN maps. With the list of 12 genes mentionned above and a default p-value cutoff of 0.05, we have a set of 8 maps or modules that are significantly enriched. The results are structured as a data frame with nine columns displaying the module name, the module size, the number of genes from the list in the module, the names of the genes that are present in the module, the size of the reference universe, the number of genes from the list that are present in the universe, the raw p-value, the p-value corrected for multiple testing and the type of test performed. The module field in the results data frame indicate the map name and the module name separated by a column character. If a complete map is significantly enriched or depleted, then only the map name is shown, without any module or column character. For instance, the third line of the results object below concern the E2F1 module of the CellCycle map.

```
library(ACSNMineR)
results <- enrichment(genes_test)
dim(results)
```

```
[1] 8 9
results[3,]
           module module_size nb_genes_in_module
V161 CellCycle:E2F1       19                12
                                                          genes_in_module
V161 ATM ATR CHEK2 CREBBP TFDP1 E2F1 EP300 HDAC1 KAT2B GTF2H1 GTF2H2
   GTF2H2B
    universe_size nb_genes_in_universe  p.value p.value.corrected test
V161         2237                 12 3.735018e-21    2.353061e-19 greater
```

The `enrichment` function can take up to nine arguments: the gene list (as a character vector), the list of maps that will be used to calculate enrichment or depletion, the type of statistical test (either the Fisher exact test or the hypergeometric test), the module minimal size for which the calculations will be done, the universe, the p-value threshold, the alternative hypothesis ("greater" for calculating over-representation, "less" for depletion and "both" for both tests) and a list of genes that should be removed from the universe (option "Remove_from_universe"). This option may be useful for instance if we know beforehand that a number of genes are not expressed in the samples considered.

Only the gene list is mandatory to call the `enrichment` function, all the other arguments have default values. The `maps` argument can either be a dataframe imported from a GMT file with the `format_from_gmt` function or a list of dataframes generated by the same procedure. The GMT format corresponds to the Broad Institute's Gene Matrix Transposed file format, a convenient and easy way to encode named sets of genes of interest in tab-delimited text files (it is not a graph or network format). By default, the function `enrichment` uses the ACSN maps previously described. The correction for multiple testing is set by default to use the method of Benjamini & Hochberg, but can be changed to any of the usual correction methods (Bonferroni, Holm, Hochberg, Holm, or Benjamini & Yekutieli [98]), or even disabled . The minimal module size represents the smallest size value of a module that will be used to compute enrichment or depletion. This is meant to remove results of low significance for module of small size. The universe in which the computation is made by default is defined by all the gene symbols contained in the maps. All the genes that were given as input and that are not present on the maps will be discarded. To keep all genes, the user can change the universe to `HUGO`, and in that case, the complete list of HUGO gene symbols will be used as the reference ($>$ 39,000 genes). The threshold corresponds to the maximal value of the corrected p-value (unless the user chose not to correct for multiple testing) that will be displayed in the result table.

It may be of interest to compare enrichment of pathways in different cohorts or

experiments. For example, enrichment of highly expressed pathways can reveal differences between two cancer types or two cell lines. To facilitate such comparisons, `ACSNMineR` provides a `multisample_enrichment` function. It relies on the `enrichment` function but takes a list of character vector genes. The name of each element of the list will be assumed to be the name of the sample for further analysis. Most of the arguments given to `multisample_enrichment` are the same as the ones passed to `enrichment`. However, the `cohort_threshold` is designed to filter out modules which would not pass the significance threshold in all samples.

Finally, to facilitate visualization of results, `ACSNMineR` integrates a representation function based on `ggplot2` syntax [99]. It allows representation of results from `enrichment` or `multisample_enrichment` with a limited number of parameters. Two types of display are available: heat-map tiles or bars. For multiple samples using a barplot representation, the number of rows used can be provided, otherwise all plots will be on the same row. For the heatmap, the color of the non-significant modules, and boundaries of the gradient for significant values can also be tuned.

We previously computed the p-value of the `genes_test` list with default parameters. The number of modules which have a p-value below 0.05 was 8, that can be compared to the 16 obtained without correction with the simple command shown below (some of the results are displayed in Table 5.2).

```
enrichment(genes_test,correction_multitest = FALSE)
```

Module	Mod. size	Genes in module	p-value	Test
CellCycle	242	ATM ATR CHEK2 CREBBP TFDP1 E2F1 EP300 HDAC1 KAT2B GTF2H1 GTF2H2 GTF2H2B	5.4×10^{-7}	greater
CellCycle:APOPTOSIS_ENTRY	10	ATM ATR CHEK2 E2F1	3.5×10^{-7}	greater
CellCycle:CYCLINB	7	ATM	0.04	greater

Table 5.2: First rows of the results from enrichment analysis without correction. Module : name of the module. Mod. size: size of the module. Genes in module: genes from input which are found in the module. p-value: uncorrected p-value. Test : null hypothesis used, greater is synonym of enrichment.

We can now plot the first six rows of the results obtained for corrected and uncorrected fisher test with heatmap format (5.1.4) or barplot (Figure 5.1.4) with the following commands:

```
# heatmap
```

```
represent_enrichment(enrichment = list(Corrected = results[1:6,],
Uncorrected = results_uncorrected[1:6,]),
                              plot = "heatmap", scale = "reverselog",
                              low = "steelblue" , high ="white", na.value =
                                  "grey")

# barplot

 represent_enrichment(enrichment = list(Corrected = results[1:6,],
                             Uncorrected = results_uncorrected[1:6,]),
                               plot = "bar", scale = "reverselog",
                               nrow = 1)
```

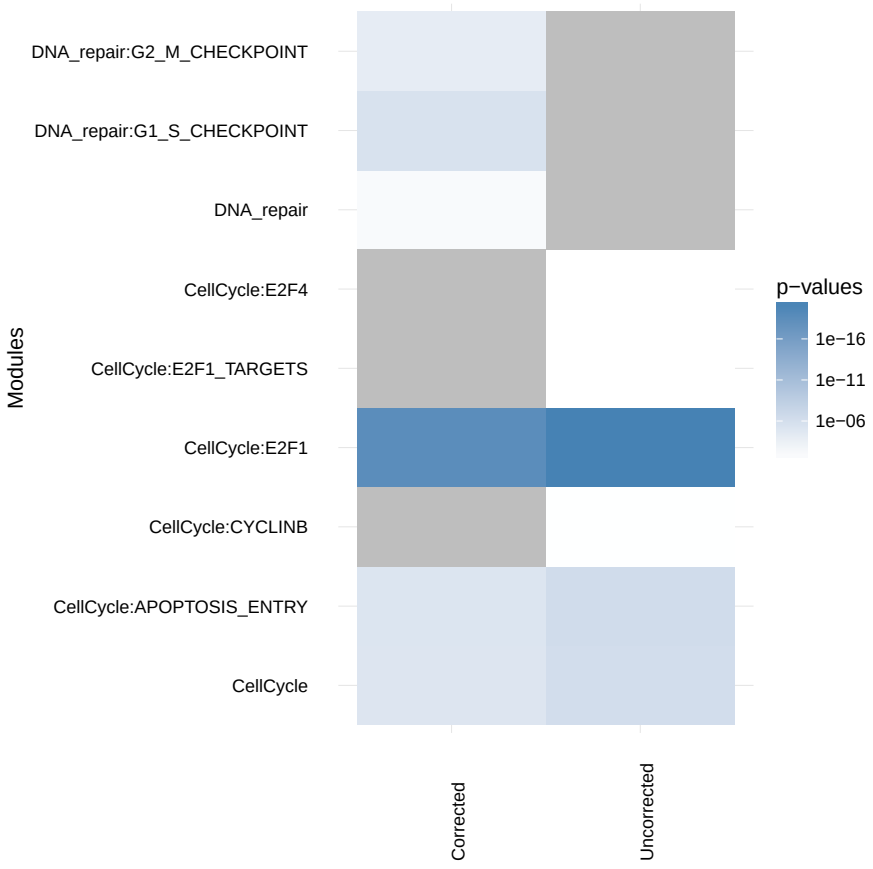


Figure 5.8: Representation of the enriched modules (first six rows for each setting), with either Bonferroni correction or no correction. Grey tiles means that the data is not available for this module in this sample. P-values of low significance are in white, whereas p-values of high significance are represented in blue.

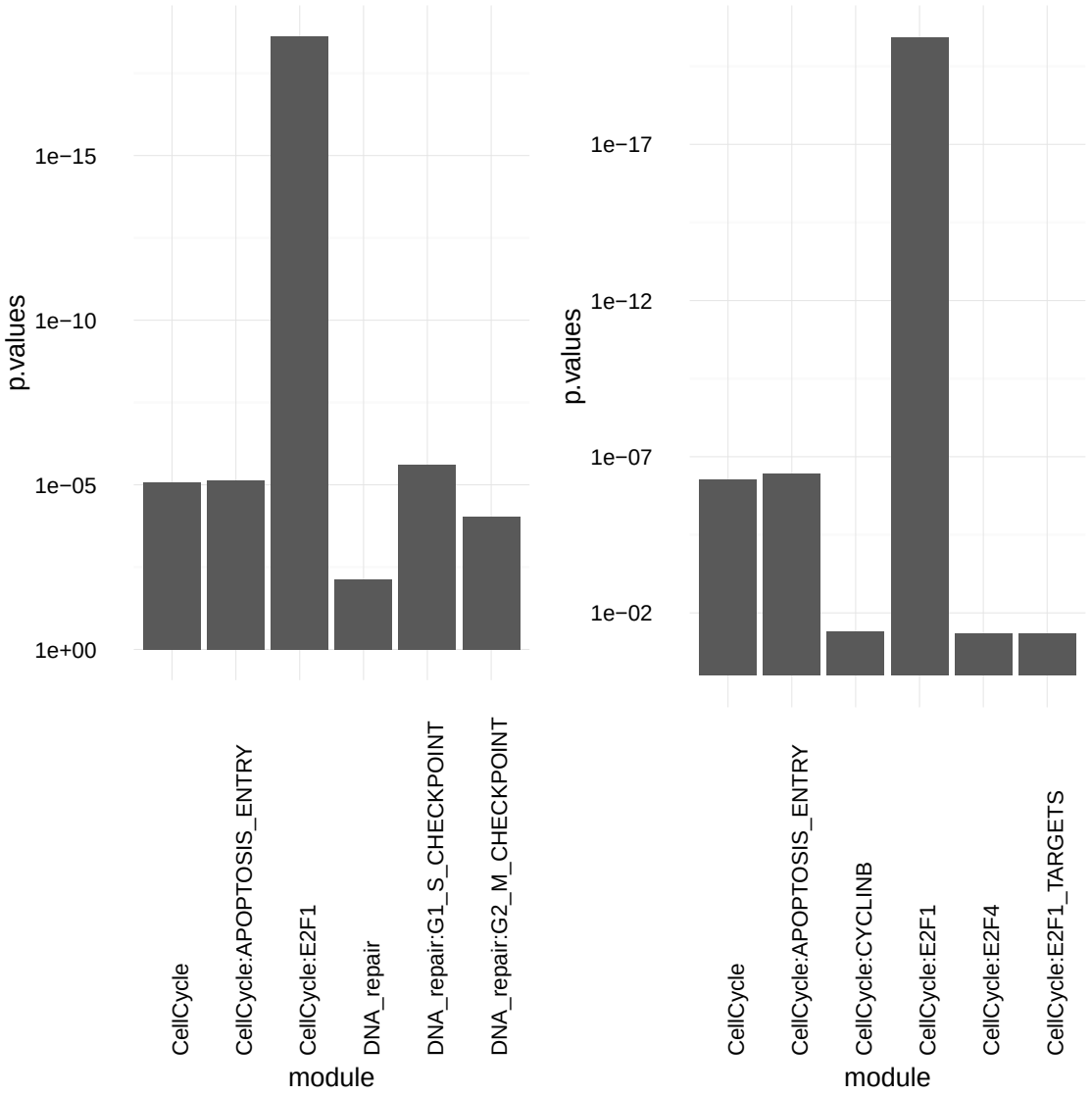


Figure 5.9: Representation of the enriched modules (first six rows for each setting), with either Bonferroni correction (left) or no correction (right). The modules are on the X axis and the p-values are on the Y axis.

5.2 A pipeline to combine enrichment and clonal reconstruction

5.2.1 Rationale and description

Previous publications have applied the same filters to variants used for the clonal reconstruction and those of biological relevance [100, 51, 61]. This approach masks the fact that the two different sets have very different purposes.

On the one hand, the clonal reconstruction set should have very low VAF dis-

persion and as few false positives as possible in order to achieve a very accurate clustering. Precisely pinpointing the cluster centers can facilitate interpretation of tumor evolution, and implies reliability of further results. Yet, those variants are not necessarily biologically relevant (passenger variants for example).

On the other hand, biologically relevant variants (variants with a known driver effect, such as ALK mutations in neuroblastoma), can be poorly covered by the sequencing. Stringent filtering of those variants would be detrimental to the understanding of the selection mechanisms that happen in the tumor.

In order to avoid compromises between the two approaches, we designed an original pipeline (Figure 5.10, adapted from 'Clonal assessment of functional variants in cancer based on a genotype-aware method for clonal reconstruction'section 9.2)

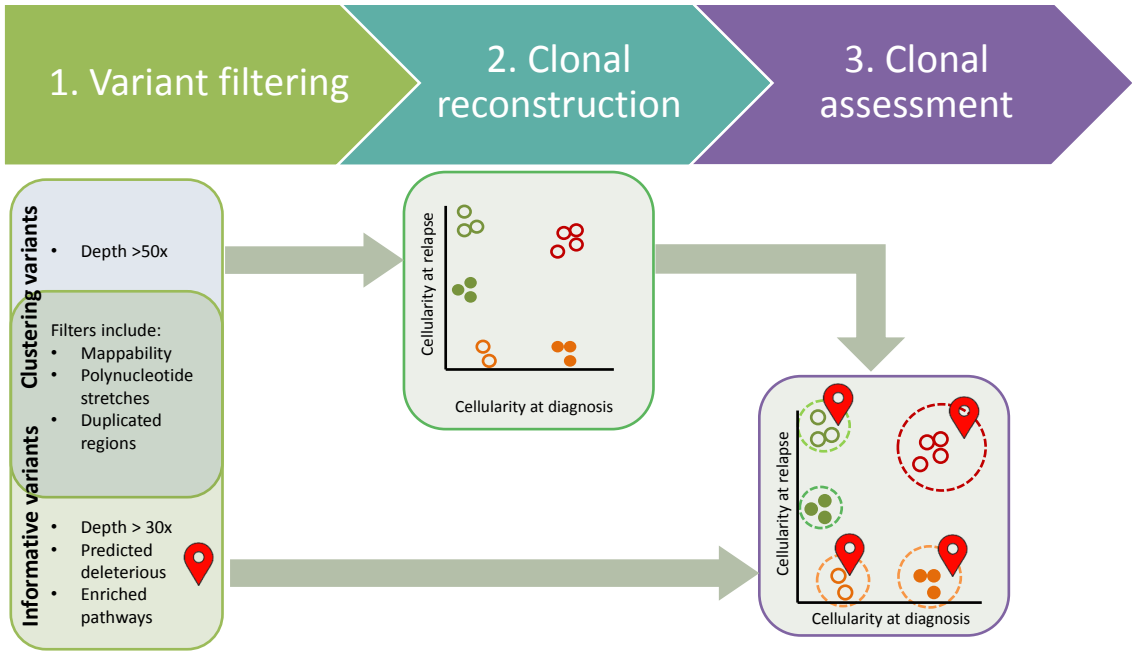


Figure 5.10: **Overview of the general clonal reconstruction workflow.** (1) Variants are filtered to remove false positive calls; stringent filters are used to produce mutations that are further employed for clonal reconstruction (step 2), tolerant filters are used to detect functional mutations. (2) Variants that pass stringent filters and have genotype information assigned to the corresponding genomic loci are used as input to QuantumClone to reconstruct clonal populations. (3) Finally, possibly damaging mutations belonging to frequently altered pathways are mapped to the reconstructed clones.

The stringent filters were previously described in subsection 4.2.3. We propose to define informative variants as variants in a pathway enriched in variants

predicted deleterious by at least one of the three algorithms (i.e. SIFT, Polyphen-2 or Funseq2). This approach would be especially benefit cancers with few genes mutated recurrently or those for which the recurrent genes are found in a small fraction of patients (e.g. neuroblastoma).

As a result of our definition, a variant will be considered deleterious in our study if at least one of the three predictive algorithm (i.e. SIFT, Polyphen-2 or Funseq2) predicts it as deleterious, and it is in a gene from a module recurrently altered, as predicted by ACSNMineR.

5.2.2 Validation of the pipeline on simulated data

We used simulations to validate the well-founded of our original pipeline design. *The following subsection is adapted from 'Clonal assessment of functional variants in cancer based on a genotype-aware method for clonal reconstruction '(see section 9.2)*

In silico validation data were generated using the QuantumCat method from package QuantumClone (version 1.0.0.3). We simulated variants coming from six clones observed in two samples per patient, with a purity of 70% for the first sample and 60% for the second. We created 150 variants that pass stringent filters, and an additional 150 variants passing tolerant filters but not stringent filters. All variants passing stringent filters were simulated in diploid regions, with a depth of coverage higher than $50\times$, whereas mutations passing permissive filters were located either in AB regions with a coverage between $30\times$ and $50\times$ (approximately $1/4$ of permissive variants), or in AAB regions with coverage $\geq 30\times$ (approximately $1/2$ of permissive variants), or in AABB regions with coverage $\geq 50\times$. We then attributed the 'driver 'characteristic to 100 variants, by sampling without replacement with probability $10/11$ to be selected from the variants passing permissive filters and probability $1/11$ to be selected from stringent filtering.

Pipelines

The 'classical'pipeline used all 300 simulated variants as input for the clonal reconstruction, using direct clustering by QuantumClone. The 'selective'pipeline used the 150 variants passing stringent filters as well as all variants qualified as drivers from the permissive filters as input for direct clustering. The 'two-step'pipeline first used the 150 stringent variants as input for direct clustering, and then attributed the variants qualified as drivers *a posteriori* to the clusters,

using the characteristics of the clones found by the initial QuantumClone clustering of high confidence variants. All three pipelines searched for two to ten clones, running with two different initializations, on four threads. Computational time was measured on a computer running Windows 10, with an Intel i7 at 2.7Gb with 8Gb of RAM, Rstudio 1.0.44 and R version 3.3.2.

Evaluation

Evaluation of the L2 error and NMI was made using only variants from the stringent and driver groups. The displayed computational time takes into account data processing, clustering and when necessary *a posteriori* attribution to the clonal structure.

Results

Comparison of the pipelines is summarized in Figure 5.11.

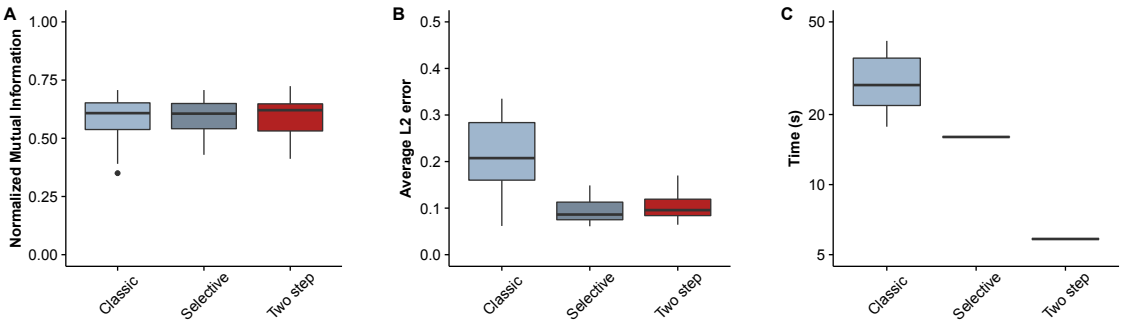


Figure 5.11: **Comparison of the three pipelines**. The pipeline aforementioned (two step), or a clustering using all variants called (classic) or a pipeline using only variants of biological interest and variants of high quality (selective) are assessed in terms of NMI **(A)**, average ℓ^2 error **(B)** or computational time **(C)**. The pipelines are evaluated on 20 simulations.

We demonstrate that the proposed two-step approach allows for a better reconstruction of the tumor, as well as an important decrease in computational time (Figure 5.11C). To test our pipeline, we compared it to two common pipelines: the first one, termed 'classic', uses all variants as input for the clustering. The second one, called 'selective', only uses variants passing the stringent filters and informative variants as input for the clustering. The third pipeline, termed 'two-step', uses *a posteriori* attribution of the putative drivers to the clones found using only variants passing stringent filters. While all three pipelines had similar outcomes when we compared the quality of reconstruction using normalized mutual

information (Figure 5.11A), the selective and two step pipelines fared significantly better than the classical pipeline in terms of ℓ^2 error ($p - value < 8 \times 10^{-6}$, one-sided Welch two-sample t-test, Figure 5.11B). In addition, the two step analysis resulted in an average 4.9 fold decrease in computational time compared to the classical pipeline and an average 2.7 fold decrease compared to the selective pipeline (Figure 5.11C). Furthermore, separating both steps facilitates iterative improvement of the clonal reconstruction. Once achieved, this reconstruction can be reused to answer questions about the evolution of different pathways separately, while previous pipelines required re-running the whole reconstruction with the new set of data.

Chapter 6

Application to neuroblastoma

> *It is only by means of the sciences of life that the quality of life can be radically changed.*
>
> — Aldous Huxley, *Foreword to Brave New World, 1947*

In this chapter we will aggregate methodology from all previous chapters in order to extract knowledge from the neuroblastoma WGS dataset.

6.1 Pathway enrichment results

The following subsection is adapted from 'Clonal assessment of functional variants in cancer based on a genotype-aware method for clonal reconstruction '(see section 9.2)

In our framework, we assumed that *functional* mutations (i.e. putative drivers) in a given cancer type should target specific signaling pathways or pathway modules (Figure 5.10, Step 2). We attributed annotated deleterious variants obtained with tolerant filters (Figure 5.10, section 9.2) to the ACSN maps and detected recurrently altered gene modules using the ACSNmineR package [72]. Overall, six general gene maps (apoptosis, cell cycle, DNA repair, EMT / cell motility, cell survival and neuritogenesis) and their 53 gene modules were found to be enriched in mutations (threshold 0.01 on the p-value, one-sided exact Fisher test, corrected to account for multiple testing with the Benjamini-Hochberg False Discovery Rate correction, corresponding to the q-value) (Table 6.1). The enrichment of pathways in ACSN was corroborated by enrichment of similar pathways from two other methods [75, 76, 73, 74] (not shown). In further analysis, deleterious mutations were annotated as *functional* when corresponding genes were included in the enriched pathways, or when such genes belonged to the Cancer Census list. The resulting number of *functional* mutations per patient varied from 2 to 147, with a median of 51.

Module	Module Size	Number of genes in module	Universe size	Number of genes in universe	p- value	p-value corrected
Apoptosis	666	132	25637	466	7.11E-86	9.38E-85
AKT_MTOR	79	24	25637	466	1.29E-22	3.88E-22
APOPTOSIS_GENES	189	51	25637	466	7.74E-43	4.26E-42
CASPASES	77	21	25637	466	6.53E-19	1.80E-18
HIF1	19	7	25637	466	2.88E-08	5.14E-08
MITOCH_METABOLISM	381	58	25637	466	1.16E-33	5.11E-33
MOMP_REGULATION	102	34	25637	466	8.11E-33	3.35E-32
TNF_RESPONSE	105	20	25637	466	8.07E-15	1.90E-14
CellCycle	239	57	25637	466	2.83E-44	1.87E-43
APC	15	4	25637	466	1.29E-04	1.74E-04
APOPTOSIS_ENTRY	10	3	25637	466	6.63E-04	8.10E-04
CYCLIND	9	3	25637	466	4.70E-04	5.86E-04
E2F1	17	3	25637	466	3.41E-03	4.02E-03
E2F4	8	2	25637	466	8.65E-03	9.68E-03
E2F1_TARGETS	129	31	25637	466	3.10E-25	1.02E-24
E2F2_TARGETS	35	4	25637	466	3.71E-03	4.29E-03
E2F3_TARGETS	51	5	25637	466	2.38E-03	2.86E-03
E2F4_TARGETS	100	22	25637	466	1.52E-17	4.01E-17
E2F5_TARGETS	6	3	25637	466	1.17E-04	1.60E-04
E2F6_TARGETS	34	5	25637	466	3.65E-04	4.63E-04
RB	12	4	25637	466	4.90E-05	6.89E-05
DNA_repair	343	80	25637	466	5.91E-60	5.57E-59
CELL_CYCLE	82	19	25637	466	7.97E-16	2.02E-15
G1_CC_PHASE	25	8	25637	466	1.05E-08	1.98E-08
G1_S_CHECKPOINT	32	13	25637	466	7.12E-15	1.74E-14
G2_M_CHECKPOINT	67	22	25637	466	1.03E-21	2.95E-21
M_CC_PHASE	24	7	25637	466	1.82E-07	3.17E-07
S_CC_PHASE	46	7	25637	466	1.98E-05	2.84E-05
S_PHASE_CHECKPOINT	44	14	25637	466	3.62E-14	8.25E-14
SPINDLE_CHECKPOINT	28	7	25637	466	5.85E-07	9.65E-07
DNA_REPAIR	169	36	25637	466	5.20E-27	1.91E-26
DR_REGULATORS	136	39	25637	466	1.86E-34	8.79E-34
HR	54	13	25637	466	1.55E-11	3.09E-11
MMR	18	3	25637	466	4.04E-03	4.59E-03
NER	54	13	25637	466	1.55E-11	3.09E-11
BER	49	10	25637	466	1.87E-08	3.43E-08
SSA	8	3	25637	466	3.18E-04	4.11E-04
A_NHEJ	18	5	25637	466	1.44E-05	2.11E-05
C_NHEJ	16	5	25637	466	7.55E-06	1.16E-05
FANCONI	41	8	25637	466	7.11E-07	1.12E-06
EMT_motility	628	167	25637	466	2.83E-128	6.23E-127
ADHERENS_JUNCTIONS	33	12	25637	466	3.77E-13	8.03E-13
CELL_CELL_ADHESIONS	107	31	25637	466	5.78E-28	2.25E-27
CELL_MATRIX_ADHESIONS	73	24	25637	466	1.51E-23	4.73E-23
CYTOSKELETON_POLARITY	153	34	25637	466	2.85E-26	9.90E-26
DESMOSOMES	29	12	25637	466	5.91E-14	1.30E-13
ECM	147	44	25637	466	1.73E-39	8.76E-39
EMT_REGULATORS	624	167	25637	466	9.03E-129	2.98E-127
GAP_JUNCTIONS	18	5	25637	466	1.44E-05	2.11E-05
TIGHT_JUNCTIONS	41	8	25637	466	7.11E-07	1.12E-06
Survival	1035	240	25637	466	3.49E-163	2.30E-161
HEDGEHOG	276	60	25637	466	6.28E-44	3.77E-43
MAPK	207	63	25637	466	7.22E-56	5.96E-55
PI3K_AKT_MTOR	293	89	25637	466	2.98E-77	3.28E-76
WNT_CANONICAL	424	81	25637	466	1.78E-53	1.30E-52
WNT_NON_CANONICAL	415	112	25637	466	1.69E-89	2.80E-88
Neuritogenesis	26	9	25637	466	5.59E-10	1.08E-09
Neuritogenesis_mutated	17	6	25637	466	3.92E-07	6.63E-07
Neuritogenesis_substrate	8	3	25637	466	3.18E-04	4.11E-04

Table 6.1: **Results from gene set enrichment analysis on the Atlas of Cancer Signalling Network.** Each component of the table is related to a map of ACSN (in bold), other lines correspond to modules of the map. **Module Size**: number of unique HUGO symbols in the module. **Number of genes in module**: number of genes from the input list in the tested module. **Universe size**: Size of the universe tested (here all HUGO symbols related to coding and non coding RNAs). **Number of genes in universe**: Number of unique symbols from the input list inside of the universe. **p-value**: p-value computed by Fisher test. **p-value corrected**: p-value after

At this step, the cell survival map registered the highest enrichment in putative drivers, and among its modules, the highest enrichment in putative driver mutations was observed for the non-canonical WNT pathway ($q-value \leq 10^{-88}$). In ad-

dition, we also detected significant enrichment in *functional* mutations of the WNT canonical and the MAPK pathways ($q-value \leq 10^{-51}$ and $\leq 10^{-54}$, respectively), and of the PI3K/AKT/mTOR and Hedgehog gene modules ($q-value \leq 10^{-75}$ and $\leq 10^{-43}$, respectively). As for the modules of other maps, genes coding for the EMT regulators were also significantly affected by the deleterious mutations in our cohort of relapsed neuroblastoma patients ($q-value \leq 10^{-126}$).

