University of Warsaw
Faculty of Mathematics, Informatics and Mechanics
Sorbonne University
Faculty of Computer Science

Probabilistic graphical models for mapping tumor clones in cancerous tissues and single cells

PhD dissertation
in COMPUTER SCIENCE

Author:
Shadi Darvish Shafighi

Supervisor:
Prof. Ewa Szczurek, University of Warsaw
Prof. Alessandra Carbone, Sorbonne Université

Warsaw, April 2023

Abstract

Spatial, genomic, and phenotypic heterogeneity are crucial for understanding cancer progression, treatment, and survival. However, identifying cancer clones and their gene expression profiles alongside their location in the tumor tissue is challenging. This thesis is devoted to comprehensive modeling of different aspects of tumor heterogeneity and builds upon three projects. In the first project, we focused on the genomic heterogeneity of the tumor and developed a probabilistic model that leverages independent genomic clustering of cells and scarce single-cell RNA sequencing data to map cells to given imperfect genotypes of tumor clones. In the second project, we explored all three aspects of heterogeneity with the main focus on spatial heterogeneity. We developed a complex probabilistic model to accurately infer the cancer clones and their localization in close to single-cell resolution by integrating pathological images, whole-exome sequencing, and spatial transcriptomics data. Expanding upon our previous project, in the third project, we focused on phenotypic heterogeneity. We proposed a probabilistic model that combines spatial transcriptomics and whole-exome sequencing data to accurately identify cancer clones and their gene expression profiles in tumor tissue. Our integrated approach provides a comprehensive understanding of the spatial, genomic, and phenotypic organization of tumors, opening new avenues to study the functional implications of tumor heterogeneity and the origins of resistance to targeted therapies.

Keywords

Probabilistic graphical models, statistical learning, cancer, genomics

Thesis domain (Socrates-Erasmus subject area codes)

11.3 Informatics, Computer Science

Subject classification

Applied computing → Life and medical sciences → Computational biology
Computing methodologies → Statistical learning → Probabilistic graphical models

Acknowledgments

I am incredibly grateful to my supervisors, Prof. Ewa Szczurek and Prof. Alessandra Carbone, for their unwavering guidance and support throughout my Ph.D. studies. Their scientific and personal support has been invaluable, and I could not have completed this thesis without them.

I am also grateful to Prof. Jens Lagergren, who served as an unofficial scientific supervisor and mentor. I learned so much from him, and his guidance was instrumental in shaping my research.

I would also like to extend my gratitude to all my colleagues and collaborators, including Dr. Cornelis A.M. van Bergen, Dr. Szymon M Kiełbasa, Prof. Dominika Nowis, Dr. Hab. Łukasz Koperski, Agnieszka Geras, Barbara Jurzysta, Alireza Sahaf-Naeini, Igor Filipiuk, Dr. Łukasz Rączkowski, Dr. Hosein Toosi, Julieta Sepúlveda-Yáñez, and Dr. Ramin Monajemi. It was an absolute pleasure working with you all, and I am grateful for the fun and productive collaborations we had together.

I would like to thank the program that has provided me with the following funding during the research and preparation of this dissertation:

- European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 766030.

I wish to extend my heartfelt appreciation to my family. My parents, Ezzatollah Darvish-Shafighi and Fahimeh Dadpoor, have always instilled in me a passion for science and taught me how to make good life choices. I am also grateful to my brother, Shahin Darvish-Shafighi, who sparked my interest in computer science.

Finally, I want to thank my husband, Alireza Sahaf-Naeini, for his patience, support, understanding, and belief in my abilities which have been an endless source of strength and inspiration.

Acknowledgment of scientific collaboration

The work presented in this thesis was a collaborative effort involving multiple colleagues and research groups.

The first project, presented in chapter 5, is a collaboration with the Leiden University Medical Center in the Netherlands. The patient samples and data were provided by Prof. Hendrik Veelken. The experiments, sample preparation, and identification of input BCR clusters were conceived and planned by Dr. Cornelis A.M. van Bergen. The preprocessing of the data, including exome and scRNA sequencing, alignment of scRNA reads, mutation calling in WES data, single-cell data sample deconvolution, and clustering of single cells to subjects was conducted by Prof. Szymon Kiełbasa, Prof. Susan Kloet, Dr. Roberta Menafra, Dr. Hailiang Mei, Dr. Ramin Monajemi, Julieta Sepúlveda Yáñez, and Davy Cats. Prof. Ewa Szczurek coordinated our work and helped with the development of the probabilistic model. The whole project was conceived with the great help of the valuable insights of Prof. Ewa Szczurek, Prof. Cornelis A.M. van Bergen, and Prof. Szymon Kiełbasa.

The second project, presented in chapter 6, is a collaboration with KTH University in Sweden. The biological experiments including extraction of the breast cancer tumor, preparing the sample, performing the DNA extraction for whole-exome sequencing of breast cancer dataset, performing spatial transcriptomics, performing the cell sorting, performing the st pipeline, and bulk DNA-seq mutation calling was done by Prof. Johan Hartman, Dr. Xinsong Chen, Dr. Kim Thrane, Dr. Camilla Engblom, Dr. Jeff Mold, and Alireza Sahaf Naeini. Analyzing the H&E images was done by Dr. Hab. Łukasz Koperski, Dr. Łukasz Rączkowski, and Igor Filipiuk. Barbara Jurzysta applied the regression on gene expression data and performed GSEA analysis. Prof. Ewa Szczurek coordinated our work. Prof. Ewa Szczurek and Prof. Jens Lagergren and Agnieszka Geras helped with the development of the probabilistic model. Analyzing and interpreting of the model results was done by the great help of Prof. Alessandra Carbone, Prof. Dominika Nowis, and Prof. Łukasz Koperski.

The third project, presented in the chapter 7, is close a collaboration with master student Barbara Jurzysta. Prof. Ewa Szczurek and Barbara Jurzysta helped with the development of the probabilistic model. Barbara Jurzysta implemented the model. Prof. Ewa Szczurek helped with the coordination of the work and conceiving the study.

Publications with results from the dissertation

The work presented in this dissertation has been published in the following research papers and preprints:

Shadi Darvish Shafighi*, Szymon M. Kiełbasa*, Julieta Sepúlveda-Yáñez, Ramin Monajemi, Davy Cats, Hailiang Mei, Roberta Menafra, Susan Kloet, Hendrik Veelken, Cornelis A.M. van Bergen and Ewa Szczurek. CACTUS: integrating clonal architecture with genomic clustering and transcriptome profiling of single tumor cells *Genome medicine* 13, Article number: 45 (2021) `https://doi.org/10.1186/s13073-021-00842-w`

Shadi Darvish Shafighi, Agnieszka Geras, Barbara Jurzysta, Alireza Sahaf Naeini, Igor Filipiuk, Łukasz Rączkowski, Hosein Toosi, Łukasz Koperski, Kim Thrane, Camilla Engblom, Jeff Mold, Xinsong Chen, Johan Hartman, Dominika Nowis, Alessandra Carbone, Jens Lagergren, Ewa Szczurek. Tumoroscope: a probabilistic model for mapping cancer clones in tumor tissues *BioRxiv* `https://doi.org/10.1101/2022.09.22.508914`

Shadi Darvish Shafighi*, Barbara Jurzysta*, Ewa Szczurek. Clonal gene expression analysis from spatial transcriptomics data *In preparation*

Other publications

Agnieszka Geras, Shadi Darvish Shafighi, Kacper Domżał, Igor Filipiuk, Łukasz Rączkowski, Hosein Toosi, Leszek Kaczmarek, Łukasz Koperski, Jens Lagergren, Dominka Nowis, Ewa Szczurek. Celloscope: a probabilistic model for marker-gene-driven cell type deconvolution in spatial transcriptomics datas *BioRxiv* `https://doi.org/10.1101/2022.05.24.493193`

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer is a complex and multifaceted disease characterized by abnormal cell growth [1]. It is caused by genetic alterations, which are changes in the genetic material (the genome) of a cell that can be transmitted to the cell's descendants [2]. These alterations can cause cells to grow and divide abnormally, leading to the formation of tumors. Tumors can be benign (not cancerous) or malignant (cancerous). There are over 200 distinct types of cancer, and it is one of the major causes of death worldwide [3]. Cancer cells can create subpopulations of tumor cells by acquiring new mutations during cell division. These mutations can cause the cells to differentiate and acquire new characteristics, leading to the formation of subpopulations. Each subpopulation with a group of cells that are genetically identical to each other is called a clone. Each clone over time grows and creates new descendant clones by gaining new mutations over the process of clonal evolution. The identical mutation profile of the cells in each clone is called genotype. This diverse genetic characteristics is called genetic heterogenity, which poses a significant challenge for cancer treatment and therapy resistance [4, 5, 6, 7, 8, 9, 10, 11, 12, 13].

