COPY-NUMBER AWARE CLONAL-TREE RECONSTRUCTION USING SINGLE-CELL RNA SEQUENCE DATA

VITTORIO ZAMPINETTI

Master of Science
Data Science and Engineering
Politecnico di Torino

July 2021

Supervisors: Prof. Jens Lagergren
Ph.D. Seong-Hwan Jun
Prof. Mauro Gasparini
Prof. Alessandra Guglielmi

Vittorio Zampinetti: *Copy-number aware clonal-tree reconstruction using single-cell RNA sequence data,* Study on the integration of copy-number variation in PhylEx, a tool for accurate reconstruction of clonal structure via combined analysis of bulk DNA-seq and single cell RNA-seq data, © July 2021

ABSTRACT

Recent advancements in DNA and RNA sequencing technologies allow for higher resolution single cell and bulk data. Interest in research for cancer-driving mutations has therefore increased. The underlying models for such biological data are usually complex, typically showing hierarchical structures and still being subject to noisy and highly variable data. Here we propose an extension to a state-of-the-art method, PhylEx, which combines allelic imbalance data of somatic mutations both from scRNA-seq data and DNA bulk data to build a clonal tree of genetic mutations. The extension of the statistical model takes into consideration also the copy-number variation throughout the whole genome which helps in better explaining the data, providing a more accurate framework for Bayesian inference procedures.
More specifically, in addition to allelic imbalance data, we try to leverage on gene expression data which has been shown to be highly correlated with copy numbers. The devised model is partially based on the model described in Clonealign, which is basically a regression model on count data. The scope of the thesis is to investigate the new model, generate synthetic data, and show that true copy number profiles maximize the likelihood of the data. In a future work, it will be possible then to implement a sampling algorithm in order to perform inference on the new parameters of the model, potentially increasing the performance of the mutation tree generation already achieved by PhylEx.

ACKNOWLEDGEMENTS

I would first like to express my gratitude to my supervisors, Professor Jens Lagergren and Dr. Seong Hwan-Jun, whose expertise and support, both academic and motivational, have been fundamental. A special thank also for having been given the chance to work on such a valuable and interesting research topic.

I would also like to thank my family and my girlfriend for all the love and support given throughout all these years and during my experience abroad, in spite of the physical distance.

Last but not least, thank you to all of my friends and fellow students that helped me with their wise counsel, motivational pushes and happy distractions.

Turin, 2021
Vittorio Zampinetti

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

CNV	Copy Number Variation
DNA	Deoxyribonucleic acid
DP	Dirichlet Process
HGSOC	High-Grade Serous Ovarian Cancer
mRNA	messenger-RNA
NGS	Next Generation Sequencing
NHGRI	National Human Genome Research Institute
RNA	Ribonucleic acid
SNV	Single Nucleotide Variation
TSSB	Tree-Structured Stick-breaking Process
UMI	Unique Molecular Identifiers
WGS	Whole Genome Sequencing
ZINB	Zero-Inflated Negative Binomial
ZIP	Zero-Inflated Poisson

GLOSSARY

allele	One of the different forms of a gene. It can be normal or mutant. Alleles of the same gene can be either dominant or recessive. If both alleles are recessive, then the expressed phenotype of the individual changes with respect to the dominant one.
bin	Region of the DNA (or RNA) in which genes are clustered. The bins size can be fixed or vary depending on the genetic content.
break-point	Position in the genome that separates two sequences that have different copy numbers.
codon	Triplet of nucleotides in DNA or RNA which codifies for a specific amino acid. E.g. the codon CAA translates to Glutamine.
dropout	Event that occurs in single-cell RNA sequencing when a gene is observed at a low or moderate expression level in one cell but is not detected in another cell of the same cell type. Dropout events occur due to the low amounts of mRNA in individual cells and inefficient mRNA capture, as well as the stochasticity of mRNA expression [28].
genome	The complete set of genetic information in an organism.
genotype	The subsequent combination of alleles of an individual for a specific gene. A particular genotype determines the phenotype of an organism.
hypermutability	The state or condition of undergoing mutation at a high rate, or of being highly prone to mutation.
library size	The count of all reads that come out of the sequencer for the used library (referred to a sequencing experiment). In the context of differential expression normalization it could simply refer to the sum of the reads over the genes.

locus	A single position in the genome. Each locus can be referenced through two coordinates, namely the chromosome where it is located and the position of its associated base.
phenotype	The sum of an organism's observable characteristics. The phenotype is influenced by the genotype. Unlike the genotype, which requires the analysis of biological assays, the phenotype can be observed simply by looking at the organism's outward features and characteristics.
stop codon	Codon which notifies the termination of the translation process when generating a protein.
transcription	Biological process in which several enzymes, including helicase and topoisomerase, unwind DNA to provide access to another enzyme known as RNA polymerase. RNA polymerase travels along the unwound DNA strand to construct the mRNA molecule until it is ready to leave the nucleus.

1 INTRODUCTION

1.1 BACKGROUND

Cancer arises when cells grow and multiply uncontrollably and, starting from one tissue, spread over other parts of the body[1]. When cells contain faulty genetic information (DNA) the normal cell division behaviour can be altered and may eventually lead to tumors. Normally, cells with defected DNA are recognized by the organism and therefore eliminated before they turn cancerous. However, the body is not always able to do so, and sometimes these cells proliferate without being stopped.

In 1976, Nowell [25] proposed that genetically unstable cells evolve producing new variants and following a selection process similar (but not identical [13]) to Darwin's evolution of species. Most cancer cells are eliminated by immune cells, as just mentioned, but occasionally some have additional selective advantage, which allow them not only to survive, but also to generate new mutations, which might be even more resistant against an individual's immune system.

Nowell's theory and the related advancements in cancer research have led the way to more specific subareas, dedicated to the development of theory and methods for interpreting tumor evolution. The hereby presented work of research belongs to the thread of tumor phylogenetics, i.e. the scientific study of the relationship between cancer clones in a clonal tree structure which characterizes the evolutionary process. Studies in the last decades [32] show that it might indeed be possible to devise computational methods to reconstruct the evolutionary processes and therefore increase the level of understanding of the initiation and development of tumors.

1.2 PROBLEM STATEMENT

One peculiar characteristic of tumor evolution is that cancer cells present mutations at higher rate with respect to species evolution. This property is also called hypermutability [22]. Variants are of many kinds and recent methods predominantly focus on single nucleotide variations (SNV) and copy number variations (CNV) [32] (more on that in Section 2.1), defining models which leverage on that information to characterize the evolution of one or multiple tumors. This work indeed wants to expand the model of a state-of-the-art method, PhylEx [18], which uses SNVs to build a clonal tree that fits the given data. More specifically, the aim is to incorporate CNVs in the model in a way that this additional information could lead to greater performance in terms of accuracy.

1 *What Is Cancer?* - National Cancer Institute

In the context of Bayesian inference methods, reconstructing a clonal-tree, which describes the evolution of a tumor in a patient, from biological data $\mathcal{D}$, can be done by finding the tree $\hat{\mathcal{T}}$ which yields the maximum probability given the data:

$$\hat{\mathcal{T}} = \text{argmax}_{\mathcal{T}}\, p(\mathcal{T}|\mathcal{D}),$$

where $p(\cdot)$ is a complex probability distribution, main target of the probabilistic modelling process.

Although PhylEx method is shown to perform better than all the other methods for clonal-tree reconstruction, it leverages only on SNVs and not on CNVs. Since both variations are valuable sources of information when studying cancer evolution, a method that is able to make inference taking into account these two kinds of mutations together is likely to achieve more accurate results. In order to build such method, a new mathematical model has to be defined on top of the existing one.

1.3 PURPOSE

The work aims at further developing the project started with PhylEx, which has been carried out by Dr. Seong-Hwan Jun and Prof. Lagergren, supervisors of this degree project, and the rest of the research team, which features members from several institutions, including KTH Royal Institute of Technology.

The high level purpose of the thesis is therefore the same as in PhylEx, i.e. improve the state-of-the-art methods in tumor phylogenetics, acquire new insights on the evolution of cancer, thus paving the way for developing innovative medical treatments in cancer therapy.

1.4 GOALS AND RESULTS

The thesis will address the following main research questions:

1. How to incorporate copy-number variations in PhylEx model?
2. Is there a preferred model among alternative formulations?
3. Given the new model, can it drive an inference algorithm to the true copy-number values?

We answer Question 1 showing that Clonealign [6] offers a suitable model of the data given the analysis presented in Section 4.1. Then, we design a new model (Section 3.4) and consider some variations of it, in order to seek the most appropriate one; we address Question 2 by predicting copy number values using a maximum likelihood approach and we observe that models accounting for zero-inflation should be preferred (Section 4.2). Lastly, we answer Question 3 by carrying out a sensitivity analysis on the copy numbers with the likelihood as the target measure (Section 4.3), which finds maximum likelihood value for the true copy number values.

1.5 SCOPE AND DELIMITATIONS

This work proposes a novel probabilistic model for tumor evolutionary tree inference which extends the original model in PhylEx by including copy number variations and gene expression data. It brings under analysis various possible models of the data and compares them. Moreover, it provides a tool for generating synthetic data according to several parameters which have been found in literature and by extensively analyzing real data. Finally, it shows the impact of copy number variations on the log-likelihood of the model through a sensitivity analysis, which might be of value for a future implementation of a sampling technique for Bayesian inference.

However, designing, implementing and testing a specific inference algorithm for the new model is out of the scope of this work. As a consequence, a comparison between the original PhylEx model and the new one in terms of accuracy in the clonal-tree reconstruction will not be addressed.

1.6 OUTLINE

In Chapter 2 I introduce all the most relevant concepts that are necessary for understanding the present study. Then, in Chapter 3 the original model and the extension proposed are described, along with the adopted tools and methodologies. In Chapter 4 I show the results obtained. Lastly I elaborate on the meaning and validity of the results in Chapter 5, concluding with a few comments and reflections in Chapter 6.

1.7 ABOUT THE THESIS

The simulation software is written in C++ on top of PhylEx and is available at

`https://github.com/junseonghwan/PhylEx/tree/copy-number`,

while all the data analysis software has been coded in R and can be found at

`https://github.com/toyo97/cn-phylex-analysis`.

2 BACKGROUND

In this section I first present a brief introduction to genomics, which is fundamental to understand the biological concepts embraced in this work. Then I elaborate on the basic principles of Bayesian methods and statistical modelling. Finally, I summarize some related methods to give a brief overview of the state-of-the-art approaches.

2.1 GENOMICS

2.1.1 *DNA and RNA*

Deoxyribonucleic acid (DNA) is a molecule composed by a long double stranded chain of base units called *nucleotides*. Each nucleotide contains one of the four nucleobases (cytosine, guanine, adenine or thymine) resulting in a sequence of alternating bases whose order defines the unique characteristics of the individual. The two separate strands are connected together according to the two base pairing rules (A with T and C with G) and form the so-called "double helix" structure. In eukaryotic cells, category to which animals, and therefore humans belong, DNA is organized in structures called *chromosomes* that are stored within the cell nucleus. In Figure 1 a schema of the DNA structure along with the base-paring rules is shown.

In human cells, each DNA molecule is formed by 23 pairs of chromosomes, with one pair being the X/Y sex chromosomes (see Figure 2). The length of a human genome is typically around 3 billion bases (or 6Gb if we consider also the complementary sequence), which constitutes an extremely large amount of genetic information. This information is used by the cells as instructions saying which protein to generate, hence letting them provide the necessary components for an organism to grow and survive.
The information required to build a specific protein is coded in a subsequence of DNA called *gene*. More than 99.9% of the human genome is common to all humans, while the rest 0.01% determines the individual traits such as eyes color, skin tone etc., but also the risk of developing diseases and the responses to medications[1]. The number of genes in the human genome is around 20 and 25 thousands, all divided into the 23 pairs of chromosomes. However, not all sequences in a gene codify for proteins: the protein-coding sequences are called *exons*, but most of the gene is constituted of sequences which contain no coding information, and these are called *introns*.

Proteins are not generated directly from the DNA, but rather from a copy of its information that is transferred to the RNA during transcription. The RNA is a single-stranded molecule which, among other

1 *Introduction to Genomics* - National Human Genome Research Institute

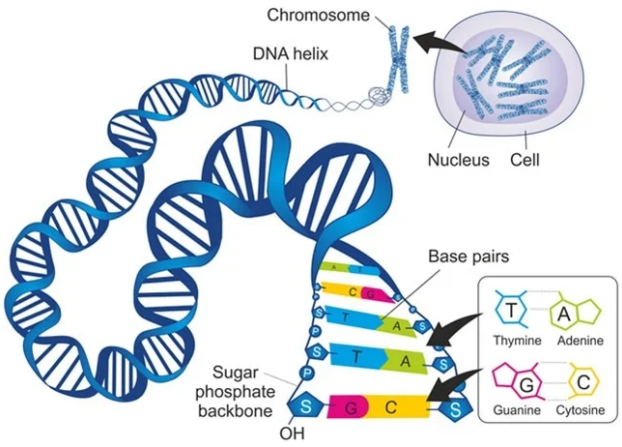


Figure 1: Schematic illustration of the DNA structure, as a double-stranded helix, forming the chromosome. Image credits to Soleil Nordic, Shutterstock

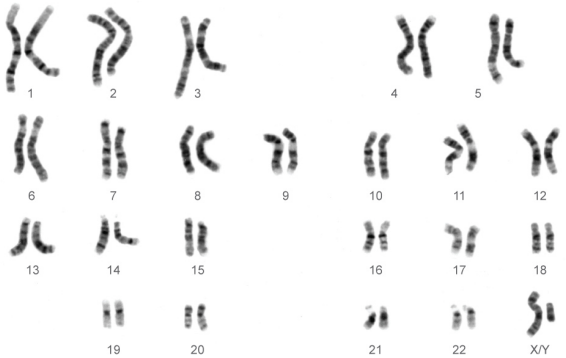


Figure 2: Karyotype of human male picturing the 22 chromosomes plus the sex chromosomes. Note that the first chromosomes are significantly larger than the last ones, therefore containing more genetic information. Image credits to NHGRI.