6.1.1 Discussion of enrichment results

General agreement between the enrichment tools validated the results from ACSN maps. However, one can wonder how from the 67 modules from ACSN, 59 items were found enriched. First, the three neuritogenesis maps and modules derived from Molenaar et al [9] are not part of the canonical ACSN, and have to be removed from the comparison, as well as the five different maps. This leaves a total of 51 modules from the 67 original found enriched, which is a rather high number.

One explanation for the high number of modules enriched in damaging variants stems from the intrinsic nature of ACSN that had been built to pick up signals from pathways deregulated in cancer. Moreover, we expect a slight bias towards cancer related genes as Funseq2 had been trained on the COSMIC database, meaning that this prediction tool could more easily detect variants biologically relevant for cancer.

6.2 Clonal structure in neuroblastoma

The following section is adapted from 'Clonal assessment of functional variants in cancer based on a genotype-aware method for clonal reconstruction '(see section 9.2)

6.2.1 Clonal reconstruction

We applied QuantumClone on *high fidelity* variants we defined using stringent filters (Figure 6.1A,). Across our cohort, we did not observe a significant association between the predicted number of clones and the number of mutations per patient (Spearman's $rho = -0.23$, $p-value = 0.35$). In addition, the number of clones at relapse was similar to that at diagnosis, even despite the fact that the relapse samples had about twice as many mutations as the diagnosis samples (number of mutation clusters varied from one to four with a median of three for

both time points).

In 79% of reconstructed clonal structures (15 out of 19 patients, we identified mutations coming from the ancestral clone (Fig. 4A), i.e. the clone that gave rise to all cells in both diagnosis and relapse samples.

Assignment of *functional* mutations to the identified clonal structure

Using the results of the mapping of *functional* mutations on the clonal structure detected for each patient by QuantumClone (Figure 5.10, Step 3), we annotated mutations as (i) those belonging to expanding clones - corresponding to a two-fold cellular prevalence increase between diagnosis and relapse, (ii) those belonging to shrinking clones - cellular prevalence halved between diagnosis and relapse, and (iii) those belonging to ancestral clones - cellular prevalence higher than 70% in both samples (Figure 6.1A). Overall, 36%, 30% and 9.6% of all *functional* mutations fell in these three categories.

Analysis of pathways enriched in *functional* mutations in shrinking and expanding clones

Assignment of mutations to clones shrinking or expanding after the treatment resulted in the identification of 336 and 400 possible driver mutations in these clone types, respectively. Expanding clones had more deleterious mutations targeting genes from all six general maps (apoptosis, cell cycle, DNA repair, EMT/cell motility, cell survival and neuritogenesis) than the shrinking clones (Fig. Figure 6.1B). Similarly, in these expanding clones, most of the corresponding gene modules (e.g., MAPK, WNT canonical or PI3K/AKT/mTOR) were also more frequently targeted. An extreme example of this behavior can be given with the neuritogenesis substrates module, the RB pathway or the E2F1 pathway in which genes are only found mutated in the expanding clones. The increase in functional variants can partly be explained by the observed doubling of variants at relapse compared to diagnosis.

We define μ the functional mutation rate in a module as the number of functional variants per high fidelity variants of the patient by number of genes in a module. The functional mutation rate across modules was significantly different between the three classes of clones according to the z-score computed as sug-

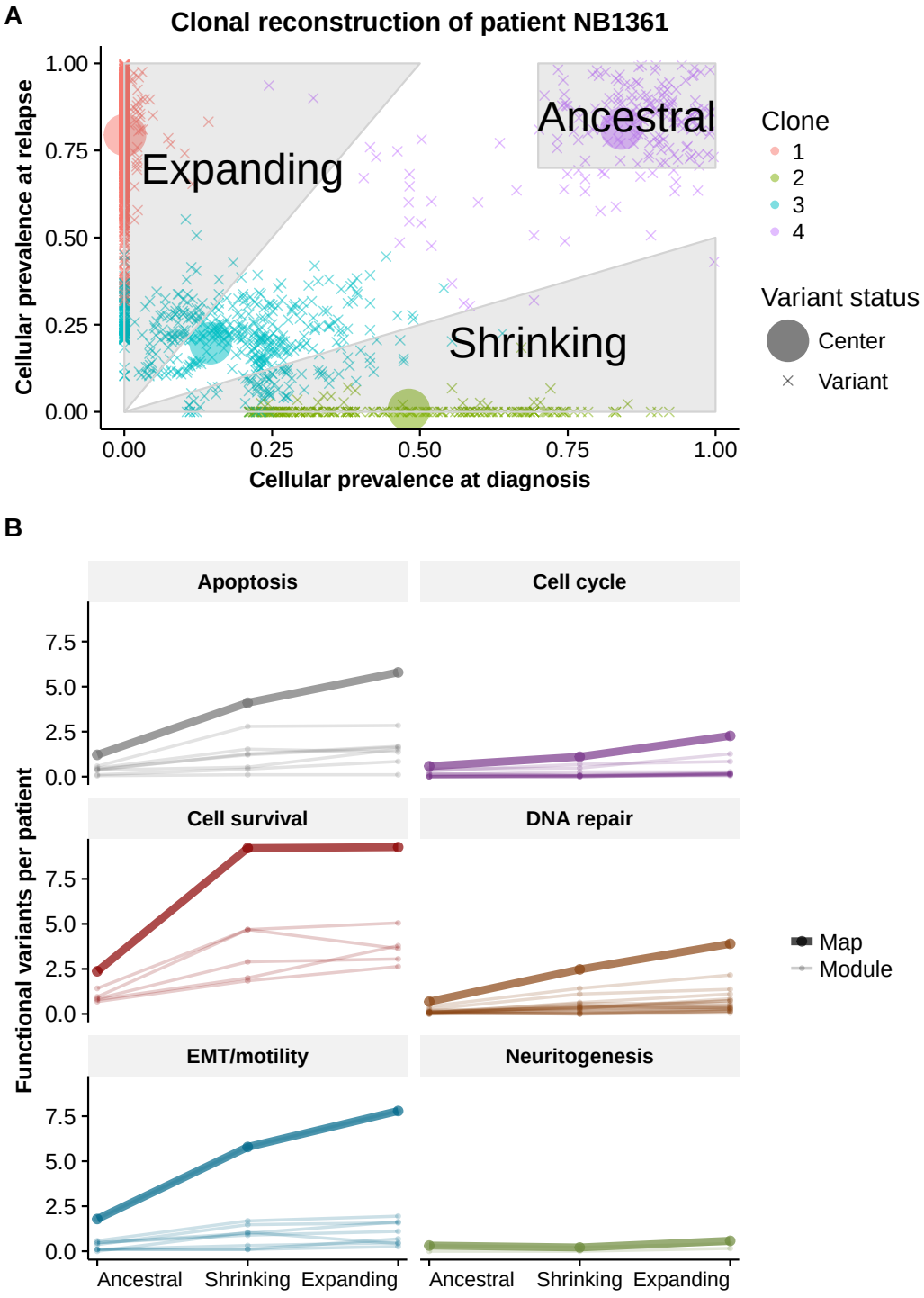


Figure 6.1: **Annotation of clones in neuroblastoma and pathway enrichment analysis. (A)** Illustration with data from patient NB1361 of the rules for assignment of variants to (i) the ancestral clone (cellular prevalence of the mutation cluster exceeds 70% both at diagnosis and relapse), (ii) clones expanding after the treatment (cellular prevalence of the mutation cluster increases at least two-fold at relapse) and (iii) shrinking clones (cellular prevalence of such mutation clusters decreases at least two-fold). **(B)** Evolution of the total number of functional variants for enriched maps and modules, across all 19 patients. The majority of modules show an increase in the number of functional variants between two time points.

gested by Paternoster et al. [101] (Figure 6.2, $p - value = 8.35 \times 10^{-5}$ between ancestral and shrinking, $p - value = 2.84 \times 10^{-3}$ between ancestral and expanding and $p - value = 4.98 \times 10^{-2}$ between expanding and shrinking). This functional mutation rate has been previously linked to the fitness of a clone [102], and it is interesting to notice that the functional mutation rate is lower in the ancestral clone ($\mu = 5.803$ functional variations per 1,000 variants per 1,000 genes in module, standard error $s.e = 1.322$), than in the shrinking clones ($\mu = 15.78$, $s.e. = 1.919$) or expanding clones ($\mu = 10.92$, $s.e. = 0.7583$). The change in functional mutation rate suggests different selection mechanisms. The fact that there are fewer functional variants in the ancestral population than in the shrinking or expanding populations and that the expanding population has a lower functional mutation rate suggests that a clone with fewer functional variants had better adaptive capabilities, as proposed by Chen et al [103].

6.2.2 Model for clonal evolution

For some of our samples, we did not succeed in uncovering an ancestral clone despite the fact that copy number breakpoints were consistent between samples, ensuring a common phylogeny [104] (Figure 4.4). Disappearance at relapse of many potential driver mutations seemingly present in the ancestral clone at diagnosis, may be due to tumor heterogeneity and the fact that biopsies were taken from different tumor sites. This situation has been termed "illusion of clonality" [105].

Previous studies have shown that the number of variants was linked to the number of divisions a cell undergoes [106]. The observed doubling of variants between diagnosis and relapse suggests that cells have undergone as many divisions between diagnosis and relapse as between cancer origin and diagnosis - with the assumption that the mutational rate remains constant. This would in particular exclude the possibility of the relapse emerging from a quiescent population.

We showed that in neuroblastoma, the functional mutation rate was significantly lower in the ancestral populations compared to the clones expanding or shrinking at relapse. Chen et al [103] have shown that wild-type cells have more adaptive capabilities than mutants, even though a mutant can appear fitter than the wild-type lineage in a specific culture condition. Applied to our results, their finding could suggest that a clone with a low level of functional variants would be more likely to adapt to environment changes during and after treatment. After this selection round and once the tumor environment has returned to physiological

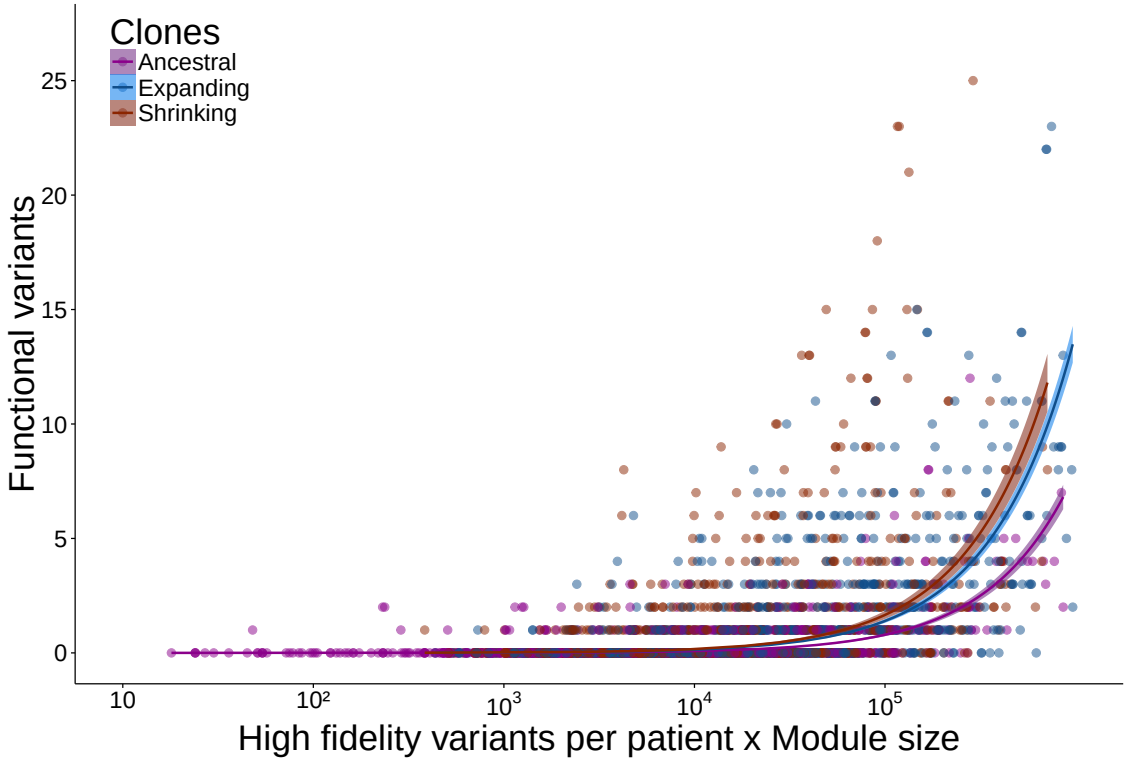


Figure 6.2: **Ancestral, shrinking and expanding clones exhibit different mutation patterns in neuroblastoma relapse tumors.** Functional mutation rate is higher in shrinking and expanding clones compared to the ancestral ones. We define the functional mutation rate as a ratio of the number of functional mutations to the number of high fidelity variants. For a given gene module the number of functional mutations in each patient is supposed to linearly depend on the product of the module size and the total number of detected variants. Therefore, we used the product of the module size and number of high fidelity variants as a covariate in a linear regression model evaluating functional mutation rate for neuroblastoma tumors. The rate was defined as the slope of the linear regression.

state, another set of functional variants would appear, giving selective advantage to the expanding clone.
A direct consequence of this assumption is that the functional mutation rate should be lower at relapse compared to diagnosis, as a period of low functional mutation rate before treatment would be followed by a period of higher functional mutation rate during disease progression (Figure 6.3). This consequence is in line with the 29% functional mutation rate decrease observed between expanding and shrinking clones in neuroblastoma.

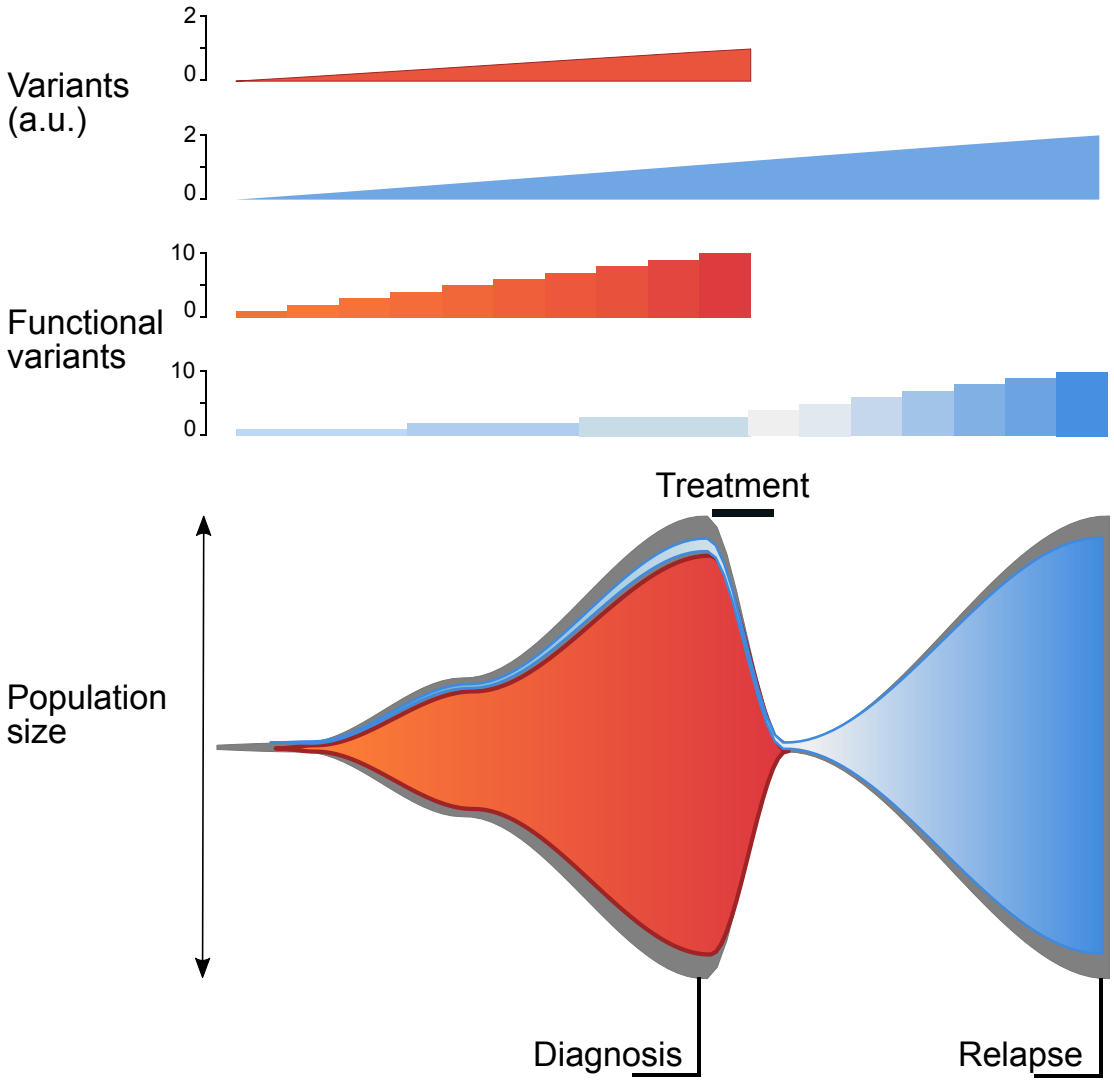


Figure 6.3: **Model for clonal evolution in neuroblastoma.** Given the differences in functional mutation rates observed in neuroblastoma relapse tumors we propose the following model for clonal selection in this type of cancer: (1) Clones with high functional mutation rate (red) disappear after the chemotherapy; lower mutational burden provides an advantage in escape from treatment; (2) lower values for functional mutation rate in clones expanding at relapse (blue) compared to the shrinking clones (red) is due to a lower frequency of functional mutations before treatment, followed by a gradual accumulation of functional mutations at relapse. From top to bottom: the number of variants in the clone, number of functional variants in the clone, and population size in the tumor; a.u. (arbitrary units). The change in color hue represents the changes in time of the genotype with the accumulation of variants.

Chapter 7

Conclusion and perspectives

> *May all your dreams but one come true, for what is life without a dream?*
>
> — David Gemmell, *The First Chronicles of Druss the Legend, 1993*

We contributed in this thesis to different fields of computational biology.

The first contribution, the development of a new clonal reconstruction method using HTS data, is at the core of this manuscript. We have demonstrated that despite the existence of many different competing algorithms, there was room for improvement both in terms of clustering quality and computation time. Our method, QuantumClone, is now available as a CRAN package as well as its source code and code to reproduce all simulations presented in this manuscript. This code diffusion is not only made in an effort of making reproducible research, but also to enable continuous improvements of existing tools.

The second contribution presented here was to the variant calling field, through participation to a DREAM challenge, in order to gather new insights on the sources of error from variant caller, and through the development of filtering pipeline to extract high confidence variants.

The third topic raised was systems biology. In the same way that QuantumClone was made public, the R package ACSNMineR is freely available both from CRAN and GitHub, allowing anyone to contribute to this effort either by adding to the code or by integrating new maps.

Finally, the last contribution presented here was to the understanding of neuroblastoma biology. In addition to the usual description of clonal architecture, we also proposed a model derived from these observations to explain different mutational rates in the ancestral clones, clones shrinking or expanding at relapse.

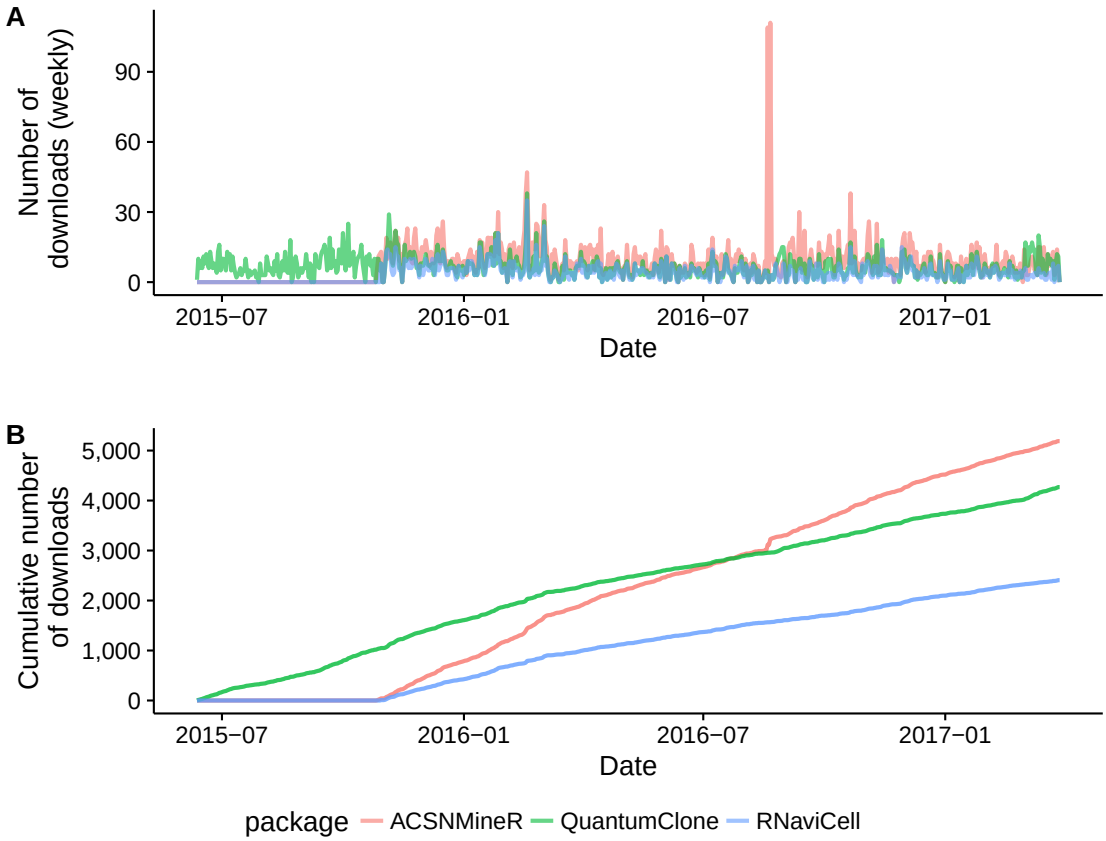


Figure 7.1: **Number of downloads of ACSNMineR and QuantumClone packages from CRAN repository.**The numbers are either displayed **weekly (A)**, or **cumulative (B)** The figures for the RNaviCell package, also described in Deveau et al[72], are given as a reference.

Hopefully, this story will not end with this conclusion. As of today, at least two other projects are being developed on clonal reconstruction and use our algorithm. The first one, in the translational unit of Fabien Reyal, focuses on evolution after treatment in breast cancer using whole exome sequencing. The second project, supported by Isabelle Janoueix and Simon Durand, aims at finding the different populations coexisting in a cell line derived from a patient. Once identified, the populations will be monitored under different treatments. The tool diffusion can also be observed by the number of downloads of each package from CRAN servers (this is not taking into account the github repository), as shown in Figure 7.1.

This figure also shows that simpler tools such as ACSNMineR can have a much broader echo in the scientific community than specialized tools. Interesetingly, we can also see that the publication in December 2016 in the R Journal of the ACSNMineR and RNaviCell packages did not increase the download rate.

In the application of our framework to neuroblastoma sequencing data, we excluded information about translocations and indels. The reason for this was that the analysis of clonal structure is based on the number of sequencing reads supporting each genetic variant. While we suppose that the number of reads with a mismatch mutation is proportional to the number of DNA molecules harboring this variant, we expect that due to read mapping issues the fraction of reads indicating an indel or a translocations will be generally lower than the actual proportion of DNA molecules with the rearrangement. Eviction of large and small SVs seemingly resulted in a decrease in sensitivity of the detection of genetic driver events. A possible way to solve this issue would be to estimate the cellular prevalence of these event using specific tools and attribute such events to the most likely clone.

The proposed framework can be applied in the future to any type of cancer. The pre-requirements are sufficient number of candidate mutations (at least 50 mutations per sample) and a minimal read depth of coverage of 50$\times$. These requirements are usually met by WGS or whole exome sequencing datasets. Our simulation results show that increasing the number of mutations used for clonal reconstruction above 50 does not improve significantly the clonal reconstruction accuracy provided that mutations specific for every clone are present in the input. This technique would suit the breakthrough of cell-free DNA (i.e. the DNA coming from apoptotic cells and carried in the bloodstream) sequencing can lead to a surge in the samples available to track the disease. Indeed, with this non-invasive technique, it should be possible to track the evolution of the tumor during treatment using only a blood sample. With this increase in sample availability from a patient, not only could the resolution of the clonal architecture be improved, but also warnings could be raised for the possible reemergence of the tumor after treatment. It could also help distinguishing variants that confer resistance to treatment to a subclonal population.