Besides the genetic heterogeneity, there are differences in the observable characteristics of cells within and between different clones. These characteristics of a cell are called phenotype and the differences in the phenotype between cells are called phenotypic heterogeneity [14]. If a cell has a phenotype that helps it survive and grow, it can pass that phenotype on to its offspring. This can lead to increased levels of heterogeneity in specific phenotypic traits [15]. In addition, the characteristics of a tumor can also vary at different locations due to the uneven distribution of various concentrations of each clone within a tissue. This is called spatial heterogeneity. The genetic, phenotypic, and spatial heterogeneity of clones are responsible for their various behaviors and localization, which introduce a variety of difficulties in treatment. Understanding these factors can help improve treatment methods [16, 17].

Determining the tumor clones and tracing their evolutionary connections can shed light on the genetic heterogeneity [18]. Furthermore, by understanding the behaviors of the cells (phenotype) within separate cancer clones as well as their location and interactions, we can gain insight into the underlying mechanisms of tumor development [19]. Overall, our goal in this thesis is to resolve the evolution of cancer and understand the different types of heterogeneity in the tumor to provide valuable information that can be used to improve cancer diagnosis and treatment [20].

For modeling and understanding such a complex environment and relations we need a powerful computational method. Probabilistic graphical models (PGMs) are a type of mathematical model that can be used to represent complex relationships between different variables. They are useful for modeling real-world scenarios where there is a lot of uncertainty, such as cancer study. They are also widely used in feature selection [21, 22], data integration

[23, 24, 25, 26], classification tasks [27, 28], image Analysis [29], social network analysis [30], recognition in speech processing, time-series modeling, and finally dealing with uncertainty in expert systems [31]. PGMs use nodes to represent random variables and graphs to represent the relationships between the variables. This makes them a powerful tool for understanding complex environments and relationships. [32, 33, 22, 34, 35].

The first important strength of PGMs is that they are interpretable. The graphical structure of PGMs provides a clear, visual representation of the variables and their dependencies, making it easier to understand the system. Secondly, they are modular and allow the modeling of complex systems by breaking them down into smaller, interconnected components that can be solved and combined independently. Thirdly, they are flexible as they can accommodate a wide range of probability distributions and can easily handle missing data or latent variables. Finally, they are efficient. PGMs make it trivial to efficiently calculate the marginal probabilities and making predictions based on the underlying graph structure [36, 37]. These strengths make PGMs an ideal choice for modeling complex systems such as cancer studies.

1.1. Research topics covered in the thesis

The primary objective of my thesis is to develop computational methods that improve our understanding of the clonal architecture of tumors and enable tracing the genetic origin, location, and phenotypic characteristics of individual cells within a tumor. To achieve this, we undertake multiple projects that integrate various sources of data that can measure different aspects of tumor tissue including bulk deoxyribonucleic acid sequencing (bulk DNA-seq), which provides aggregated genomics information of millions of cells in a tumor sample [38], single-cell ribonucleic acid sequencing (scRNA-seq) that examines the gene expression level of individual cells [39], B-cell Receptor (BCR) sequence which is a protein complex on the surface of B-cells (a type of white blood cell) that binds to antigens (a foreign molecule that stimulates an immune response) [40], hematoxylin and eosin (H&E) stained images that provide general layout and distribution of cells [41], and spatially resolved tumor transcriptomic (ST) that capture the transcriptomics data alongside their coordinates (spots) in the tissue [42].

The first project is focused on exploring the genetic heterogeneity in the tumor. In this project, we develop a probabilistic model, called CACTUS, which accurately identify the clonal origin of tumor cells, as well as corrects any errors in the given phylogeny tree from bulk DNA-seq and scRNA-seq data. In this project, we additionally utilize the clustering of cells based on their B-cell receptor (BCR) sequences to improve the accuracy of the cell-to-clone assignment. BCR sequences refer to the genetic information that codes for the proteins on the surface of B-cells, which are a type of immune cell. We applied this model to newly generated follicular lymphoma single-cell data. This project is a collaboration with the Leiden University Medical Center in the Netherlands and has been published in *Genome Medicine* [43].

The second project is focused on the spatial heterogeneity. In this project, we design a probabilistic model, called Tumoroscope, that is able to localize the cancer clones within the tumor tissue. It integrates ST and bulk DNA-seq data to deconvolute the mixture of cells in each ST spot, containing a mini-batch of cells, into the cells coming from different clones. Using simulated data, we validate the model's performance and demonstrate its ability to accurately infer the fraction of the cells in the ST spots coming from the specific cancer clone. Additionally, we apply the model to both a previously published prostate cancer dataset [44] and a newly generated breast cancer dataset, showing the generalizability of our approach. We also use a regression model to estimate the gene expression profiles of the clones by taking into consideration the gene expression of the spots and the inferred proportion of the clones

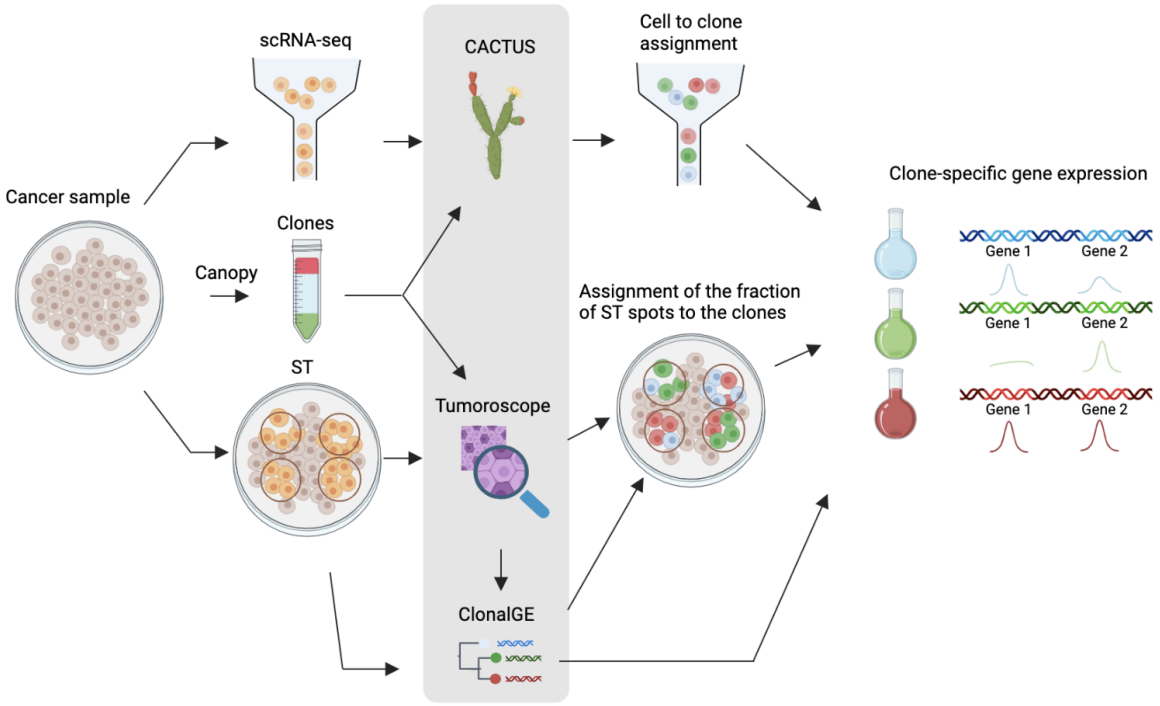


Figure 1.1: **Overview of the three projects presented in the thesis.** Image created by BioRender.

in each spot by Tumoroscope. This project is a collaboration with KTH University in Sweden and has been submitted to a journal and is available in BioRxiv [45].

The third project is focused on the phenotypic heterogeneity. In this project, we present a model, called clonalGE, which is built upon the Tumoroscope model by incorporating the inference of clone-specific gene expression alongside the clonal composition of the tumor across the tissue. For this, we introduce additional data sources of information to the model, containing the gene expression in each ST spot, which is the original output of ST data. We use both synthetic data and real data to demonstrate the improvement in the estimation of clone-specific gene expression and apply the improved model to the previously used prostate cancer dataset [44].

Throughout these projects, we are combining information from various sources of data, formalizing this information in the language of probabilistic graphical models, and devising efficient methods for statistical inference 1.1. In the following section, I will describe the major challenges that we faced and our solution to these challenges.

1.2. Challenges and our solutions

1.2.1. Mixture deconvolution

In all the projects in this thesis, we use bulk DNA-seq data for inferring the clones. Bulk DNA-seq technology mixes many genetically distinct cells in each sample, which must then be computationally deconvolved as different clones with similar genotypes. This problem refers to as the deconvolution of a mixture, which denotes a process of separating individual components of a mixed sample. In the second project and third projects, we face the same problem, having the mixtures of aggregated reads coming from the cells belong to different clones with

the same coordinates in the ST spots. This deconvolution can be a challenging task because the individual components may have similar characteristics and may be difficult to distinguish from one another. Additionally, the process of mixing the components can introduce noise or other confounding factors that can make it difficult to accurately deconvolute the mixture. This challenge is often addressed by using mathematical models such as Gaussian Mixture Model (GMM) [46], Dirichlet distributions [47], generative models [48], Categorical distributions [49] and matrix factorization methods [50, 51].