... CAT CAT CAT CAT ...
... CAT CAT C**C**T CAT ...

Figure 3: Example of SNV, the adenine nucleotide in the original sequence has been replaced by a cytosine nucleotide.

functions, is responsible for the generation of the appropriate proteins. In fact, the RNA (more precisely, the mRNA) filters out all the non-coding part of DNA while keeping only the exons, and then it is translated into amino acids by the ribosome[2].

2.1.2 *Mutations: SNV and CNV*

Somatic mutations can occur in various forms. The simplest mutation is the single nucleotide variation (SNV) and consists of an alteration of a nucleotide base (e.g. see Figure 3). Sometimes, a certain SNV can be more common and qualify as SNP (single nucleotide polymorphism), if it is present in at least 1% of the population. Since particular sequences of bases codify for specific amino acids, and therefore proteins, an SNV might result in a different instruction for the ribosome. More specifically, an SNV can lead to one of the following cases:

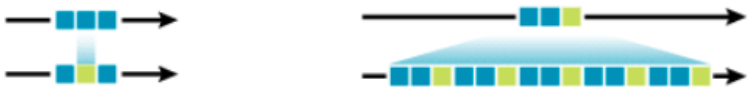
- the SNV produces a sequence which codifies for the same amino acid as before, having no effect during the translation process (e.g. both the codons CGC and CGA codify for Arginine). This is called *synonymous change*;
- the SNV leads to a codon that codifies for a different amino acid, then it is called *non-synonymous change*. Even in this case, the mutation might not result in a pathogenic variant;
- the SNV produces a stop codon, prematurely stopping the translation process. This is called *nonsense variant*.

While SNVs consist of a change of a single base, a mutation can generally involve more nucleotides at once, and this is not restricted to a change of the nucleotide, but it can also consist of an insertion (or deletion) of one or more bases. These mutations are also called *indels*. When an indel is larger than 50b, it is called *structural variation*. In particular, a structural variation in which a portion of DNA (or RNA) is duplicated or deleted) is known as *copy number variation*.

Recently, copy number variations have been widely studied along with SNVs as it has been generally accepted that these variations are highly present in tumors [5, 37] and they have been shown to affect the gene expression [35].

According to the tumor evolution theory introduced by Nowell [25], cancer cells arise from healthy cells that accumulate mutations. In this setting, it is possible to construct an evolutionary tree in which each node represents a *clone*. A clone is a population of cells that share

2 *What Is RNA?* - News Medical Life Sciences



(a) SNV (b) CNV

Figure 4: Comparison between SNV and CNV at DNA sequence level. Note that in CNVs (b) the repeated sequence is composed by a large number of nucleotides, while SNVs (a) involve only one nucleotide.

some characteristics, such as a set of mutations, and the root node is generally identified as the healthy population. Also, the prevalence of each clone in the individual indicates its growth and fitness. Knowing the clonal population structure of a tumor is therefore a great source of information for cancer diagnosis and treatment.

2.2 BIOLOGICAL DATA

Collecting genomic data is a notably expensive and complex task; besides, obtaining high-quality data is even more challenging. Nevertheless, in the recent years, next generation sequencing (NGS) technologies has led to faster and higher quality data collection, making large-scale whole genome sequencing (WGS) accessible and practical for the average researcher [2].
Through DNA sequencing it is possible to extract the sequence of nucleotides in the DNA and therefore recognize eventual SNVs, CNVs and other mutations. There are several methods and technologies that perform sequencing and they usually follow a complex pipeline which starts from the treatment of the tissue containing the genetic information, which then passes through an amplification process and is matched with a sequencing library to eventually obtain the *reads*. A detailed description of the sequencing process is beyond the scope of the present work.

Sequencing can be performed both on DNA and RNA. As mentioned in Section 2.1, the DNA is situated, unchanged, in all cells of an organism, while RNA molecules are transcriptions of DNA containing the information used by the cells. This means that DNA sequencing provides us with a static picture of the cells behaviour, whereas RNA sequencing is more likely to give us information about what the cells are actually doing. None of them is a better source of information when analyzing the state of a tumor, in fact these two processes are dependent upon and inform each other.
Moreover, sequenced data can be obtained either from sets of different cells put all together (*bulk data*) or from single cells taken separately (*single-cell data*). Again, each of the two data types has its strengths and weaknesses: bulk data are generally of higher resolution and less noisy while single-cell data present regulatory elements specific to one cell type useful to determine how genes are differentially expressed across cells, although they feature high rates of

ID	b	d	M	m
s0	73	150	7	0
s1	101	198	4	3
s2	109	369	4	3

(a) bulk DNA

gene_id	c1	c2	c4	c5
ENSG00000223972	0	0	0	0
ENSG00000227232	0	11	16	5
ENSG00000241860	0	9	23	3

(b) scRNA

Figure 5: First lines of the (a) bulk DNA allelic-imbalance dataset and the (b) scRNA gene expression dataset. In (b) only four out of the 360 cells are shown.

dropout[3] and few reads per cells [14]. After sequencing DNA (or RNA) several types of information can be extracted from the genomic sequence. Here I present the two data-types that have been taken under analysis in this work, namely allelic-imbalance of somatic mutations and gene expression data. The former, which I will just refer to as allelic imbalance data, deal with SNVs, and present, for each single position in the genome, the number of reads of the variant nucleotide against the total number of reads of that position. The latter feature the number of reads for each gene regardless the variations, which represents how much a gene is expressed.

2.2.1 *The datasets*

The datasets available for the presented work have been obtained from a culture of three ovarian cancer cell-lines, all originating from the same patient. For more details on the origin of the datasets and the preparation process, see PhylEx description [18].
The main datasets are three: bulk DNA allelic imbalance data, single-cell RNA allelic imbalance data, and single-cell RNA gene expression data.

Figure 5 shows an example of two of the above mentioned datasets. The third, scRNA allelic imbalance dataset, is equivalent to the bulk DNA with the additional distinction between different cells.
The allelic imbalance dataset features four attributes for each position in the genome n (from now on, also called *locus*): the variant reads count b_n, the total reads count d_n, the major copy number M_n and the minor copy number m_n. The ID of each locus maps to the coordinates in the genome, which are written in terms of chromosome and base position (e.g. SNV s0 is located in chromosome 1 at position 33282970). The mappings between SNV IDs and coordinates are stored in a separate file.
The single-cell RNA dataset is instead in the form of a matrix (sometimes referred to as UMI-count matrix) with the genes row-wise and cells column-wise. The gene IDs refer to the *Ensembl* genes database, one of the most used gene databases. In particular, the IDs belong to the Genome Reference Consortium Human Build 37 (GRCh37),

3 see Glossary for definition of *dropout*

also called simply *hg19*[4]. Each Ensembl gene ID is associated to a set of coordinates. Unlike the SNV IDs the mapping is not relative to the dataset, but refers to the standard just mentioned, and since each gene is a sequence of multiple nucleotides, the coordinates are now three: chromosome, start position and end position (e.g. the gene ENSG00000227232 is located in chromosome 1 between positions 14363 and 29806).

2.3 BAYESIAN INFERENCE

In the last two decades, many methods have been developed and tested in tumor phylogenetics for estimating the parameters of the evolutionary trajectories of cancer cells, such as combinatorial optimization methods (like ILP), maximum likelihood, etc. [32]. More recently, the field moved towards more sophisticated probabilistic methods like Bayesian sampling which better handles the noisy nature of the data and gives more insights on the uncertainty in trees inference, although they can be more computationally demanding compared to the other approaches.

In this section I briefly introduce the Bayesian approach for inference since PhylEx, and other methods upon which this work is based, perform machine learning with probabilistic models.

2.3.1 *Maximum A Posteriori (MAP)*

In any machine learning method, we can identify three main concepts: a model, the data and the learning algorithm. The goal is then to design an algorithm with which it is possible to automatically find some parameters $\hat{\theta}$ of a model $\mathcal{M}$ such that it fits the given dataset $\mathcal{D}$. In parametric Bayes the model is a probabilistic model, and the learning process, also known as inference, is performed using Bayes' rule:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

where

- $p(\theta|\mathcal{D})$ is called the posterior distribution;
- $p(\mathcal{D}|\theta)$ is the likelihood of the data given the parameters;
- $p(\theta)$ is the prior distribution over the parameters;
- $p(\mathcal{D})$ is the marginal distribution for the data.

Most of the times the marginal distribution $p(\mathcal{D})$ is not considered as its computation is typically intractable. In fact, we can also simply write

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta).$$

4 Human Genome GRCh37, Ensembl archive

In the case of *maximum a posteriori* (MAP) inference, the goal is then to find the point estimate of the parameters that maximizes the posterior distribution, hence

$$\hat{\theta}_{MAP} = \text{argmax}_\theta \, p(\theta|\mathcal{D}) \tag{1}$$
$$= \text{argmax}_\theta \, p(\mathcal{D}|\theta)p(\theta) \tag{2}$$

which is, intuitively, the most likely parametrization that the model $\mathcal{M}$ would have when provided with the dataset $\mathcal{D}$.

2.3.2 *Sampling methods*

There are several ways or finding the point estimate $\hat{\theta}$ in MAP inference. Sometimes it is possible to derive a closed form solutions (exact inference), especially if the prior distribution $p(\theta)$ is conjugate for the likelihood distribution $p(\mathcal{D}|\theta)$, which means that the posterior distribution belongs to the same family of the prior.
When exact inference is intractable, it is necessary to resort to some kind of approximation. One way of obtaining this approximation is by *numerical sampling*.

Sampling methods, also known as *Monte Carlo* techniques, aim at finding an approximation of the posterior distribution by drawing a large amount of samples from it. This way it is possible, for instance, to compute the expected value of the posterior distribution simply by taking the average of the samples. To obtain an approximation of the MAP parameters it is enough to draw a sufficiently large amount of samples from the posterior and then take the one that reaches maximum value of $p(\theta|\mathcal{D})$.
However, it is worth to mention that numerical sampling allows us to learn *all* information about the posterior, not just obtaining a point estimate. For example, one could also plot an histogram of the samples, obtaining an approximation of the density function of the posterior distribution.

2.4 NONPARAMETRIC BAYES

In parametric Bayesian inference, when we deal with distributions over structured data, we need to fix the structure in advance. For example, in the case of distribution over an evolutionary tree, we would have to define first what is the size and topology of the tree, and then assign prior probabilities to each node of the tree, which means a limited number of parameters. Fixing the number of parameters is possible if the tree structure can be observed, however it is not always the case.
Indeed, in tumor phylogenetics inferring the unobserved structure of the evolution process is one of the main goals. To make Bayesian inference possible in this case, models have to adopt a *nonparametric* Bayesian approach to obtain more flexible priors, with infinite-dimensional structures. Due to their flexibility, nonparametric models have gained considerable popularity in the field of machine learning,

especially in *unsupervised learning*.
The idea is to define not just a distribution over the data, but a distribution over distributions (or distributions over probability measures). The structure is then not specified *a priori*, but rather it is determined from the data. Note however that the fact it is called "nonparametric" does not mean that the model lacks parameters.
To make a simple example, an histogram is a nonparametric estimate of a probability distribution: although we have to define the width of each bin in advance (which is a parameter of the model), the number of bins for an histogram depends on the nature of the data[5].

2.4.1 *Dirichlet process*

The Dirichlet process, introduced by Ferguson in 1973 [10], is one of the models used in Bayesian nonparametric statistics.
The name comes from the Dirichlet distribution, a family of continuous multivariate distributions parametrized by $\boldsymbol{\alpha}$ which describes $K \geqslant 2$ variables $X_1, ..., X_2$ such that each $x_i \geqslant 0$ and $\sum_i x_i = 1$. In fact, the Dirichlet process is a generalization of this family of distributions in which $K \to \infty$. Given this last definition, it might seem unclear how it is possible to describe a distribution on an infinite dimensional space. Instead of defining a density function, it is enough to define an algorithm for drawing samples from the distribution.