Chapter 8

Annexes

8.1 Computation of the exact gradient

Starting from the equation 3.1, we can compute the exact gradient of the partial log-likelihood:

$$\ell = \sum_{i \in variant} \sum_{k \in clones} \sum_{s \in samples} \sum_{p \in possibilities(i)} \omega_{(i,p)} t_{(i,k,p)} \log\left(P\left(Alt_{i,s,p} | \dot{\varsigma}_{k,s}\right)\right)$$

With:

- Alt : the number of alternative reads;
- $\omega_{(i,p)}$: the weight of possibility p, so that for a variant the sum of possibilities is 1;
- $t_{(i,k,p)}$: the contribution of possibility p to cluster k computed during the E-step;
- $\dot{\varsigma}_{k,s}$: the cellular prevalence of cluster k in sample s

In addition, in the case of a binomial model, we have:

$$P\left(Alt_{i,s,p} | \dot{\varsigma}_{k,s}\right) = \left(\alpha_{i,s,p} \dot{\varsigma}_{k,s}\right)^{Alt_{i,s}} \left(1 - \alpha_{i,s,p} \dot{\varsigma}_{k,s}\right)^{Ref_{i,s}}$$

With Ref the number of reads supporting the reference allele and alpha:

$$\alpha_{i,s,p} = \frac{N_{i,s,p}}{G_{i,s}}$$

Where N is the number of copies of the variant i in possibility p, and G the number of copies of the locus.

$$
\begin{aligned}
\Rightarrow \frac{\partial \ell}{\partial \dot{\varsigma}_{k,s}} &= \frac{\partial}{\partial \dot{\varsigma}_{k,s}} \left(\sum_{i \in variant} \sum_{k \in clones} \sum_{s \in samples} \sum_{p \in possibilities(i)} \omega_{(i,p)} t_{(i,k)} \log\left(P\left(Alt_{i,s,p} | \dot{\varsigma}_{k,s}\right)\right) \right) \\
&= \sum_{i \in variant} \sum_{p \in possibilities(i)} \omega_{(i,p)} t_{(i,k)} \frac{\partial}{\partial \dot{\varsigma}_{k,s}} \left(\log\left(P\left(Alt_{i,s,p} | \dot{\varsigma}_{k,s}\right)\right)\right) \\
&= \sum_{i \in variant} \sum_{p \in possibilities(i)} \omega_{(i,p)} t_{(i,k)} \frac{\partial}{\partial \dot{\varsigma}_{k,s}} \left(\log\left(\left(\alpha_{i,s,p}\dot{\varsigma}_{k,s}\right)^{Alt_{i,s}} \left(1 - \alpha_{i,s,p}\dot{\varsigma}_{k,s}\right)^{Ref_{i,s}}\right)\right) \\
&= \sum_{i \in variant} \sum_{p \in possibilities(i)} \omega_{(i,p)} t_{(i,k)} \frac{\partial}{\partial \dot{\varsigma}_{k,s}} (Alt_{i,s}(\log(\alpha_{i,s,p}) + \log(\dot{\varsigma}_{k,s}) \\
&\quad + Ref_{i,s} \log(1 - \alpha_{i,s,p}\dot{\varsigma}_{k,s}))) \\
&= \sum_{i \in variant} \sum_{p \in possibilities(i)} \omega_{(i,p)} t_{(i,k)} \left(\frac{Alt_{i,s}}{\dot{\varsigma}_{k,s}} - \frac{Ref_{i,s} \times \alpha_{i,s,p}}{1 - \alpha_{i,s,p} \times \dot{\varsigma}_{k,s}}\right) \\
&= \sum_{i \in variant} \sum_{p \in possibilities(i)} \omega_{(i,p)} t_{(i,k)} \frac{Alt_{i,s} - \alpha_{i,s}\left(Alt_{i,s} + Ref_{i,s}\right)\dot{\varsigma}_{k,s}}{\dot{\varsigma}_{k,s}\left(1 - \alpha_{i,s}\dot{\varsigma}_{k,s}\right)}
\end{aligned}
$$

In addition, for diploid or haploid cases - where $\alpha_{i,s}$ is independant of i and s- the solutions for $\nabla \ell = 0$ are :

$$\sum_{i \in variant} \sum_{p \in possibilities(i)} \omega_{(i,p)} t_{(i,k)}$$

8.2 Dissimilarity matrix and weighted average initialization

For two variants i and j, characterized by:

- their number of reads supporting each (Alt_i and Alt_j);
- their total number of reads overlapping the position (D_i and D_j);
- their number of copies of the locus (NL_i and NL_j);
- their tested number of copies of the variant (N_i and N_j).

We can compute a normalized number of alternative reads (which would be the number of alternative reads expected for a diploid variant):

$$Alt_{diplo} = round(\frac{Alt \times NL}{2 \times NC}$$

In addition, we define the observed probability for each variant:

$$p_x = \frac{Alt_{diplo,x}}{Depth_x}$$

As well as the weighted probability for both variants:

$$p = \frac{Alt_{diplo,i} + Alt_{diplo,j}}{Depth_i + Depth_j}$$

Then

$$z = \frac{(p_i - p_j)^2}{p \times (1-p)} \times \left(\frac{1}{Depth_i} + \frac{1}{Depth_j} \right)$$

We can then obtain the p-value associated with such score:

$$p - value = 2 \times pnorm(-\sqrt{z})$$

With pnorm the distribution of the Gaussian of mean 0 and standard deviation 1.

After hierarchical clustering, we obtain n clusters. The initialization weights are simply the ratio of the number of variants in a cluster divided by the total number of variants. The center of each cluster in a sample is defined by:

$$\dot{c}_{k,s} = 2 \times \frac{\sum_{i \in variants} Alt_{i,normalized}}{\sum_{i \in variants} Depth_i} \tag{8.1}$$

Where $Alt_{i,normalized}$ is the number of alternative reads that should be observed if the variant was in a diploid locus.

Bibliography

[1] Fred Harding. *Breast Cancer: Cause, Prevention, Cure*. Tekline Publishing, February 2007.

[2] Douglas Hanahan and Robert A. Weinberg. The Hallmarks of Cancer. *Cell*, 100(1):57–70, January 2000.

[3] Douglas Hanahan and Robert A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, March 2011.

[4] Peter Kaatsch. Epidemiology of childhood cancer. *Cancer Treatment Reviews*, 36(4):277–285, June 2010.

[5] Manrong Jiang, Jennifer Stanke, and Jill M. Lahti. The Connections Between Neural Crest Development and Neuroblastoma. *Current topics in developmental biology*, 94:77–127, 2011.

[6] John M Maris, Michael D Hogarty, Rochelle Bagatell, and Susan L Cohn. Neuroblastoma. *The Lancet*, 369(9579):2106–2120, June 2007.

[7] P. F. Ambros, I. M. Ambros, G. M. Brodeur, M. Haber, J. Khan, A. Nakagawara, G. Schleiermacher, F. Speleman, R. Spitz, W. B. London, S. L. Cohn, A. D. J. Pearson, and J. M. Maris. International consensus for neuroblastoma molecular diagnostics: report from the International Neuroblastoma Risk Group (INRG) Biology Committee. *British Journal of Cancer*, 100(9):1471–1482, 2009.

[8] Angela Bellini, Virginie Bernard, Quentin Leroy, Thomas Rio Frio, Gaelle Pierron, Valérie Combaret, Eve Lapouble, Nathalie Clement, Herve Rubie, Estelle Thebaud, Pascal Chastagner, Anne Sophie Defachelles, Christophe Bergeron, Nimrod Buchbinder, Sophie Taque, Anne Auvrignon, Dominique Valteau-Couanet, Jean Michon, Isabelle Janoueix-Lerosey, Olivier Delattre, and Gudrun Schleiermacher. Deep Sequencing Reveals Occurrence of Subclonal ALK Mutations in Neuroblastoma at Diagnosis. *Clinical Cancer Research*, 21(21):4913–4921, January 2015.

[9] Jan J. Molenaar, Jan Koster, Danny A. Zwijnenburg, Peter van Sluis, Linda J. Valentijn, Ida van der Ploeg, Mohamed Hamdi, Johan van Nes, Bart A. Westerman, Jennemiek van Arkel, Marli E. Ebus, Franciska Haneveld, Arjan Lakeman, Linda Schild, Piet Molenaar, Peter Stroeken, Max M. van Noesel, Ingrid Øra, Evan E. Santo, Huib N. Caron, Ellen M. Westerhout,

and Rogier Versteeg. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature*, 483(7391):589–593, March 2012.

[10] John M. Maris, Matthew J. Weiss, Yael Mosse, George Hii, Chun Guo, Peter S. White, Michael D. Hogarty, Tamar Mirensky, Garrett M. Brodeur, Timothy R. Rebbeck, Margrit Urbanek, and Suzanne Shusterman. Evidence for a Hereditary Neuroblastoma Predisposition Locus at Chromosome 16p12–13. *Cancer Research*, 62(22):6651–6658, November 2002.

[11] Yaël P. Mossé, Marci Laudenslager, Luca Longo, Kristina A. Cole, Andrew Wood, Edward F. Attiyeh, Michael J. Laquaglia, Rachel Sennett, Jill E. Lynch, Patrizia Perri, Geneviève Laureys, Frank Speleman, Cecilia Kim, Cuiping Hou, Hakon Hakonarson, Ali Torkamani, Nicholas J. Schork, Garrett M. Brodeur, Gian P. Tonini, Eric Rappaport, Marcella Devoto, and John M. Maris. Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature*, 455(7215):930–935, October 2008.

[12] Lindi Chen, Angharad Humphreys, Lisa Turnbull, Angela Bellini, Gudrun Schleiermacher, Helen Salwen, Susan L. Cohn, Nick Bown, Deborah A. Tweddle, Lindi Chen, Angharad Humphreys, Lisa Turnbull, Angela Bellini, Gudrun Schleiermacher, Helen Salwen, Susan L. Cohn, Nick Bown, and Deborah A. Tweddle. Identification of different ALK mutations in a pair of neuroblastoma cell lines established at diagnosis and relapse. *Oncotarget*, 7(52):87301–87311, November 2016.

[13] D. R. Lohmann, B. Brandt, W. Höpping, E. Passarge, and B. Horsthemke. The spectrum of RB1 germ-line mutations in hereditary retinoblastoma. *American Journal of Human Genetics*, 58(5):940–949, May 1996.

[14] K C Sippel, R E Fraioli, G D Smith, M E Schalkoff, J Sutherland, B L Gallie, and T P Dryja. Frequency of somatic and germ-line mosaicism in retinoblastoma: implications for genetic counseling. *American Journal of Human Genetics*, 62(3):610–619, March 1998.

[15] James William Tutt. *British Moths*. G. Routledge, 1896.

[16] Bruno Canard and Robert S. Sarfati. DNA polymerase fluorescent substrates with reversible 3'-tags. *Gene*, 148(1):1–6, October 1994.

[17] H. Thorvaldsdottir, J. T. Robinson, and J. P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, March 2013.

[18] Paramvir Dehal and Jeffrey L. Boore. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLOS Biology*, 3(10):e314, September 2005.

[19] Hui Shen, Jian Li, Jigang Zhang, Chao Xu, Yan Jiang, Zikai Wu, Fuping Zhao, Li Liao, Jun Chen, Yong Lin, Qing Tian, Christopher J. Papasian, and Hong-Wen Deng. Comprehensive Characterization of Human Genome

Variation by High Coverage Whole-Genome Sequencing of Forty Four Caucasians. *PLoS ONE*, 8(4), April 2013.

[20] Daniel C. Koboldt, David E. Larson, and Richard K. Wilson. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, 44:15.4.1–15.4.17, December 2013.

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

[22] Christopher T. Saunders, Wendy S. W. Wong, Sajani Swamy, Jennifer Becq, Lisa J. Murray, and R. Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817, July 2012.

[23] Ronald A Fisher. On the interpretation of $\chi 2$ from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, pages 87–94, 1922.

[24] Bruno Zeitouni, Valentina Boeva, Isabelle Janoueix-Lerosey, Sophie Loeillet, Patricia Legoix-né, Alain Nicolas, Olivier Delattre, and Emmanuel Barillot. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, 26(15):1895–1896, August 2010.

[25] Ken Chen, John W. Wallis, Michael D. McLellan, David E. Larson, Joelle M. Kalicki, Craig S. Pohl, Sean D. McGrath, Michael C. Wendl, Qunyuan Zhang, Devin P. Locke, Xiaoqi Shi, Robert S. Fulton, Timothy J. Ley, Richard K. Wilson, Li Ding, and Elaine R. Mardis. BreakDancer: An algorithm for high resolution mapping of genomic structural variation. *Nature methods*, 6(9):677–681, September 2009.

[26] Tom M. Mitchell. *Machine Learning*. McGraw-Hill series in computer science. McGraw-Hill, New York, 1997.

[27] Walter W. Piegorsch. *Statistical data analytics: foundations for data mining, informatics, and knowledge discovery*. John Wiley & Sons Inc, Chichester, West Sussex, 2015.

[28] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[29] William H Wolberg and Olvi L Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the national academy of sciences*, 87(23):9193–9196, 1990.

[30] Olvi L Mangasarian, R Setiono, and WH Wolberg. Pattern recognition via linear programming: Theory and application to medical diagnosis. *Large-scale numerical optimization*, pages 22–31, 1990.

[31] Kristin P Bennett and Olvi L Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization methods and software*, 1(1):23–34, 1992.

[32] Olvi L Mangasarian, W Nick Street, and William H Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, 1995.

[33] Jianping Zhang. Selecting typical instances in instance-based learning. In *Proceedings of the Ninth International Machine Learning Conference*, pages 470–479, 1992.

[34] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[35] Stéphane Tuffery. *Data Mining et Statistique Decisionnelle: L'intelligence des données*. Technip, Paris, 2012. OCLC: 912492491.

[36] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982.

[37] Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Wiley series in probability and mathematical statistics. Wiley, Hoboken, N.J, 2005.

[38] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987.

[39] Chao Wang, Raghu Machiraju, and Kun Huang. Breast cancer patient stratification using a molecular regularized consensus clustering method. *Methods*, 67(3):304–312, June 2014.

[40] Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P. Butler, Carlos Caldas, Helen R. Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A. Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinsk, Natalie Jäger, David T. W. Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R. Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C. Munshi, Hiromi Nakamura, Paul A. Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V. Pearson, Xose S. Puente, Keiran Raine, Manasa Ramakrishna, Andrea L. Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N. Schumacher, Paul N. Span, Jon W. Teague, Yasushi Totoki, Andrew N. J. Tutt, Rafael Valdés-Mas, Marit M. van Buuren, Laura van 't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R.

Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, Icgc PedBrain, Jessica Zucman-Rossi, P. Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M. Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M. Pfister, Peter J. Campbell, and Michael R. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, August 2013.

[41] Loredana Martignetti, Laurence Calzone, Eric Bonnet, Emmanuel Barillot, and Andrei Zinovyev. ROMA: Representation and Quantification of Module Activity from Target Expression Data. *Frontiers in Genetics*, 7, February 2016.

[42] Motoo Kimura. The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations. *Genetics*, 61(4):893–903, April 1969.

[43] Fumio Tajima. Infinite-allele model and infinite-site model in population genetics. *Journal of Genetics*, 75(1):27, April 1996.

[44] V. Boeva, T. Popova, K. Bleakley, P. Chiche, J. Cappo, G. Schleiermacher, I. Janoueix-Lerosey, O. Delattre, and E. Barillot. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, 28(3):423–425, February 2012.

[45] Layla Oesper, Ahmad Mahmoody, and Benjamin J Raphael. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biology*, 14(7):R80, 2013.

[46] Gavin Ha, Andrew Roth, Jaswinder Khattra, Julie Ho, Damian Yap, Leah M. Prentice, Nataliya Melnyk, Andrew McPherson, Ali Bashashati, Emma Laks, Justina Biele, Jiarui Ding, Alan Le, Jamie Rosner, Karey Shumansky, Marco A. Marra, C. Blake Gilks, David G. Huntsman, Jessica N. McAlpine, Samuel Aparicio, and Sohrab P. Shah. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research*, 24(11):1881–1893, January 2014.

[47] F. Favero, T. Joshi, A. M. Marquard, N. J. Birkbak, M. Krzystanek, Q. Li, Z. Szallasi, and A. C. Eklund. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology*, 26(1):64–70, January 2015.

[48] Christopher A. Miller, Brian S. White, Nathan D. Dees, Malachi Griffith, John S. Welch, Obi L. Griffith, Ravi Vij, Michael H. Tomasson, Timothy A. Graubert, Matthew J. Walter, Matthew J. Ellis, William Schierding, John F. DiPersio, Timothy J. Ley, Elaine R. Mardis, Richard K. Wilson, and Li Ding. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Comput Biol*, 10(8):e1003665, August 2014.

[49] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah. PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398, April 2014.

[50] Iman Hajirasouliha, Ahmad Mahmoody, and Benjamin J. Raphael. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, 30(12):i78–i86, June 2014.

[51] Yi Qiao, Aaron R. Quinlan, Amir A. Jazaeri, Roeland GW Verhaak, David A. Wheeler, and Gabor T. Marth. SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. *Genome Biology*, 15(8):443, August 2014.

[52] Niko Beerenwinkel, Roland F. Schwarz, Moritz Gerstung, and Florian Markowetz. Cancer Evolution: Mathematical Models and Computational Inference. *Systematic Biology*, 64(1):e1–e25, January 2015.

[53] Wei Jiao, Shankar Vembu, Amit G Deshwar, Lincoln Stein, and Quaid Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, 15(1):35, 2014.

[54] Habil Zare, Junfeng Wang, Alex Hu, Kris Weber, Josh Smith, Debbie Nickerson, ChaoZhong Song, Daniela Witten, C. Anthony Blau, and William Stafford Noble. Inferring Clonal Composition from Multiple Sections of a Breast Cancer. *PLoS Comput Biol*, 10(7):e1003703, July 2014.

[55] Francesco Strino, Fabio Parisi, Mariann Micsinai, and Yuval Kluger. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Research*, 41(17):e165, September 2013.

[56] Andrej Fischer, Ignacio Vazquez-García, Christopher J. R. Illingworth, and Ville Mustonen. High-Definition Reconstruction of Clonal Composition in Cancer. *Cell Reports*, 7(5):1740–1752, June 2014.

[57] E. Purdom, C. Ho, C. S. Grasso, M. J. Quist, R. J. Cho, and P. Spellman. Methods and challenges in timing chromosomal abnormalities within cancer samples. *Bioinformatics*, 29(24):3113–3120, December 2013.

[58] Chris D. Greenman, Erin D. Pleasance, Scott Newman, Fengtang Yang, Beiyuan Fu, Serena Nik-Zainal, David Jones, King Wai Lau, Nigel Carter, Paul A.W. Edwards, P. Andrew Futreal, Michael R. Stratton, and Peter J. Campbell. Estimation of rearrangement phylogeny for cancer genomes. *Genome Research*, 22(2):346–361, February 2012.

[59] Roland F. Schwarz, Anne Trinh, Botond Sipos, James D. Brenton, Nick Goldman, and Florian Markowetz. Phylogenetic Quantification of Intra-tumour Heterogeneity. *PLoS Computational Biology*, 10(4):e1003535, April 2014.

[60] Eric Letouzé, Yves Allory, Marc A Bollet, François Radvanyi, and Frédéric Guyon. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome Biology*, 11(7):R76, 2010.

[61] Li Ding, Timothy J. Ley, David E. Larson, Christopher A. Miller, Daniel C. Koboldt, John S. Welch, Julie K. Ritchey, Margaret A. Young, Tamara Lamprecht, Michael D. McLellan, Joshua F. McMichael, John W. Wallis, Charles Lu, Dong Shen, Christopher C. Harris, David J. Dooling, Robert S. Fulton, Lucinda L. Fulton, Ken Chen, Heather Schmidt, Joelle Kalicki-Veizer, Vincent J. Magrini, Lisa Cook, Sean D. McGrath, Tammi L. Vickery, Michael C. Wendl, Sharon Heath, Mark A. Watson, Daniel C. Link, Michael H. Tomasson, William D. Shannon, Jacqueline E. Payton, Shashikant Kulkarni, Peter Westervelt, Matthew J. Walter, Timothy A. Graubert, Elaine R. Mardis, Richard K. Wilson, and John F. DiPersio. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510, January 2012.

[62] Peter Eirew, Adi Steif, Jaswinder Khattra, Gavin Ha, Damian Yap, Hossein Farahani, Karen Gelmon, Stephen Chia, Colin Mar, Adrian Wan, Emma Laks, Justina Biele, Karey Shumansky, Jamie Rosner, Andrew McPherson, Cydney Nielsen, Andrew J. L. Roth, Calvin Lefebvre, Ali Bashashati, Camila de Souza, Celia Siu, Radhouane Aniba, Jazmine Brimhall, Arusha Oloumi, Tomo Osako, Alejandra Bruna, Jose L. Sandoval, Teresa Algara, Wendy Greenwood, Kaston Leung, Hongwei Cheng, Hui Xue, Yuzhuo Wang, Dong Lin, Andrew J. Mungall, Richard Moore, Yongjun Zhao, Julie Lorette, Long Nguyen, David Huntsman, Connie J. Eaves, Carl Hansen, Marco A. Marra, Carlos Caldas, Sohrab P. Shah, and Samuel Aparicio. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518(7539):422–426, February 2015.

[63] Karin M. Hardiman, Peter J. Ulintz, Rork Kuick, Daniel H. Hovelson, Christopher M. Gates, Ashwini Bhasi, Ana Rodrigues Grant, Jianhua Liu, Andi K. Cani, Joel Greenson, Scott Tomlins, and Eric R. Fearon. Intratumor Genetic Heterogeneity in Rectal Cancer. *Laboratory investigation; a journal of technical methods and pathology*, 96(1):4–15, January 2016.

[64] Noemi Andor, Trevor A. Graham, Marnix Jansen, Li C. Xia, C. Athena Aktipis, Claudia Petritsch, Hanlee P. Ji, and Carlo C. Maley. Pan-cancer analysis of the extent and consequences of intra-tumor heterogeneity. *Nature medicine*, 22(1):105–113, January 2016.

[65] Philippe Lamy, Iver Nordentoft, Karin Birkenkamp-Demtröder, Mathilde Borg Houlberg Thomsen, Palle Villesen, Søren Vang, Jakob Hedegaard, Michael Borre, Jørgen Bjerggaard Jensen, Søren Høyer, Jakob Skou Pedersen, Torben F. Ørntoft, and Lars Dyrskjøt. Paired Exome Analysis Reveals Clonal Evolution and Potential Therapeutic Targets in Urothelial Carcinoma. *Cancer Research*, 76(19):5894–5906, October 2016.

[66] Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006.

[67] Thomas F. Eleveld, Derek A. Oldridge, Virginie Bernard, Jan Koster, Leo Colmet Daage, Sharon J. Diskin, Linda Schild, Nadia Bessoltane Bentahar, Angela Bellini, Mathieu Chicard, Eve Lapouble, Valérie Combaret, Patricia Legoix-Né, Jean Michon, Trevor J. Pugh, Lori S. Hart, JulieAnn Rader, Edward F. Attiyeh, Jun S. Wei, Shile Zhang, Arlene Naranjo, Julie M. Gastier-Foster, Michael D. Hogarty, Shahab Asgharzadeh, Malcolm A. Smith, Jaime M. Guidry Auvil, Thomas B. K. Watkins, Danny A. Zwijnenburg, Marli E. Ebus, Peter van Sluis, Anne Hakkert, Esther van Wezel, C. Ellen van der Schoot, Ellen M. Westerhout, Johannes H. Schulte, Godelieve A. Tytgat, M. Emmy M. Dolman, Isabelle Janoueix-Lerosey, Daniela S. Gerhard, Huib N. Caron, Olivier Delattre, Javed Khan, Rogier Versteeg, Gudrun Schleiermacher, Jan J. Molenaar, and John M. Maris. Relapsed neuroblastomas show frequent RAS-MAPK pathway mutations. *Nature Genetics*, 47(8):864–871, August 2015.

[68] James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, January 2011.

[69] Fabio Vandin, Patrick Clay, Eli Upfal, and Benjamin J. Raphael. Discovery of mutated subnetworks associated with clinical data in cancer. In *Biocomputing 2012*, pages 55–66. WORLD SCIENTIFIC, December 2011.

[70] Matan Hofree, John P. Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature Methods*, 10(11):1108–1115, November 2013.

[71] Marine Le Morvan, Andrei Zinovyev, and Jean-Philippe Vert. NetNorM: capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. December 2016.

[72] Paul Deveau, Emmanuel Barillot, Valentina Boeva, Andrei Zinovyev, and Eric Bonnet. Calculating biological module enrichment or depletion and visualizing data on large-scale molecular maps with ACSNMineR and RNaviCell packages. *The R Journal*, 8(2):293–306, December 2016.

[73] Paul D. Thomas, Michael J. Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Research*, 13(9):2129–2141, January 2003.

[74] Huaiyu Mi, Qing Dong, Anushya Muruganujan, Pascale Gaudet, Suzanna Lewis, and Paul D. Thomas. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Research*, 38(suppl 1):D204–D210, January 2010.

[75] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009.

[76] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, January 2009.

[77] Chava Kimchi-Sarfaty, Jung Mi Oh, In-Wha Kim, Zuben E. Sauna, Anna Maria Calcagno, Suresh V. Ambudkar, and Michael M. Gottesman. A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity. *Science*, 315(5811):525–528, January 2007.

[78] Gary Loughran, Ming-Yuan Chou, Ivaylo P. Ivanov, Irwin Jungreis, Manolis Kellis, Anmol M. Kiran, Pavel V. Baranov, and John F. Atkins. Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Research*, 42(14):8928–8938, August 2014.

[79] Tanja Kalstrup and Rikard Blunck. Reinitiation at non-canonical start codons leads to leak expression when incorporating unnatural amino acids. *Scientific Reports*, 5:11866, July 2015.

[80] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(8):1073–1081, June 2009.

[81] Pauline C. Ng and Steven Henikoff. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, July 2003.

[82] Ivan Adzhubei, Daniel M. Jordan, and Shamil R. Sunyaev. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, 0 7:Unit7.20, January 2013.

[83] Michal Levo and Eran Segal. In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics*, 15(7):453–468, July 2014.

[84] Ekta Khurana, Yao Fu, Vincenza Colonna, Xinmeng Jasmine Mu, Hyun Min Kang, Tuuli Lappalainen, Andrea Sboner, Lucas Lochovsky, Jieming Chen, Arif Harmanci, Jishnu Das, Alexej Abyzov, Suganthi Balasubramanian, Kathryn Beal, Dimple Chakravarty, Daniel Challis, Yuan Chen, Declan Clarke, Laura Clarke, Fiona Cunningham, Uday S. Evani, Paul Flicek, Robert Fragoza, Erik Garrison, Richard Gibbs, Zeynep H. Gümüş, Javier Herrero, Naoki Kitabayashi, Yong Kong, Kasper Lage, Vaja Liluashvili, Steven M. Lipkin, Daniel G. MacArthur, Gabor Marth, Donna Muzny, Tune H. Pers, Graham R. S. Ritchie, Jeffrey A. Rosenfeld, Cristina Sisu, Xiaomu Wei, Michael Wilson, Yali Xue, Fuli Yu, Emmanouil T. Dermitzakis, Haiyuan Yu, Mark A. Rubin, Chris Tyler-Smith, and Mark Gerstein. Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science (New York, N.Y.)*, 342(6154):1235587, October 2013.