1.2.2. Error correction

In this thesis, we use DNA and RNA sequencing for genome analysis, which is plagued by potential sources of error, including limitations in the techniques used for sequencing, which is leading to inaccuracies in the number of reads and nucleotides identified in the sequences. Another significant source of error is the use of different bioinformatics tools or methods for the steps of pre-processing the data. Each step can introduce different biases and inaccuracies in the analysis of the sequences [52]. To improve the accuracy and reliability of genome analysis results, various solutions can be employed, including advancements in sequencing technologies [53], the implementation of quality control metrics [54, 55], the use of error correction algorithms [56], data normalization techniques [57, 58], the integration of data from multiple sources, and the application of machine learning and statistical methods for estimating the error [59].

1.2.3. Feature allocation

In the second and third projects, we need to resolve the mixture of the clones in the ST spots. The problem involves assigning a subset of clones (features) with specific genotypes (properties of the features) to the spots (samples) in the data. Each spot can belong to more than one feature. This problem is refer to as feature allocation, which is a generalization of the clustering problem, where each sample can belong to more than one cluster, called features. It is useful for modeling data that have multiple attributes or aspects, such as consumers' preferences for genres or document topics. There are many approaches to this problem including Latent Dirichlet Allocation (LDA) [60], Hierarchical Dirichlet Process (HDP) [61], Indian Buffet Process (IBP) [62, 63], Gause-Poisson Process (GPP) [64], and Beta-Bernoulli Process (BBP) [65].

This problem can be challenging due to the number of factors. One factor is the high dimensionality of the feature space. The large number of properties belong to the features (mutations) can make it difficult to identify the relevant features and assign them to the samples. Another factor is that the problem of feature allocation can be further complicated by the presence of dependencies and interactions among the features and samples. For example, the expression of certain clones and therefore, the present mutations may be influenced by the location of the sample or the presence of other clones.

1.2.4. Bias

In ST data, which is used in the second and third projects, the number of reads per mutation in each spot can be biased due to the interplay between cell number and gene expression levels in that region of the tissue. This same issue applies to single-cell data where the gene expression levels vary across the tissue. These issues refer to as biases that can compromise the accuracy of the data and must be considered when interpreting results. Bias, generally refers to a systematic error or deviation from the true value in a measurement or estimate.

In data analysis, bias can result from various sources such as measurement errors, sampling issues, or incorrect assumptions in the analysis process. Bias reduction can be accomplished through various methods that are tailored to the research question and type of data, such as normalization techniques [66], cell number correction, or the use of statistical models that consider the relationship between cell number and gene expression [67, 68, 69].

1.2.5. Lack of ground truth

In all my projects, we do not have the ground truth of the characteristic of the clones. Without knowing the true clones, it is difficult to determine the accuracy or performance of the clonal inference methods. Besides, in the first project, we do not have the ground truth of the BCR clusters and yet we are correcting these clusters in our model. Also, we do not have the true clones for the cells to verify if we infer them correctly. In my second and third projects, not only we do not know the present clones in each spot and the fraction of cells that belong to them, but also we do not have the ground truth of the clone-specific gene expression which makes it hard to verify if we estimate them correctly. This problem is referred to as the lack of ground truth, which is the absence of true or known values. One example of this problem in machine learning is comparing two classifiers without labels, which is a challenging task, as labels are usually needed to measure the accuracy, precision, recall, and other metrics of the classifiers [70]. However, there are some possible ways to compare two classifiers without labels, such as using synthetic or simulated datasets, where the true labels are known or generated, and applying the classifiers to these datasets to compare their performance. Another solution is using unsupervised or semi-supervised methods, such as clustering or dimensionality reduction to group or label the data based on some features or criteria, and then compare the classifiers based on these groups or labels. Finally, we can use domain knowledge, expert opinions, or external sources, such as literature and databases, to provide some labels or information for the data, and then to compare the classifiers based on these labels or information [71, 72].

Chapter 2

Cancer biology

In this chapter, our aim is to establish a basic understanding of cancer and its properties. First, we start with the explanation of cell normal functionality. Then, we define cancer cells and explore the role of genes and mutations in the development of cancer. We also explain the general hallmarks of this complex disease alongside the clonal evolution and tumor heterogeneity. After obtaining a fundamental comprehension of cancer, we provide a brief characterization of the specific types of cancer that our project focuses on, including prostate cancer, breast cancer, and follicular lymphoma.

2.1. Cell cycle

A cell is the smallest functional unit that can live on its own and compose all the body's tissues. The organs of the body are made of these tissues [73]. Cells replicate themselves into two daughter cells in a tightly controlled replication process that is called mitosis. Mitosis for the cells is controlled by a series of organized processes called the cell cycle (see fig. 2.1). It is a cycle since this series of events will start for each of the daughters after mitosis. The cell cycle for the cells that have a nucleus (eg. eukaryotic cells) divides into two stages: interphase, for the cell growth and replication of its DNA, and the mitotic (M) phase, for separating the DNA and cytoplasm into two cells.
The interphase includes three phases.

- **G_1**: in this stage, the cell grows, it also makes copies of organelles and the necessary molecular building blocks.
- **S**: in this stage, the cell duplicates DNA and a microtubule-organizing structure called the centrosome which is helpful in separating the DNA during the M phase.
- **G_2**: the cell grows more, produces proteins and organelles, and gets prepared for mitosis.

The M phase includes two stages:

- **mitosis** in this stage, the duplicated DNAs condense into two different DNAs, each with two chromosomes, and locate on two sides of the cell, ready for the splitting of the cell.
- **cytokinesis** in this stage, the cell physically gets split into two cells.

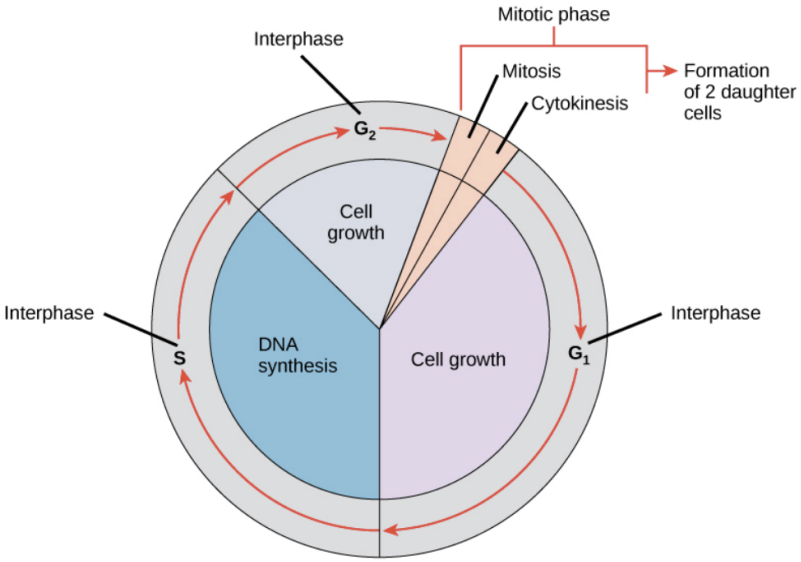


Figure 2.1: **Cell cyle.** Image credit: by OpenStax College, Biology (CC BY 3.0).

There are some cells that do not want to divide again, or they divide very slowly. In this case, the cell enters the resting phase called G-0. The cell may stay in this phase forever or again enter the G-1 phase for dividing [74].

2.2. Cancer development and properties

2.2.1. What is cancer generally?

Cancer is the leading cause of death in the world. Breast, lung, colorectal, and prostate cancer are the most common types of cancer (see Figure 2.2) [75, 76, 1]. Cancer develops via the uncontrolled growth of cells and all types of it can be grouped into four main branches: carcinoma, sarcoma, leukemia, and lymphoma. The first type, carcinoma, happens in epithelium tissues containing cells on the internal and external surface of the body, such as prostate cancer, breast cancer, lung cancer, and colorectal cancer. The second type is sarcoma which is rare cancer that happens in connective tissue found in bones, blood vessels, nerves, tendons, cartilage, muscle, and fat. The third type is leukemia, which is a cancer of the blood; originating in the bone marrow, including acute myeloid leukemia (AML) and chronic lymphocytic leukemia (CLL) that occur most frequently. The last type is lymphoma, which is a cancer of the lymph system.