More specifically, following the constructive approach to the Dirichlet process (DP) by Sethuraman [34], denoting by Θ the continuous parameter space on which we define the DP, we can draw a random probability measure G with parameters H, α using a sequence of beta random variables[6]. The base measure H is a distribution over the same space Θ that serves like a mean $\mathbb{E}[G(A)] = H(A)$ (e.g. a Gaussian over the real line), and the strength parameter $\alpha > 0$ serves like inverse variance $\mathrm{Var}[G(A)] = \frac{H(A)(1-H(A))}{\alpha+1}$, for A as any subset of Θ. Furthermore, it can be proved that any finite partition $(A_1, A_2, ..., A_n)$ of the parameter space Θ is such that $(G(A_1), G(A_2), ..., G(A_n))$ is Dirichlet distributed.
Sethuraman's approach can be viewed as a way of breaking a stick of unitary length (hence, the alternative name *stick-breaking process*) in an unlimited number of pieces and is described as follows:

1. draw a sample from a Beta distribution $\nu_i \sim \mathrm{Beta}(1, \alpha)$;
2. compute the actual length of the stick normalizing over the previous Beta variables

$$\pi_i = \nu_i \prod_{i'=1}^{i-1} (1 - \nu_{i'})$$

5 A more detailed description of histograms in the context of nonparametric Bayesian inference can be found in [3]

6 The Beta distribution is continuous and describes a r.v. with support $[0, 1]$ given two shape parameters

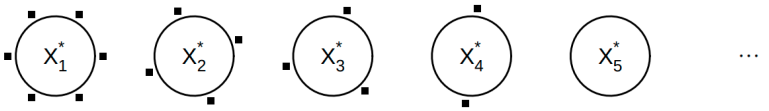


Figure 6: Illustration of the *Chinese restaurant process*. A new person arrives and either sits at a table with people or sits at a new table. The probability of sitting at a table is proportional to the number of people at the table.

if $i = 1$, then $\pi_i = \nu_i$;

3. draw a value $\theta_i \sim H$, with each θ_i being i.i.d;
4. repeat infinitely many times.

The outcome of this process is the infinite discrete distribution over the continuous space Θ:

$$G := \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}.$$

Note that, while H is a continuous distribution and thus the probability that any two samples are equal is precisely zero, G is made up of countably infinite number of masses, therefore there is a positive probability that two samples collide. This property implies the creation of an implicit clustering on samples drawn from G.

An effective way of representing a DP can be viewed in the popular analogy of the so-called "Chinese restaurant process" [1] in which the N^{th} customer is assigned to the table j with probability $N_j/(N+\alpha-1)$ or assigned to a new (empty) table with probability $\alpha/(N+\alpha-1)$, where N_j is the number of customers at table j, which implies $N = 1 + \sum_j N_j$. Each customer is a sample drawn from the Dirichlet process, the number of tables is potentially infinite and each table corresponds to one the θ_i *atoms* that build the G distribution.

In practice, when applying this model to a finite set of data, each datum is then assigned to its cluster according to the *urn-based* scheme just described, i.e. sequentially positioning the data in the segments according to a rule. Equivalently, for each datum, the cluster can be sampled from G using the π_i probabilities, segments of the unitary length stick in the *stick-breaking process*. Once all data have been assigned, the result is an implicit grouping (or clustering) of the data where the number of clusters is not defined in advance and, in case of data streams, could potentially grow indefinitely.

2.4.2 *Tree-structured stick-breaking process*

The *tree-structured stick-breaking process* [12], or TSSB, is an extension of the Dirichlet process in which the structure is not linear but takes the form, indeed, of a tree. The cluster are then related through an ancestry. This is achieved by interleaving two parallel stick-breaking

processes as shown in Figure 7 (b): the first determines the size of the partition reserved to each node, as a function of depth, the second determines the branching probabilities.
The process therefore makes use of two sequences of Beta variates: denoting with $|\epsilon|$ the depth of node ϵ, $\nu_\epsilon \sim \text{Beta}(1, \alpha(|\epsilon|))$ is the relative (unnormalized) length of the stick that is reserved to the node itself, that is the proportion of the probability mass allocated to ϵ versus the mass allocated to the descendants, while $\psi_\epsilon \sim \text{Beta}(1, \gamma)$ defines the splitting of the probability mass between the children of node ϵ.

Similarly to a DP, we can describe this process with the following steps:

1. ν-break: draw $\nu_\epsilon \sim \text{Beta}(1, \alpha(|\epsilon|))$ which defines the portion of the current stick that is allocated to node ϵ, while $(1 - \nu_\epsilon)$ is the portion of the stick that will be split between its children;

2. ψ-break: draw $\psi_{\epsilon 1} \sim \text{Beta}(1, \gamma)$ which selects the portion of the previous $(1 - \nu_\epsilon)$ stick that goes down to the first child. The remaining part is recursively split with other ψ-breaks, generating more siblings.

3. Given the current node ϵ, compute the probabilities of a certain sequence of children with the formula

$$\varphi_{\epsilon\epsilon_i} = \psi_{\epsilon\epsilon_i} \prod_{j=1}^{\epsilon_i - 1} (1 - \psi_{\epsilon j})$$

   where $\epsilon\epsilon_i$ denotes the i^{th} children of ϵ;

4. compute the absolute probability measure assigned to the current node as

$$\pi_\epsilon = \nu_\epsilon \varphi_\epsilon \prod_{\epsilon' \prec \epsilon} \varphi_{\epsilon'} (1 - \nu_{\epsilon'})$$

   where $\epsilon' \prec \epsilon$ denote the ancestors of ϵ. In case of the root node, $\pi_0 = \nu_0$;

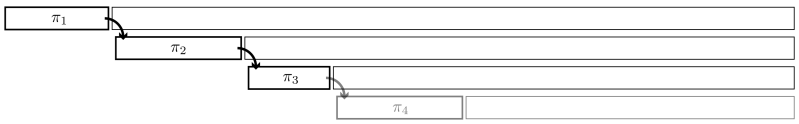
5. repeat infinitely many times.

Again, the process goes on indefinitely generating a tree of unlimited depth and width. The actual tree structure induced by a finite set of data is created by assigning each datum to the nodes of the tree. This can be achieved through a slight modification of the Chinese restaurant process introduced in Section 2.4.1.

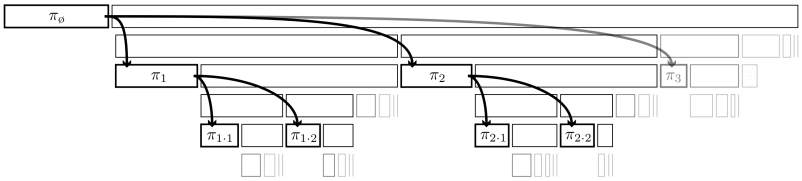
In particular, a datum (i.e. a customer, following the metaphor) is assigned to node ϵ with probability

$$\frac{N_\epsilon + 1}{N_\epsilon + N_{\epsilon\prec\cdot} + \alpha(|\epsilon|) + 1} \tag{3}$$

where N_ϵ is the number of data already assigned to node ϵ and $N_{\epsilon\prec\cdot}$ the number of data that came down this path but did not stop at ϵ, i.e. the sum over all descendants.
If the datum does not stop at ϵ, it goes down the tree, choosing



(a) Dirichlet process stick breaking



(b) Dirichlet process stick breaking

Figure 7: Illustration and comparison of (a) the Dirichlet process with linear partitioning and (b) its tree-structured extension seen as *stick-breaking* processes. In (b) rows 1, 3 and 5 are ν-breaks (depth); rows 2 and 4 are ψ-breaks (width). Image credits to [12]

the child (and therefore the path) with a classic Chinese restaurant process where the previous customers are only the data who have reached this point. Precisely, it descends to child $\epsilon\epsilon_i$ with probability

$$\frac{N_{\epsilon\epsilon_i} + N_{\epsilon\epsilon_i \prec \cdot}}{N_{\epsilon \prec \cdot} + \gamma}$$

or creates a new child with probability $\gamma/(N_{\epsilon \prec \cdot} + \gamma)$.

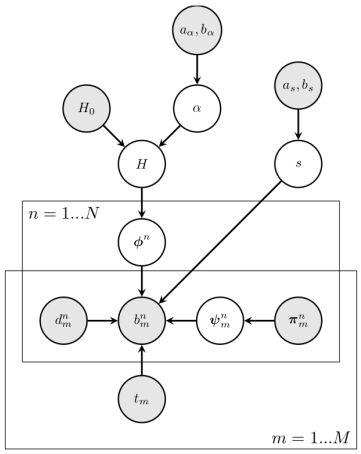
Note that unlike the DP, the strength parameter α is a function of the depth. More specifically, $\alpha(\cdot) : \mathbb{N} \to \mathbb{R}^+$ and must satisfy $\sum_{j=1}^{\infty} \ln(1 + 1/\alpha(j-1)) = +\infty$ [15]. With this additional flexibility, it is possible to put most of the probability mass at an intermediate depth; this is done, for instance, with a function like $\alpha(j) = \lambda^j \alpha_0$, where $\lambda \in (0, 1]$ is the decay parameter, regulating the depth of the tree.

2.5 RELATED METHODS

In this section I present a few methods developed in the past and attempt to give an overview of the state-of-the-art tools that perform inference with tumor phylogenetics data.

PYCLONE PyClone [31] is a statistical tool that performs inference of clonal population structures in cancer. It models bulk DNA allelic imbalance data (as described in Section 2.2) in order to cluster SNVs in different evolution clones and estimate their cellular prevalence, that is the fraction of cancer cells.

Cellular prevalences (portion of cells belonging to a clone) are modeled with a Dirichlet process so to let the number of clones be inferred along with the other parameters. However, the DP does not build a tree-structured phylogeny, unlike the TSSB process in PhylEx (see Section 3.2).



$$
\begin{aligned}
\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) \\
H_0 &= \text{Uniform}([0,1]^M) \\
H|\alpha, H_0 &\sim \text{DP}(\alpha, H_0) \\
\phi^n|H &\sim H \\
\psi_m^n|\pi_m^n &\sim \text{Categorical}(\pi_m^n) \\
\psi_m^n &= (g_{m,\text{N}}^n, g_{m,\text{R}}^n, g_{m,\text{V}}^n) \\
&\text{either} \\
b_m^n|d_m^n, \psi_m^n, \phi_m^n, t_m &\sim \text{Binomial}(d_m^n, \xi(\psi_m^n, \phi_m^n, t_m)) \\
&\text{or} \\
s|a, b &\sim \text{Gamma}(a_s, b_s) \\
b_m^n|d_m^n, \psi_m^n, \phi_m^n, t_m, s &\sim \text{BetaBinomial}(d_m^n, \xi(\psi_m^n, \phi_m^n, t_m), s) \\
&\text{where} \\
\xi(\psi, \phi, t) &= \tfrac{(1-t)c(g_\text{N})}{Z}\mu(g_\text{N}) + \tfrac{t(1-\phi)c(g_\text{R})}{Z}\mu(g_\text{R}) + \\
&\quad \tfrac{t\phi c(g_\text{V})}{Z}\mu(g_\text{V}) \\
Z &= (1-t)c(g_\text{N}) + t(1-\phi)c(g_\text{R}) + \\
&\quad t\phi c(g_\text{V})
\end{aligned}
$$

Figure 8: Directed graphical model of PyClone. Image credits to [31].

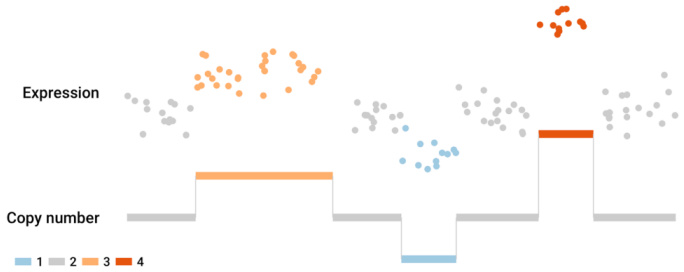


Figure 9: An illustration of the dependency between gene expression of single cells and their copy number signal. Image credits to [6].

CLONEALIGN Clonealign [6] models single-cell RNA expression data along with copy number data in order to perform inference on the cell assignments to the tumor clones. The idea behind the model is that an increase in the copy number for a certain gene, will lead to a proportional increase in that gene's expression level, property that has been also shown in other studies [8, 24, 35].

The model is based on a formula that links the expression value y_{cg} of a gene g in cell c to the copy number λ_{gv} of that gene in the clone v to which the cell belongs. We can define each cell assignment as $\zeta_c = v$ and write Clonealign formula as follows[7]

$$\mathbb{E}[y_{cg}|\zeta_c = v] = s_c \frac{\mu_g \times \lambda_{gv} \times e^{w_{cg}}}{\sum_{g'=1}^{G} \mu_{g'} \times \lambda_{g'v} \times e^{w_{cg'}}} \tag{4}$$

where s_c is the total read depth size of cell c, i.e. the total number of reads, which tells how much information there is for a single cell; μ_g is the *per-copy expression* for gene g, that is how much a certain gene expression depends on the copy number. w_{cg} is simply an additional factor that accounts both for the structured noise (not all expression values are necessarily explained in terms of the copy number) and for

7 The presented formula is slightly different from the original as it lacks some details that are not relevant for the purpose of this summary.

the known covariates (if any).
The gene expression is assumed to follow a negative binomial distribution, i.e.

$$y_{cg} \sim \text{NegativeBinomial}(m_{cg}, r_g)$$

with parametrization by means of the mean m_{cg} (defined in Eq. (4)) and the gene-specific inverse of dispersion parameter r_g.