[85] Yao Fu, Zhu Liu, Shaoke Lou, Jason Bedford, Xinmeng Jasmine Mu, Kevin Y. Yip, Ekta Khurana, and Mark Gerstein. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology*, 15(10), 2014.

[86] Fabio Vandin, Patrick Clay, Eli Upfal, and Benjamin J. Raphael. DISCOVERY OF MUTATED SUBNETWORKS ASSOCIATED WITH CLINICAL DATA IN CANCER. pages 55–66. WORLD SCIENTIFIC, December 2011.

[87] Fabio Vandin, Eli Upfal, and Benjamin J. Raphael. Algorithms for Detecting Significantly Mutated Pathways in Cancer. *Journal of Computational Biology*, 18(3):507–522, March 2011.

[88] W. Qiu, J. Wu, E. M. Walsh, Y. Zhang, C.-Y. Chen, J. Fujita, and Z.-Xj Xiao. Retinoblastoma protein modulates gankyrin–MDM2 in regulation of p53 stability and chemosensitivity in cancer cells. *Oncogene*, 27(29):4034–4043, March 2008.

[89] H. Wu and G. Lozano. NF-kappa B activation of p53. A potential mechanism for suppressing cell growth in response to stress. *Journal of Biological Chemistry*, 269(31):20067–20074, May 1994.

[90] Patricia M. Flatt, Luo Jia Tang, Caroline D. Scatena, Suzanne T. Szak, and Jennifer A. Pietenpol. p53 Regulation of G2 Checkpoint Is Retinoblastoma Protein Dependent. *Molecular and Cellular Biology*, 20(12):4210–4223, June 2000.

[91] Xin Liu, Ying Zhang, Man Tong, Xiu-ying Liu, Guan-zheng Luo, Dongfang Xie, Shao-fang Ren, Dong-hui Bai, Liu Wang, Qi Zhou, and Xiu-jie Wang. Identification of a small molecule 1,4-bis-[4-(3-phenoxy-propoxy)-but-2-ynyl]-piperazine as a novel inhibitor of the transcription factor p53. *Acta Pharmacologica Sinica*, 34(6):805–810, June 2013.

[92] Huaiyu Mi, Anushya Muruganujan, John T. Casagrande, and Paul D. Thomas. Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, 8(8):1551–1566, August 2013.

[93] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, page gkr988, 2011.

[94] Darryl Nishimura. BioCarta. *Biotech Software & Internet Report*, 2(3):117–120, June 2001.

[95] I Kuperstein, E Bonnet, HA Nguyen, D Cohen, E Viara, L Grieco, S Fourquet, L Calzone, C Russo, M Kondratova, et al. Atlas of cancer signalling network: a systems biology resource for integrative analysis of cancer data with google maps. *Oncogenesis*, 4(7):e160, 2015.

[96] Ronald A Fisher. *Statistical methods for research workers*. Oliver and Boyd, 1934.

[97] Isabelle Rivals, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, 2007.

[98] Anat Reiner, Daniel Yekutieli, and Yoav Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, 2003.

[99] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[100] Sohrab P. Shah, Andrew Roth, Rodrigo Goya, Arusha Oloumi, Gavin Ha, Yongjun Zhao, Gulisa Turashvili, Jiarui Ding, Kane Tse, Gholamreza Haffari, Ali Bashashati, Leah M. Prentice, Jaswinder Khattra, Angela Burleigh, Damian Yap, Virginie Bernard, Andrew McPherson, Karey Shumansky, Anamaria Crisan, Ryan Giuliany, Alireza Heravi-Moussavi, Jamie Rosner, Daniel Lai, Inanc Birol, Richard Varhol, Angela Tam, Noreen Dhalla, Thomas Zeng, Kevin Ma, Simon K. Chan, Malachi Griffith, Annie Moradian, S.-W. Grace Cheng, Gregg B. Morin, Peter Watson, Karen Gelmon, Stephen Chia, Suet-Feung Chin, Christina Curtis, Oscar M. Rueda, Paul D. Pharoah, Sambasivarao Damaraju, John Mackey, Kelly Hoon, Timothy Harkins, Vasisht Tadigotla, Mahvash Sigaroudinia, Philippe Gascard, Thea Tlsty, Joseph F. Costello, Irmtraud M. Meyer, Connie J. Eaves, Wyeth W. Wasserman, Steven Jones, David Huntsman, Martin Hirst, Carlos Caldas, Marco A. Marra, and Samuel Aparicio. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403):395–399, June 2012.

[101] Raymond Paternoster, Robert Brame, Paul Mazerolle, and Alex Piquero. USING THE CORRECT STATISTICAL TEST FOR THE EQUALITY OF REGRESSION COEFFICIENTS. *Criminology*, 36(4):859–866, November 1998.

[102] C. D. McFarland, K. S. Korolev, G. V. Kryukov, S. R. Sunyaev, and L. A. Mirny. Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences*, 110(8):2910–2915, February 2013.

[103] Guangbo Chen, Wahid A. Mulla, Andrei Kucharavy, Hung-Ji Tsai, Boris Rubinstein, Juliana Conkright, Scott McCroskey, William D. Bradford, Lauren Weems, Jeff S. Haug, Chris W. Seidel, Judith Berman, and Rong Li. Targeting the Adaptability of Heterogeneous Aneuploids. *Cell*, 160(4):771–784, February 2015.

[104] Marc A. Bollet, Nicolas Servant, Pierre Neuvial, Charles Decraene, Ingrid Lebigot, Jean-Philippe Meyniel, Yann De Rycke, Alexia Savignoni, Guillem Rigaill, Philippe Hupé, Alain Fourquet, Brigitte Sigal-Zafrani, Emmanuel Barillot, and Jean-Paul Thiery. High-Resolution Mapping of DNA Breakpoints to Define True Recurrences Among Ipsilateral Breast Cancers. *Journal of the National Cancer Institute*, 100(1):48–58, January 2008.

[105] Elza C. de Bruin, Nicholas McGranahan, Richard Mitter, Max Salm, David C. Wedge, Lucy Yates, Mariam Jamal-Hanjani, Seema Shafi, Nirupa Murugaesu, Andrew J. Rowan, Eva Grönroos, Madiha A. Muhammad, Stuart Horswell, Marco Gerlinger, Ignacio Varela, David Jones, John Marshall,

Thierry Voet, Peter Van Loo, Doris M. Rassl, Robert C. Rintoul, Sam M. Janes, Siow-Ming Lee, Martin Forster, Tanya Ahmad, David Lawrence, Mary Falzon, Arrigo Capitanio, Timothy T. Harkins, Clarence C. Lee, Warren Tom, Enock Teefe, Shann-Ching Chen, Sharmin Begum, Adam Rabinowitz, Benjamin Phillimore, Bradley Spencer-Dene, Gordon Stamp, Zoltan Szallasi, Nik Matthews, Aengus Stewart, Peter Campbell, and Charles Swanton. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*, 346(6206):251–256, October 2014.

[106] Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81, January 2015.

Chapter 9

Publications

9.1 ACSNMineR

ACSNMineR is an R package that can be used to compute statistical enrichment of mutations in a pathway. By default it relies on the maps from the Atlas of Cancer Signaling Network (ACSN), and was published together with RNaviCell in the R Journal in the December 2016 volume.

As of March, 15^{th} 2017, it has been downloaded 5075 on CRAN, which corresponds to 224 downloads a month on average.

Calculating Biological Module Enrichment or Depletion and Visualizing Data on Large-scale Molecular Maps with ACSNMineR and RNaviCell Packages

by Paul Deveau, Emmanuel Barillot, Valentina Boeva, Andrei Zinovyev and Eric Bonnet

Abstract Biological pathways or modules represent sets of interactions or functional relationships occurring at the molecular level in living cells. A large body of knowledge on pathways is organized in public databases such as the KEGG, Reactome, or in more specialized repositories, the Atlas of Cancer Signaling Network (ACSN) being an example. All these open biological databases facilitate analyses, improving our understanding of cellular systems. We hereby describe **ACSNMineR** for calculation of enrichment or depletion of lists of genes of interest in biological pathways. **ACSNMineR** integrates ACSN molecular pathways gene sets, but can use any gene set encoded as a GMT file, for instance sets of genes available in the Molecular Signatures Database (MSigDB). We also present **RNaviCell**, that can be used in conjunction with **ACSNMineR** to visualize different data types on web-based, interactive ACSN maps. We illustrate the functionalities of the two packages with biological data taken from large-scale cancer datasets.

Introduction

Biological pathways and networks comprise sets of interactions or functional relationships, occurring at the molecular level in living cells (Adriaens et al., 2008; Barillot et al., 2012). A large body of knowledge on cellular biochemistry is organized in publicly available repositories such as the KEGG database (Kanehisa et al., 2011), Reactome (Croft et al., 2014) and MINT (Zanzoni et al., 2002). All these biological databases facilitate a large spectrum of analyses, improving our understanding of cellular systems. For instance, it is a very common practice to cross the output of high-throughput experiments, such as mRNA or protein expression levels, with curated biological pathways in order to visualize the changes, analyze their impact on a network and formulate new hypotheses about biological processes. Many biologists and computational biologists establish list of genes of interest (e.g. a list of genes that are differentially expressed between two conditions, such as normal vs disease) and then evaluate if known biological pathways have significant overlap with this list of genes.

We have recently released the Atlas of Cancer Signaling Network (ACSN), a web-based database which describes signaling and regulatory molecular processes that occur in a healthy mammalian cell but that are frequently deregulated during cancerogenesis (Kuperstein et al., 2015). The ACSN atlas aims to be a comprehensive description of cancer-related mechanisms retrieved from the most recent literature. The web interface for ACSN is using the NaviCell technology, a software framework dedicated to web-based visualization and navigation for biological pathway maps (Kuperstein et al., 2013). This environment is providing an easy navigation of maps through the use of the Google Maps JavaScript library, a community interface with a web blog system, and a comprehensive module for visualization and analysis of high-throughput data (Bonnet et al., 2015).

In this article, we describe two packages related to ACSN analysis and data visualization. The package **ACSNMineR** is designed for the calculation of gene enrichment and depletion in ACSN maps (or any user-defined gene set via the import function), while **RNaviCell** is dedicated to data visualization on ACSN maps. Both packages are available on the Comprehensive R Archive Network (`https://cran.r-project.org/web/packages/ACSNMineR/` and `https://cran.r-project.org/web/packages/RNaviCell/`), and on the GitHub repository (`https://github.com/sysbio-curie/ACSNMineR` and `https://github.com/sysbio-curie/RNaviCell`). For the remainder of this article, we describe the organization of each package and illustrate their capacities with several concrete examples demonstrating their capabilities.

Packages organization

ACSNMineR

Currently, ACSN maps cover signaling pathways involved in DNA repair, cell cycle, cell survival, cell death, epithelial-to-mesenchymal transition (EMT) and cell motility. Each of these large-scale

molecular maps is decomposed in a number of functional modules. The maps themselves are merged into a global ACSN map. Thus the information included in ACSN is organized in three hierarchical levels: a global map, five individual maps, and several functional modules. Each ACSN map covers hundreds of molecular players, biochemical reactions and causal relationships between the molecular players and cellular phenotypes. ACSN represents a large-scale biochemical reaction network of 4,826 reactions involving 2,371 proteins (as of today), and is continuously updated and expanded. We have included the three hierarchical levels in the **ACSNMineR** package, in order to be able to calculate enrichments at all three levels. The calculations are made by counting the number of occurences of gene symbols (HUGO gene names) from a given list of genes of interest in all ACSN maps and modules. Table 1 is detailling the number of gene symbols contained in all the ACSN maps.

Table 1: ACSN maps included in the **ACSNMineR** package. Map: map name, Total: total number of gene symbols (HUGO) used to construct the map, Nb mod.: number of modules, Min: mimimum number of gene symbols in the modules, Max: maximum number of gene symbols in the modules, Mean: average number of gene sybols per module. N.B.: one gene symbol may be present in several modules of the map.

Map	Total	Nb mod.	Min	Max	Mean
ACSN global	2239	67	2	629	79
Survival	1053	5	208	431	328
Apoptosis	667	7	19	382	136
EMT & Cell motility	634	9	18	629	137
DNA repair	345	21	3	171	45
Cell cycle	250	25	2	130	20

The statistical significance of the counts in the modules is assessed by using either the Fisher exact test (Fisher, 1922, 1934) or the hypergeometric test, which are equivalent for this purpose (Rivals et al., 2007).

The current ACSN maps are included in the **ACSNMineR** package, as a list of character matrices.

```
> length(ACSN_maps)
[1] 6
> names(ACSN_maps)
[1] "Apoptosis"    "CellCycle"    "DNA_repair"   "EMT_motility" "Master"
[6] "Survival"
```

For each matrix, rows represent a module, with the name of the module in the first column, followed by a description of the module (optional), and then followed by all the gene symbols of the module. The maps will be updated according to every ACSN major release.

The main function of the **ACSNMineR** package is the `enrichment` function, which is calculating over-representation or depletion of genes in the ACSN maps and modules. We have included a small list of 12 Cell Cycle related genes in the package, named `genes_test` that can be used to test the main enrichment function and to get familiar with its different options.

```
> genes_test
 [1] "ATM"     "ATR"     "CHEK2"   "CREBBP"  "TFDP1"   "E2F1"    "EP300"
 [8] "HDAC1"   "KAT2B"   "GTF2H1"  "GTF2H2"  "GTF2H2B"
```

The example shown below is the simplest command that can be done to test a gene list for over-representation on the six included ACSN maps. With the list of 12 genes mentionned above and a default p-value cutoff of 0.05, we have a set of 8 maps or modules that are significantly enriched. The results are structured as a data frame with nine columns displaying the module name, the module size, the number of genes from the list in the module, the names of the genes that are present in the module, the size of the reference universe, the number of genes from the list that are present in the universe, the raw p-value, the p-value corrected for multiple testing and the type of test performed. The module field in the results data frame indicate the map name and the module name separated by a column character. If a complete map is significantly enriched or depleted, then only the map name is shown, without any module or column character. For instance, the third line of the results object below concern the E2F1 module of the CellCycle map.

```
> library(ACSNMineR)
> results <- enrichment(genes_test)
> dim(results)
```

```
[1] 8  9
> results[3,]
            module module_size nb_genes_in_module
V161 CellCycle:E2F1          19                 12
                                                                genes_in_module
V161 ATM ATR CHEK2 CREBBP TFDP1 E2F1 EP300 HDAC1 KAT2B GTF2H1 GTF2H2 GTF2H2B
     universe_size nb_genes_in_universe      p.value p.value.corrected    test
V161          2237                  12 3.735018e-21      2.353061e-19 greater
```

The `enrichment` function can take up to nine arguments: the gene list (as a character vector), the list of maps that will be used to calculate enrichment or depletion, the type of statistical test (either the Fisher exact test or the hypergeometric test), the module minimal size for which the calculations will be done, the universe, the p-value threshold, the alternative hypothesis ("greater" for calculating over-representation, "less" for depletion and "both" for both tests) and a list of genes that should be removed from the universe (option "Remove_from_universe"). This option may be useful for instance if we know beforehand that a number of genes are not expressed in the samples considered.

Only the gene list is mandatory to call the `enrichment` function, all the other arguments have default values. The `maps` argument can either be a dataframe imported from a GMT file with the `format_from_gmt` function or a list of dataframes generated by the same procedure. The GMT format corresponds to the Broad Institute's Gene Matrix Transposed file format, a convenient and easy way to encode named sets of genes of interest in tab-delimited text files (it is not a graph or network format). By default, the function `enrichment` uses the ACSN maps previously described. The correction for multiple testing is set by default to use the method of Benjamini & Hochberg, but can be changed to any of the usual correction methods (Bonferroni, Holm, Hochberg, Holm, or Benjamini & Yekutieli (Reiner et al., 2003)), or even disabled . The minimal module size represents the smallest size value of a module that will be used to compute enrichment or depletion. This is meant to remove results of low significance for module of small size. The universe in which the computation is made by default is defined by all the gene symbols contained in the maps. All the genes that were given as input and that are not present on the maps will be discarded. To keep all genes, the user can change the universe to `HUGO`, and in that case, the complete list of HUGO gene symbols will be used as the reference (> 39,000 genes). The threshold corresponds to the maximal value of the corrected p-value (unless the user chose not to correct for multiple testing) that will be displayed in the result table.

It may be of interest to compare enrichment of pathways in different cohorts or experiments. For example, enrichment of highly expressed pathways can reveal differences between two cancer types or two cell lines. To facilitate such comparisons, **ACSNMineR** provides a `multisample_enrichment` function. It relies on the `enrichment` function but takes a list of character vector genes. The name of each element of the list will be assumed to be the name of the sample for further analysis. Most of the arguments given to `multisample_enrichment` are the same as the ones passed to `enrichment`. However, the `cohort_threshold` is designed to filter out modules which would not pass the significance threshold in all samples.

Finally, to facilitate visualization of results, **ACSNMineR** integrates a representation function based on **ggplot2** syntax (Wickham, 2009). It allows representation of results from `enrichment` or `multisample_enrichment` with a limited number of parameters. Two types of display are available: heat-map tiles or bars. For multiple samples using a barplot representation, the number of rows used can be provided, otherwise all plots will be on the same row. For the heatmap, the color of the non-significant modules, and boundaries of the gradient for significant values can also be tuned.

We previously computed the p-value of the `genes_test` list with default parameters. The number of modules which have a p-value below 0.05 was 8, that can be compared to the 16 obtained without correction with the simple command shown below (some of the results are displayed in table 2).

```
enrichment(genes_test,correction_multitest = FALSE)
```

We can now plot the first six rows of the results obtained for corrected and uncorrected fisher test with heatmap format (Figure 1) or barplot (Figure 2) with the following commands:

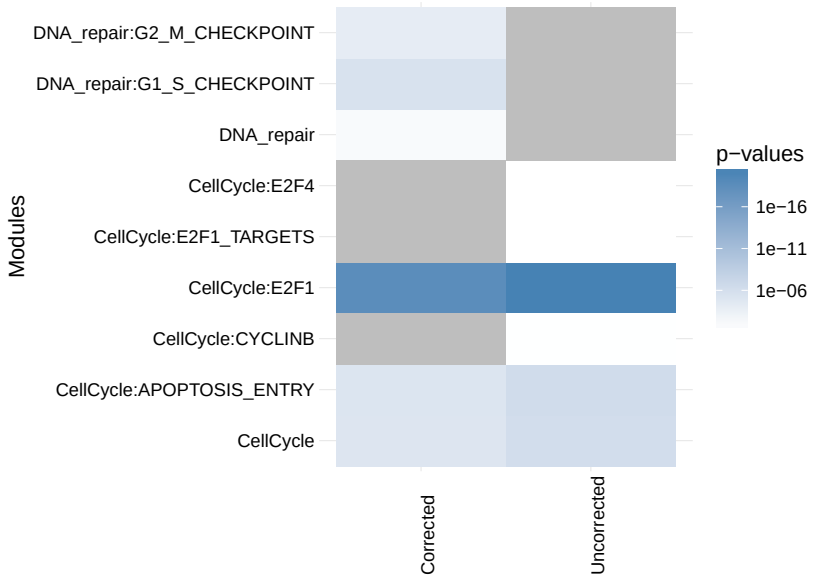
```
# heatmap

represent_enrichment(enrichment = list(Corrected = results[1:6,],
Uncorrected = results_uncorrected[1:6,]),
                                plot = "heatmap", scale = "reverselog",
                                low = "steelblue" , high ="white", na.value = "grey")

# barplot
```

Table 2: First rows of the results from enrichment analysis without correction. Module : name of the module. Mod. size: size of the module. Genes in module: genes from input which are found in the module. p-value: uncorrected p-value. Test : null hypothesis used, greater is synonym of enrichment.

Module	Mod. size	Genes in module	p-value	Test
CellCycle	242	ATM ATR CHEK2 CREBBP TFDP1 E2F1 EP300 HDAC1 KAT2B GTF2H1 GTF2H2 GTF2H2B	5.4×10^{-7}	greater
CellCycle:APOPTOSIS_ENTRY	10	ATM ATR CHEK2 E2F1	3.5×10^{-7}	greater
CellCycle:CYCLINB	7	ATM	0.04	greater

```
represent_enrichment(enrichment = list(Corrected = results[1:6,],
                            Uncorrected = results_uncorrected[1:6,]),
                              plot = "bar", scale = "reverselog",
                              nrow = 1)
```

Figure 1: Representation of the enriched modules (first six rows for each setting), with either Bonferroni correction or no correction. Grey tiles means that the data is not available for this module in this sample. P-values of low significance are in white, whereas p-values of high significance are represented in blue.

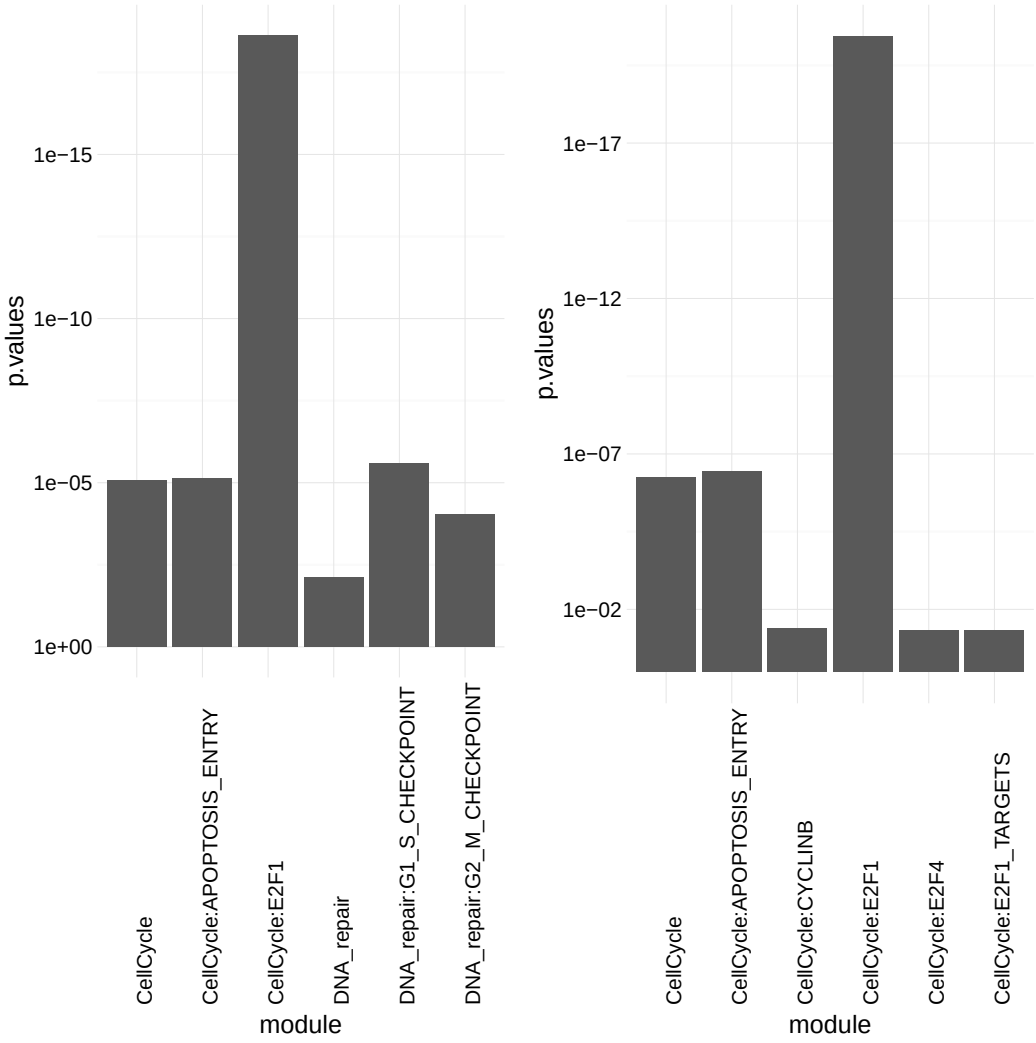


RNaviCell

The NaviCell Web Service provides a server mode, which allows automating visualization tasks and retrieving data from molecular maps via RESTful (standard http/https) calls. Bindings to different programming languages are provided in order to facilitate the development of data visualization workflows and third-party applications (Bonnet et al., 2015). RNaviCell is the R binding to the NaviCell Web Service. It is implemented as a standard R package, using the R object-oriented framework known as Reference Classes (Wickham, 2015). Most of the work done by the user using graphical point-and-click operations on the NaviCell web interface are encoded as functions in the library encapsulating http calls to the server with appropriate parameters and data. Calls to the NaviCell server are performed using the library RCurl (Lang and the CRAN team, 2015), while data encoding/decoding in JSON format is performed with the RJSONIO library (Lang, 2014).

Once the RNaviCell library is installed and loaded, the first step is to create a NaviCell object and launch the browser session. This will automatically create a unique session ID with the NaviCell server. Once the session is established, various functions can be called to send data to the web session, set graphical options, visualize data on a map or get data from the map. There are 125

Figure 2: Representation of the enriched modules (first six rows for each setting), with either Bonferroni correction (left) or no correction (right). The modules are on the X axis and the p-values are on the Y axis.

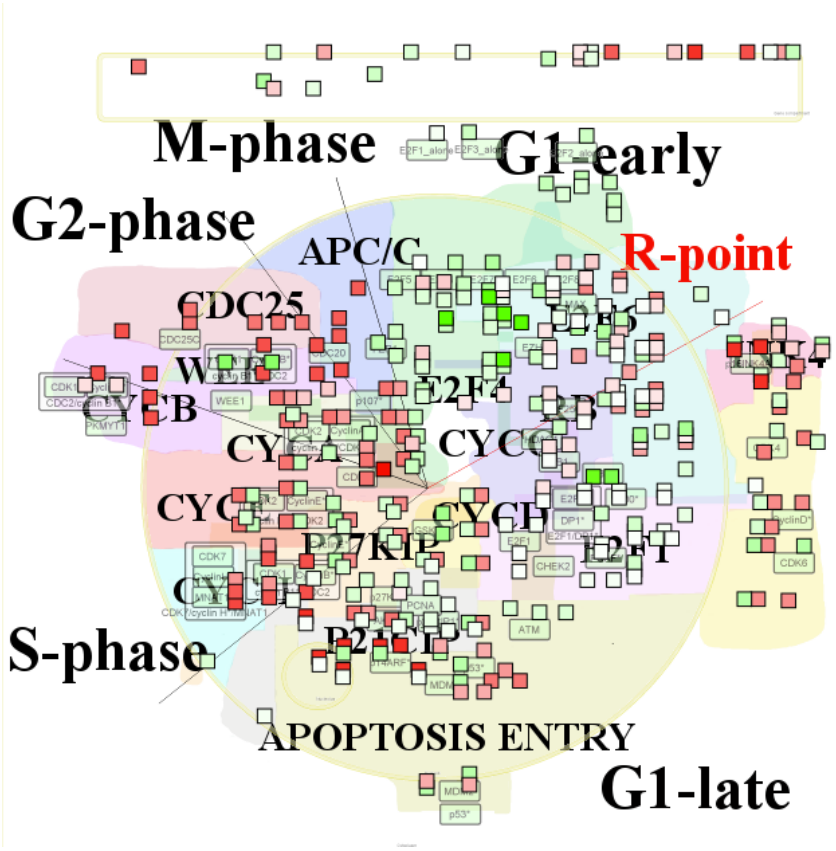


functions available in the current version of RNaviCell. All of them are described with their different options in the RNaviCell documentation, and we provide a tutorial on the GitHub repository wiki (https://github.com/sysbio-curie/RNaviCell/wiki/Tutorial).