2.2.2. Role of genes and mutations in causing cancer

DNA is a double-stranded molecule, composed of two chains constructed by nucleotides that coil around each other to form a double-helix [77]. The genetic instructions for the development, functioning, growth, and reproduction of all known organisms and many viruses are coded by DNA. The set of DNA instructions found in a cell is called the genome. Around 99.5 percent of the genome is the same in all humans—only the 0.5 percent variations in the genome account for our individuality. A region of DNA with possible different lengths that

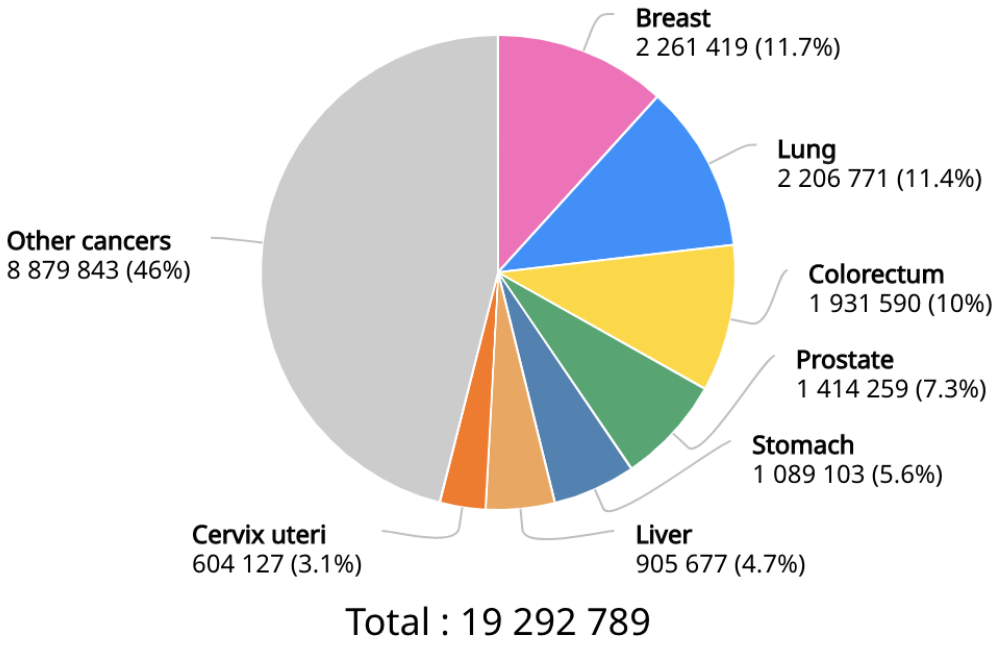


Figure 2.2: **The distribution of cancer types across patients.** Data source: GLOBOCAN 2020
Graph production: global cancer observatory `https://gco.iarc.fr/`
International Agency for Research on Cancer 2023

encodes a function is called a gene. There exist around 25000 genes in the human genome, each taking part in a function happening in the body [78].

There are two main gene types that have the potential to cause cancer. The first category is the oncogenes that regulate cell proliferation and growth together with controlling the cell cycle and apoptosis [79]. The second one is the tumor suppressor gene (TSG) or anti-oncogenes that encodes a protein that regulates and checks cell division. Detrimental genetic variations called mutation on the oncogenes and anti-oncogenes are the critical reasons for developing cancer. There are two types of mutations:

- **Activating mutations** change the function of the gene or alternatively the time and level of the expression. These mutations in oncogenes possibly cause cancer.
- **Inactivating mutations** reduce functionality of the gene. These mutations in tumor suppressors possibly cause cancer [80].

2.2.3. Factors influencing genetic mutations

Mutations are classified into two classes: germline and somatic mutations. Germline mutations are the ones that we inherit from the egg and sperm cells during conception. These mutations are the reason why we are not exactly like our parents. On the other hand, somatic mutations are the ones that happen after conception to all the cells other than the egg and sperm. The somatic mutations will be passed to the cells formed by the division (daughter cells) but not to the offspring. Cancer can be caused by somatic mutations possibly happening by chance during different cell division steps or environmental causes. Therefore, the main risk for cancer is aging. It also can be caused by inherited germline mutations.

2.2.4. Hallmarks of cancer

The hallmarks of cancer are a concept in the field of oncology, describing the fundamental properties that enable cancer cells to evade normal biological constraints such as cell proliferation, cell survival, and cell communication and afterward drive the development and progression of cancer. These hallmarks include replicative immortality, genome instability, evasion of growth suppressor signals, resistance to death, sustained proliferation, altered metabolism, avoiding immune destruction, and tumor-promoting inflammation. In this section, we will provide an overview of the hallmarks of cancer, including the molecular and cellular processes that enable cancer cells to bypass normal biological controls and contribute to cancer development.

Hallmark number 1: replicative immortality

The normal human cell is only able to divide for a finite number of times, since there are parts at the end of the chromosomes called telomeres, which are shortening after each division. The normal cell will go to the G0 phase after reaching the limit. On the other hand, the cancer cells are able to greatly exceed these limits using an enzyme called telomerase. This enzyme is found in most cancer cells and enables them to get divided infinitely [81].

Hallmark number 2: genome instability

If a change happens to a normal cell DNA during replication, the cell will notice it during G1 or G2 phases (gap phases) and stop the cell cycle to repair the DNA and then continue the cell cycle afterward. DNA repair mechanisms are controlled by tumor suppressor genes (TSGs) and genome instability often occurs due to a malfunction in the regulation of any of these mechanisms. Therefore, in cancer cells, notable gene alterations are observed, such as point mutations, the deletion of regions of chromosomes, and loss of heterozygosity (LOH) [82].

Hallmark number 3: evasion of growth suppressor signals

Damage to the suppressor signals might be caused by a mutation in a TSG, which leads to uncontrolled growth. For example, a TSG called retinoblastoma (Rb) does not allow the normal cell to go through some restriction point in the G1 cell cycle phase. Mutations on this gene caused by genome instability, disrupt this mechanism and allow the cell to pass that phase even if it is not meeting the conditions. Another example is a TSG called p53, which is responsible for cell death after finding damage to the DNA. P53 alterations will let the damaged cell live and proliferate. Therefore, a failure of the suppressor signals might allow cells to bypass necessary restrictions, which can contribute to the development of cancer [83].

Hallmark number 4: resistance to death

Apoptosis is a programmed cell death, which means that the aged or DNA-damaged cells undergo the death process to naturally be removed from the body. Cancer cells by deregulating the apoptotic signaling such as activation of anti-apoptotic systems, escape from apoptosis, which leads to uncontrolled proliferation. For example, cancer cells use pro-cell-survival proteins, such as the Bcl-2 family to avoid apoptosis. Therefore, the deregulation of apoptosis, or programmed cell death, is a crucial hallmark in the development and progression of cancer [84].

Hallmark number 5: sustained proliferation

Normal cells control their growth and proliferation process by balancing and tightly controlling the signals that promote cell division. But cancer cells ignore these controls and become unstoppable. Loosening this control happens when cancer cells knock down the tumor suppressor genes and over-activate oncogenes such as RAS and SARK. This is debatably the pivotal property of cancer cells and involves their ability to sustain chronic proliferation [85].

Hallmark number 6: altered metabolism

The normal cells take the glucose and change it to CO2 for metabolism and gaining energy. On the other hand, tumors have increased metabolic demands compared to normal cells and require a constant supply of nutrients for growth and survival. Therefore, to meet these demands, tumors alter the way energy is produced and used in the body in order to support their growth and development. For example, some of them take another path by turning glucose into lactate to advance survival and proliferation. Both cancer-causing proteins and non-coding RNAs control this process. Non-coding RNAs are small molecules that do not encode proteins but can regulate gene expression and play a role in tumor metabolism by fine-tuning the metabolic pathways in tumors. This altered energy metabolism is a crucial aspect of tumorigenesis, as it provides the necessary nutrients to support tumor growth and survival [86].

Hallmark number 7: avoiding immune destruction

The immune system in the body is responsible for identifying and destroying harmful invaders using immune cells such as T-cells. There are signaling pathways that activate and deactivate T-cells for attacking and avoiding foreign cells. For example, Programmed Death 1 and 2 Ligands (PD-L1 and PD-L2) play important roles in deactivating T-cells. Cancer cells can exploit these proteins to suppress T-cell activation and evade the immune response, a process known as avoiding immune destruction. This is known as immune evasion [87].

Hallmark number 8: tumor-promoting inflammation

Inflammation is a critical component of the immune system and serves as the body's initial response to injury, infection, or tissue damage. It helps to recruit immune cells to the affected area, remove harmful motive of the inflammation, and initiate the healing process. This means attracting immune cells to the site of injury or infection, where they can help to clear away damaged tissue, pathogens, and other harmful motives. This process also increases blood flow to the affected area, bringing with it oxygen and nutrients that are needed for repair, and triggers the release of growth factors and cytokines that promote tissue regeneration and remodeling [88]. If inflammatory cells stay too long, it may lead to chronic inflammation. There are multiple shreds of evidence that suggested that chronic inflammation can promote the tumor. In fact, cancer cells hijack the immune system mechanisms for inflammation to promote their own proliferation and survival by the extra oxygen and nutrients and even spreading to another part of the body, which is called metastasis [89, 90, 91, 85].