Inference in Clonealign is performed with Variational Bayes, a statistical method which approximates the posterior through the minimization of the evidence lower bound (ELBO) [4]; nevertheless, since the model is not conjugate, a sampling method is still required to compute the ELBO. The overall inference machine takes the form of a Variational Auto-Encoder [27]. The optimization process finds estimates of the per-copy expression μ and the cell assignments to the clones ζ.

COPYKAT CopyKAT [11] is a recent statistical tool that infers copy number profiles from scRNA-seq data by integrating a Bayesian and hierarchical clustering methods.
Briefly, it filters the relevant cells and genes in the input UMI-count matrix and smooths the expression values. Then cells are grouped in several clusters through hierarchical clustering with ward linkage. The cluster with minimum variance is identified as the *confident normal cell cluster*. This cluster is then used as a baseline for determining the relative gene expression of cancer cells, i.e. those cells that belong to the other clusters.

CopyKAT is shown to perform better than previous tools like inferCNV [26]. In fact, previous methods have been designed for first-generation scRNA-seq data with lower cell throughput, and they are not suitable for data acquired using new high-throughput scRNA-seq technologies.

3

METHODS

3.1 OVERVIEW

In this chapter I present the main method related to this work, that is the original PhylEx model; then I describe the tools used and the steps performed for the exploratory data analysis; after that, I define two alternative extensions of the PhylEx model, illustrating how they are related to each other; lastly, I report the process for synthetic data generation that has been developed.

3.2 EXISTING METHOD

PhylEx [18] is a Bayesian statistical tool that combines scRNA-seq and DNA bulk data to reconstruct a tumor evolution clonal-tree and estimate cell assignments to the clones.

The tool has been shown to outperform other current state-of-the-art methods in clonal-tree reconstruction, namely Canopy [17] and PhyloWGS [9]. This is mainly due to the fact that it incorporates single-cell data in the probabilistic model in order to better separate the sub-clonal lines and therefore infer non-linear trees. Since each cell in a clone inherits every SNV from the ancestor clones, and not those of the other clones, co-occurrence of mutations in single cells helps separating clones with different sets of SNVs. For example, looking at Figure 10, we can deduce that the cells in the light-blue clone share some mutations with both blue and yellow clones, but they also have some new mutations that distinguish them from the others; at the same time, the yellow clone cells might have mutations not present in the cells of the light-blue one. This distinction cannot be achieved with only bulk data because we can only see that the set of mutations assigned to the light-blue node is less frequent than the others, and therefore should belong to a child node.

3.2.1 *Probabilistic model*

The model of PhylEx is built around two datasets, namely the bulk DNA and the scRNA-seq allelic imbalance datasets. The former is defined as a set of SNVs data $\mathbf{B} = \{(b_n, d_n, M_n, m_n)\}_{n=1}^{N}$ where b_n and d_n are, respectively, the variant and total reads of the SNVs at locus n; M_n, m_n are the major and minor copy numbers and N the total number of SNVs in the dataset. Single-cell data $\mathbf{S} = \{\{b_{cn}, d_{cn}\}_{n=1}^{N}\}_{c=1}^{C}$ contains the variant and total reads for each locus n and for each single cell c, from a set of C cells.

To better explain the meaning of major and minor copy numbers, let us consider an example. If an SNV in locus n is associated to

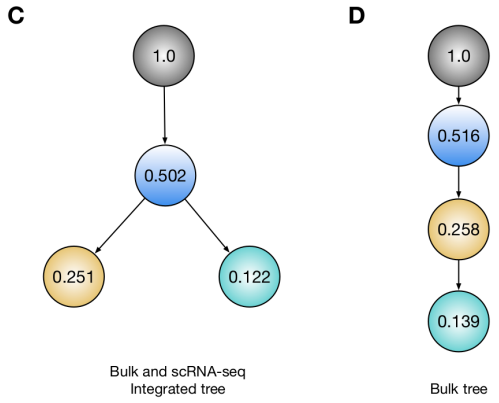


Figure 10: Non-linear (**C**) and linear (**D**) evolutionary tree example. Cellular prevalences of the clones are written inside the nodes. Image credits to [18].

$M_n = 2$ and $m_n = 1$, then there are three genotypes compatible with it: two in which the variant belongs to the major copy $(BB, A), (BA, A)$ and one with the variant in the minor copy (AA, B). The major and minor copy numbers therefore tell us that $(AAA, -)$ and $(BBB, -)$ are not compatible, even though the total copy number is still three. Note that the tuple notation represents the maternal and paternal copies; however, we have no information on whether the paternal copy is the major and the maternal is the minor or vice versa.

The target variables for inference in PhylEx are the clonal-tree structure T, the cellular prevalence for each node v of the tree $\boldsymbol{\phi} = \{\phi_v\}_{v=1}^{V}$, and the SNVs assignment to the nodes $\boldsymbol{z} = \{z_n\}_{n=1}^{N}$, where $z_n \in [V] = \{1, ..., V\}$. The likelihood of the bulk and single-cell data is assumed to be conditionally independent given the aforementioned variables:

$$l(\mathbf{B}, \mathbf{S}|T, z, \boldsymbol{\phi}) = l(\mathbf{B}|T, z, \boldsymbol{\phi}) l(\mathbf{S}|T, z, \boldsymbol{\phi}). \tag{5}$$

Then, the posterior distribution can be expressed as follows

$$\pi(T, z, \boldsymbol{\phi}|\mathbf{B}, \mathbf{S}) \propto l(\mathbf{B}|T, z, \boldsymbol{\phi}) l(\mathbf{S}|T, z, \boldsymbol{\phi}) \pi_0(\boldsymbol{\phi}|T, z) \pi_0(z|T) \pi_0(T), \tag{6}$$

where $\pi_0(\cdot)$ denotes the prior distribution. More specifically, the prior distribution on the tree T is given by the tree-structured stick-breaking process (TSSB) (explained in Section 2.4.2). Please note that, due to this prior, the number of nodes V is not pre-defined, but rather it is a random variable related to T.
The prior on the SNV assignments under T follows the urn-scheme introduced in the same section of Chapter 2, in particular an SNV is assigned to a node with probability given by Eq. (3) with $\alpha(v) = \alpha_0 \lambda^{|v|}$.
The prior distribution on the cellular prevalences is expressed through a Dirichlet distribution on a set of derived variables η_v called *clone fractions* and defined as

$$\eta_v = \phi_v - \sum_{v' \in \kappa(v')} \phi_{v'}, \tag{7}$$

where $\kappa(v)$ is the set of children nodes of v. This leads to a multivariate $\boldsymbol{\eta} = \{\eta_v\}_{v=1}^{V}$ such that $\sum_v \eta_v = 1$ and $\eta_v \geqslant 0$, upon which it is possible to place a Dirichlet prior, conditioned on the tree T.

Assuming site independence given the tree and the assignments of the SNVs, the joint likelihood of the bulk data factorizes as

$$l(\mathbf{B}|T, z, \boldsymbol{\phi}) \propto \prod_{n=1}^{N} P(b_n|T, z_n, \phi_{z_n}, d_n, M_n, m_n), \tag{8}$$

where the likelihood of each site is obtained marginalizing over the possible genotypes $\mathcal{G}(M_n, m_n) \subset \{A, B, AA, AB, BB, AAA, ...\}$ that are compatible with M_n and m_n. Precisely we have

$$P(b_n|T, z_n, \phi_{z_n}, d_n, M_n, m_n) = \sum_{g_n \in \mathcal{G}(M_n, m_n)} P(b_n|d_n, g_n, \phi_n) P(g_n), \tag{9}$$

where the genotype follows a Binomial distribution over the variant and the total copy number, normalized over all the genotypes with at least one variant, and the variant reads given the genotype is also modeled with a Binomial distribution

$$b_n|d_n, g_n, \phi_n \sim \text{Binomial}(d_n, \psi(g_n, \phi_{z_n}, \epsilon)), \tag{10}$$

with d_n as the number of trials of a Bernoulli experiment with probability of success (i.e. probability of seeing a variant) equal to $\psi(g_n, \phi_{z_n}, \epsilon)$. If we denote by ϵ the sequencing error probability, with $\nu(g_n)$ the variant copy number in genotype g_n and with $c(g_n)$ the total copy number, then $\psi(\cdot)$ is defined by:

$$\psi(g_n, \phi_{z_n}, \epsilon) = \begin{cases} \epsilon & \text{if } \nu(g_n) = 0 \\ \phi_{z_n}(1-\epsilon) + (1-\phi_{z_n})\epsilon & \text{if } \nu(g_n) = c(g_n) \\ \phi_{z_n} \frac{\nu(g_n)}{c(g_n)} + (1-\phi_{z_n})\epsilon & \text{otherwise.} \end{cases} \tag{11}$$

The meaning of this definition is straightforward: if the genotype does not contain any variant, then the variant read is recorded when a sequencing error occurs, therefore with probability ϵ; if the genotype only contains variant copies, either the read comes, in case of no sequencing error, from the population that has cellular prevalence ϕ_{z_n}, or it comes from another population of cells when an error occurs; in all the other cases, i.e. $0 < \nu(g_n) < c(g_n)$, the variant read probability is proportional to the ratio of variant and total copies in the genotype (in absence of errors).

The likelihood of the single-cell data $\mathbf{S}$, under the assumption of conditionally independence over cell c and locus n given the tree T, the SNV assignments to the nodes z and the cell-to-clone membership $\zeta = \{\zeta_c\}_{c=1}^{C}$, factorizes as:

$$l(\mathbf{S}|T, z, \boldsymbol{\phi}, \zeta) \propto \prod_{c=1}^{C} \prod_{n=1}^{N} P(b_{cn}|T, z, \zeta_c, d_{cn}). \tag{12}$$

The data is modelled with a mixture of two Beta-Binomial distributions:

$$b_{cn} \sim \begin{cases} \begin{matrix}(1-\delta_{cn})\text{BetaBinomial}(d_{cn}, \alpha_0, \beta_0) \\ +\delta_{cn}\text{BetaBinomial}(d_{cn}, \alpha_n, \beta_n)\end{matrix} & \text{if } \mu_{cn} = 1 \\ \text{BetaBinomial}(d_{cn}, \epsilon, 1-\epsilon) & \text{otherwise} \end{cases} \tag{13}$$

where μ_{cn} is the mutation status of cell c in locus n, that is 1 when the cell contains a variant in that locus, 0 otherwise; this status can be seen as a function of T, z and ζ_c since it is determined by the SNVs that are assigned to the node to which the cell belongs.
The Beta-Binomial distribution is chosen due to the noisy nature of scRNA-seq data. In fact, the probability of the underlying Binomial cannot be assessed with confidence through the ratio of variant and total reads due to sequencing errors. The Beta-Binomial model gives more flexibility as the unknown probability becomes a Beta-distributed random variable, on which marginalization is performed. The mixture of the two Beta-Binomial accounts, again, for the sparsity of the scRNA-seq data: one distribution models the monoallelic expression (dropout event, $\delta_{cn} = 0$) and the other models the biallelic distribution ($\delta_{cn} = 1$). In absence of mutations, the reads are still modeled with a Beta-Binomial distribution, in which the Beta distribution modelling the probability of a variant read is more skewed towards 0 as ϵ decreases.
The parameters of the underlying Beta variates, $\alpha_0, \beta_0, \alpha_n, \beta_n$ are hyperparameters of the model, obtained as a pre-processing of the data.

Lastly, the joint likelihood of the single-cell data as presented in Eq. (5) is obtained through a marginalization over the cell-to-node assignments:

$$l(\mathbf{S}|T, z, \boldsymbol{\phi}) = \prod_{c=1}^{C} \sum_{\zeta_c} \prod_{n=1}^{N} P(b_{cn}|T, z, \zeta_c, d_{cn}) P(\zeta_c) \tag{14}$$

where a Uniform distribution is placed on the cell assignments, having $P(\zeta_c) = 1/V$.

3.3 DATASETS

The datasets used for this work are the same described in Section 2.2.1, that is bulk DNA and scRNA-seq data from the HGSOC cell-line. In particular, for the extension of the model mainly the scRNA-seq gene expression data has been considered.

Additionally, for the purpose of data analysis and the study on gene expression to copy number correlation, as well as for the synthetic data generation, also clone-specific copy number data has been taken into account.
This datatype consists of a total copy number value for each clone and for each bin of the genome. Here, *bin* refers to a sequence of the DNA (or RNA), typically larger than a gene, that is used to summarize the properties of genes that are close together.
The dataset comes as a result of the analysis carried out by Laks et al. [20] and it has been considered as *de facto ground truth* for the scope of the present work.

Figure 11 shows an extract of this dataset. It has to be specified, though, that the copy number data is only available for the leaf nodes of the evolutionary tree (see Figure 12). This is of course a limitation for the purpose of this work since only a few cells of the PhylEx

clone_id	chr	start	end	total_cn
E	1	1	500000	4
E	1	500001	1000000	4
E	1	1000001	1500000	4
		...		
A	10	1	500000	2
		...		