In the simple example detailed below, we create a NaviCell session, then load an expression data set from a local (tab-delimited) file. The data represent gene expression measured in a prostate cancer cell line resistant to hormonal treatment (agressive), and is taken from the Cell Line Encyclopedia project (Barretina et al., 2012). We visualize the data values on the Cell Cycle map (the default map), using heat maps. With this visualization mode, gene expression values are represented as a color gradient (green to red) in squares positioned next to the entities where the gene has been mapped (Figure 3). Note that the map is displayed in a browser and is *interactive*, i.e. users can zoom in to display more information and for example look in what reactions are involved the genes selected to be displayed, and lots of other informations (see Bonnet et al. (2015) and Kuperstein et al. (2015) for more details).

```
# a short RNaviCell script example

# load RNaviCell library

library(RNaviCell)

# create a NaviCell object and launch a server session
```

Figure 3: Gene expression values from a prostate cancer cell line visualized on the cell cycle map as heat map plots. The figure is a screenshot of the NaviCell map browser, with the map set at the top (the less detailed) zoom level. The essential phases of the cell cycle are indicated on the map (G1/S/G2/M). Note that on the web browser the map is interactive and the user can zoom in and out, change the graphical parameters, import additional data and perform functional analysis.




```
# this will automatically open a browser on the client

navicell <- NaviCell()
navicell$launchBrowser()

# import a gene expression matrix and
# send the data to the NaviCell server
# NB: the data_matrix object is a regular R matrix

data_matrix <- navicell$readDatatable('DU145_data.txt')
navicell$importDatatable("mRNA expression data", "DU145", data_matrix)

# set data set and sample for heat map representation

navicell$heatmapEditorSelectSample('0','data')
navicell$heatmapEditorSelectDatatable('0','DU145')
navicell$heatmapEditorApply()
```

Case studies

Analysis of breast cancer expression data

In a study published in 2008, Schmidt and colleagues analyzed gene expression patterns of 200 breast cancer patients not treated by systemic therapy after surgery using discovery approach to reveal additional prognostic motifs (Schmidt et al., 2008). Estrogen receptor (ER) expression and proliferative

activity of breast carcinomas are well-known and described prognostic markers. Patients with ER-positive carcinomas have a better prognosis than those with ER-negative carcinomas, and rapidly proliferating carcinomas have an adverse prognosis. Knowledge about the molecular mechanisms involved in the processes of estrogen-dependent tumor growth and proliferative activity has led to the successful development of therapeutic concepts, such as antiendocrine and cytotoxic chemotherapy.

The dataset corresponding to this study is available as a Bioconductor package. The code shown below is creating a list of differentially expressed genes between ER positive and ER negative samples, and calculates the enrichment in ACSN maps from this list of genes. As seen in Table 3, there is one map (DNA repair) and seven modules (belonging to the Cell Cycle, DNA repair and Apoptosis maps) enriched.

```
# load all necessary packages
library(breastCancerMAINZ)
library(Biobase)
library(limma)
library(ACSNMineR)
library(hgu133a.db)
library(RNaviCell)

# load data and extract expression and phenotype data
data(mainz)
eset <- exprs(mainz)
pdat <- pData(mainz)

# Create list of genes differentially expressed between ER positive and
# ER negative samples using moderated t-test statistics
design <- model.matrix(~factor(pdat$er == '1'))
lmFit(eset, design) -> fit
eBayes(fit) -> ebayes
toptable(ebayes, coef=2,n=25000) -> tt
which(tt$adj < 0.05) -> selection
rownames(tt[selection,]) -> probe_list
mget(probe_list, env = hgu133aSYMBOL) -> symbol_list
symbol_list <- as.character(symbol_list)

# calculate enrichment in ACSN maps

enrichment(symbol_list) -> results

dim(results)
[1] 8 9
```

Table 3: ACSN maps enrichment for genes differentially expressed between ER positive and ER negative samples in breast cancer. Module : name of the map/module. Mod. size: size of the module. Nb genes: number of genes from input which are found in the module. pval: raw p-value. Cor. pval: corrected p-value.

Module	Mod. size	Nb genes	pval	Cor. pval
Apoptosis:AKT_MTOR	79	47	0.00043	0.0068
CellCycle:E2F2_TARGETS	35	22	0.0055	0.043
CellCycle:E2F3_TARGETS	51	31	0.0023	0.025
CellCycle:E2F4_TARGETS	100	60	5.8×10^{-5}	0.0037
DNA_repair	346	172	0.00038	0.0068
DNA_repair:CELL_CYCLE	82	49	0.00029	0.0068
DNA_repair:G1_CC_PHASE	25	18	0.0013	0.016
DNA_repair:S_CC_PHASE	46	28	0.0036	0.033

The Molecular Signatures Database (MSigDB) is one of the most widely used repository of well-annotated gene sets representing the universe of biological processes (Liberzon et al., 2011). We downloaded the canonical pathways set, counting more than 1,300 gene sets representing canonical pathways compiled by domain experts. The dataset is encoded with the GMT format, and can be imported within ACSNMineR with the `format_from_gmt` function. We calculate the enrichment for the

breast cancer differentially expressed gene list, simply specifying the MSigDB data we just imported as the maps option. Table 4 is displaying the pathways having a corrected p-value < 0.05. The prefix is indicating the database source, so we see that we have pathways from the KEGG, Reactome and PID databases. Consistent with our previous results, most of the enriched pathways are related to the cell cycle regulation.

```
# Import MSigDB canonical pathways and calculate enrichment on this database

mtsig <- format_from_gmt('c2.cp.v5.0.symbols.gmt')
enrichment(symbol_list, maps = mtsig)
```

Table 4: MSigDB canonical pathway database enrichment for genes differentially expressed between ER positive and ER negative samples in breast cancer. This table presents the 10 modules with lowest p-value out of 125 with corrected p-value lower than 0.05. Module : name of the module. Mod. size: size of the module. Nb genes: number of genes from input which are found in the module. Cor. pval: corrected p-value.

Pathway	Mod. size	Nb genes	Cor. pval
KEGG_CELL_CYCLE	128	76	3.9×10^{-8}
REACTOME_CELL_CYCLE_MITOTIC	325	159	3.9×10^{-8}
REACTOME_DNA_REPLICATION	192	98	4.9×10^{-6}
PID_FOXM1PATHWAY	40	29	3.1×10^{-5}
REACTOME_MITOTIC_M_M_G1_PHASES	172	87	3.1×10^{-5}
REACTOME_CELL_CYCLE	421	182	5×10^{-5}
REACTOME_MITOTIC_G1_G1_S_PHASES	137	71	9×10^{-5}
PID_AURORA_B_PATHWAY	39	27	0.0002
REACTOME_S_PHASE	109	58	0.00024
PID_SYNDECAN_1_PATHWAY	46	30	0.00026

At last, we visualize the mean expression values for ER negative samples for all genes differentially expressed on the ACSN master (global) map using RNaviCell commands to create heatmaps.

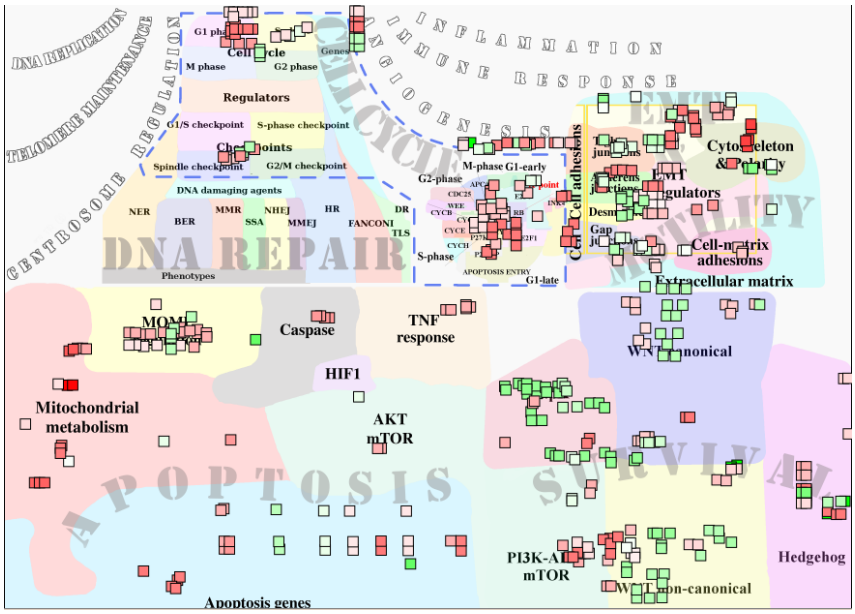
```
# Select ER negative samples and calculate mean expression values

apply(eset[probe_list,pdat$er == 0],1,mean) -> er_minus_mean
names(er_minus_mean) <- symbol_list
er_minus_mean <- as.matrix(er_minus_mean)
colnames(er_minus_mean) <- c('exp')

# create a NaviCell session, import the expression matrix on the map and create
# heatmaps to represent the data points.

navicell <- NaviCell()
navicell$proxy_url <- "https://acsn.curie.fr/cgi-bin/nv_proxy.php"
navicell$map_url <- "https://acsn.curie.fr/navicell/maps/acsn/master/index.php"

navicell$launchBrowser()
navicell$importDatatable("mRNA expression data", "GBM_exp", er_minus_mean)
navicell$heatmapEditorSelectSample('0','exp')
navicell$heatmapEditorSelectDatatable('0','GBM_exp')
navicell$heatmapEditorApply()
```

The Figure 4 is displaying the map for genes having a corrected p-value < 0.05.

Analysis of glioblastoma mutation frequencies

Recent years have witnessed a dramatic increase in new technologies for interrogating the activity levels of various cellular components on a genome-wide scale, including genomic, epigenomic, transcriptomic, and proteomic information (Hawkins et al., 2010). Integrating these heterogeneous datasets provides more biological insights than performing separate analyses. For instance, international consortia such as The Cancer Genome Atlas (TCGA) have launched large-scale initiatives to characterize

Figure 4: Mean expression values for ER negative differentially expressed genes in breast cancer visualized as heatmaps on the ACSN master map.



multiple types of cancer at different levels on hundreds of samples. These integrative studies have already led to the identification of novel cancer genes (McLendon et al., 2008).

Malignant gliomas, the most common subtype of primary brain tumors, are aggressive, highly invasive, and neurologically destructive tumors considered to be among the deadliest of human cancers. In its most aggressive manifestation, glioblastoma (GBM), median survival ranges from 9 to 12 months, despite maximum treatment efforts (Maher et al., 2001). In this study we have analyzed whole-genome mutation data generated by the TCGA project on hundreds of patients. More specifically, we parsed the MAF (Mutation Annotation Format) GBM files produced by different sequencing centers to count and calculate gene mutation frequencies. We kept the mutations having a status likely to disturb the target protein's function (i.e. Frame_Shift_Del, Nonstop_Mutation, In_Frame_Del, In_Frame_Ins, Missense_Mutation, Nonsense_Mutation, Splice_Site, Translation_Start_Site). In total, we collected mutations for more than 13,000 genes in a total of 379 mutated samples. In order to retain the most frequently mutated genes, we calculated frequencies across all mutated samples, and kept genes having a frequency greater than 1% (3,293 genes). We further labelled genes having a frequency greater than 1% and less than 5% as "1" and genes highly mutated (frequency higher than 5%) as "2".

We loaded the data as a matrix in R and calculated the enrichment in ACSN maps with the ACSNMineR function `enrichment`. The results are displayed in table 5. There are 6 modules significantly enriched in the DNA repair and EMT motility maps. Cell matrix adhesions and ECM (extra cellular matrix), part of the EMT motility map, are the modules with highest significance. The EMT motility map is significantly enriched at the global map level (second line in the table).

Table 5: ACSN maps enrichment for frequently mutated glioblastoma genes. Module : name of the module. Mod. size: size of the module. Nb genes: number of genes from input which are found in the module. Cor. pval: corrected p-value.

module	Mod. size	Nb genes	Cor. pval
DNA_repair:S_PHASE_CHECKPOINT	45	19	0.008
EMT_motility	635	181	0.0002
EMT_motility:CELL_MATRIX_ADHESIONS	73	45	3.73e-12
EMT_motility:CYTOSKELETON_POLARITY	154	47	0.022
EMT_motility:DESMOSOMES	29	15	0.002
EMT_motility:ECM	147	69	9.77e-11
EMT_motility:EMT_REGULATORS	629	178	0.0002

Visualization of the list of glioblastoma mutated genes is shown on figure 5. This figure was generated with the ACSNMineR commands detailed below. Results of the enrichment test correlate well with the visualization on the map, with a high density of low and high frequency mutated genes

in the EMT motility and DNA repair regions (maps) of the global ACSN map. Although they are not statistically significant, quite high densities can also be seen in other regions of the map.

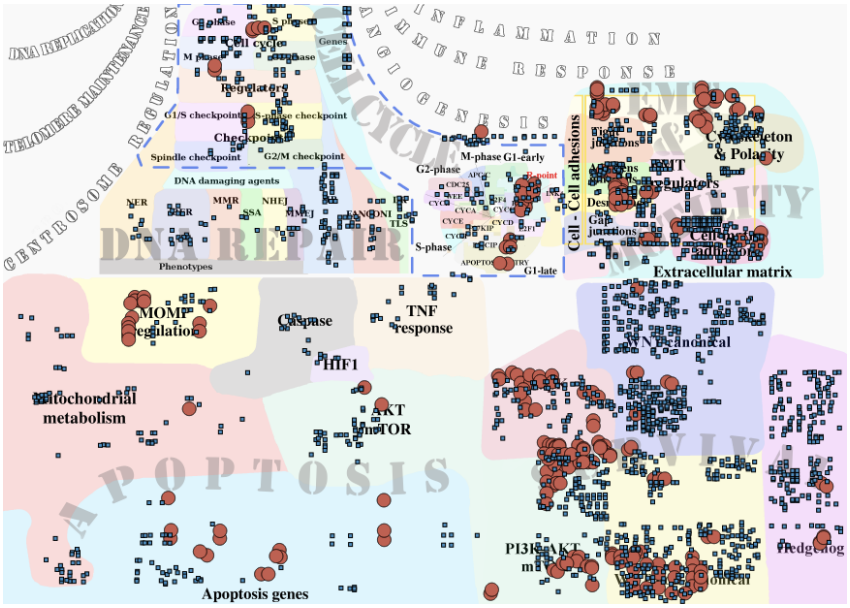
```
library(RNaviCell)

# Create a NaviCell object, point it to the ACSN master map and launch
# a session.

navicell <- NaviCell()
navicell$proxy_url <- "https://acsn.curie.fr/cgi-bin/nv_proxy.php"
navicell$map_url <- "https://acsn.curie.fr/navicell/maps/acsn/master/index.php"
navicell$launchBrowser()

# Read the GBM data file and import it into the session.

mat <- navicell$readDatatable('gbm.txt')
navicell$importDatatable("Mutation data", "GBM", mat)

# set datatable and sample names for the glyph editor

navicell$drawingConfigSelectGlyph(1, TRUE)
navicell$glyphEditorSelectSample(1, "categ")
navicell$glyphEditorSelectShapeDatatable(1, "GBM")
navicell$glyphEditorSelectColorDatatable(1, "GBM")
navicell$glyphEditorSelectSizeDatatable(1, "GBM")
navicell$glyphEditorApply(1)

# set color, shape and size parameters for glyphs

navicell$unorderedConfigSetDiscreteShape("GBM", "sample", 0, 1)
navicell$unorderedConfigSetDiscreteShape("GBM", "sample", 1, 5)
navicell$unorderedConfigApply("GBM", "shape")

navicell$unorderedConfigSetDiscreteColor("GBM", "sample", 0, "398BC3")
navicell$unorderedConfigSetDiscreteColor("GBM", "sample", 1, "CC5746")
navicell$unorderedConfigApply("GBM", "color")

navicell$unorderedConfigSetDiscreteSize("GBM", "sample", 0, 4)
navicell$unorderedConfigSetDiscreteSize("GBM", "sample", 1, 14)

navicell$unorderedConfigApply("GBM", "size")
```

Summary and perspectives

In this work, we presented the R package **ACSNMineR**, a novel package for the calculation of p-values for enrichment or depletion of genes in biological pathways. The package includes the six large-scale molecular maps and 67 functional modules of the Atlas of Cancer Signaling Network (ACSN) . Enrichment can be calculated for those maps and modules with several options to play with, but can also be calculated for other databases of molecular pathways, that can be imported from GMT formated files.

We also describe in this work the **RNaviCell** package, a R package convenient to use with **ACSNMineR**. This package is dedicated to create web-based and interactive data visualization on ACSN maps. Users can use this tools to represent genes of interest that have been shown to be related to the maps by calculating enrichment with **ACSNMineR**. Creating maps with the graphical user interface of the ACSN website can be a tedious task if the user has multiple samples or gene lists, and wants to compare their representations on ACSN maps. The **RNaviCell** package can be used to automate the process of creating the graphical representations automatically. The maps are displayed in a browser and are interactive, with the possibility for the user to zoom in and out, search for genes or molecular species, and see the details of the molecular reactions (what partners are involved, what is the state of a given species, etc.). For more details on how to use the interface and the different possibilities, see Kuperstein et al. (2013), Bonnet et al. (2015) and Kuperstein et al. (2015). We have shown how the packages **ACSNMineR** and **RNaviCell** can be combined to analyze expression data from breast cancer samples, and also to analyze the frequency of mutated genes in glioblastoma cancer samples.

Figure 5: Glioblastoma gene mutation frequency categories represented as glyphs on the ACSN global cancer map. High frequency mutated genes are pictured as large red circles, while low frequency mutated genes are depicted as small blue squares.



Of course, **ACSNMineR** is not the only R package for enrichment calculations. For instance, **GOstats** (Falcon and Gentleman, 2007) is probably one of the first packages that was created to calculate enrichment for Gene Ontology categories. **GOstats** can also be used to calculate enrichment for other biological pathways categories, such as KEGG pathways (by using an instance of the class `KEGGHyperGParams`) or PFAM protein families (using `PFAMHyperGParams`). However, its usage might not be as straightforward as **ACSNMineR**, and it does not seem possible to test user-defined biological pathways. Furthermore, other authors have pointed out that the KEGG database used by this package has not been updated since 2012. **clusterProfiler** is a recent R package released for enrichment analysis of Gene Ontology and KEGG with either hypergeometric test or Gene Set Enrichment Analysis (GSEA) (Yu et al., 2012). Via other packages, support for analysis of Disease Ontology and Reactome Pathways is possible. Interestingly, this package also offers the possibility to import user-defined gene set, through tab-delimited pairwise definition files. Other notable packages for enrichment calculations are **ReactomePA** for Reactome molecular pathways (Yu and He, 2016), **miRNApath** for microRNA pathways (Cogswell et al., 2008) and **gage** (Luo et al., 2009). We believe that the main advantages of ACSNMineR compared to other packages are a direct access to the full set of ACSN maps (updated on a regular basis) and an easy way to test MSigDB gene sets or any user-defined gene set formatted appropriately.

In order to improve **ACSNMineR**, we may in the near future try to improve the speed of calculation, which might be a problem if a very large number of samples or experiments have to be analyzed rapidly. For instance, we could use the `foreach` and `%dopar%` operator to parallelize the most computationally demanding operations. It could also be useful to implement more sensitive methods of gene set enrichment measures, such as the Gene Set Enrichment Analysis (GSEA) method.

RNaviCell relies on standard HTTP calls to provide informations and calculations, and we have developped a number of bindings for popular programming languages such as R, Java and Python (Bonnet et al., 2015). This open architecture is designed to facilitate the development of utilities by other programmers and to facilitate the integration of ACSN maps in existing frameworks. The development of such services, sometimes called “microservices" (Fowler, 2014) is in expansion. Furthermore, this kind of open architecture could clear the way for a more unified and general access to reaction networks database, including for example WikiPathways (Kelder et al., 2012), Reactome (Croft et al., 2014) and other databases. The PSICQUIC project is a successfull example of such an architecture (Aranda et al., 2011). It is an effort of the HUPO proteomics standard initiative to standardize the access to molecular interaction databases programmatically, based on the specification of web services (using REST and SOAP calls) and a common query language (MIQL).

Acknowledgements

This work was supported by a grant “Projet Incitatif et Collaboratif Computational Systems Biology Approach for Cancer" from Institut Curie. The authors would like to thank Pierre Gestraud for his comments on early versions of the **ACSNMineR** package and Eric Viara for guidance and assistance on the development of the **RNaviCell** package.

Bibliography

M. E. Adriaens, M. Jaillard, A. Waagmeester, S. L. Coort, A. R. Pico, and C. T. Evelo. The public road to high-quality curated biological pathways. *Drug Discovery Today*, 13(19):856–862, 2008. [p293]

B. Aranda, H. Blankenburg, S. Kerrien, F. S. Brinkman, A. Ceol, E. Chautard, J. M. Dana, J. De Las Rivas, M. Dumousseau, E. Galeota, et al. Psicquic and psiscore: accessing and scoring molecular interactions. *Nature Methods*, 8(7):528–529, 2011. [p303]

E. Barillot, L. Calzone, P. Hupe, J.-P. Vert, and A. Zinovyev. *Computational systems biology of cancer*. CRC Press, 2012. [p293]

J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012. [p297]

E. Bonnet, E. Viara, I. Kuperstein, L. Calzone, D. P. Cohen, E. Barillot, and A. Zinovyev. Navicell web service for network-based data visualization. *Nucleic Acids Research*, page gkv450, 2015. [p293, 296, 297, 302, 303]

J. P. Cogswell, J. M. Ward, I. A. Taylor, M. Waters, Y. Shi, B. Cannon, K. Kelnar, J. Kemppainen, D. Brown, C. Chen, R. K. Prinjha, J. C. Richardson, A. M. Saunders, A. D. Roses, and C. A. Richards. Identification of mirna changes in alzheimer’s disease brain and csf yields putative biomarkers and insights into disease pathways. *Journal of Alzheimer’s Disease*, 14:27–41, 2008. [p303]

D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, et al. The reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1):D472–D477, 2014. [p293, 303]

S. Falcon and R. Gentleman. Using gostats to test gene lists for go term association. *Bioinformatics*, 23 (2):257–258, 2007. [p303]

R. A. Fisher. On the interpretation of $\chi 2$ from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, pages 87–94, 1922. [p294]

R. A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, 1934. ISBN 0-05-002170-2. [p294]

M. Fowler. *Microservices*, 2014. URL http://martinfowler.com/articles/microservices.html. Microservices. [p303]

R. D. Hawkins, G. C. Hon, and B. Ren. Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, 11(7):476–486, 2010. [p300]

M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, page gkr988, 2011. [p293]

T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, and A. R. Pico. Wikipathways: building research communities on biological pathways. *Nucleic Acids Research*, 40 (D1):D1301–D1307, 2012. [p303]

I. Kuperstein, D. P. Cohen, S. Pook, E. Viara, L. Calzone, E. Barillot, and A. Zinovyev. Navicell: a web-based environment for navigation, curation and maintenance of large molecular interaction maps. *BMC systems biology*, 7(1):100, 2013. [p293, 302]

I. Kuperstein, E. Bonnet, H. Nguyen, D. Cohen, E. Viara, L. Grieco, S. Fourquet, L. Calzone, C. Russo, M. Kondratova, et al. Atlas of cancer signalling network: a systems biology resource for integrative analysis of cancer data with google maps. *Oncogenesis*, 4(7):e160, 2015. [p293, 297, 302]

D. T. Lang. *RJSONIO: Serialize R objects to JSON, JavaScript Object Notation*, 2014. URL http://CRAN.R-project.org/package=RJSONIO. R package version 1.3-0. [p296]

D. T. Lang and the CRAN team. *RCurl: General Network (HTTP/FTP/...) Client Interface for R*, 2015. URL http://CRAN.R-project.org/package=RCurl. R package version 1.95-4.7. [p296]

A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011. [p299]

Luo, Weijun, Friedman, Michael, Shedden, Kerby, Hankenson, Kurt, Woolf, and Peter. Gage: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10:161, 2009. [p303]

E. A. Maher, F. B. Furnari, R. M. Bachoo, D. H. Rowitch, D. N. Louis, W. K. Cavenee, and R. A. DePinho. Malignant glioma: genetics and biology of a grave matter. *Genes & development*, 15(11):1311–1333, 2001. [p301]

R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, G. M. Mastrogianakis, J. J. Olson, T. Mikkelsen, N. Lehman, K. Aldape, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008. [p301]

A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, 2003. [p295]

I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier. Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, 2007. [p294]

M. Schmidt, D. Böhm, C. von Törne, E. Steiner, A. Puhl, H. Pilch, H.-A. Lehr, J. G. Hengstler, H. Kölbl, and M. Gehrmann. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer research*, 68(13):5405–5413, 2008. [p298]

H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN 978-0-387-98140-6. URL http://had.co.nz/ggplot2/book. [p295]

H. Wickham. *Reference classes*, 2015. URL http://adv-r.had.co.nz/R5.html. Advanced R. [p296]

G. Yu and Q.-Y. He. Reactomepa: an r/bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, page accepted, 2016. doi: 10.1039/C5MB00663E. URL http://pubs.rsc.org/en/Content/ArticleLanding/2015/MB/C5MB00663E. [p303]

G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287, 2012. [p303]

A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni. Mint: a molecular interaction database. *FEBS Letters*, 513(1):135–140, 2002. [p293]

Paul Deveau
Computational Systems Biology of Cancer, Institut Curie
Genetics and Biology of Cancers, Institut Curie
26, rue d'Ulm 75248 Paris
France
paul.deveau@curie.fr

Emmanuel Barillot
Computational Systems Biology of Cancer, Institut Curie
26, rue d'Ulm 75248 Paris
France
emmanuel.barillot@curie.fr

Valentina Boeva
Institut Cochin, Inserm U1016, CNRS UMR 8104, Université Paris Descartes UMR-S1016
22, Rue Mechain 75014 Paris
France
valentina.boeva@curie.fr

Andrei Zinovyev
Computational Systems Biology of Cancer, Institut Curie
26, rue d'Ulm 75248 Paris
France
andrei.zinovyev@curie.fr

Eric Bonnet
Centre National de Génotypage, Institut de Génomique, CEA
2, rue Gaston Crémieux, 91057 Evry
France
eric.bonnet@cng.fr

9.2 QuantumClone

QuantumClone is an R package to reconstruct and visualize clonal populations from HTS data.