2.2.5. Tumor heterogeneity and evolution

Tumor evolution is a complex and dynamic process that occurs over time. As it is established, the generation of the tumor begins with the transformation of a single cell in the normal tissue,

which undergoes genetic and epigenetic changes that allow it to evade the body's natural mechanisms of cell growth regulation and apoptosis. This aberrant cell then multiplies and expands to form a mass of cells, which is the beginning of a cancerous tumor.

As the tumor grows and develops, the clonal lineages within it diverge and acquire different genetic and epigenetic alterations, resulting in the formation of distinct subpopulations of cells. This phenomenon is known as intratumor heterogeneity (ITH). The distinct subpopulations of cells within the tumor can have different biological properties, including differing responses to treatment, metastatic potential, and immune evasion [92, 93, 94]. Different aspects of the ITH can be categorized and studied in three different types of heterogeneity: genomic, phenotypic, and spatial heterogeneity.

Genomic heterogeneity

Genomic heterogeneity refers to the genetic differences that exist between different cells in a tissue. It can manifest in different ways, including variations in DNA sequence, epigenetic modifications, or chromosome number and structure. In cancer, genomic heterogeneity is a common feature, where cells within a tumor and between the clones can have different genetic mutations. Genomic heterogeneity can also exist between individuals, as people can have variations in their DNA sequences that affect their susceptibility to diseases or response to drugs. For example, some people may have genetic variations that make them more likely to develop certain types of cancer, while others may be more resistant to infections [94].

Understanding genomic heterogeneity is important for developing personalized medicine approaches, where treatments can be tailored to an individual's unique genetic makeup. It also has a significant effect in defeating drug resistance by identifying the resistant clone in advance [92].

Phenotypic heterogeneity

Phenotypic heterogeneity refers to the variation in observable traits or characteristics between different cells or individuals. This can include differences in gene expression, morphology, behavior, metabolism, or response to environmental stimuli. Phenotypic heterogeneity can be caused by both genetic and non-genetic factors, including epigenetic modifications, environmental influences, and stochastic processes [93].

Spatial heterogeneity

Spatial heterogeneity refers to the variation in biological properties or behaviors that exist across different locations within a tissue or organism. This includes differences in cell density, nutrient availability, oxygen concentration, or cellular interactions. Spatial heterogeneity can be important for various biological processes, such as development, tissue regeneration, and immune response.

These three types of heterogeneity are interconnected and can influence each other. For example, genomic variation can lead to phenotypic and spatial heterogeneity, while environmental factors can modulate both genomic and phenotypic heterogeneity. Understanding these different types of heterogeneity is important for advancing our knowledge of biology and developing effective diagnostic and therapeutic strategies [95, 96].

2.3. Focal types of cancer in our projects

As the focus of this thesis is on the genomic, phenotypic, and spatial heterogeneity of tumors, we have chosen to investigate three types of cancer that are known to exhibit high levels of heterogeneity [97, 98, 99]. In the following sections, we provide a short description of each of these cancer types.

2.3.1. Prostate cancer

Prostate cancer is a type of cancer that affects the prostate gland in men that locates below the bladder. It can progress from a localized form with a good prognosis to an aggressive and lethal form. Advances in genomics have allowed the identification of genes responsible for the development and progression of the disease. This cancer is characterized by high levels of genomic instability, including the overexpression of the androgen receptor, also known as *NR3C4*, and mutations in genes such as *PTEN*, *HOXB13*, and *TP53* [100, 101]. Studies have shown that genetic abnormalities in DNA repair pathways are involved in the development and prognosis of prostate cancer and that genes such as *BRCA1* and *BRCA2* are associated with increased risk and severity of the disease [102]. Recent advances in genomic research have led to the discovery of new therapeutic targets for advanced cases of prostate cancer that are resistant to traditional treatments [103, 104].

2.3.2. Breast cancer

Breast cancer is the abnormal growth of malignant cells in the breast and it occurs mostly in women and in some cases in men. Several factors that increase the risk of breast cancer have been identified, including hormonal and lifestyle factors, but the most significant one is family history [105, 106]. The discovery of risk genes for hereditary breast cancer progressed in the 90s through linkage analyses and candidate gene approach, leading to the identification of *BRCA1, BRCA2, TP53, STK11, CDH1, PDL1, PIK3CA* and *PTEN* [107, 108, 109, 110, 111, 112, 113]. Treatment approaches for breast cancer often utilize these breakthroughs to predict or treat the disease. One recent example is using *PDL1* as a predictor for response to immunotherapy. Also, genes like *BRCA1*, *BRCA2*, and *PIK3CA* are used to classify breast cancer and the latter is used to determine systemic treatment. It is anticipated that more genes will be utilized in the future to inform treatment decisions [114].

2.3.3. Follicular lymphoma (FL)

Follicular lymphoma is a cancer of the blood and immune system, which is characterized by the abnormal growth of B-cells, in the follicles (small rounded masses) of lymphatic tissue. B-cells are a type of white blood cell involved in the immune response and the B-cell receptor (BCR) is important in their function and regulation. Specifically, the BCR plays a crucial role in the recognition and response to foreign antigens, as it allows the B-cells to bind to specific antigens and initiate the immune response. Abnormalities in the BCR can contribute to the development of follicular lymphoma. The growth of follicular lymphoma is typically slow, but it can still spread to other parts of the body over time. Symptoms can include swollen lymph nodes, fatigue, and weight loss, but many people with the condition have no symptoms [115].

Chapter 3

Data

3.1. Data characterization

3.1.1. Whole exome sequencing

Whole exome sequencing (WES) is a technique that determines the sequence of all protein-coding regions in the human genome, which makes up about 1-2% of the total genome. This technique typically requires a tissue sample, which can be obtained through a biopsy or other surgical procedure. The number of cells in a tissue sample can vary widely, depending on the size and type of tissue collected. In general, biopsy samples from solid tumors can contain millions of cells, while fluid-based samples, such as blood or cerebrospinal fluid, may contain only a few hundred to a few thousand cells [116, 117].

The data generated from WES can be used to identify genetic mutations, including single nucleotide variations (SNVs), insertions, deletions, and structural variations. These identified mutations can include those that contributed to the development and progression of cancer. Analysis of those mutations can give insights into the evolution of the tumor and the genetic basis of cancer [118, 119]. In all of my projects, we use WES to call the somatic mutations, infer the tumor evolution and identify the genotype of the clones.

3.1.2. Spatial transcriptomics

Spatial transcriptomics (ST) is an impactful molecular profiling method that enables the analysis of gene expression patterns within a tissue or organ, providing a spatially resolved view of the molecular landscape of the sample. It consists of spots across the tissue with specific coordinates. Each spot includes multiple probes and each probe is capturing the expression of the specific gene. Besides, all the probes in one spot have unique molecular barcodes corresponding to the coordinates of that spot. Using this method, we can measure the gene activity in tissue and record its coordinates [42]. This provides information on the gene expression patterns within a tissue section, allowing for a highly detailed view of the molecular and cellular composition of the tissue. The resulting data can be analyzed to identify patterns of gene expression that are associated with specific cell types, tissue structures, or disease mechanisms [120]. In my second and third projects, we use the reads over the mutations captured in the ST data for deconvoluting the clonal composition of the spots across the tissue. We also utilized the aggregated expression of the genes in each spot to estimate the clone-specific gene expression profiles.

3.1.3. Hematoxylin and Eosin stained images

Hematoxylin and Eosin (H&E) staining is a commonly used technique in histology, the study of tissue structure, to visualize and distinguish different cellular and tissue structures. Hematoxylin is a basic dye that stains nuclei blue, and eosin is an acidic dye that stains cytoplasm and extracellular matrix pink to red. The H&E staining procedure involves first treating tissue sections with hematoxylin to stain the nuclei, followed by treatment with eosin to stain the cytoplasm and other tissue components. To produce H&E images, a thin slice of tissue is first fixed in formalin, embedded in paraffin wax, and then cut into thin sections using a microtome. These sections are placed on slides and subjected to H&E staining. Finally, the stained slides are examined under a microscope, and images are captured using a camera or a scanner [41, 121].

H&E makes it possible to see nuclei and the shape of the cells. Besides, using these images, pathologists are able to distinguish cancerous cells from normal cells and evaluate the overall architecture of the tumor tissue. We use H&E images to annotate the cancerous spots in the ST data and count the number of cells inside each spot [122, 123].

3.1.4. Single cell RNA-sequencing

Single-cell RNA-sequencing (scRNA-seq) is a technology for transcriptional profiling (the set of all RNA molecules, including mRNA, in a single cell) of individual cells. The process of generating scRNA-seq data typically involves isolating individual cells from a tissue sample, lysing the cells to release their RNA, and reverse transcribing the RNA into complementary DNA (cDNA). The cDNA is then amplified, fragmented, and sequenced using high-throughput sequencing technologies, such as Illumina or Oxford Nanopore. The resulting data is then processed and analyzed to produce a profile of gene expression in each individual cell [124, 125].