Figure 11: Clonal copy number data. Chromosome number, start and end positions define the coordinates of the bin and the clone ID, a letter from A to I, refers to a specific leaf node in the true phylogeny of the HGSOC cell-line delineated in [20].

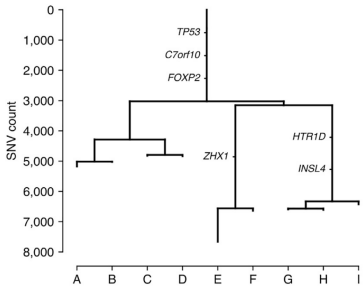


Figure 12: Phylogenetic tree with branch lengths calculated as counts of SNVs originating on each branch. Leaf nodes are single letters from A to I, all internal nodes are denoted by the string union of the descendant clones, e.g. the node right before the branching around 4000 SNV count is ABCD. Image credits to [20].

dataset have been assigned to the leaf node, thus the copy number of many cells is not known and would have to be imputed in alternative ways.

Along with the copy number dataset, the cell-to-node assignments obtained with the original PhylEx software as output of the inference process, have been used to relate clonal copy numbers and single cell gene expressions and considered as a ground truth.

3.4 PROBABILISTIC MODELS

Let us extend PhylEx model adding an extra set of variables in the single-cell dataset, namely the expression counts for each cell and each gene. Then we have $\mathbf{S} = \{\{b_{cn}, d_{cn}\}_{n=1}^{N}, \{y_{cg}\}_{g=1}^{G}\}_{c=1}^{C}$.

The posterior distribution now can be factorized as

$$\pi(T, z, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\mu} | \mathbf{B}, \mathbf{S}) \propto \\ l(\mathbf{B}, \mathbf{S} | T, z, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \pi_0(\boldsymbol{\phi} | T, z) \pi_0(\boldsymbol{\mu}) \pi_0(z | T) \pi_0(\boldsymbol{\lambda} | T) \pi_0(T). \tag{15}$$

Here we introduce two new variables, $\boldsymbol{\lambda} = \{\lambda_{bv} | b \in [B], v \in [V]\}$ and $\boldsymbol{\mu} = \{\mu_g\}_{g=1}^{G}$. Both are new target variables of the inference process. The first is the matrix of the bin-specific copy number values for each clone. We assume, as in Clonealign, that the copy number profile is shared across all cells in the same clone. Then $\boldsymbol{\mu}$ is the so called *per-copy expression* for each bin, describing the dependency between copy number and gene expression.

The prior distribution over $\boldsymbol{\mu}$ is the same as in [6], therefore a Normal distribution.

Copy numbers, instead, are modeled through an Hidden Markov Model (HMM), which is a full probabilistic model whose aim is to determine an unknown sequence of states based upon a sequence of observations. In this setting, copy numbers are the states and expression counts are the obervations. This approach is commonly adopted when modeling CNVs (e.g. [7, 33, 36]) and usually it is composed by a six states HMM, while transition and emission probability calculation strategies can vary from one method to another.

Note that the dependence built using HMM is for modelling convenience. From a biological standpoint, the copy number of a gene doesn't depend on a "previous" gene in a causative manner, although they are correlated.

However, the number of genes (therefore the number of copy number values for each clone) is excessively large (around 15k expressed genes) and the gene length is highly variable. For this reason, the HMM model is applied not at the gene level but rather at bin level. This approach allows smoother copy number signal predictions and, since typically copy number changes do not occur precisely at the end of a gene, is even more correct from a biological point of view.

In fact, unlike in Clonealign, here we denote the gene-specific copy numbers by $\tilde{\boldsymbol{\lambda}} = \{\tilde{\lambda}_{gv} | g \in [G], v \in [V]\}$ and we assume that one copy of gene g is effectively expressed with probability δ_g. To make an example, given a copy number $\lambda_{bv} = 4$ we have that for each gene inside that bin we can observe a value of $\tilde{\lambda}_{gv} \leqslant 4$ depending on δ_g; this value is the copy number that actually influences the expression data, also referred to as *active copy number*.

The complete likelihood, assuming again independence between bulk and single-cell data given the rest, is equal to

$$l(\mathbf{B}, \mathbf{S} | T, z, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \zeta) = l(\mathbf{B} | T, z, \boldsymbol{\phi}) l(\mathbf{S} | T, z, \boldsymbol{\mu}, \boldsymbol{\lambda}). \tag{16}$$

Keeping the idea of the marginalization over the cell assignments, we modify the modelling of the allelic imbalance data using the information given by the copy number.

Let us denote with $b(g)$ the bin containing gene g and with $\lambda_{b(g)\nu}$ the total copy number of the bin at a certain node ν of the clonal tree. The joint likelihood for the single cell data is as follows:

$$l(\mathbf{S}|T, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \prod_{c=1}^{C} \sum_{\zeta_c} \prod_{g=1}^{G} P(y_{cg}, b_{cn}|T, \mathbf{z}, \boldsymbol{\lambda}, \boldsymbol{\mu}, d_{cn}, \zeta_c) P(\zeta_c). \quad (17)$$

Like in PhylEx model we have $P(\zeta_c) = 1/V$, which places a Uniform distribution on the cell assignments as the prior. The joint probability of the gene counts and the variant reads given $T, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\lambda}, d_{cn}$ and ζ_c is given by the formula

$$P(y_{cg}, b_{cn}|T, \mathbf{z}, \boldsymbol{\lambda}, \boldsymbol{\mu}, d_{cn}, \zeta_c) = \quad (18)$$
$$\sum_{e=1}^{\lambda_{b(g)\zeta_c}} q(e|\lambda_{b(g)\zeta_c}, \delta_g) P(y_{cg}|T, \mu_g, e) \sum_{\nu=0}^{e} \binom{e}{\nu} q(b_{cn}|d_{cn}, \nu/e),$$

where we marginalize both over the active copy numbers $\tilde{\lambda}_{g\zeta_c}$, which take values denoted by e to avoid cluttered notation, and over the variant copy number ν.
The active copy number is assumed to be drawn from a Binomial distribution (for which the $q(\cdot)$ distribution stands), and so is the variant reads count b_{cn}. With this framework the Beta-Binomial distribution can be replaced by a simple Binomial since variant and total copy numbers are given and we can estimate the probability of recording a variant read as the ratio between the variant and total copy numbers. Also, the variant copies can be any combination over the total copies, and this leads to a multiplication factor of $\binom{e}{\nu}$.

$$\tilde{\lambda}_{g\nu}|\lambda_{b(g)\nu}, \delta_g \sim \text{Binomial}(\lambda_{b(g)\nu}, \delta_g) \quad (19)$$
$$b_{cn}|d_{cn}, \nu/e \sim \text{Binomial}(d_{cn}, \nu/e). \quad (20)$$

As for the gene expression counts, we model them in various ways and compare the different models.

First we consider a simpler version of the model shown in [6] in which the expected value of the expression count is given by

$$\mathbb{E}[y_{cg}|\zeta_c = \nu] = s_c \frac{\mu_g \tilde{\lambda}_{g\nu}}{\sum_{g'=1}^{G} \mu_{g'} \tilde{\lambda}_{g'\nu}}, \quad (21)$$

where s_c is the library size of cell c.

Then, since y_{cg} is a count variable, a Poisson distribution can be appropriate. However, many studies (such as [29]) adopt a Negative Binomial distribution instead, which gives more flexibility to handle the over-dispersion that is typically observed in scRNA-seq data. In fact, while in a Poisson distribution the variance is equal to the mean, in a Negative Binomial distribution we allow the variance to be larger.

Moreover, given the sparse nature of the data some studies (again, such as [29]) adopt a zero-inflated distribution, which accounts for frequent zero valued observations. This leads to distributions like Zero-Inflated Poisson (ZIP) or Zero-Inflated Negative Binomial (ZINB).

Given these premises, in this work we consider all four distributions to further analyze the advantages and drawbacks of each one.

In the case of ZINB, the most complex yet flexible one, we can therefore define the gene expression data as follow:

$$y_{cg}|T, \mu_g, \lambda_{g\zeta_c} \sim \text{ZINB}(m_{g\zeta_c}, r_g, \rho_{cg}), \tag{22}$$

where $m_{g\zeta_c}$ is the mean given by Eq. (21), r_g the gene-specific inverse of dispersion parameter for the underlying Negative Binomial and ρ_{cg} the zero-inflation probability for each cell and gene.

Precisely, the probability mass function is defined as

$$\begin{aligned} P(y_{cg}|T, \mu_g, \lambda_{\zeta_c}) &= f_{\text{ZINB}}(y_{cg}; m_{g\zeta_c}, r_g, \rho_{cg}) && (23)\\ &= \rho_{cg}\, \mathbb{1}(y_{cg} = 0) + (1 - \rho_{cg}) f_{\text{NB}}(y_{cg}; m_{g\zeta_c}, r_g), && (24) \end{aligned}$$

with $\mathbb{1}(\cdot)$ being the indicator function, i.e. equal to 1 when the condition in the argument is true, 0 otherwise. The Negative Binomial distribution here is parametrized by the mean and the inverse of dispersion parameters, instead of the typical number of failures and probability of success parameters. Given mean m and inverse dispersion r, the associated variance and probability of success are

$$\sigma^2 = m + \frac{m^2}{r}, \qquad p = \frac{m}{\sigma^2}. \tag{25}$$

Then, letting r be real-valued, the *pmf* of the negative binomial can be written as such

$$f_{\text{NB}}(y; m, r) = \frac{\Gamma(r + y)}{\Gamma(y + 1)\Gamma(r)} \left(\frac{r}{r + m}\right)^r \left(\frac{m}{m + r}\right)^y. \tag{26}$$

It is worth mentioning that the NB distribution boils down to a Poisson when $r \to +\infty$. Also, it can be viewed as a mixture of Poisson and Gamma distribution: the unknown rate parameter is described by a Gamma prior with parameters α and β, resulting in a NB of parameters $r = \beta$ and $p = \frac{\alpha}{\alpha+1}$.

We notice that the expression mean in Eq. (21) is obtained through normalization over all genes. This might lead to complications in terms of computation time and could be replaced by a even simpler model. In particular, we want to compare the Clonealign model for the mean with

$$\mathbb{E}[y_{cg}|\zeta_c = v] = \exp\left(s_c \times \mu_g \times \tilde{\lambda}_{gv}\right). \tag{27}$$

Here we remove the normalization factor and we model the log-mean of the Poisson (or NB, ZIP, ZINB) distribution since the natural link function for the Poisson is the log-link.

From now on, I will refer to the model given by Eq. (21) as the

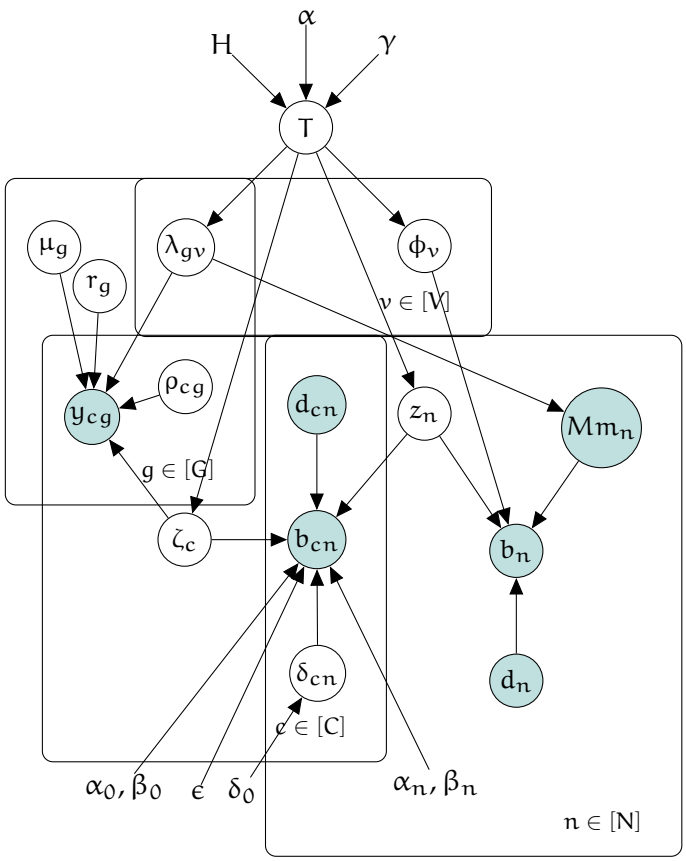


Figure 13: Directed graphical model of PhylEx with copy number extension

clonealign model and the one given by Eq. (27) as the *log-link model*. Each of the two models can be associated to all of the four distributions taken under analysis.

Although a detailed inference procedure has not been designed in the present work due to time constraints, some parameters have to be jointly inferred with the target variables, such as the copy number expression probability δ_g, while others can be computed from the data. In fact, size factors can be obtained with tools such as the one explained in Section 4.1 The inverse dispersion parameters r_g can be inferred in several ways as described in [21]; one is, for instance, to estimate the inverse dispersion parameters using empirical Bayes with *edgeR* as proposed in [30].

To conclude the section, in Figure 13 I summarize the variable dependencies drawing the graphical model of PhylEx with the proposed extension (in particular the one described by the ZINB distribution). Please note that, in order to avoid confusion, I omit some of the parameters which would over-complicate the graph without adding any relevant information.