As of March, 15^{th} 2017, it has been downloaded 4164 times, which corresponds to 221 downloads a month on average.

Clonal assessment of functional variants in cancer based on a genotype-aware method for clonal reconstruction

Paul Deveau[1-3], Leo Colmet Daage[2], Derek Oldridge[4-6], Virginie Bernard[7], Angela Bellini[2], Mathieu Chicard[2], Nathalie Clement[2], Eve Lapouble[8], Valérie Combaret[9], Anne Boland[10], Vincent Meyer[10], Jean-François Deleuze[10], Isabelle Janoueix-Lerosey[11], Emmanuel Barillot[1], Olivier Delattre[11], John Maris[4-6], Gudrun Schleiermacher[2,12,†,*] and Valentina Boeva[1,13,†,*]

[1]Institut Curie, PSL Research University, Mines Paris Tech, INSERM U900, 75005, Paris, France

[2]Institut Curie, PSL Research University, INSERM U830, Laboratoire RTOP (Recherche Translationelle en Oncologie Pédiatrique), Département de recherche translationnelle, 75005, Paris, France

[3]Univ. Paris-Sud, Orsay, France

[4]Division of Oncology, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA

[5]Center for Childhood Cancer Research, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA

[6]Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, USA

[7]Institut Curie, PSL Research University, ICGex, 75005, Paris, France

[8]Institut Curie, PSL Research University, Unité de Génétique Somatique, 75005, Paris, France

[9]Centre Léon-Bérard, Laboratoire de Recherche Translationnelle Lyon, France

[10]Centre National de Génotypage, Institut de Génomique, CEA, Evry, 91057, France.

[11]Institut Curie, PSL Research University, INSERM U830, Paris, 75005, France

[12]Institut Curie, PSL Research University, Département de Pédiatrie, Paris, 75005, France

[13]Institut Cochin, INSERM U1016, CNRS UMR 8104, Université Paris Descartes UMR-S1016, 75014, Paris, France

[†]These authors jointly supervised this work.

[*]Correspondance should be addressed to valentina.boeva@inserm.fr or gudrun.schleiermacher@curie.fr

Running title: Framework for clonal reconstruction in cancer

Keywords: Clonal inference, Cancer, Algorithms, Neuroblastoma, Whole genome sequencing

Abstract

In cancer, clonal evolution is assessed based on information coming from single nucleotide variants and copy number alterations. Nonetheless, previous methods often fail to accurately combine information from both sources to truthfully reconstruct clonal populations in a given tumor sample or in a set of tumor samples coming from the same patient. Moreover, existing methods detect clones from a single set of variants. As a result, compromises have to be done between stringent variant filtering (reducing VAF dispersion) and using all biologically relevant variants. Here, we present a framework for defining cancer clones using most reliable variants of high depth of coverage and assigning functional mutations to the detected clones. The key element of our framework is QuantumClone, a method for variant clustering into clones based on VAFs, genotypes of corresponding regions and information about tumor purity. We validated QuantumClone and our framework on simulated data. We then applied our framework to whole genome sequencing data for 19 neuroblastoma trios each including constitutional, diagnosis and relapse samples. In this cohort, we confirmed an enrichment of damaging variants within such pathways as MAPK, neuritogenesis, epithelial-mesenchymal transition, cell survival and DNA repair. Most pathways had more damaging variants in the expanding clones compared to shrinking ones, which can be explained by the increased total number of variants between these two populations. Functional mutational rate varied for ancestral clones and clones shrinking or expanding upon treatment, suggesting changes in clone selection mechanisms at different time points of tumor evolution.

Introduction

The principal cause of cancer is believed to be the accumulation of somatic variants and structural variations (SVs) in the genome. Recently, many efforts have focused on the identification of driver mutations; nonetheless, passenger variants, although they are not directly linked to the disease, may provide additional evidence from which to infer the phylogeny of a tumor and so help uncover the basis for its proliferative activity (Marusyk et al., 2014). Indeed high confidence passenger set of variants shared by a clonal population should be observed at the same cellular prevalence at any given point in time, allowing statistical models to cluster variants together and define a clone.

To understand the role that driver mutations play in clonal expansion and cancer progression, it is essential to accurately reconstruct the clonal structure and assign functional variants to it. We define a clone as a cell population that harbors a unique pattern of mutations and SVs. Such clones are related to each other and share a common ancestor. A hierarchical phylogenetic tree, which represents the ancestry of clones, can be constructed to reflect the order of appearance of new sets of mutations defining each clone. Each such set of mutations is expected to contain at least one driver mutation or SV giving a selective advantage to the clone compared to its ancestry. A clone can thus have a different behavior from its ancestral clone when facing the same stimuli. With accumulation of driver mutations, clones are likely to gain hallmarks of cancer such as evading growth suppressors and activating invasion and metastasis (Hanahan and Weinberg, 2011).

High-Throughput Sequencing (HTS) of bulk tumor tissues has allowed uncovering genetic differences at the clonal level in primary and relapse/metastatic tumors. Modern computational methods provide ways to reconstruct the structure of the phylogenetic tree from variant allele frequencies (VAFs) in sequenced reads, where VAF is a proportion of reads supporting each given variant among all reads spanning the position of interest (Fischer et al., 2014; Jiao et al., 2014; Kepler, 2013; Malikic et al., 2015; Miller et al., 2014; Qiao et al., 2014; Schwarz et al., 2014). However, existing methods for clonal reconstruction often neglect information about the genotype of each position, which refers to the paternal or maternal inheritance of a locus and the number of copies of each allele. Accounting for the genotype information is especially crucial in the case of hyper-diploid cancers and cancers with highly rearranged genomes, as the cellular prevalence – measured as the proportion of cancer cells carrying a variant – is linked to VAF through such parameters as copy number of the locus and the number of chromosome bearing the mutation. Computationally, we can detect different clones based on the clustering of VAF values (Miller et al., 2014; Roth et al., 2014; Qiao et al., 2014). However, identifying the correct hierarchical tree is a complex task, and this problem often does not have a

unique solution. Therefore, in this paper, clones and variant clusters are considered as synonyms.

Here we show that by combining the genotype and VAF information it is possible to correctly cluster variants and assign them to specific clones, thus reconstructing the clonal architecture of an individual cancer. This may be done with our novel method, QuantumClone, designed to reconstruct clones based on both VAF and genotype information; so we call it "genotype-aware". We demonstrate that our algorithm accurately clusters variants on simulated data, even when cancer is hyper-diploid or contaminated by normal cells. We also propose a general framework based on QuantumClone to detect driver mutations of clonal evolution. This general approach is applied to 19 neuroblastoma cases; each case includes whole genome sequencing (WGS) data from a sample at diagnosis and relapse. We show that mutations possibly affecting the expression level or the structure of the protein (here called damaging or deleterious) in neuroblastoma accumulate at relapse in specific pathways such as cell motility (e.g., cell-matrix adhesion and regulation of epithelial–mesenchymal transition, EMT) and cell survival (e.g., PI3K/AKT/mTOR, MAPK or noncanonical Wnt pathways).

Results

The QuantumClone method presented here applies an expectation-maximization (EM) algorithm and allows for accurate inference of clonal structure using VAFs from one or several tumor samples sequenced using WGS. It can analyze variants coming from highly rearranged and hyper-diploid cancer genomes. We extensively validated QuantumClone on simulated data, where we compared it with recently published methods (Miller et al., 2014; Roth et al., 2014). We complement QuantumClone with a robust framework for the functional assessment of mutations based on signaling pathway analysis combined with the assignment of functional variants to the reconstructed clones.

This framework was applied to WGS neuroblastoma datasets: 19 patients' primary and relapse samples including 7 new triplets. Novel and previously published samples (Eleveld et al., 2015) have an average sequencing depth of $\sim 100\times$ (Methods). Application of the QuantumClone-based framework allowed us to discover pathways recurrently altered by mutations in neuroblastoma at diagnosis and relapse.

Assessment of clonal reconstruction accuracy of QuantumClone

For clonal reconstruction using VAFs, we developed an approach that applies an EM algorithm (Methods). QuantumClone utilizes genotype information and assigns variants to clones providing the most likely values of cellular prevalence (Methods).

Comparison of QuantumClone with existing methods

Using *in silico* data, we compared the performance of QuantumClone, sciClone (Miller et al., 2014) and pyClone (Roth et al., 2014) in order to infer the clonal structure of a set of tumors derived from the same patient. sciClone is based on variational Bayesian Mixture Models, while pyClone relies on a hierarchical Bayes statistical model. Similarly to QuantumClone, pyClone leverages copy number information to better infer clonal architecture.

We generated a phylogenetic tree for each simulated tumor, which was used to compute observed alternative allele read count given the cell fraction of the clone, the ploidy, and the depth of coverage at this position (Methods). In our simulation experiments, the following parameters varied within realistic ranges: depth of sequencing (100× to 1000×), fraction of contamination by normal cells (from 0 to 70%), number of variants used for the clonal reconstruction (from 50 to 200), number of tumor samples used for each patient (from 1 to 5) and number of distinct clones per cancer type (from 2 to 10) (Fig. 1).

For each set of parameters, we performed and analyzed 50 independent simulation experiments (Methods). The accuracy of clonal reconstruction was assessed by evaluating the normalized mutual information (NMI) (Manning et al., 2008) and the average error in distance between the estimated cellularity of a clone and its theoretic value. Perfect variant clustering would result in a $L2$ (or Euclidean distance) mean error of 0, and a NMI value of 1, which would correspond to an identification of the exact number of clones and correct assignment of all the variants of a clone to the same cluster.

Our analysis showed that QuantumClone is equivalent to or better than the best published algorithm in clustering quality (Fig. 1A) for diploid genomes. In terms of NMI QuantumClone showed similar performances compared to pyClone. However, QuantumClone generally outcompeted sciClone for NMI ($p-value < 2.2 \times 10^{-16}$, one-sided Welch two-sample t-test). In samples with 50% contamination by normal cells QuantumClone drastically outperformed sciClone ($p-value = 3.6 \times 10^{-10}$ Welch one-sided two-sample

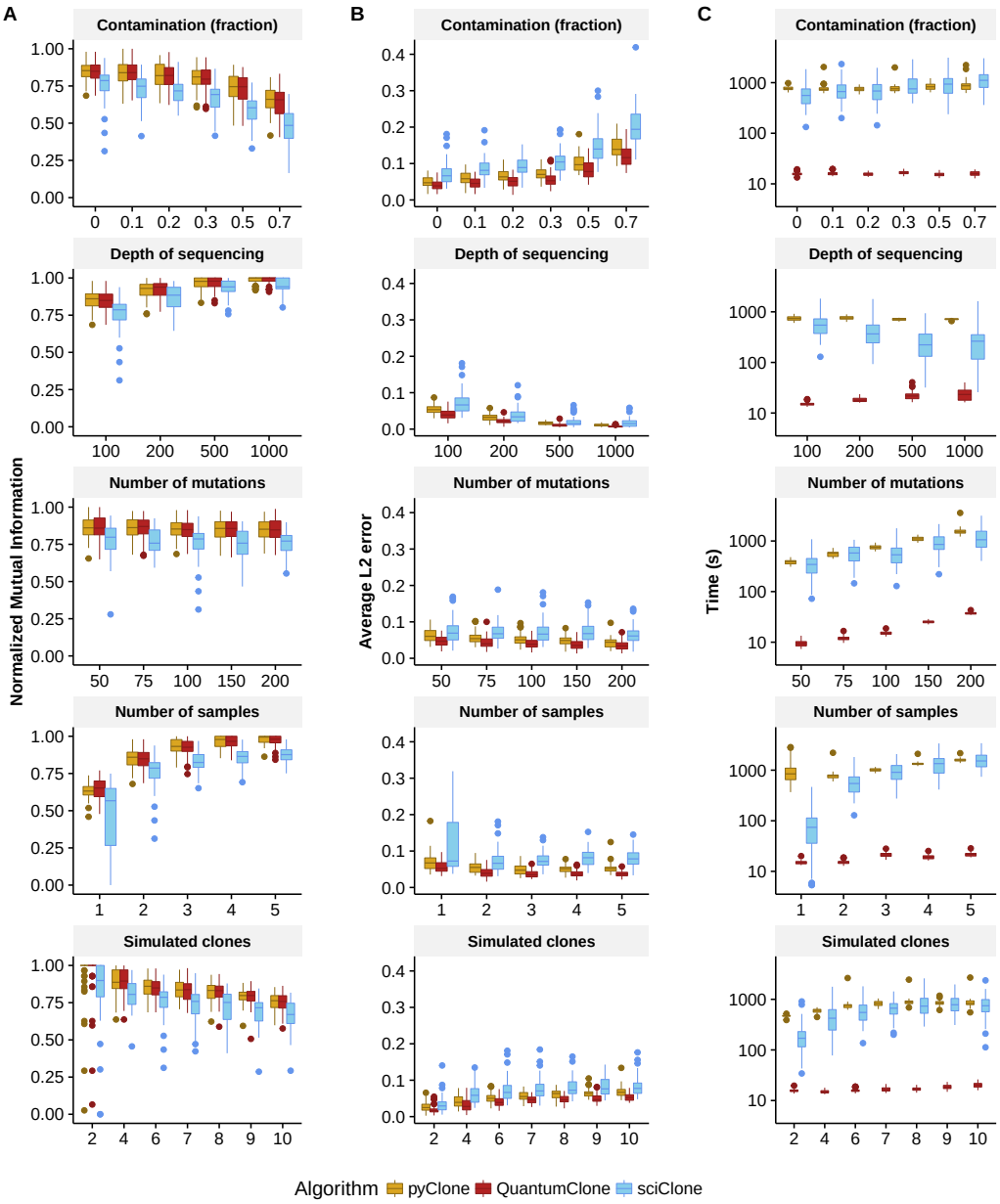


Figure 1: **Comparison of QuantumClone to existing methods. (A) Normalized Mutual Information (NMI)** is used to assess the quality of variant clustering on simulated data, with a single parameter varying in each test. This measure evaluates correct assignment of two variants to the same cluster. QuantumClone (red) shows equivalent performance to the best tool in each settings. **(B) L2 average error** is used to assess the error for each clustered variants between its simulated position and its reconstructed position. **(C) Computational time** necessary to complete the clustering with each algorithm. Default parameters: two tumor samples without contamination sequenced at 100×; 6 clones; 100 mutations used for clustering.

t-test). On average, QuantumClone decreased the $L2$ mean error by 69% compared to sciClone and 22% compared to pyClone, significantly improving predictions compared to both methods ($p-value < 2.2 \times 10^{-16}$). At high values of sequencing depth, all methods accurately estimated prevalence of variants (Fig 1B, $L2$ mean error < 0.059 at $1000\times$ for all methods). However, a sequencing depth of $100\times$, which is the depth of sequencing currently used for the majority of WES and WGS experiments, QuantumClone consistently gave better predictions than pyClone ($p - value = 1.5 \times 10^{-6}$,Welch one-sided two-sample t-test) and sciClone ($p - value = 4.9 \times 10^{-9}$). In addition, compared to the other methods, QuantumClone took the best advantage of data when multiple samples were provided for the analysis ($p-value = 2.4 \times 10^{-10}$ and $< 2.2 \times 10^{-16}$ for pyClone and sciClone respectively, Welch one-sided two-sample t-test, for simulated tumors with five samples).

Also, the average computational time was significantly decreased using QuantumClone compared to sciClone (median 35 fold improvement), or pyClone (median 46 fold improvement, Figure 1C). In the case of highly heterogeneous tumors (e.g. tumors with 10 simulated clones), the gain in computational time was of 41 fold ($p-value < 2.2 \times 10^{-16}$) compared to sciClone and 44 fold ($p-value < 2.2 \times 10^{-16}$) compared to pyClone. Similarly, when five samples were provided, we observed a 74.1 fold decrease ($p-value < 2.2 \times 10^{-16}$) compared to pyClone and 74.2 fold decrease($p - value < 2.2 \times 10^{-16}$) compared to sciClone.

Assessment of clonal reconstruction accuracy in hyper-diploid cancers or cancers with highly rearranged genomes

We expect that in addition to the parameters discussed above, the degree of genome rearrangement and chromosome duplication significantly affects the quality of the mutation clustering and consecutive clonal reconstruction. Indeed, values of cellular prevalence are linked to VAF values through the parameters representing the number of copies of the variant and the number of copies of the reference allele. Given an observed VAF value, a variant occurring in a high copy number locus has more possibilities for values of cellular prevalence: a variant with an observed allele frequency of 25% can only be linked to a cellular prevalence of 50% in a AB locus, while this variant can arise from cellular prevalence values of 33.3%, 50% or 100% if the genotype at this locus is AAAB (Methods).

In order to validate QuantumClone on diploid and hyper-diploid genomes, we simulated variants in loci of genotype AB, AAB, AABB, and in a nearly diploid genome, where all possible genotypes can be observed (Fig. 2). In addition to QuantumClone, we tested the performance of pyClone (Methods). We excluded

sciClone from this experiment as it cannot use variants from non-diploid regions.

In all types of regions, QuantumClone and pyClone performed equally in terms of NMI (Fig. 2A), but QuantumClone outperformed pyClone in terms of mean L2 error with an improvement of 31% (Fig. 2B, $p - value = 5.7 \times 10^{-11}$). In addition, QuantumClone without parallelization was faster than pyClone in three out of four settings (from 6.3 fold slower to 61.5 faster; 15.6 times faster on average), while the distributed algorithm outcompeted pyClone in all settings (average computational time decreased by a 43 fold compared to pyClone, Fig. 2C).

In addition, in the majority of cases QuantumClone correctly assumed the exact number of copies of a variant in polyploid regions (average accuracy = 68.9%, $p - value < 2.2 \times 10^{-16}$, Supplementary Figure 1).

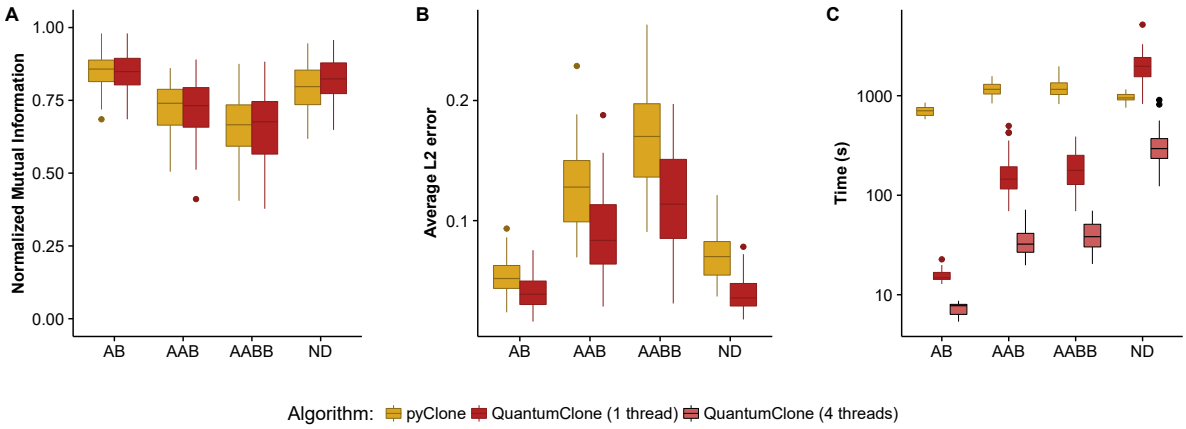


Figure 2: **Quality of clonal reconstruction for mutations located in regions of altered copy number. (A) Normalized Mutual Information** shows equivalent performances of pyClone and QuantumClone in diploid, triploid and tetraploid tumors, or nearly diploid (ND) tumors, whereas the **average L2 error (B)** shows significantly better performance of QuantumClone. **(C)** Parallel computing implemented in QuantumClone allows it to significantly decrease computational time and makes QuantumClone remarkably faster than pyClone.

We demonstrated that when a mutation can have a single or multiple copy status (AAB and AABB regions), QuantumClone performed better than the other methods. This validated our computational strategy for hyper-diploid cancers.

Overall, validation on simulated data showed that (1) QuantumClone can be applied to cancer samples with hyperploid or rearranged genomes and (2) QuantumClone generally performs better than its peers in difficult settings, for example when the number of clones is higher than or equal to six, or when the contam-

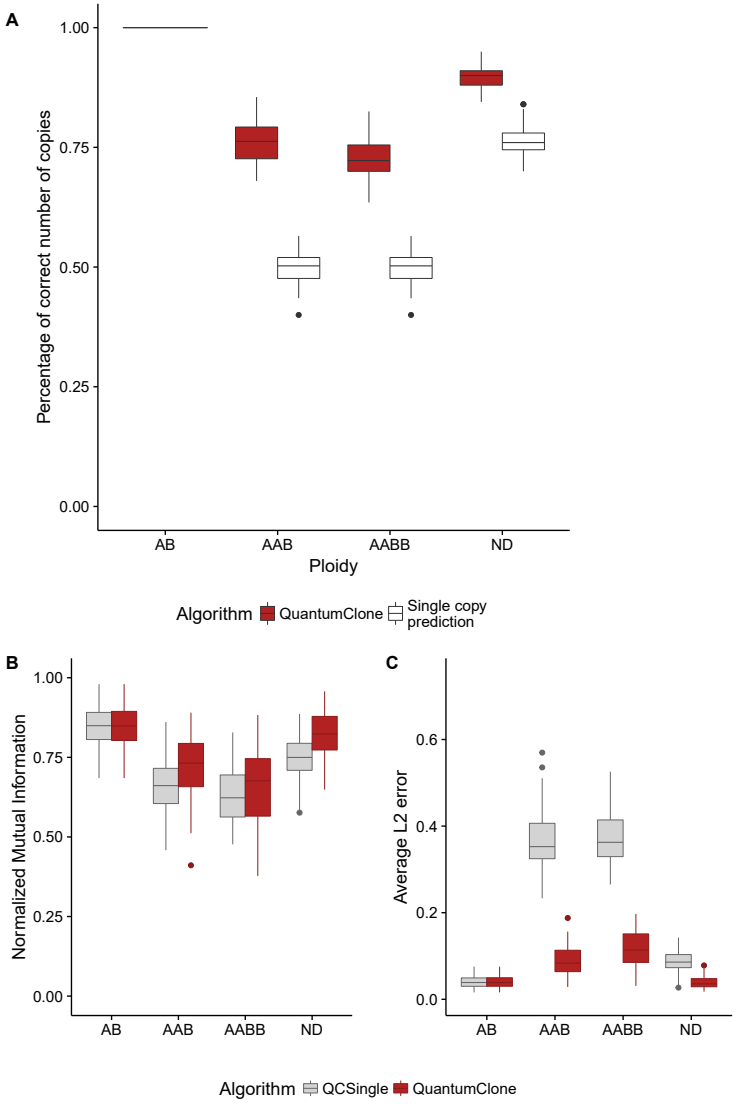


Figure S1: **Importance of the accurate assignment of variants to correct number of chromosomal copies by QuantumClone.** QuantumClone is compared to a predictor (QCSingle) that assumes that all variants are present in a single chromosomal copy. For AB regions, both approaches correctly assign variants to a single copy. Therefore, both predictors achieve similar accuracy. **(A)** Proportion of cases when the number of chromosomal copies with a variant is assessed correctly. In over-diploid regions, the maximal number of copies is determined by the number of copies of the A-allele. **(B-C)** Effect of choosing the correct number of chromosomal copies with a variant on the clonal reconstruction accuracy. ND, near diploid genome.

ination by normal cells is higher than or equal to 30%.

Creating a robust framework for clonal assignment of functional mutations

We proposed a novel concept of reconstruction of the clonal architecture in cancer. Our method combines the identification of clones, using high confidence variants, with the attribution of functional variants (potential drivers) to identified clones (Fig.3). The approach is based on the different usage of *'functional'* variants which can potentially affect cell phenotype and *'high fidelity'* variants that are used to define clones. *High fidelity* variants can be either drivers or passengers; however, they should have high depth of coverage ($> 50\times$ in our implementation), have no strand bias and should not coincide with annotated single-nucleotide polymorphisms (SNPs). As we showed in the simulation studies (Fig. 1), 50 high fidelity variants are sufficient for an accurate clonal reconstruction (Methods). *High fidelity* variants, because they have a lower dispersion of observed VAF compared with other variants, are applied to define clones, i.e., *high fidelity* variants serve as input to QuantumClone or to an alternative method. *Functional* mutations are defined here as variants that can possibly alter protein function as predicted by commonly used annotation tools (Adzhubei et al., 2013; Khurana et al., 2013; Ng and Henikoff, 2003) and that can affect either genes reported in the Cancer Census List (Futreal et al., 2004) or genes from gene modules/signaling pathways that are enriched in deleterious variants (Methods). At the last step of our framework, functional variants are mapped to the clonal structure inferred from high fidelity variants based on the likelihood values.

Here we have demonstrated that having the proposed two-step approach allows for a better reconstruction of the tumor, as well as an important decrease in computational time (Fig. 3D). To test our pipeline, we compared it to two common pipelines: the first one, termed 'classic', uses all variants as input for the clustering. The second one, called 'selective', only uses variants passing the stringent filters and informative variants as input for the clustering. The third pipeline, termed 'two-step', uses *a posteriori* attribution of the putative drivers to the clones found using only variants passing stringent filters. While all three pipelines had similar outcomes when we compared the quality of reconstruction using normalized mutual information (Figure 3B), the selective and two step pipelines fared significantly better than the classical pipeline ($p - value < 8 \times 10^{-6}$, one-sided Welch two-sample t-test, Figure 3C). In addition, the two step analysis resulted in an average 4.9 fold decrease in computational time compared to the classical pipeline and an average 2.7 fold decrease compared to the selective pipeline (Figure 3D). Furthermore, separating both steps eases iterative improvement of the clonal reconstruction. Once achieved, this reconstruction can be reused to answer questions about the evolution of different pathways separately, while previous pipelines required re-running the whole reconstruction with the new set of data.