ScRNA-seq is a powerful tool for studying cellular heterogeneity, uncovering new cell types, and understanding gene regulation and cellular pathways at the single-cell level [126]. In our CACTUS project, we use the reads over the mutations captured in the scRNA-seq data for mapping the cells to the clones.

3.2. Extraction of relevant features

3.2.1. Variant calling

The first step in genome analysis is identifying genetic variations through variant calling. This is achieved by comparing a genome of a sample (e.g. probed from a patient's tumor) to a reference genome and detecting variations by counting altered read counts at differing positions. However, this approach does not take into account factors such as noise and error. The current methods such as GATK [127], VarScan [128, 129], VarDict[130], and Strelka2 [131] utilize probabilistic, statistical, or machine learning-based techniques to infer variants more robustly and precisely [132].

Variant calling is a crucial step in cancer research as it enables us to identify and quantify the somatic changes in the DNA sequences and understand the evolution of the cell population in a tumor. By analyzing these genetic variants, we can gain insights into the accumulation of genetic changes. We used Varscan (v2.3.9 [129]) for the variant calling in CACTUS and Vardict [130] for variant calling in Tumoroscope and ClonalGE.

3.2.2. Copy number alterations

Copy number alterations (CNAs) refer to changes in the number of copies of specific DNA segments within a genome. These changes can be either a gain (amplifications) or a loss (deletions) and can affect one or multiple genes. CNAs can have a significant impact on cellular processes and are associated with various diseases, including cancer, its development, and progression [133, 134, 135, 136]. CNA regions often contain genes, which leads to differing levels of gene expression, so they may play a major role in normal and abnormal phenotypic variation [137, 138, 139].

The general CNA calling involves determining the overall copy number of a specific DNA segment in a sample, typically compared to a reference genome or a normal control sample. Allele-Specific CNA calling, on the other hand, goes a step further by determining the copy number of each individual allele (one of two or more versions of DNA sequence at a given genomic location) in a sample. This can be useful in the case of heterozygous (having two or more different alleles) DNA segments, where a change in copy number of one allele may have different implications than a change in copy number of all alleles [140, 141].

There are several methods used for general or allele-specific calling CNAs from next-generation sequencing (NGS) data including window-based methods, segmentation-based methods, and Bayesian approaches such as Hidden Markov Model (HMM)-based methods. In Tumoroscope and ClonalGE, we used Falcon-X [142] for allele-specific CNA calling.

Chapter 4

Statistical models

The content of this chapter mostly follows and summarizes the material that is in much more detail introduced in the book "Probabilistic Graphical Models: Principles and Techniques" [143].

4.1. Probabilistic graphical models and sampling methods

Real-world problems involve a significant amount of uncertainty, which arises because of limitations in observing the world and modeling it, as well as because of its innate non-determinism. For solving these problems we need a reasoning system that can reach a conclusion using the available information. Due to the huge uncertainty, the reasoning system should consider different possibilities alongside their probabilities. Such a complex system is characterized by different interconnected aspects of the domain. These are formally called random variables, denoted $X = (X_1, ..., X_N)$. Random variables describe important properties of the world. We construct the joint distribution over this set of random variables, $P(\chi_1, ..., \chi_N)$, to describe the state of the system. We aim to reason probabilistically about one variable given the observations of some other variables.

Probabilistic graphical models (PGMs) are a class of statistical models that represent complex relationships among variables using graph structures. The nodes in the graph represent variables in our domain and the edges represent the probabilistic interaction between them. They provide a way to compactly represent complex joint probability distributions and efficiently make inferences about the variables based on observed data. One can interpret a PGM from two points of view. Firstly, it can be interpreted as a set of independencies in the distribution. Secondly, the graph could be a representation of the high-dimensional distribution that we can break up into smaller factors and write the joint probability as a product of these factors. Both perspectives are mathematically equivalent. For more details, please refer to [143].

Two main families of graphical models are Bayesian networks and Markov networks. Bayesian network is a directed acyclic graph, in which the edges have a source and a target. Markov network is an undirected graph, in which the edges show the dependencies of the connected nodes without any direction. These models are widely used in a variety of fields, such as machine learning, artificial intelligence, and cognitive science. They are powerful tools for making inferences, predictions, and decisions in uncertain environments. In my thesis, we focus on the Bayesian networks to model the causal relationship of the variables using domain knowledge.

The framework of the PGMs has three main advantages that are critical components in

constructing intelligent systems: representation, inference, and learning. Firstly, the graphical model is an accurate and understandable reflection of a real-world, complicated system, usually constructed based on domain expert knowledge. Secondly, there were inference algorithms proposed for computing the posterior probability of the random variables (see below), which makes it possible to answer possible queries about the system, using the distributions. Thirdly, although some rough guidelines are defined by the domain expert, PGMs support the data-driven approach to learning by fitting the model to the data.

4.1.1. The Bayesian networks representation

Let us consider that we are given a set of binary variables ($X = (\chi_1, ..., \chi_N)$). One general representation of a complex system is using the joint distribution over all the variables: $\mathbb{P}(\chi_1, ..., \chi_N)$. This representation for large number of variables is unmanageable from different perspectives. Firstly, we have 2^N different states, which is huge to store in memory. Secondly, it is cognitively impossible to be acquired by an expert. Thirdly, there are huge amount of parameters, which needs huge amount of data to be learned.

Using graphical models representation, we take advantage of the independence properties and reduce the number of parameters drastically.

Naive Bayes

One example of a common and easy to explain graphical model is the naive Bayes model (Fig. 4.1). The model includes K class variables $C = (C_1, C_2, ..., C_K)$ and N observed variables $X_1, X_2, ..., X_N$. In this model, the observed variables given the instance class are conditionally independent of each other. This conditional independence reduces the number of parameters necessary for the joint distribution representation to the linear scale.

$$\mathbb{P}(C, X_1, ..., X_N) = \mathbb{P}(C) \prod_{i=1}^{N} \mathbb{P}(X_i|C)$$

This model is an easy to use model for clustering and calculating the confidence of assignment of a sample to a specific class.

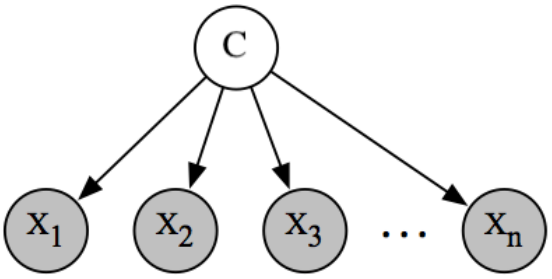


Figure 4.1: **The naive Bayes graphical model.** White node represents the hidden class variable and gray nodes represent the observed variables.

Bayesian networks

Bayesian networks are directed graphs, meaning that the arrows in these models show the direct influence of one node on another. Also, they are acyclic, meaning that they do not

include any cycle of dependencies. These models, such as the naive Bayes model, take advantage of conditional independencies between subgroups of variables, which makes the inference and learning efficient, with a reasonable number of parameters to be learned.

Independencies in Bayesian networks

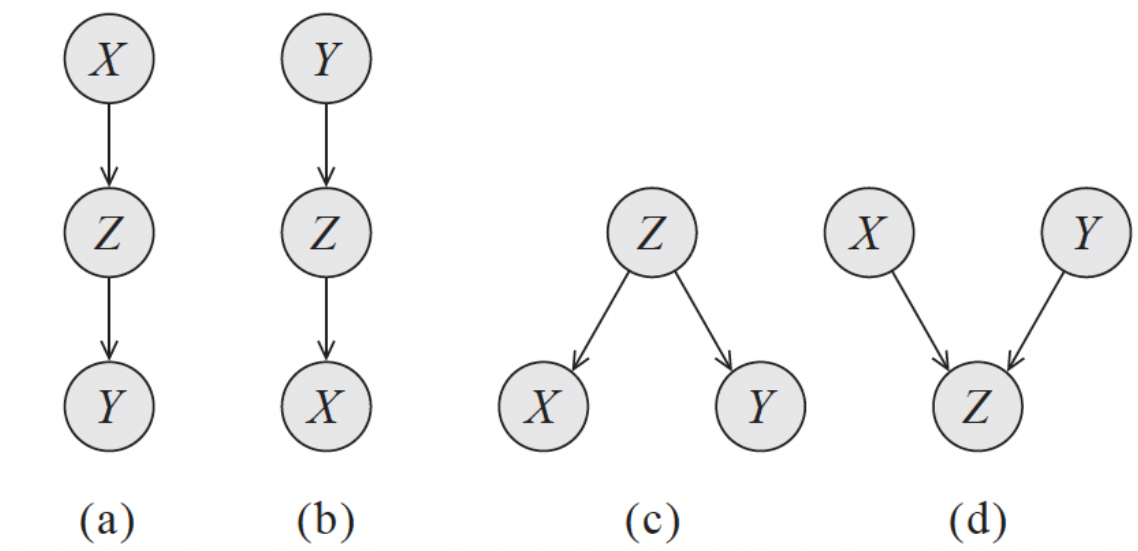


Figure 4.2: **Four possible trail from X to Y via Z.** Figure taken from the book "Probabilistic graphical models: principles and techniques" [143].