3.5 SYNTHETIC DATA

In order to test statistical models, it is common practice to devise algorithms and tools for synthetic data generation. Assuming that the synthetic data is close enough to the real data, this framework provides extremely useful additional information such as the true parameters of the underlying distributions that are unknown in real-data applications.

In this section I present the process of synthetic data generation designed for gene expression data and copy number evolution, which has been built upon the already existing simulation software provided with PhylEx (see Supplementary material of [18]).

3.5.1 *Copy number evolution*

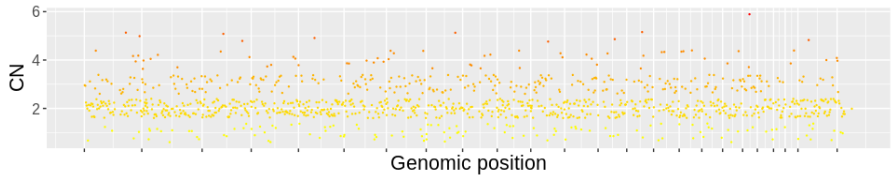
Copy numbers evolution is sampled with a simplified *birth-death* process. If we denote by λ_{max} the maximum copy number allowed, given the tree T, the copy number at bin b is sampled using a transition matrix $\mathbb{P} \in \mathbb{R}^{\lambda_{max} \times \lambda_{max}}$ which provides the probability mass function for $\lambda_{b\nu}$ given $\lambda_{b\nu'}$ where ν' is the parent node of ν. The $\mathbb{P}$ matrix is obtained through matrix exponentiation of the rate matrix $\mathbb{Q}$, which is in turn built from the birth and death rates.
A normal birth-death process would require each copy in the bin to evolve independently, therefore having different birth and death rates depending on the number of copies. However, the presented approximation, which is the one adopted in the original PhylEx study, can be considered enough for the purpose of the synthetic data generation.

Copy number changes in PhylEx are simulated by evolving copy numbers in each locus independently one locus from the other. As shown in Figure 20a of Section 4.1, although copy numbers in different bins are not dependent one from each other, they show some correlation. More specifically, a copy number in a bin is likely to be the same of the previous bin. Thus, simulating the evolution of whole genome copy numbers should take into account this property.

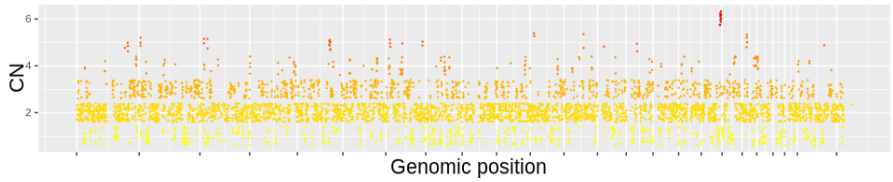
In order to replicate this pattern, two strategies have been tested, namely a HMM and a break-point sampling approach.

The HMM evolves the copy number of the bin with the simplified birth-death process introduced before, but the new copy number is accepted only with some fixed probability p, otherwise the copy number of the previous bin is taken. This approach leads to a smoother pattern than independent evolution, but due to the simplistic transition event (with fixed probability) it doesn't reflect real data variability in the length of the sequences with equal consecutive copy numbers.

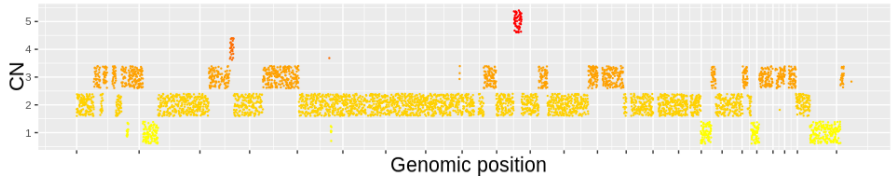
For this reason, a break-point sampling approach based on the data analysis presented above, precisely in Figure 20b, better suits the task. Specifically, a copy number variation $\Delta\lambda = \lambda_{b\nu} - \lambda_{b\nu'}$, with ν' being the parent node of ν, is sampled by subtracting the current copy number value to the new copy number obtained with the same transition



(a) Independent evolution



(b) HMM



(c) Break-point sampling

Figure 14: Different sampling approaches to copy number evolution. (a) Individual evolution between bins, (b) HMM, (c) break-point sampling. The whole genome is displayed on the abscissa with ticks at each chromosome change.

matrix P. The difference is then summed to all the subsequent copy numbers until the next break-point. The length of a sequence between two consecutive break-points, measured in terms of number of bins, is drawn from a Negative Binomial distribution with mean and variance estimated from the empirical distribution of the ground truth copy number data.

In Figure 14 a comparison between the different simulations is shown. Eventually, break-point sampling seems to replicate the real data quite accurately.

3.5.2 *Gene expression reads*

Genes are uniformly sampled from the complete set of genes stored in a *hg19* annotation file. This provides both the Ensembl gene ID and the position coordinates. Then genes are binned together in fixed size bins and, given the sampled copy numbers for each clone and given the cell assignments to the nodes, the gene expression count is sampled from one of the four distributions considered (Poisson, ZIP, NB, ZINB).

More specifically, first the size factor is sampled from a uniform distribution $\mathrm{U}(0, 2)$ and then the gene expression is sampled with mean computed given either formula (21) or (27), depending on the parameters of the simulation.

All the parameters such as distribution parameters, number of cells, number of genes, bin size etc. are specified in a separate configuration file.

3.6 EVALUATION

A direct comparison between the new model and the original PhylEx model is not possible since an inference algorithm based on the new model has still to be designed and implemented. However, to evaluate the proposed framework, two experiments have been carried out.

The first consists of generating synthetic data and predict the copy number of each gene using a maximum likelihood approach, i.e. by choosing the one which maximizes the partial likelihood given the expression counts and all the other parameters and variables (i.e. tree T, cell assignments to the nodes ζ, per-copy expressions μ, etc.).
The performance of the model is then measured with accuracy (ratio of correct predictions), mean squared error (MSE) and mean absolute deviation (MAD).The last two score measures are defined as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{28}$$

where y_i is the true value of copy number and $\hat{y}_i$ is the prediction. Both measures account for the same error, that is the mean distance between the true and predicted values. However, mean squared error stresses larger mistakes.

The prediction is done for each gene and for each cell, hence $n = G \times C$. The experiment allows to compare the alternative models, i.e. the *clonealign model* and the *log-link model*, each one in all four distributions, by simulating a simple likelihood-based copy-number inference process.

The second experiment consists of a sensitivity analysis with the likelihood as target measure. This helps understanding how the likelihood is affected by changes in the copy number signal. The analysis is performed by generating synthetic data and by sampling m perturbations of the true copy number signals. Each perturbation is created first by uniformly sampling the clone whose copy numbers will change, then sampling between 5% and 10% of the copy numbers for that clone (again uniformly). For each of these copy numbers, the new perturbed value is determined randomly, with higher probability given to smaller changes.

In order to distinguish different perturbations, we define the distance between the true copy number signal and its perturbation as the Euclidean distance. More formally, let $\tilde{\lambda}_v$ be the vector of length G of all copy numbers in clone v, then the *copy-number distance* is given by

$$d(\tilde{\lambda}_v, \tilde{\lambda}'_v) = \|\tilde{\lambda}_v - \tilde{\lambda}'_v\| = \sqrt{\sum_{g=1}^{G}(\tilde{\lambda}_{gv} - \tilde{\lambda}'_{gv})^2}. \tag{29}$$

Intuitively, the more a copy-number signal is far from the true value, the larger would be the decrease in the likelihood.

The results of the experiments are reported in the dedicated chapter (Chapter 4), and then further discussed in Chapter 5.

4 RESULTS

4.1 EXPLORATORY DATA ANALYSIS (EDA)

The purpose of the exploratory data analysis is two-fold:

1. visualize the marginal distribution of the data, especially scRNA-seq gene expression data which are known to have a sparse and noisy nature, for the purpose of the model definition and the synthetic data generation,

2. analyze the correlation between the gene expression and copy number data to confirm the assumptions made in the Clonealign model [6], which can then be used to extend PhylEx.

Part of the exploration and data pre-processing has been performed following the work-flow for scRNA-seq data described in [23], in particular using tools provided by *R* libraries such as *scater*, *scran* and *SingleCellExperiment*.

The single-cell gene expression dataset consists of a matrix of 63677 rows (genes) and 360 columns (cells). Each element of the matrix is the number of reads for that gene in the single cell and this value is shown to have large variance, especially for some genes. This is partially shown in Figure 15 where it is also clear that variability increases along with the gene expression values. Of course, genes with high variance are more likely to give valuable information about the origin of the cells (i.e. the clones to which they are assigned), while genes that are expressed almost at the same rate are probably not significant. Besides, many genes are under-expressed, meaning that only few reads, or even none, are present. In fact, as a pre-processing step, it is necessary to filter out uninteresting genes: although some information might get lost, filtering helps removing technical noise due to amplification biases during sequencing [16].

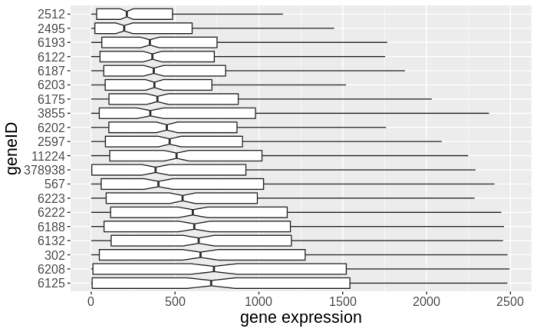


Figure 15: Box-plot showing the 20 most expressed genes in the scRNA-seq gene expression data.

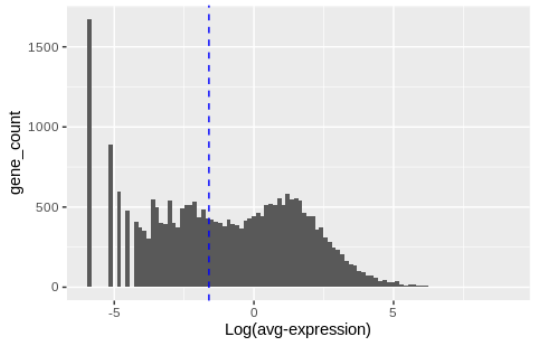


Figure 16: Low-abundance mean-based genes filtering removes the genes whose mean expression value across the cells is lower than a predefined threshold (0.2 in the figure).

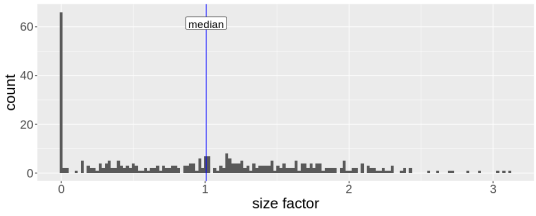


Figure 17: Cells size factor histogram.

Filtering has been performed simply by removing low-abundance gene with a mean-based filter (see Figure 16). This has reduced the number of genes to roughly 15 thousands significant genes.

As for the cells, these should not be filtered since the inference process ideally assigns each cell to a node. However, looking at Figure 17 it is possible to detect that around 60 cells contain very few reads. The histogram shows the distribution of the reads in terms of the size factor which is simply another way of expressing the library size.
In this case, the size factor has been computed with a deconvolution strategy tailored for scaling normalization of sparse count data [19]. This normalization technique, which pools similar cells together before deconvolution, is shown to better eliminate cell-specific bias, especially when dealing with sparse data. As a result, the obtained size factors are more robust in distinguishing cells from their gene expression counts (see Figure 18).

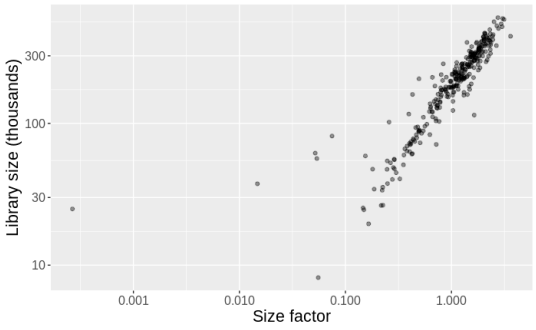


Figure 18: Size factors from deconvolution plotted against library size for all the cells.

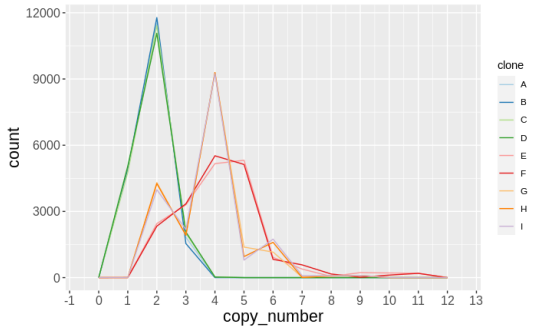


Figure 19: Histogram of the copy numbers across bins in the leaf nodes of the evolutionary tree, for each clone.