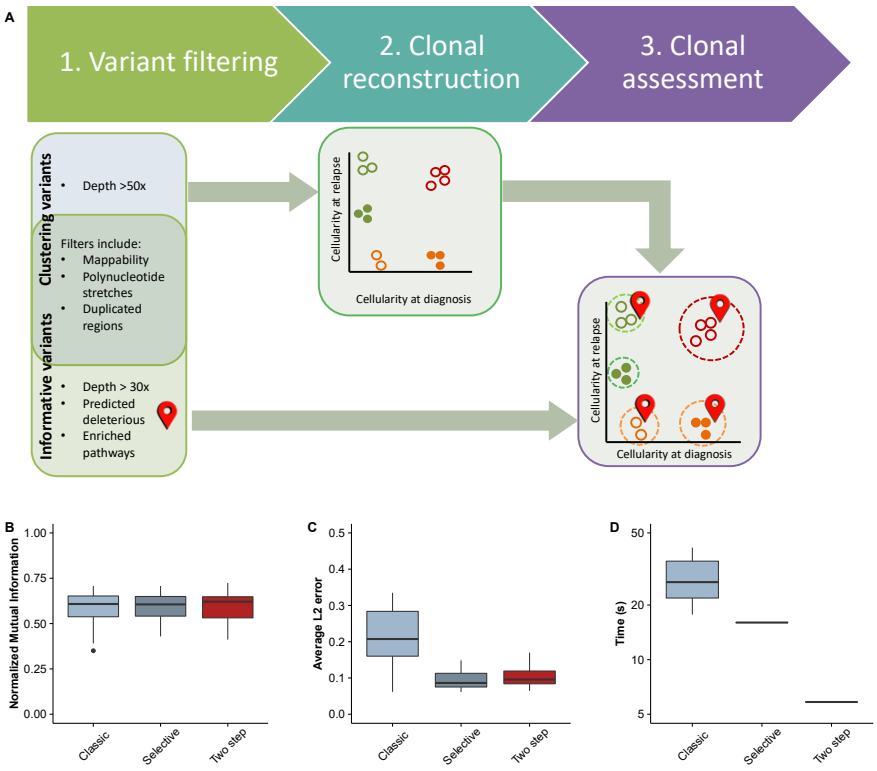


Figure 3: **Assessment of the pipeline**. **(A) Overview of the general clonal reconstruction workflow**: steps 1-3. (1) Variants are filtered to remove false positive calls; stringent filters are used to produce mutations that are further employed for clonal reconstruction (step 2), tolerant filters are used to detect functional mutations. (2) Variants that pass stringent filters and have genotype information assigned to the corresponding genomic loci are used as input to QuantumClone to reconstruct clonal populations. (3) Finally, possibly damaging mutations belonging to frequently altered pathways are mapped to the reconstructed clones. **Quality of reconstruction**. The pipeline aforementioned (two step), or a clustering using all variants called (classic) or a pipeline using only variants of biological interest and variants of high quality (selective) are assessed in terms of NMI **(B)**, average $L2$ error **(C)** or computational time **(D)**. The pipelines are evaluated on 20 simulations (Methods).

Characterization of neuroblastoma clonal evolution from diagnosis to relapse: application of the QuantumClone-based framework

We applied our framework to investigate the clonal composition of neuroblastoma primary and relapse tumors and to study their clonal evolution. We performed WGS of constitutive DNA, diagnosis and relapse tumor samples for each patient with an average depth of coverage of $\sim 100\times$. Datasets for 15 out of 22 patients came from a previously published study (Eleveld et al., 2015). Sequencing was carried out using either Illumina HiSeq 2500 and Complete Genomics platforms. Reads were mapped to the reference hg19 genome using BWA-aln (Li and Durbin, 2009) (Illumina reads) and the internal Complete Genomics mapping tool (Complete Genomics reads). Variant calling was performed using Varscan2 version 2.3.6 (Koboldt et al., 2013).

The level of contamination by normal cells (i.e. non-tumoral cells from the patient) varied from 0% to 90%, and three samples from three distinct patients had contamination levels higher than 70%. These patients were excluded from the clonal reconstruction analysis. (clinical data available in Suppl. Table 1). Consequently, we characterized clonal structure of tumors of 19 out of 22 neuroblastoma patients.

Application of filters unifies variant call numbers across different sequencing platforms

In order to remove false positive variant calls, we used a set of stringent filters (Fig. 3, Methods). The initial number of variants in the Varscan2 output was highly dependent on the sequencing technology and platform (Suppl. Fig. 2). The number of variants called for samples sequenced by the Beijing Genomics Institute (BGI) sequencing platform was an order of magnitude higher than the number of mutations called for samples processed by the Centre National de Génotypage (CNG). Application of a set of stringent filters based on read depth of coverage, read mappability, and annotated repetitive regions (listed in Fig. 3A and Methods) allowed us to remove platform bias for further analysis. After all filters, the number of variants per sample correlated with the age of the patient (Suppl. Fig. 3, Spearman's $rho = 0.93$, $p - value = 3.4 \times 10^{-6}$) which was consistent with previous studies (Molenaar et al., 2012).

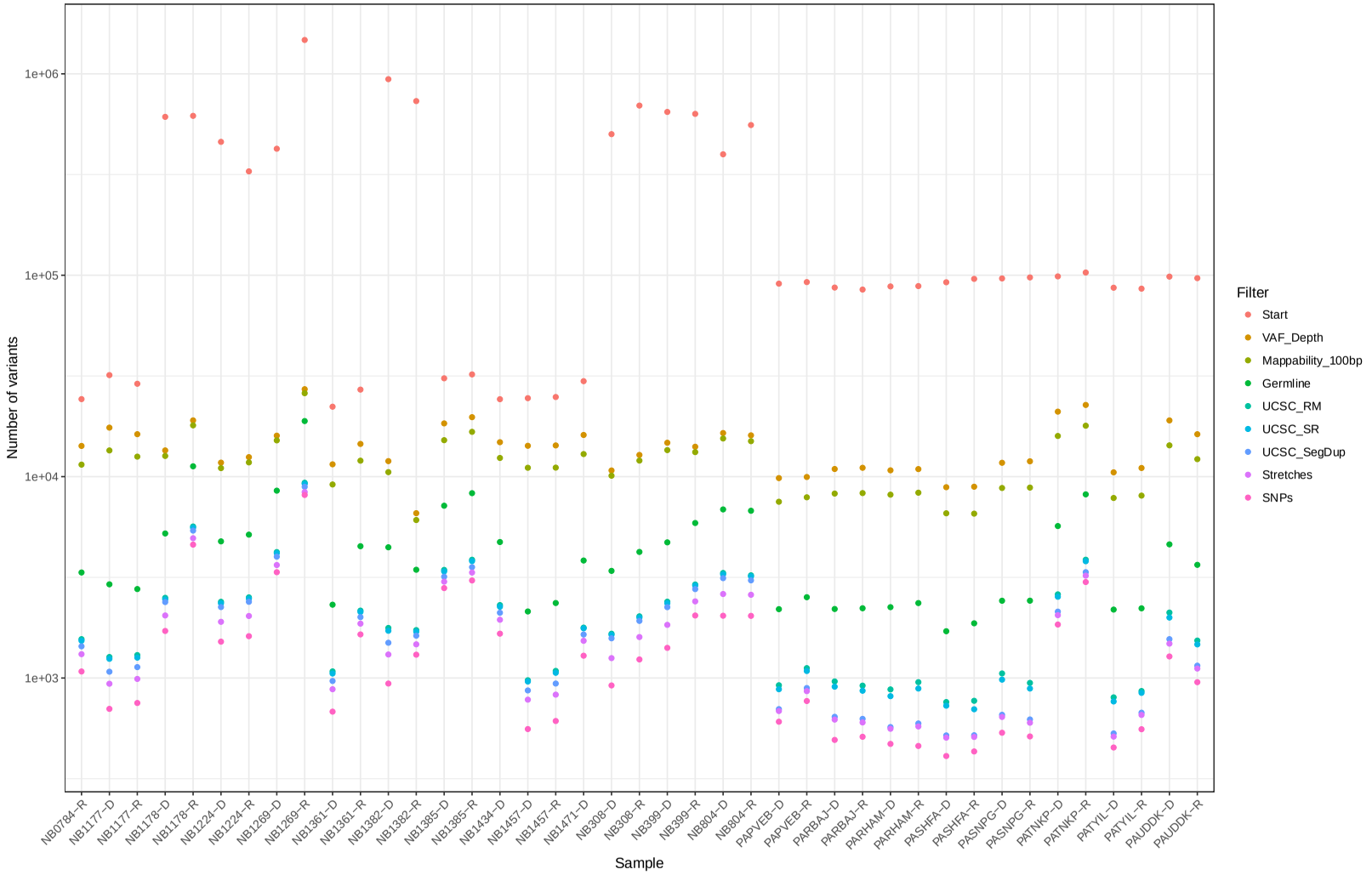


Figure S2: **Effect of filtering on the number of somatic variants called.** The initial number of variants corresponds to the raw output from Varscan2; the "VAF_Depth" filter selects variants with at least 10% of the reads supporting each variant and at least 50 reads mapping at the position of each variant; the "Mappability_100bp" filter requires the mappability of a 100 bp DNA sequence to be 1 at the position of the variant (the UCSC genome browser 100 bp mappability track); the "Germline" filter removes variants found at germline level in other patients of the cohort; the "Stretches" filter removes variants that would create a homopolymer of three or more identical bases; the "SNPs" filter removes variants that are present in the SNP databases at a frequency higher than 1% except when the variant is listed in the COSMIC database; finally, the "UCSC_RM","UCSC_SM","UCSC_SegDup" filter removes variants that are present respectively in the repeat masker, simple repeat and segmental duplication regions from the corresponding UCSC genome browser track.

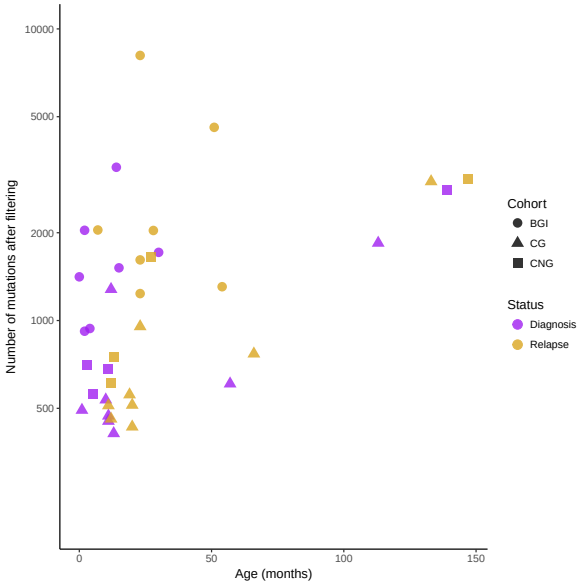


Figure S3: **Statistics on numbers of somatic variants called using stringent filters for diagnosis and relapse samples from 19 neuroblastoma patients.** The number of somatic variants is correlated with the age of the patient at the time of the biopsy (Spearman correlation test $rho = 0.93$, $p - value = 3.4 \times 10^{-6}$). Final variant numbers after the filtering step do not depend on the sequencing center or sequencing technology used.

Clonal reconstruction

We applied QuantumClone on *high fidelity* variants we defined using stringent filters (Fig. 3A, Methods). Across our cohort, we did not observe a significant association between the predicted number of clones and the number of mutations per patient (Spearman's $rho = -0.23$, $p - value = 0.35$). In addition, the number of clones at relapse was similar to that at diagnosis, even despite the fact that the relapse samples had about twice as many mutations as the diagnosis samples (number of mutation clusters varied from one to four with a median of three for both time points).

In 79% of reconstructed clonal structures (15 out of 19 patients, we identified mutations coming from the ancestral clone (Fig. 4A), i.e. the clone that gave rise to all cells in both diagnosis and relapse samples.

Annotation of functional mutations in each sample based on the global pathway enrichment analysis

In our framework, we assumed that *functional* mutations (i.e. putative drivers) in a given cancer type should target specific signaling pathways or pathway modules (Fig. 3, Step 2). We attributed annotated deleterious variants obtained with tolerant filters (Fig. 3, Methods) to the ACSN maps and detected recurrently altered

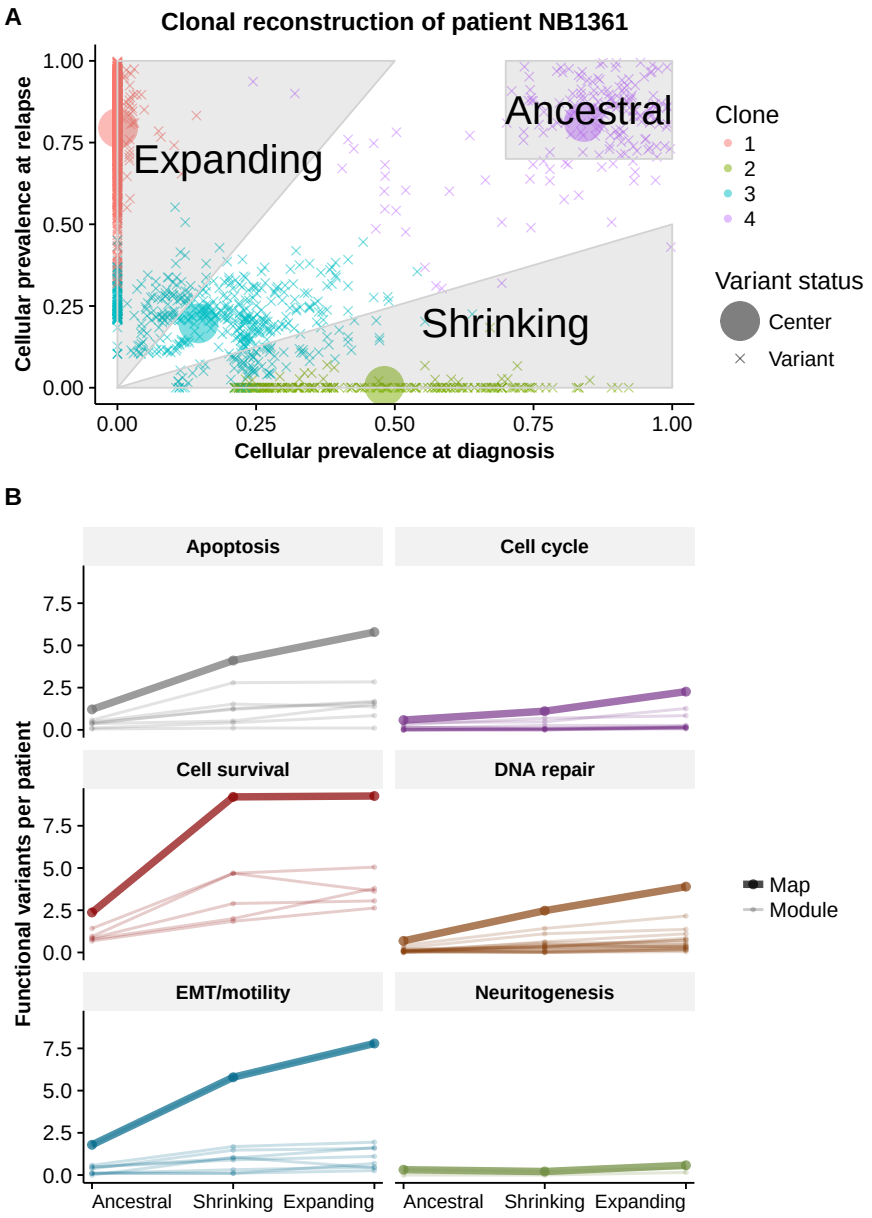


Figure 4: **Annotation of clones in neuroblastoma and pathway enrichment analysis. (A)** Illustration with data from patient NB1361 of the rules for assignment of variants to (i) the ancestral clone (cellular prevalence of the mutation cluster exceeds 70% both at diagnosis and relapse), (ii) clones expanding after the treatment (cellular prevalence of the mutation cluster increases at least two-fold at relapse) and (iii) shrinking clones (cellular prevalence of such mutation clusters decreases at least two-fold). **(B)** Evolution of the total number of functional variants for enriched maps and modules, across all 19 patients. The majority of modules show an increase in the number of functional variants between the two time points.

gene modules using the ACSNmineR package (Deveau et al., 2016). Overall, six general gene maps (apoptosis,cell cycle, DNA repair, EMT / cell motility, cell survival and neuritogenesis) and their 53 gene modules were found to be enriched in mutations (threshold 0.01 on the p-value, one-sided exact Fisher test, corrected to account for multiple testing with the Benjamini-Hochberg False Discovery Rate correction, corresponding to the q-value) (Supp. Table 2). The enrichment of pathways in ACSN was corroborated by enrichment of similar pathways from two other methods (Huang et al., 2009b,a; Thomas et al., 2003; Mi et al., 2010) (Supp. Table 3 and 4). In further analysis, deleterious mutations were annotated as *functional* when corresponding genes were included in the enriched pathways, or when such genes belonged to the Cancer Census list. The resulting number of *functional* mutations per patient varied from 2 to 147, with a median of 51.

At this step, the cell survival map registered the highest enrichment in putative drivers, and among its modules, the highest enrichment in putative driver mutations was observed for the non-canonical WNT pathway ($q - value \leq 10^{-88}$). In addition, we also detected significant enrichment in *functional* mutations of the WNT canonical and the MAPK pathways ($q - value \leq 10^{-51}$ and $\leq 10^{-54}$, respectively), and of the PI3K/AKT/mTOR and Hedgehog gene modules ($q - value \leq 10^{-75}$ and $\leq 10^{-43}$, respectively). As for the modules of other maps, genes coding for the EMT regulators were also significantly affected by the deleterious mutations in our cohort of relapsed neuroblastoma patients ($q - value \leq 10^{-126}$).

Assignment of *functional* mutations to the identified clonal structure

Using the results of the mapping of *functional* mutations on the clonal structure detected for each patient by QuantumClone (Fig. 3A, Step 3), we annotated mutations as (*i*) those belonging to expanding clones - corresponding to a two-fold cellular prevalence increase between diagnosis and relapse, (*ii*) those belonging to shrinking clones - cellular prevalence halved between diagnosis and relapse , and (*iii*) those belonging to ancestral clones - cellular prevalence higher than 70% in both samples (Fig. 4A). Overall, 36%, 30% and 9.6% of all *functional* mutations fell in these three categories.

Analysis of pathways enriched in *functional* mutations in shrinking and expanding clones

Assignment of mutations to clones shrinking or expanding after the treatment resulted in the identification of 336 and 400 possible driver mutations in these clone types, respectively. Expanding clones had more deleterious mutations targeting genes from all six general maps (apoptosis, cell cycle, DNA repair, EMT/cell motility, cell survival and neuritogenesis) than the shrinking clones (Fig. 4B). Similarly, in these expanding

clones, most of the corresponding gene modules (e.g., MAPK, WNT canonical or PI3K/AKT/mTOR) were also more frequently targeted. An extreme example of this behavior can be given with the neuritogenesis substrates module, the RB pathway or the E2F1 pathway in which genes are only found mutated in the expanding clones. The increase in functional variants can partly be explained by the observed doubling of variants at relapse compared to diagnosis. We define μ the functional mutation rate in a module as the number of functional variants per high fidelity variants of the patient by number of genes in a module. The functional mutation rate across modules was significantly different between the three classes of clones according to the z-score computed as suggested by Paternoster et al. (1998) and described in Methods (Fig. 5A, $p - value = 8.35 \times 10^{-5}$ between ancestral and shrinking, $p - value = 2.84 \times 10^{-3}$ between ancestral and expanding and $p - value = 4.98 \times 10^{-2}$ between expanding and shrinking). This functional mutation rate has been previously linked to the fitness of a clone (McFarland et al., 2013), and it is interesting to notice that the functional mutation rate is lower in the ancestral clone ($\mu = 5.803$ functional variations per 1,000 variants per 1,000 genes in module, standard error $s.e = 1.322$), than in the shrinking clones ($\mu = 15.78$, $s.e. = 1.919$) or expanding clones ($\mu = 10.92$, $s.e. = 0.7583$). The change in functional mutation rate suggests different selection mechanisms.

Overall, we proposed a new method to reconstruct clonal populations. We applied it to neuroblastoma samples obtained at diagnosis and relapse. The fact that there are fewer functional variants in the ancestral population than in the shrinking or expanding populations and that the expanding population has a lower functional mutation rate suggests that a clone with fewer functional variants had better adaptive capabilities, as proposed by Chen et al. (2015).

Discussion

Here we propose a pathway-based framework to detect functional mutations in cancer samples and associate the mutations to their corresponding clonal structure. The central part of our framework is represented by the QuantumClone method, which allows reconstruction of clonal populations based on both variant allele frequencies and genotype information. QuantumClone showed stable results on simulated data significantly outperforming other methods in difficult settings such as highly contaminated samples, heterogeneous tumors and relatively low depth of sequencing coverage.

The central idea of our analysis framework is to use high fidelity variants to reconstruct the clonal structure of tumor samples; then, map low coverage functional mutations (with high variance in VAFs) onto the

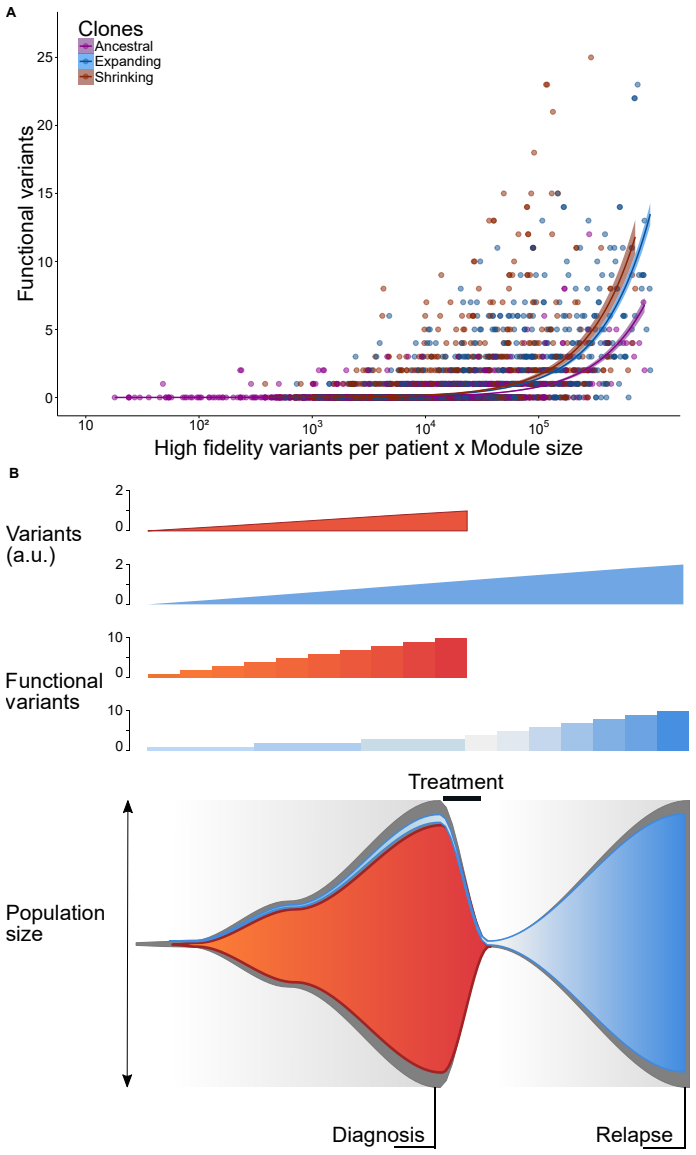


Figure 5: **Ancestral, shrinking and expanding clones exhibit different mutation patterns in neuroblastoma relapse tumors.** **(A)** Functional mutation rate is higher in shrinking and expanding clones compared to the ancestral ones. We define the functional mutation rate as a ratio of the number of functional mutations to the number of high fidelity variants. For a given gene module the number of functional mutations in each patient is supposed to linearly depend on the product of the module size and the total number of detected variants. Therefore, we used the product of the module size and number of high fidelity variants as a covariate in a linear regression model evaluating functional mutation rate for neuroblastoma tumors. The rate was defined as the slope of the linear regression. **(B)** Given the differences in functional mutation rates observed in neuroblastoma relapse tumors we propose the following model for clonal selection in this type of cancer: (1) Clones with high functional mutation rate (red) disappear after the chemotherapy; lower mutational burden provides an advantage in escape from treatment; (2) lower values for functional mutation rate in clones expanding at relapse (blue) compared to the shrinking clones (red) is due to a lower frequency of functional mutations before treatment, followed by a gradual accumulation of functional mutations at relapse. From top to bottom: the number of variants in the clone, number of functional variants in the clone, and population size in the tumor; a.u. (arbitrary units).

inferred clonal structure. Also, we suggest limiting the set of functional mutations to those in genes known to be associated with cancer (e.g., Cancer Census genes) or to those in genes from gene modules/pathways that are frequently disrupted in a given cancer type (Fig. 3).

We apply the proposed analysis framework to decipher clonal structure in neuroblastoma and assign to clones possible driver mutations. We detect 53 modules as being altered by mutations in neuroblastoma. We identify genes associated with DNA repair, cell motility, apoptosis and survival to be enriched in functional mutations in neuroblastoma. For relapsed neuroblastoma samples, we recover the previously reported enrichment of mutations in the MAPK signaling pathway (Eleveld et al., 2015), while complementing this knowledge with discovery of accumulation of functional mutations at the relapse in such functional gene modules as PI3K/AKT/mTOR, WNT, Hedgehog signaling and modules consisting of genes responsible for cell-matrix adhesion and epithelial–mesenchymal transition (EMT).

For some of our samples, we did not succeed in uncovering an ancestral clone despite the fact that copy number breakpoints were consistent between samples, ensuring a common phylogeny (Bollet et al., 2008) (Supp. Fig 4). Disappearance at relapse of many potential driver mutations seemingly present in the ancestral clone at diagnosis, may be due to tumor heterogeneity and the fact that biopsies were taken from different tumor sites. This situation has been termed "illusion of clonality" (Bruin et al., 2014). It should be noted that our framework does not intend to reconstruct the exact phylogeny of the tumor, but focuses on the architecture of the samples. If needed, possible phylogenetic trees can be obtained using existing methods based on the pre-clustered VAFs (Hajirasouliha et al., 2014).

In the application of our framework to neuroblastoma sequencing data, we excluded information about translocations and indels. The reason for this was that the analysis of clonal structure is based on the number of sequencing reads supporting each genetic variant. While we suppose that the number of reads with a mismatch mutation is proportional to the number of DNA molecules harboring this variant, we expect that due to read mapping issues the fraction of reads indicating an indel or a translocations will be generally lower than the actual proportion of DNA molecules with the rearrangement. Eviction of large and small SVs seemingly resulted in a decrease in sensitivity of the detection of genetic driver events. A possible way to solve this issue would be to estimate the cellular prevalence of these event using specific tools and attribute such events to the most likely clone.

The proposed framework can be applied in the future to any type of cancer. The pre-requirements are

sufficient number of candidate mutations (at least 50 mutations per sample) and a minimal read depth of coverage of 50×. These requirements are usually met by WGS or whole exome sequencing datasets. Our simulation results show that increasing the number of mutations used for clonal reconstruction above 50 does not improve significantly the clonal reconstruction accuracy provided that mutations specific for every clone are present in the input.

Previous studies have shown that the number of variants was linked to the number of divisions a cell undergoes (Tomasetti and Vogelstein, 2015). The observed doubling of variants between diagnosis and relapse suggests that cells have undergone as many divisions between diagnosis and relapse as between cancer origin and diagnosis - with the assumption that the mutational rate remains constant. This would in particular exclude the possibility of the relapse emerging from a quiescent population.