Consider two variables X and Y in a Bayesian network. They can be connected directly and therefore have an influence on each other. But they may be connected via a trail of variables. We consider the simplest situation where X and Y are connected via one node, Z. We have two arrows between them and each arrow have two possible directions, which leads to four different states (Fig. 4.2).

- (a) Causal trail: In this case, if we have not observed Z, Y is dependent on X. But if we observe Z, we do not need any information from X for knowing about Y. Therefore $Y \perp\!\!\!\perp X|Z$.

- (b) Evidential trail: This case is similar to (a) as dependence is a symmetric notion. Therefore, we have $Y \perp\!\!\!\perp X \mid Z$.

- (c) Common cause: In this case, if we do not know about Z, X and Y have an influence on each other. But if we observe Z, then they both only get influence from observed Z, and not each other. Therefore, we have $Y \perp\!\!\!\perp X|Z$ again.

- (d) Common effect: In this case, Both X and Y are having an influence on Z without having any correlation with each other. But if we know about Z, we can guess about Y and Y together. Therefore $Y \not\perp\!\!\!\perp X \mid Z$.

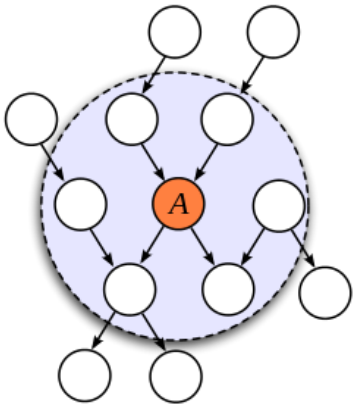


Figure 4.3: **Markov boundary of a node in a Bayesian network.** Figure taken from the Wikipedia [144].

Markov Blanket

In Bayesian networks, for a query about variable $X_i \in S = (X_1, ..., X_N)$, we only need a subset of variables (S_1), which have all the useful information about X_i. Therefore, X_i is independent from all the other variables in S given S_1,

$$X_i \perp\!\!\!\perp S \backslash S_1 \mid S_1.$$

This subset S_1 is called Markov blanket of variable X_i. Also, we can call it Markov boundary if we can not remove any variable from this subset. It can be proved that in the Bayesian networks, the Markov boundary of a node contains all its parents, children and the parents of all of its children (Fig. 4.3).

4.1.2. Probabilistic inference and learning

In graphical models, the inference is the process of computing probabilities or expectations for some variables given some other variables [145]. There are three known inference queries that are used in several applications: likelihood, conditional probability, and most probable assignment.

- **Likelihood** In probability theory, evidence refers to the available information that can be used to determine the probability of an event. The simplest query in Bayesian networks is computing the probability of the evidence given the model and parameters, which is called the likelihood.

  $$\mathscr{L}(\theta) = \mathbb{P}(x|\theta).$$

  Where here $\mathscr{L}(\theta)$ is the likelihood function, $\mathbb{P}(x|\theta)$ is the probability density of the observed data x given the hidden variable θ.

- **Posterior** Assume now that θ is a hidden variable with some prior distribution $\mathbb{P}(\theta)$. The evidence (observed variables) tells us some information about the hidden variable. Based on this information, we can calculate how probable different values of the variable are. This calculation is the conditional probability distribution of the variable given the evidence, which is called the posterior. Calculating the posterior has different applications such as prediction, diagnosis, and learning under partial observation.

  $$\mathbb{P}(\theta|x) = \frac{\mathbb{P}(x|\theta)\mathbb{P}(\theta)}{\mathbb{P}(x)}.$$

Where here x is observed data and θ is hidden variable. Also, it can be understood that the posterior is proportional to the product of the likelihood $\mathbb{P}(x|\theta)$ and the prior probability $\mathbb{P}(\theta)$ [146].

- **Most probable assignment** In this query, we are interested to find an assignment of values for a subset of variables that maximizes their posterior given the evidence. The most important application of this query is the classification in which we find the most likely label given the evidences.

In this work, we are interested in an approximation of the conditional probability. Computing the posterior in a graphical model for general DAGs is an NP-hard problem. It means that we do not have a general efficient way to solve it. There are many approaches to solve the inference problem, and it is classified in two categories of exact inference algorithms and approximate inference techniques.

- Exact inference algorithms can compute the exact answer to any query, but they are often computationally expensive and impractical for large or complex models.

- Approximate inference techniques can provide an approximate answer to a query, but they are usually faster and more scalable than exact inference algorithms. Approximate inference techniques include sampling methods such as Monte Carlo methods, variational methods such as mean field approximation, or neural network-based methods such as deep generative models [147, 143].

Since in this thesis we are designing complex graphical models, we are using efficient Monte Carlo methods including Markov Chain Monte Carlo (MCMC) and Metropolis-Hastings (MH) to approximate the inference.

4.1.3. Markov Chain Monte Carlo (MCMC)

A Markov process is a type of stochastic model, which describes a series of possible outcomes, such that each of them are dependent only on the previous one. In this process, the present event determines the probability of the next event [148, 149]. Markov chain Monte Carlo (MCMC) is a method of sampling from a given probability distribution using a Markov process, and it includes different methods such as Metropolis-Hastings and Gibbs sampler.

Metropolis-Hastings (MH)

The Metropolis-Hastings (MH) algorithm is a popular MCMC method for sampling from complex probability distributions, including those represented by PGMs. The algorithm is used to generate a sequence of samples from the distribution of interest referred to as the desired distribution. The obtained samples can then be used to estimate various quantities of interest, such as marginal probabilities or expected values.

The basic idea behind the MH algorithm is to construct a Markov Chain whose stationary distribution is the same as the desired distribution. The MH algorithm starts with an initial state and then generates a new state by proposing a move from the current state and accepting or rejecting the move based on an acceptance ratio (Algorithm 1).

Algorithm 1 Metropolis-Hastings Algorithm

Choose an initial state x_0
for $t = 0$ to $T - 1$ **do**
    Sample $y \sim Q(y|x_t)$ from the proposal distribution
    Compute the acceptance ratio $\alpha = \frac{f(y)Q(x_t|y)}{f(x_t)Q(y|x_t)}$, where f is proportional to the desired distribution
    Sample $u \sim U(0, 1)$ from a uniform distribution
    if $u < \alpha$ **then**
        Accept the proposal and set $x_{t+1} = y$
    else
        Reject the proposal and set $x_{t+1} = x_t$
    end if
end for

In Algorithm 1, f is a function that is proportional to the desired density function, and Q is the proposal distribution, which should be easy to sample. The variance of the proposal distribution determines the step-size of the sampling, indicating the distance of the proposed sample y from the previously accepted one, x_t . In this algorithm x_{t+1} is corresponding to the next accepted sample. If y gets accepted, x_{t+1} will be set to y. The new state is accepted or rejected based on the acceptance ratio. This ratio is defined as the ratio of the desired distribution at the new state to the desired distribution at the current state, multiplied by the ratio of the proposal distribution at the current state to the proposal distribution at the new state

$$\alpha = \frac{f(y)Q(x_t|y)}{f(x_t)Q(y|x_t)}.$$

The Metropolis-Hastings algorithm has the advantage of being relatively easy to implement and is widely used in practice. It can be inefficient if the proposal distribution is not well-matched to the desired distribution, resulting in low acceptance rates or getting stuck in the local minima. Besides, it can be sensitive to the choice of tuning parameters, such as the variance of the proposal distribution or the number of iterations [150].

Gibbs sampler

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm and is a MCMC method that make it possible to sample complex posterior probability distributions. The basic idea behind the Gibbs sampler is to iteratively sample from the conditional distributions of the variables given the current values of the other variables (see Algorithm 2).

Algorithm 2 Gibbs Sampling Algorithm

Choose an initial state $(x_1^{(0)}, \ldots, x_d^{(0)})$
for $t = 0$ to $T - 1$ **do**
    for $i = 1$ to d **do**
        Sample $x_i^{(t+1)} \sim p(x_i|x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_d^{(t)})$ from the conditional distribution
    end for
end for

The Gibbs sampler has the advantage of being relatively easy to implement, particularly for models with a simple graphical structure. Additionally, it can converge faster than other

MCMC methods, particularly when the conditional distributions are easy to sample from. However, it can be sensitive to the choice of initial values and can be less efficient than other methods when the variables are highly correlated.

Metropolis-Within-Gibbs (MWG)

The basic idea behind the MWG sampler is to use the Gibbs sampler with the modification, that the MH step is used to update some of the variables, rather than sampling directly from their conditional distribution. It is used when some of the conditional distributions in the normal Gibbs sampling are not easy to sample from.

However, using the MH algorithm within the Gibbs sampler can also increase the computational cost, as the calculation of the acceptance probability in the MH algorithm can be quite expensive. Therefore, it is important to carefully evaluate the trade-offs before deciding to use the MWG sampler.