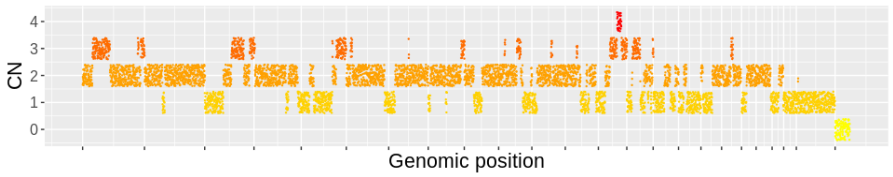
As for the copy number dataset, a first insight on its distribution can be given by the plot in Figure 19 which exhibits significant similarity between some clones. More specifically, three groups of nodes with similar copy number profile can be drawn: ABCD, EF, GHI. This is not totally unexpected since in Figure 12 we can already notice that those three groups differ substantially from each other in terms of SNVs, while all their children nodes only differ from those groups by a small amount of mutations.
This observation is quite useful as it allows to relate the known copy number profile of one of the leaf nodes to the cells belonging to the parent nodes up to those groups (whose copy numbers are not available) as a preliminary approximation.

In order to understand how copy numbers are distributed along the genome, and more specifically where a break-point can be found, it is possible to look at Figure 20a which shows the copy number for each bin in the whole genome. The number of break-points in each clone is around 90.
Moreover, all clones share similar mean over the length of the sequences between two consecutive break-points, which is between 55 and 74 bins. This information is further exploited in the synthetic data generation process which is discussed in Section 3.5.

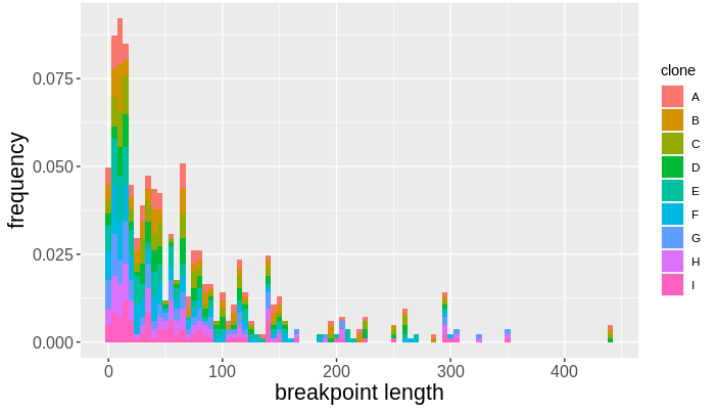
Lastly, gene expression data and copy number profiles can be analyzed together to verify the assumption made in Clonealign paper [6], which is that an increase in the copy number leads to highly expressed genes.

In order to link the two datasets, it is necessary to first map the *Ensembl* gene IDs to genome position coordinates. This has been achieved through *biomaRt* package.

The plot of the copy numbers against the gene expression values should result in a dependency as shown in Figure 21a, where the expression median across all cells and genes corresponding to a certain copy number manifests an increase. However, as already mentioned before, copy number profiles of the internal nodes are not available for this dataset and the only clone for which both cells and copy numbers are available is clone C. The support in terms of cells of this clone is very low (only 6 cells) therefore any correlation between the expression values and the copy number of those cells would be statistically



(a) Copy number for each bin plotted over the whole genome for clone C. Values are jittered in order to be able to visualize the size of a bin compared to the genome length.

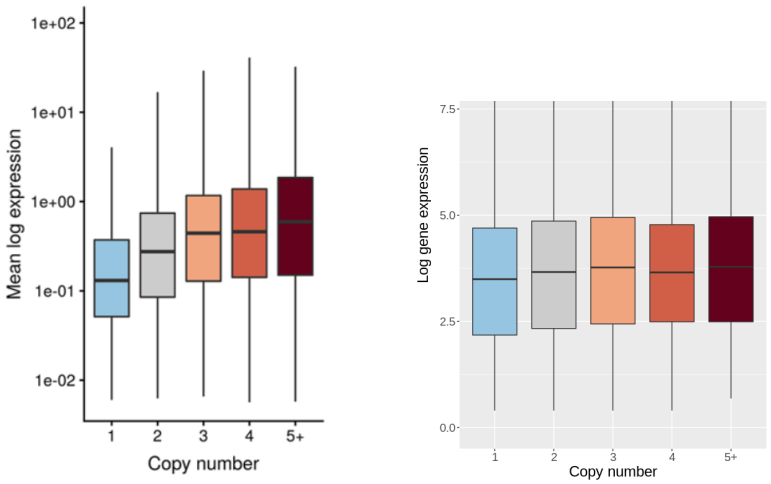


(b) Stacked histogram of the distance between consecutive break-points for each clone.

Figure 20: Copy number plots. (a) refers only to clone C, while (b) compares the distribution of all clones' copy number break-points.

irrelevant.
To try to overcome this issue and still be able to obtain some relevant indications on the data, cells assigned to internal nodes up to the three groups mentioned above (ABCD, EF, GHI) have been assigned to leaf nodes descendants of the relative groups. This approximation is justified for the reasons already discussed and leads to a higher support of cells. Nevertheless, even with a larger number of cells (precisely, 152), the resulting plot shows a dependency not as evident as in Clonealign plot (see Figure 21b). This might be due to two reasons: first, because of the approximation made by assigning cells to leaf nodes even when they belong to internal nodes with similar characteristics; second, because the nodes above those three groups in the tree are not present at all in the plot, therefore the plot lacks information that might be less noisy being closer to the healthy cells root node.

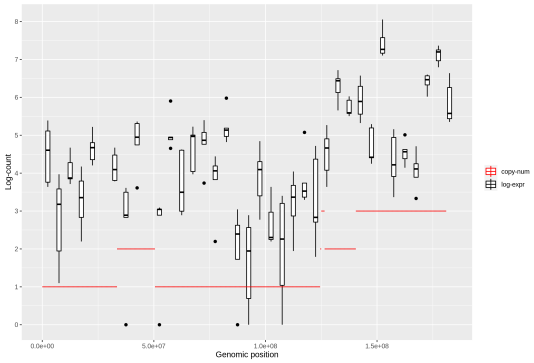
A zoomed-in picture of the dependency between copy number and gene expression is shown in Figure 22. There is definitely some correlation, although it might not be the case for all genes. This being said, it is worth trying out the Clonealign model as well as other models or variations of the same. This is addressed in the next section.



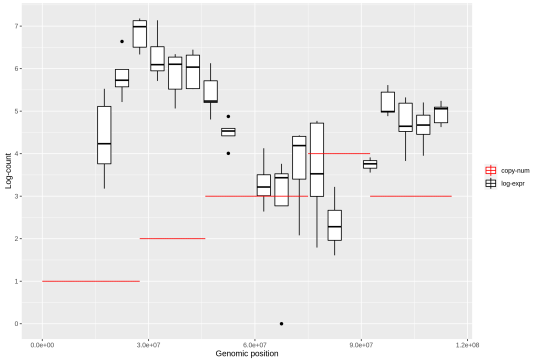
(a) Mean log expression as a function of the copy number across all clones in Clonealign. Image credits to [6].

(b) Mean log expression as a function of the copy number across clones up to ABCD, EF, GHI in the HGSOC dataset with copy numbers by [20].

Figure 21: Comparison of copy number and gene expression dependency between (a) Clonealign original study and the (b) present HGSOC cell-line study. While our dataset was obtained according to Smart-seq 3 sequencing protocol, Clonalign dataset was obtained through Chromium v3 sequencing protocol.



(a) Chromosome 5 of clone C cells, where copy number clearly reflects gene expression data.



(b) Chromosome 13 of clone C cells. Here the two trends are not related with each other in any way.

Figure 22: Log gene expression and copy number along (a) chromosome 5 and (b) chromosome 13 of cells in clone C. This shows two opposite cases: in the first one the correlation is quite evident, in the second one the gene expression does not reflect the copy number.

4.2 COPY NUMBER PREDICTION

As a way of evaluating the models accuracy and compare their performances in a quantitative manner, a copy number prediction experiment has been performed in the way it has been defined in Section 3.6. The prediction has been performed over two different kinds of datasets, one with low zero-inflation (i.e. low ratio of zero values) and one with high zero-inflation, therefore having more sparse data. More specifically, the first dataset is sampled using the ZINB distribution for gene expression values, with zero-inflation parameters ρ_{cg} sampled from a Beta distribution $\text{Beta}(1, 15)$, while the second dataset has zero-inflation ratios $\rho_{cg} \sim \text{Beta}(7, 5)$.

The results of the experiments are shown in 3a and 4a.

The performances with low zero-inflation are similar for all four distributions of the gene expression data. However, the log-link model seems to perform slightly better than the clonealign model. In both cases, even though the accuracy is not very high, the MSE and MAD connote an average error of less than 1, which means that most of the wrong predictions are one copy far from the true value. This kind of mistake could be partially fixed by smoothing the copy number prediction with a simple filter such as the median filter.

In 4a it is clear that simple Poisson and Negative-Binomial are not flexible enough to model sparse data such as scRNA-seq data. This shows that the zero-inflation model is necessary when the zero-inflation is high. In particular, around 70% of the gene expression values in the real dataset presented in Section 4.1 consists of 0s, and this ratio is better represented by the high zero-inflation dataset. Also, performances of the *clonealign model* are slightly better compared to the *log-link model* when dealing with sparse data.

	Accuracy (%)	MSE	MAD
Poisson	57.0 ± 8	0.70 ± 0.15	0.49 ± 0.1
NegBin	57.8 ± 8	0.67 ± 0.15	0.49 ± 0.1
ZIP	55.0 ± 7	0.96 ± 0.20	0.58 ± 0.1
ZINB	57.5 ± 7	0.67 ± 0.15	0.50 ± 0.1

(a) clonealign model

	Accuracy (%)	MSE	MAD
Poisson	62.7 ± 4.5	0.68 ± 0.10	0.46 ± 0.05
NegBin	63.6 ± 4.5	0.67 ± 0.10	0.45 ± 0.05
ZIP	63.7 ± 5	0.78 ± 0.15	0.47 ± 0.05
ZINB	66.0 ± 5	0.63 ± 0.10	0.42 ± 0.05

(b) log-link model

Table 3: Copy number prediction results over low zero-inflation data for ((a)) clonealign model and ((b)) log-link model with standard deviation over ten experiments.

	Accuracy (%)	MSE	MAD
Poisson	26.2 ± 7	2.16 ± 0.85	1.13 ± 0.25
NegBin	26.2 ± 7	2.13 ± 0.85	1.12 ± 0.25
ZIP	51.8 ± 8	1.15 ± 0.20	0.66 ± 0.10
ZINB	54.3 ± 9	0.85 ± 0.20	0.57 ± 0.15

(a) clonealign model

	Accuracy (%)	MSE	MAD
Poisson	12.3 ± 2	3.58 ± 0.55	1.62 ± 0.15
NegBin	12.0 ± 2	3.59 ± 0.55	1.63 ± 0.15
ZIP	46.1 ± 7	1.60 ± 0.30	0.81 ± 0.15
ZINB	47.6 ± 7	1.32 ± 0.30	0.74 ± 0.15

(b) log-link model

Table 4: Copy number prediction results over high zero-inflation data for a clonealign model and b log-link model with standard deviation over ten experiments.

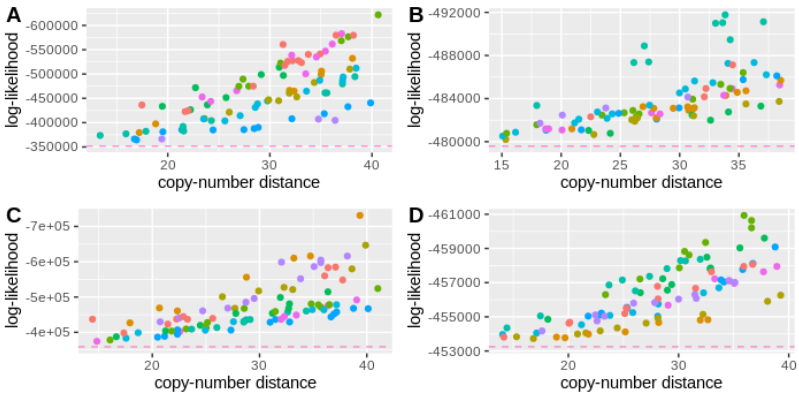


Figure 23: Scatter-plot showing copy number distance (as defined in Eq. (29)) against the log-likelihood value for the *clonealign model*. The four plots show the results for **A** Poisson, **B** Negative-Binomial, **C** ZIP, and **D** ZINB distributions. The dashed red line indicates the likelihood for the true copy number values. Different clones are shown in different colors. The number of samples is $m = 100$.

4.3 SENSITIVITY ANALYSIS

To test the behaviour of the likelihood obtained with different copy numbers we perform sensitivity analysis, simulating variations of the true copy number values along the genome and computing the relative single-cell data likelihood. The methods adopted are presented in Section 3.6.

The sensitivity analysis is performed both on the *clonealign model* and the *log-link model*, for all four count distributions. Nevertheless, the results shown in Figure 23 are relative only to the *clonealign model* since the *log-link model* does not give any additional evidence useful for comparison.