In addition, the functional mutation rate was significantly lower in the ancestral populations compared to the clones expanding or shrinking at relapse. Chen et al. (2015) have shown that wild-type cells have more adaptive capabilities than mutants, even though a mutant can appear fitter than the wild-type lineage in a specific culture condition. Applied to our results, their finding could suggest that a clone with a low level of functional variants would be more likely to adapt to environment changes during and after treatment. After this selection round and once the tumor environment has returned to physiological state, another set of functional variants would appear, giving selective advantage to the expanding clone.

A direct consequence of this assumption is that the functional mutation rate should be lower at relapse compared to diagnosis, as a period of low functional mutation rate before treatment would be followed by a period of higher functional mutation rate during disease progression (Fig. 5B). This consequence is in line with the 29% functional mutation rate decrease observed between expanding and shrinking clones.

Study of the clonal evolution and its processes can be highly relevant for drug design. We described a framework and an algorithm that performed better than previously published methods, which should allow for a better analysis of existing data. In addition, we showed that the same processes are at play throughout the course of the disease in our neuroblastoma cohort, targeting similar pathways in diagnosis and relapse.

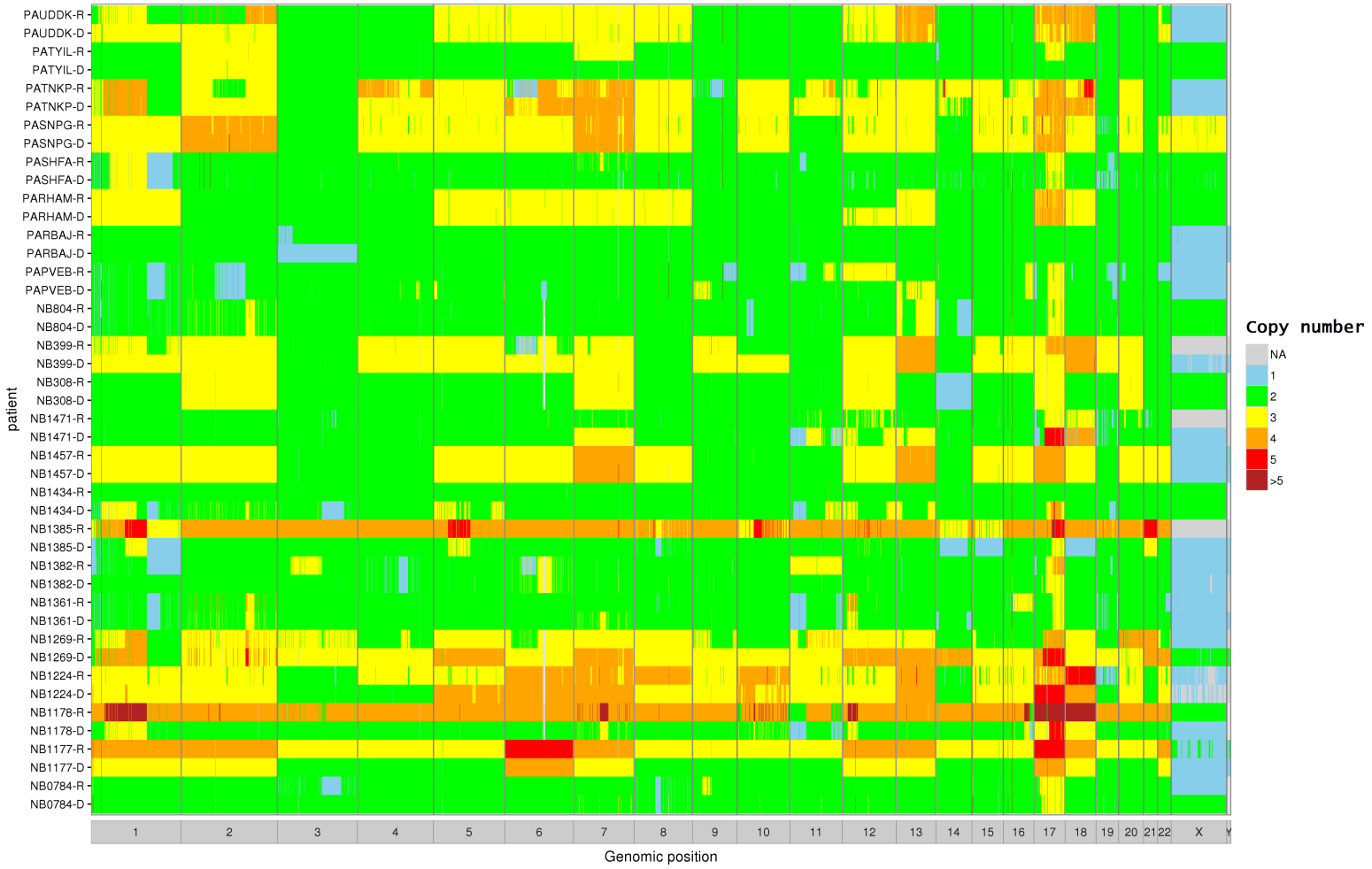


Figure S4: Summary of the copy number profiles for all samples determined by Control-FREEC. The copy number is represented by a color, with the genomic position in abscissa and the sample on the y-axis.

Methods

Datasets

Patient inclusion criteria and collection of tumor samples. The inclusion criteria for this study were histopathological confirmation of neuroblastoma at original diagnosis and the presence of biopsy material from a subsequent relapse specimen. Patients were included in this study after an informed consent was obtained from parents or guardians, with oversight from the ethics committees 'Comité de Protection des Personnes Sud-Est IV', reference L07-95/L12-171, and 'Comité de Protection des Personnes Ile-de-France', reference 0811728 in France, the review board at the Children's Hospital of Philadelphia and review boards at other Children's Oncology Group sites that submitted samples for patients on this study in the United States. In total we obtained material for 22 neuroblastoma patients (tumor tissue at diagnosis, relapse and constitutional DNA, Suppl. Table 1).

Whole-genome sequencing of neuroblastoma samples. In the framework of this study, we carried out Illumina paired-end sequencing for 7 novel neuroblastoma patients (the total of 21 patients corresponding to a tumor at diagnosis and relapse, and a matching blood sample per patient). Data for 15 patients were taken and reanalyzed from the previous study (Eleveld et al., 2015). DNA from 7 patients from the previous study and 7 new ones have been sequenced using Illumina HiSeq 2500 instruments to an average depth of coverage of 80× by Beijing Genomics Institute (BGI) and the Centre National de Génotypage (CNG) respectively. For 8 patients out of 15 previously reported, whole-genome sequencing was performed by Complete Genomics with an average read depth of coverage of 50×. DNA material for each patient (lymphocytes, primary tumors and relapse tumors) was in each case sequenced using the same sequencing platform (see Suppl. Table 1 for more detail).

Data processing. Sequenced reads were mapped to the human genome hg19 using BWA and the internal Complete Genomics tools for Illumina and Complete Genomics datasets respectively. Reads from datasets sequenced using the Illumina platform were realigned around indels with the Genome Analysis ToolKit (GATK) (McKenna et al., 2010), followed by a base recalibration. Due to the inherent structure of Complete Genomics reads, which contain an effective deletion relative to their corresponding genomic library, the indel realignment step was skipped for the Complete Genomics samples.

Variant calling and filtering

Mutations were called using Varscan2 (Koboldt et al., 2013). Two sets of variants were created for each patient (Fig. 3A) using tolerant and stringent filtering options. The 'high confidence' set of variants obtained using stringent filters was further used for clonal reconstruction, while the set of variants obtained with tolerant filters was used for inference of recurrently altered pathways.

Tolerant filters for somatic variants included those on minimal depth of coverage (30×), minimal percentage of reads supporting the mutation (10%). In addition, variants were required to be located in regions of high local mappability (based on the 100 bp mappability track), and outside of repeat and duplicated genomic regions. The latter was assessed using the UCSC repeat masker, simple repeat, and segmental duplication regions. We further filtered variants that created a stretch of four or more identical nucleotides. Finally, we only kept mutations located in regions where the genotype evaluated by Control-FREEC was available.

To obtain a set of high confidence variants, in addition to the aforementioned filters, we required the minimal depth of coverage of 50×. We filtered out variants corresponding to polymorphisms present in more than 1% of the population (snp138, 1000Genomes, esp6500) except if it was a known cancer related variant (COSMIC database for coding and non-coding mutations).

Copy number analysis

Copy number alterations in patients were detected using the Control-FREEC method (Boeva et al., 2012) (version 7.2) (Suppl. Fig. 4). We selected the main ploidy value so that the predicted copy number and B-allele frequency profiles were consistent. Control-FREEC also provided estimations of the level of contamination by normal cells, which, after manual confirmation, was further used for the clonal reconstruction.

Out of 40 tumor samples, three had estimated proportion of contamination by normal cells higher than 70% were excluded from the further analysis (NB0784:diagnosis, NB1434:relapse and NB1471:relapse).

Comparison of clonal reconstruction between QuantumClone and existing methods

Data simulation. *In silico* validation data were generated using the QuantumCat method from package QuantumClone (version 1.0.0.3). QuantumCat simulates genomic variants, copy number alterations and corresponding VAFs. It relies on the following set of rules:

1. A binary phylogenetic tree is created to simulate the clonal architecture of the tumor. The mutation cellular prevalence values correspond to the nodes and leaves of the phylogenetic tree.

2. Cellular prevalence values of mutations from each clone are independent across tumor samples. However, the cellular prevalence of each clone should always be coherent with the phylogenetic tree.

3. The allelic copy number of all mutation loci was set to AB in the tests carried out to compare QuantumClone, sciClone and pyClone (Fig. 1). For QuantumClone validation on triploid, tetraploid and nearly diploid genomes (Fig. 2), the number of chromosomal copies bearing each mutation was randomly assigned between one and the number of A-alleles for the locus considered. Generation of the genotype, number of chromosomal copies, normal contamination and cellular prevalence of a mutation allows for the computation of the exact VAF, which is the cellular prevalence (taking into account the contamination by normal cells) multiplied by the number of copies of the mutations and then divided by the number of copies of the locus in each cell. On the other hand, the observed VAF is determined by the ratio of the number of reads supporting the mutations divided by the read depth of coverage. Local depth of coverage at each given position was generated by the negative binomial distribution centered on the target depth of sequencing, fitted on experimental data. The number of reads supporting a mutation was simulated from the binomial distribution with the probability of success equal to the exact VAF.

Program versions and parameters. We used SciClone version 1.1.0 with the following changes to the default parameters: maximal number of clusters was set to 10 and the minimal depth of coverage was set to 0.

PyClone version 0.13.0 was used with the following parameters : 10,000 iterations of the Markov chain Monte Carlo, alpha and beta parameters in the Beta base measure for Dirichlet Process set to 1, concentration prior shape set to 1 and the rate parameter in the Gamma prior on the concentration parameter set to 0.001. We used the default Beta binomial distribution with precision parameter set to 1000, prior shape set to 1, rate to 0.0001, and proposal precision set to 0.01.

We used an implementation of the k-medoids algorithm provided by the R package "fpc", version 2.1.10, with a range of clusters between 2 and 10.

For clonal reconstruction, QuantumClone version 1.0.0.3 was used with default parameters except for the the maximal number of clusters, which was set to 10. For clonal reconstruction of neuroblastoma data, variants used to compute centers of clusters (corresponding to clones) were selected using the stringent set of filters. Copy number information from Control-FREEC (version 7.2) was also passed to the QuantumClone algorithm as well as the predicted value of contamination by normal cells.

For simulated data, quality of clustering was assessed by using Normalized Mutual Information (NMI) (Manning et al., 2008), which is given for a group of clones Ω and a group of reconstructed clusters $\mathbb{C}$:

$$NMI_{(\Omega,\mathbb{C})} = -2 \times \frac{\sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} log\left(\frac{N \times |\omega_k \cap c_j|}{|\omega_k||c_j|}\right)}{\sum_k \frac{|\omega_k|}{N} log\left(\frac{|\omega_k|}{N}\right) + \sum_j \frac{|c_j|}{N} log\left(\frac{|c_j|}{N}\right)}, \quad (1)$$

where N is the number of variants observed, $|\omega_k|$ the number of variants in clone k , and $|c_j|$ the number of variants attributed to cluster j.

Reproducibility. All scripts necessary to reproduce the results can be found at: `https://github.com/DeveauP/QuantumClone/tree/master/tests/reproducible_testing/Comparison_other_methods`

Pipeline comparison

Data simulation. *In silico* validation data were generated using the QuantumCat method from package QuantumClone (version 1.0.0.3). We simulated variants coming from six clones observed in two samples per patient, with a purity of 70% for the first sample and 60% for the second. We create 150 variants that pass stringent filters, and an additional 150 variants passing tolerant filters but not stringent filters. All variants passing stringent filters were simulated in diploid regions, with a depth of coverage higher than 50×, whereas mutations passing permissive filters were located either in AB regions with a coverage between 30× and 50× (approximately 1/4 of permissive variants), or in AAB regions with coverage $\geq$ 30× (approximately 1/2 of permissive variants), or in AABB regions with coverage $\geq$ 50×. We then attributed the 'driver' characteristic to 100 variants, by sampling without replacement with probability 10/11 to be selected from the variants passing permissive filters.

Pipelines The 'classical' pipeline used all 300 simulated variants as input for the clonal reconstruction,

using direct clustering by QuantumClone. The 'selective' pipeline used the 150 variants passing stringent filters as well as all variants qualified as drivers from the permissive filters as input for direct clustering. The 'two-step' pipeline first used the 150 stringent variants as input for direct clustering, and then attributed the variants qualified as drivers *a posteriori* to the clusters, using the characteristics of the clones found by the initial QuantumClone clustering of high confidence variants. All three pipelines searched for two to ten clones, running with two different initializations, on four threads. Computational time was measured on a computer running Windows 10, with an Intel i7 at 2.7GHz with 8Gb of RAM, Rstudio 1.0.44 and R version 3.3.2.

Evaluation Evaluation of the L2 error and NMI was made using only variants from the stringent and driver groups. The displayed computational time takes into account data processing, clustering and when necessary *a posteriori* attribution to the clonal structure.

Reproducibility. All results shown in figure 3 can be reproduced using the Figure3.Rmd file that can be downloaded from `https://github.com/DeveauP/QuantumClone/tree/master/tests/reproducible_testing/Rscript`.

Clonal reconstruction

In this section, we describe QuantumClone, a method we have developed for the clonal reconstruction of a tumor. QuantumClone performs clustering of cellular prevalence values θ of variants defined by:

$$\theta = \frac{VAF \times N_{Ch}}{NC \times P}, \tag{2}$$

where N_{Ch} is the number of copies of the corresponding locus, NC the (*a priori* unknown) number of chromosomal copies bearing the variant, and P the tumor purity. For instance, the cellular prevalence is equal to $2 \times VAF$ only in the case of a purely diploid tumor without loss of heterozygosity (LOH) regions, with no contamination of the sample by normal cells. The latter assumption has been frequently used in cancer studies (Schramm et al., 2015; Williams et al., 2016). As we do not have information about the number of chromosomal copies bearing a variant, our approach was to compute each possible value of cellular prevalence associated with the variant allele frequency. For example, a mutation can have a VAF of $1/3$ in a locus of genotype AAB when it is present in 100% of tumor cells on a single chromosome copy and when it is present in 50% of tumor cells on two chromosomes. Yet the latter case is rather unlikely. Each mutation thus corresponds to several possible values of cellular prevalence; each solution is associated with a value of NC. In order to address the problem of non-uniqueness of a solution, we use an EM algorithm based on the probability to observe a specific number of reads confirming a mutation given the number of reads

overlapping the position, the contamination and the cellularity of a clone. In more detail, we attribute to each possibility a probability $P(f|\theta)$ to observe f reads supporting the variant given that the latter belongs to a clone of cellular prevalence θ, based on a binomial distribution:

$$P(f|\theta) = \binom{d}{f} \left(\frac{\theta \times NC\,(1-c)}{N_{Ch}} \right)^{f} \times \left(\frac{1-(\theta \times NC\,(1-c))}{N_{Ch}} \right)^{d-f}, \tag{3}$$

where

- d the depth of coverage of the variation;
- f the number of reads supporting the variant;
- c the sample contamination by normal cells.

We can then write the log likelihood function to maximize:

$$L = \sum_{i \in mutations} \sum_{k \in clones} \sum_{s \in samples} \sum_{p \in possibilities(i)} \omega_{(i,p)} t_{(i,k)} log\left(P_{i,s,p}\left(f_{i,s,p}|\theta_{k,s}\right)\right), \tag{4}$$

where $\omega_{i,p}$ are weights of the possibility computed for a corresponding genotype $xAyB$ (major allele A is present x times and the minor allele B is present y times):

$$\omega_{i,p} = \prod_{s \in samples} \frac{\binom{x_s}{NC_{i,s,p}} + \binom{y_s}{NC_{i,s,p}}}{2^{N_{Ch_s}}}.$$

By adding weights that, for each variant, sum up to one, we include in our model the fact that variants in low copy number regions bear more information than those in hyper diploid regions. Each variant is then attributed to its most likely possibility, which is the possibility with highest probability to belong to a clone. In the situation described above (a variant in a AAB region with the VAF of 1/3), this approach would assign probabilities of 2/3 and 1/2 to the presence of the mutation in 100% and 50% of cells respectively. However, if there is a second mutation present, for example, in a locus of genotype AB with a VAF of 1/2 and thus having unambiguously cellular prevalence of 100%, the first mutation will have a high density of probability for a cellular prevalence of 100% and our approach will assign both mutations to the same cluster corresponding to the same cellular prevalence (100%).

The number of clones is determined by minimization of the Bayesian Information Criterion (BIC). Priors can be provided by the user, randomly generated, determined by the k-medoids clustering on mutations in A and AB sites when the latter contain enough mutations, or using a hierarchical clustering based on the

probability of two variants to belong to the same distribution (default).

Analysis of mutation enrichment in signaling pathways and gene modules

ACSNMineR (Deveau et al., 2016) version 0.17.1.6 was used to detect gene modules and pathways enriched in deleterious mutations. Gene modules included by default in ACSNMineR come from the manually curated Atlas of Cancer Signalling Networks (ACSN) (Kuperstein et al., 2015). In addition to the ACSN modules, we calculated variant enrichment in a set of neuritogenesis genes frequently mutated in neuroblastoma (Molenaar et al. (2012), Suppl. Table 8). We called deleterious mutations all stop-gain mutations or variants that were predicted to be possibly damaging or deleterious by SIFT (Ng and Henikoff, 2003), PolyPhen-2 (Adzhubei et al., 2013), or FunSeq2 (Khurana et al., 2013).

To get a list of genes to use as an input to ACSNMineR, we pooled mutations from all neuroblastoma patients; genes mutated at least once were included in the final list. Modules with a p-value lower than 0.01 after Benjamini-Hochberg correction were considered as enriched.

Statistical comparison of regression parameters

Regression parameters were found using a linear model in R. The z-score was calculated as recommended by (Paternoster et al., 1998):

$$Z = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

With μ the number of functional variant per module per variant and σ the standard error of the regression. The p-value is then computed with a two-sided option:

$$p = 2 \times P_{(X \leq |Z|)}$$

Where P is the probability of the normal distribution of mean 0 and standard deviation 1.

Data access

The whole-genome sequencing data have been deposited at the European Genome-phenome Archive (EGA) under accession number EGAS00001001184 for the French cases sequenced at BGI and under accession num-

ber EGAS00001001825 for the French cases sequenced at CNG. Sequence data for the US cases are available in the database of Genotypes and Phenotypes (dbGaP) under accession number phs000467.

QuantumClone is available at `https://github.com/DeveauP/QuantumClone/` and can be downloaded as an R package from the CRAN repository.

Acknowledgments

The authors would like to thank Elodie Girard for developing the variant calling pipeline and Pierre Gestraud for his help with statistical analysis of the data.
GS and her team were supported by the Annenberg Foundation and the Nelia and Amadeo Barletta Foundation. Funding was also obtained from SiRIC/INCa (Grant INCa-DGOS-4654) and from the CEST of Institut Curie. This study was also funded by the Associations Enfants et Santé, Association Hubert Gouin Enfance et Cancer, Les Bagouz à Manon, Les amis de Claire. VB and her team were supported by the ATIP-Avenir Program, the ARC Foundation (grant ARC - RAC16002KSA - R15093KS), Worldwide Cancer Research Foundation (grant WCR16-1294 R16100KK) and the "Who Am I?" laboratory of excellence ANR-11-LABX-0071 funded by the French Gouvernement through its “Investissement d’Avenir” program operated by The French National Research Agency (ANR) (grant ANR-11-IDEX-0005-02). EB was supported by the ABS4NGS project of the French Program ‘Investissement d’Avenir’. Sequencing of French samples was carried out in a collaboration of Institut Curie with CEA/IG/CNG financed by France Génomique infrastructure, as part of the program "Investissements d’Avenir" from the ANR (grant ANR-10-INBS-09).
JM and his team were supported in part by US National Institutes of Health grants RC1MD004418 to the TARGET consortium, and CA98543 and CA180899 to the Children’s Oncology Group. In addition, this project was funded in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views of policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

DISCLOSURE DECLARATION

We have no conflict of interest to declare.

References

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R., 2013. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, **0 7**:Unit7.20.

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O., and Barillot, E., 2012. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, **28**(3):423–425.

Bollet, M. A., Servant, N., Neuvial, P., Decraene, C., Lebigot, I., Meyniel, J.-P., Rycke, Y. D., Savignoni, A., Rigaill, G., Hupé, P., *et al.*, 2008. High-Resolution Mapping of DNA Breakpoints to Define True Recurrences Among Ipsilateral Breast Cancers. *Journal of the National Cancer Institute*, **100**(1):48–58.

Bruin, E. C. d., McGranahan, N., Mitter, R., Salm, M., Wedge, D. C., Yates, L., Jamal-Hanjani, M., Shafi, S., Murugaesu, N., Rowan, A. J., *et al.*, 2014. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*, **346**(6206):251–256.

Chen, G., Mulla, W. A., Kucharavy, A., Tsai, H.-J., Rubinstein, B., Conkright, J., McCroskey, S., Bradford, W. D., Weems, L., Haug, J. S., *et al.*, 2015. Targeting the Adaptability of Heterogeneous Aneuploids. *Cell*, **160**(4):771–784.

Deveau, P., Barillot, E., Boeva, V., Zinovyev, A., and Bonnet, E., 2016. Calculating biological module enrichment or depletion and visualizing data on large-scale molecular maps with ACSNMineR and RNaviCell packages. *The R Journal*, **8**(2):293–306.

Eleveld, T. F., Oldridge, D. A., Bernard, V., Koster, J., Daage, L. C., Diskin, S. J., Schild, L., Bentahar, N. B., Bellini, A., Chicard, M., *et al.*, 2015. Relapsed neuroblastomas show frequent RAS-MAPK pathway mutations. *Nature Genetics*, **47**(8):864–871.

Fischer, A., Vázquez-García, I., Illingworth, C. J. R., and Mustonen, V., 2014. High-Definition Reconstruction of Clonal Composition in Cancer. *Cell Reports*, **7**(5):1740–1752.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R., 2004. A census of human cancer genes. *Nature Reviews Cancer*, **4**(3):177–183.

Hajirasouliha, I., Mahmoody, A., and Raphael, B. J., 2014. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, **30**(12):i78–i86.

Hanahan, D. and Weinberg, R. A., 2011. Hallmarks of Cancer: The Next Generation. *Cell*, **144**(5):646–674.

Huang, D. W., Sherman, B. T., and Lempicki, R. A., 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, **37**(1):1–13.

Huang, D. W., Sherman, B. T., and Lempicki, R. A., 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4**(1):44–57.

Jiao, W., Vembu, S., Deshwar, A. G., Stein, L., and Morris, Q., 2014. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, **15**(1):35.

Kepler, T. B., 2013. Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. *F1000Research*, .

Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., *et al.*, 2013. Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science (New York, N.Y.)*, **342**(6154):1235587.

Koboldt, D. C., Larson, D. E., and Wilson, R. K., 2013. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, **44**:15.4.1–15.4.17.

Kuperstein, I., Bonnet, E., Nguyen, H.-A., Cohen, D., Viara, E., Grieco, L., Fourquet, S., Calzone, L., Russo, C., Kondratova, M., *et al.*, 2015. Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis*, **4**(7):e160.

Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**(14):1754–1760.

Malikic, S., McPherson, A. W., Donmez, N., and Sahinalp, C. S., 2015. Clonality Inference in Multiple Tumor Samples using Phylogeny. *Bioinformatics*, **31**(9):1349–1356.

Manning, C. D., Raghavan, P., and Schütze, H., 2008. *Introduction to information retrieval*. Cambridge University Press, New York.

Marusyk, A., Tabassum, D. P., Altrock, P. M., Almendro, V., Michor, F., and Polyak, K., 2014. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature*, **514**(7520):54–58.

McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R., and Mirny, L. A., 2013. Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences*, **110**(8):2910–2915.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.*, 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**(9):1297–1303.

Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., and Thomas, P. D., 2010. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Research*, **38**(suppl 1):D204–D210.

Miller, C. A., White, B. S., Dees, N. D., Griffith, M., Welch, J. S., Griffith, O. L., Vij, R., Tomasson, M. H., Graubert, T. A., Walter, M. J., *et al.*, 2014. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Comput Biol*, **10**(8):e1003665.

Molenaar, J. J., Koster, J., Zwijnenburg, D. A., van Sluis, P., Valentijn, L. J., van der Ploeg, I., Hamdi, M., van Nes, J., Westerman, B. A., van Arkel, J., *et al.*, 2012. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature*, **483**(7391):589–593.

Ng, P. C. and Henikoff, S., 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, **31**(13):3812–3814.

Paternoster, R., Brame, R., Mazerolle, P., and Piquero, A., 1998. USING THE CORRECT STATISTICAL TEST FOR THE EQUALITY OF REGRESSION COEFFICIENTS. *Criminology*, **36**(4):859–866.

Qiao, Y., Quinlan, A. R., Jazaeri, A. A., Verhaak, R., Wheeler, D. A., and Marth, G. T., 2014. SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. *Genome Biology*, **15**(8):443.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., and Shah, S. P., *et al.*, 2014. PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, **11**(4):396–398.

Schramm, A., Köster, J., Assenov, Y., Althoff, K., Peifer, M., Mahlow, E., Odersky, A., Beisser, D., Ernst, C., Henssen, A. G., *et al.*, 2015. Mutational dynamics between primary and relapse neuroblastomas. *Nature Genetics*, **47**(8):872–877.

Schwarz, R. F., Trinh, A., Sipos, B., Brenton, J. D., Goldman, N., and Markowetz, F., 2014. Phylogenetic Quantification of Intra-tumour Heterogeneity. *PLoS Computational Biology*, **10**(4):e1003535.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A., 2003. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Research*, **13**(9):2129–2141.

Tomasetti, C. and Vogelstein, B., 2015. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, **347**(6217):78–81.

Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A., and Sottoriva, A., 2016. Identification of neutral tumor evolution across cancer types. *Nature Genetics*, **48**(3):238–244.