As expected, the more a copy number profile is distant from the true one, the lower is the likelihood. Also, some clones present larger sensitivity to the variation of the copy number. This is explained by the fact that clones that are closer to the leaf nodes present a more variable copy number signal and therefore the likelihood is more affected by it.

5 DISCUSSION

The probabilistic models defined in this work gather information and results from the most recent and advanced research papers in the field of tumor phylogenetics, such as Clonealign [6], ZINB-WaVE [29] and, most importantly, PhylEx [18]. The model extends an already existing tool for inference, incorporating copy number information alongside single nucleotide variations, leading to a novel method.

It should be noted that the problem that PhylEx, and therefore this work, is trying to solve is generally considered as a complex task. Reconstructing clonal-trees in tumor phylogenetics is a recent and difficult research goal, also because of the lack of a methodology in developing solutions to such problem. There are several outstanding problems in tumor phylogenetics that have still to be solved, for instance the fact that data sources are heterogeneous and new sequencing technologies are developed every year [32].

This method has not been fully tested in this work and therefore further development should follow in future works. However, the results reported in the previous section still allow for a valuable discussion.
More specifically it has been shown that one alternative of the model (the *clonealign model*) should be preferred to the other, and also that due to the sparsity of scRNA-seq data, the count distribution modelling the gene expression data should account for zero-inflation when inferring copy-numbers.
Moreover, the sensitivity analysis points out that the true copy number profiles actually lead to the maximum likelihood. It is therefore likely that an inference algorithm based on this model would estimate the true copy-number correctly given a sufficient amount of time.

5.1 LIMITATIONS

The main limitations of this work are two: computational complexity due to large number of variables, and lack of complete ground truth data.

The number of variables introduced in Section 3.4 is quite large, especially considering the fact that some parameters are specific for each gene, and that the number of genes is typically around 10^4. Some parameters of the model can be estimated from the data, but others have to be jointly inferred with the target variables. In order to avoid issues of undetermined parameters and excessively long execution time, some different strategies might have to be adopted, such as approximations of computationally expensive tasks (e.g. likelihood computation) and the definition of tailored, efficient sampling techniques.

The available real dataset, as already mentioned in Section 3.3, is not complete: the copy-number data is in fact partial, representing only the leaf nodes of the evolution tree. For this reasons, only a few experiments have been conducted on real data while most of the results have been based on synthetic datasets. This is of course a major limitation, since manipulation of real data is the final objective and typically presents peculiar aspects that synthetic data is not able to capture. The results are therefore to be considered as an indication on which path to follow first when further developing the work, and not on what is best: we can say that the *clonealign model* is more likely to achieve better results on real data, but we cannot exclude completely the *log-link model*.

5.2 FUTURE WORK

Future works may consider to further develop the method, define sampling techniques for inference to be applied to the model and test these procedures on real data. Eventually, additional ground truth data, including internal nodes copy number profiles, will be available, hence more accurate real data analysis could be performed.

More specifically, the implementation of a working inference algorithm for reconstructing the clonal-tree while inferring clonal copy-numbers would allow a direct comparison with the original PhylEx model. This will eventually lead to the discovery of whether copy number information can actually improve the accuracy of the tree with respect to a model only based on SNVs such as PhylEx.

6 CONCLUSION

This work extends the probabilistic model of a new method that reconstructs tumor phylogenetic tree combining bulk DNA and scRNA-seq data. PhylEx has been shown to achieve state-of-the-art accuracy, and by incorporating also copy number variations along with single nucleotide variation, we inject additional information in the inference process, aiming at even better performances.

In this research, we initially performed an exploratory data analysis on real genomic data, acquiring information about their nature. Then we defined more than one alternatives of the novel probabilistic model based on PhylEx. Furthermore, we built a simulation software based on the devised models and showed that synthetic data actually reflect real data. Lastly, we evaluated and compared the models with tailored experiments over simulated data, observing that models accounting for zero-inflation has to be preferred, and we motivate this fact with the sparse nature of scRNA-seq data.
The analysis that has been conducted shows that the devised model might be used as a reference for the development of specific inference procedures, which have not been addressed in the current work.
Although this study undergoes some limitations, it provides several insights on the topic under analysis and offers a well-defined starting point for other future works.

Fight against cancer through numerical computation is still far away from success, and faces several complex challenges. Nevertheless, research in this field is moving forward with increasing interest and this thesis can be viewed as another small step towards that direction.

BIBLIOGRAPHY

[1] David J. Aldous. "Exchangeability and related topics." In: *École d'Été de Probabilités de Saint-Flour XIII — 1983*. Ed. by P. L. Hennequin. Vol. 1117. Berlin, Heidelberg: Springer Berlin Heidelberg, 1985, pp. 1–198. ISBN: 9783540152033 9783540393160. DOI: 10.1007/BFb0099421. URL: http://link.springer.com/10.1007/BFb0099421.

[2] Sam Behjati and Patrick S Tarpey. "What is next generation sequencing?" In: *Archives of Disease in Childhood. Education and Practice Edition* 98.6 (Dec. 2013), pp. 236–238. ISSN: 1743-0585. DOI: 10.1136/archdischild-2013-304340.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. eng. 1st ed. Information Science and Statistics. OCLC: 1031889079. New York, NY: Springer New York, 2016. ISBN: 9781493938438. Reprint.

[4] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians." en. In: *Journal of the American Statistical Association* 112.518 (Apr. 2017), pp. 859–877. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2017.1285773.

[5] Jan Budczies et al. "Pan-cancer analysis of copy number changes in programmed death-ligand 1 (PD-L1, CD274) – associations with gene expression, mutational load, and survival." en. In: *Genes, Chromosomes and Cancer* 55.8 (2016), pp. 626–639. ISSN: 1098-2264. DOI: https://doi.org/10.1002/gcc.22365.

[6] Kieran R. Campbell et al. "clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers." In: *Genome Biology* 20.1 (Mar. 2019), p. 54. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1645-z.

[7] Stefano Colella, Christopher Yau, Jennifer M. Taylor, Ghazala Mirza, Helen Butler, Penny Clouston, Anne S. Bassett, Anneke Seller, Christopher C. Holmes, and Jiannis Ragoussis. "QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data." en. In: *Nucleic Acids Research* 35.6 (Mar. 2007), pp. 2013–2025. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkm076.

[8] Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, et al. "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups." eng. In: *Nature* 486.7403 (Apr. 2012), pp. 346–352. ISSN: 1476-4687. DOI: 10.1038/nature10983.

[9] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. "PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors." en. In: *Genome Biology* 16.1 (Dec. 2015), p. 35. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0602-8.

[10] Thomas S. Ferguson. "A Bayesian Analysis of Some Nonparametric Problems." In: *The Annals of Statistics* 1.2 (Mar. 1973), pp. 209–230. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176342360.

[11] Ruli Gao et al. "Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes." en. In: *Nature Biotechnology* 39.5 (May 2021), pp. 599–608. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-020-00795-2.

[12] Zoubin Ghahramani, Michael Jordan, and Ryan P Adams. "Tree-Structured Stick Breaking for Hierarchical Data." In: *Advances in Neural Information Processing Systems*. Ed. by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta. Vol. 23. Curran Associates, Inc., 2010.

[13] Mel Greaves and Carlo C. Maley. "Clonal evolution in cancer." In: *Nature* 481.7381 (Jan. 2012), pp. 306–313. ISSN: 0028-0836. DOI: 10.1038/nature10762.

[14] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines." en. In: *Experimental & Molecular Medicine* 50.8 (Aug. 2018), pp. 1–14. ISSN: 2092-6413. DOI: 10.1038/s12276-018-0071-8.

[15] Hemant Ishwaran and Lancelot F James. "Gibbs Sampling Methods for Stick-Breaking Priors." en. In: *Journal of the American Statistical Association* 96.453 (Mar. 2001), pp. 161–173. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/016214501750332758.

[16] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. "Quantitative single-cell RNA-seq with unique molecular identifiers." en. In: *Nature Methods* 11.2 (Feb. 2014), pp. 163–166. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.2772.

[17] Yuchao Jiang, Yu Qiu, Andy J. Minn, and Nancy R. Zhang. "Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing." en. In: *Proceedings of the National Academy of Sciences* 113.37 (Sept. 2016), E5528–E5537. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1522203113.

[18] Seong-Hwan Jun et al. "PhylEx: Accurate reconstruction of clonal structure via integrated analysis of bulk DNA-seq and single cell RNA-seq data." In: *bioRxiv* (2021). DOI: 10.1101/2021.02.16.431009. eprint: https://www.biorxiv.org/content/early/2021/02/17/2021.02.16.431009.full.pdf.

[19] Aaron T. L. Lun, Karsten Bach, and John C. Marioni. "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts." en. In: *Genome Biology* 17.1 (Dec. 2016), p. 75. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0947-7.

[20] Emma Laks et al. "Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing." en. In: *Cell* 179.5 (Nov. 2019), 1207–1221.e22. ISSN: 00928674. DOI: 10.1016/j.cell.2019.10.026.

[21] Luis Leon-Novelo, Claudio Fuentes, and Sarah Emerson. "Bayesian Estimation of Negative Binomial Parameters with Applications to RNA-Seq Data." In: *arXiv:1512.00475 [stat]* (Dec. 2015). arXiv: 1512.00475.

[22] Lawrence A. Loeb. "Mutator Phenotype May Be Required for Multistage Carcinogenesis." en. In: *Cancer Research* 51.12 (June 1991), pp. 3075–3079. ISSN: 0008-5472, 1538-7445.

[23] Aaron T.L. Lun, Davis J. McCarthy, and John C. Marioni. "A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor." en. In: *F1000Research* 5 (Oct. 2016), p. 2122. ISSN: 2046-1402. DOI: 10.12688/f1000research.9501.2.

[24] Iain C. Macaulay et al. "G&T-seq: parallel sequencing of single-cell genomes and transcriptomes." eng. In: *Nature Methods* 12.6 (June 2015), pp. 519–522. ISSN: 1548-7105. DOI: 10.1038/nmeth.3370.

[25] P. C. Nowell. "The clonal evolution of tumor cell populations." eng. In: *Science (New York, N.Y.)* 194.4260 (Oct. 1976), pp. 23–28. ISSN: 0036-8075. DOI: 10.1126/science.959840.

[26] A. P. Patel et al. "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma." en. In: *Science* 344.6190 (June 2014), pp. 1396–1401. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1254257.

[27] Dang Pham and Tuan Le. "Auto-Encoding Variational Bayes for Inferring Topics and Visualization." en. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 5223–5234. DOI: 10.18653/v1/2020.coling-main.458.

[28] Peng Qiu. "Embracing the dropouts in single-cell RNA-seq analysis." en. In: *Nature Communications* 11.1 (Dec. 2020), p. 1169. ISSN: 2041-1723. DOI: 10.1038/s41467-020-14976-9.

[29] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. "Publisher Correction: A general and flexible method for signal extraction from single-cell RNA-seq data." en. In: *Nature Communications* 10.1 (Dec. 2019), p. 646. ISSN: 2041-1723. DOI: 10.1038/s41467-019-08614-2.

[30] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." en. In: *Bioinformatics* 26.1 (Jan. 2010), pp. 139–140. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btp616.

[31] 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." eng. In: *Nature Methods* 11.4 (Apr. 2014), pp. 396–398. ISSN: 1548-7105. DOI: 10.1038/nmeth.2883.

[32] Russell Schwartz and Alejandro A. Schäffer. "The evolution of tumour phylogenetics: principles and practice." en. In: *Nature Reviews Genetics* 18.4 (Apr. 2017), pp. 213–229. ISSN: 1471-0064. DOI: 10.1038/nrg.2016.170.

[33] Eric L. Seiser and Federico Innocenti. "Hidden Markov Model-Based CNV Detection Algorithms for Illumina Genotyping Microarrays." en. In: *Cancer Informatics* 13s7 (Jan. 2014), CIN.S16345. ISSN: 1176-9351, 1176-9351. DOI: 10.4137/CIN.S16345.

[34] Jayaram Sethuraman. "A Constructive Definition of Dirichlet Priors." In: *Statistica Sinica* 4.2 (1994), pp. 639–650. ISSN: 10170405, 19968507.

[35] Xin Shao, Ning Lv, Jie Liao, Jinbo Long, Rui Xue, Ni Ai, Donghang Xu, and Xiaohui Fan. "Copy number variation is highly correlated with differential gene expression: a pan-cancer study." In: *BMC Medical Genetics* 20.1 (Nov. 2019), p. 175. ISSN: 1471-2350. DOI: 10.1186/s12881-019-0909-5.

[36] Lingyang Xu, Yali Hou, Derek Bickhart, Jiuzhou Song, and George Liu. "Comparative Analysis of CNV Calling Algorithms: Literature Survey and a Case Study Using Bovine High-Density SNP Data." en. In: *Microarrays* 2.3 (June 2013), pp. 171–185. ISSN: 2076-3905. DOI: 10.3390/microarrays2030171.

[37] Chenhao Zhou et al. "Integrated Analysis of Copy Number Variations and Gene Expression Profiling in Hepatocellular carcinoma." en. In: *Scientific Reports* 7.1 (Sept. 2017), p. 10570. ISSN: 2045-2322. DOI: 10.1038/s41598-017-11029-y.