4.2. Measures

4.2.1. Dunn Index

Dunn Index (DI) is a measure of cluster validity that evaluates the compactness and separation of clusters. It is defined as the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance, with higher values indicating better clustering.

$$D = \frac{\min\limits_{1 \leq i < j \leq k} d_{ij}}{\max\limits_{1 \leq l \leq k} D_l} \tag{4.1}$$

where k is the number of clusters, d_{ij} is the distance between the i^{th} and j^{th} clusters, and D_l is the diameter of the l^{th} cluster, defined as the maximum distance between any two points in the cluster[151, 152].

4.2.2. Connectivity

The connectivity measurement for clusters is typically defined as the average distance between each data point and its nearest neighbor in the same cluster. This measurement is also known as intra-cluster connectivity or cohesion. The formula for intra-cluster connectivity is given as:

$$\text{connectivity}(C) = \frac{1}{n} \sum_{i=1}^{n} \min_{j \in C_i} \|x_i - x_j\|_2 \tag{4.2}$$

where C is the set of clusters, n is the total number of data points, C_i is the cluster that contains the i-th data point, x_i is the i-th data point, and $|x_i - x_j|_2$ is the Euclidean distance between the i-th and j-th data points.

In words, this formula calculates the average of the minimum Euclidean distances between each data point and its nearest neighbor within the same cluster. A higher intra-cluster connectivity indicates that the data points in each cluster are closer to each other, and therefore the clustering is more compact and well-separated [153].

4.2.3. RMSSTD

The Root Mean Squared Standard Deviation (RMSSTD) is a measure of clustering tightness, commonly used in data analysis and machine learning. It measures the amount of dispersion or spread of data points within each cluster. The formula for RMSSTD is:

$$\text{RMSSTD} = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{k}(x_i - c_j)^2}{N}}$$

where N is the total number of data points, k is the number of clusters, x_i is the i-th data point, and c_j is the centroid of the j-th cluster. The summation is taken over all data points and clusters. A lower RMSSTD indicates that the clusters are more compact and well-separated, while a higher RMSSTD indicates that the clusters are more diffuse and overlapping. Therefore, RMSSTD is used as a measure to evaluate the quality of clustering algorithms and to select the optimal number of clusters for a given dataset[151, 154].

4.2.4. Calinski-Harabasz Index

The Calinski-Harabasz Index (CH) is a measure of clustering quality that evaluates the ratio of between-cluster variance to within-cluster variance. It is calculated as follows:

$$CH = \frac{B(k)}{W(k)} \cdot \frac{n-k}{k-1}$$

where n is the total number of data points, k is the number of clusters, $B(k)$ is the between-cluster variance, and $W(k)$ is the within-cluster variance.

The between-cluster variance is the sum of squared distances between the cluster centroids and the overall data mean, weighted by the number of data points in each cluster. It is calculated as:

$$B(k) = \sum_{j=1}^{k} n_j \left|c_j - c\right|_2^2$$

where n_j is the number of data points in the j-th cluster, c_j is the centroid of the j-th cluster, c is the overall data mean, and $\left|c_j - c\right|_2$ is the Euclidean distance between the j-th cluster centroid and the overall data mean.

The within-cluster variance is the sum of squared distances between each data point and its cluster centroid, weighted by the number of data points in the cluster. It is calculated as:

$$W(k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left|x_i - c_j\right|_2^2$$

where x_i is the i-th data point, C_j is the j-th cluster, c_j is the centroid of the j-th cluster, and $\left|x_i - c_j\right|_2$ is the Euclidean distance between the i-th data point and the centroid of the j-th cluster.

In words, the CH index measures the ratio of the between-cluster variance to the within-cluster variance, adjusted for the number of clusters and the number of data points. A higher CH value indicates that the clusters are more compact and well-separated [155].

4.2.5. Entropy

Entropy is a measure of the randomness or uncertainty of a distribution. It is defined as:

$$H(p) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

where p_i is the probability of the i-th outcome, and n is the total number of possible outcomes. The logarithm base 2 is commonly used, in which case the units of entropy are "bits".

In words, the formula for entropy calculates the negative sum of the probability of each outcome times the logarithm base 2 of that probability. The resulting value is a measure of the amount of information or uncertainty in the distribution. A higher entropy value indicates greater randomness or uncertainty [155].

4.2.6. Gini Index

The Gini Index is a measure of cluster inequality that calculates the relative difference between the actual distribution of data points across clusters and a hypothetical equal distribution. It is defined as:

$$Gini = 1 - \sum_{i=1}^{k} p_i^2$$

where p_i is the proportion of the i-th category or class, and k is the total number of categories or classes. The proportion of the i-th category or class is the fraction of observations in a given class, which is used to compute the contribution of that class to the overall inequality measure.

$$p_i = \text{Proportion of class } i = \frac{\text{Number of observations in class } i}{\text{Total number of observations in all classes}}$$

The Gini Index ranges from 0 (perfect equality) to 1 (perfect inequality). The formula for the Gini Index calculates the difference between 1 and the sum of the squares of the proportions of each category or class. A higher Gini Index value indicates greater inequality or concentration of the distribution [156].

4.2.7. Mean Absolute Error

Mean Absolute Error (MAE) is a measure of the average magnitude of errors between predicted and actual values. It is calculated as the average absolute difference between the predicted values and the actual values. The formula for calculating MAE is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where n is the number of observations, y_i is the actual value for observation i, and $\hat{y}_i$ is the predicted value for observation i.

The MAE is a useful metric for evaluating regression models because it is easy to interpret and provides a measure of the average error in the units of the response variable. In contrast to other metrics like the root mean squared error (RMSE), the MAE is not as sensitive to outliers and is less influenced by large errors [157].

Chapter 5

CACTUS: integrating clonal architecture with genomic clustering and transcriptome profiling of single tumor cells

Tumor heterogeneity and clonal evolution present a major challenge for cancer therapy [4]. Tumor cells carry founder and subsequently acquired driver mutations that cause transformation of the healthy cell into an expanding population of malignant cells. Continuous acquisition of mutations creates populations of tumor cells with divergent mutational profiles. Diverging cells with acquired driver mutations result in preferential clonal expansion leading to intraclonal diversity. Given that distinct genotypes induce key phenotypic differences between the clones [158], gene expression variation between the clones is expected. Measuring the phenotypes of tumor clones, however, is challenged by the difficulties in resolving the clonal genotype-to-phenotype maps in tumors [159].

Follicular lymphoma (FL) is a common type of malignant B-cell lymphoma with characteristics of normal germinal center (GC) B-cells. FL cells maintain the typical follicle-like structure of normal GC reactions in response to pathogens. FL pathogenesis is founded by the paradigmatic translocation (14;18)(q32;q21) that places BCL-2 under transcriptional control of the IGH@ locus enhancer. Secondary drivers affect genetic modifiers that enhance germinal center (GC) formation, reduce B-cell differentiation and freeze FL cells in the GC stage [160, 161]. Despite commonly observed pathogenic genomic events, clinical behaviour of FL is unpredictable and ranges from spontaneous remission over long-term stable disease to transformation to aggressive B-cell lymphoma.

In addition, FL cells are continuously exposed to a physiological mutator mechanism, i.e. expression and action of activation induced cytidine deamidase (AID) [162]. AID focuses on B-cell receptor (BCR) loci and results in highly mutated BCR heavy and light chain genes in FL [163]. Whereas BCR mutations intrinsically may lead to a proliferative signal by acquisition of N-linked glycosylation [164], preferential expansion of clones with identical BCR can also be explained by co-acquisition of underlying driver mutations that enhance their proliferation. In addition to grouping of individual cells into evolutionary clones by exome-wide mutations and structural variants, single FL cells can also be clustered based on the expression of identical BCR sequences. BCR mutations can therefore be considered events in clonal evolution in FL and present suitable markers that may allow a more accurate reconstruction of clonal evolution than based on exome mutations only.

Elucidation of tumor evolution and reconstruction of the tumor clonal architecture is possible from bulk DNA sequencing [165, 166, 167, 168] and from single cell (sc) DNA sequencing data [169, 170, 171, 172]. The outcome of such evolutionary analysis is a set of tumor clones, defined by their genotypes and frequencies. The genotype indicates which mutations are present in each clone, and the frequency indicates the fraction of cells from that clone in the entire tumor cell population. The task of identifying the tumor clones and their genotypes is computationally very difficult [168], and thus the tumor clone genotypes inferred from DNA sequencing alone are likely to be imperfect.

Recent efforts into the direction of mapping genotypes to phenotypes in tumors include characterizing gene expression profiles of tumor clones based on matching the single cell RNA sequencing (scRNA-seq) readouts to copy number variants in the clones [173, 174, 175]. Poirion et al. [176] proposed a linear model detecting association of single nucleotide variants from scRNA-seq with gene expression. This approach, however, ignores the evolutionary history of the tumor, which can be resolved to determine the genotypes of the tumor clones. Such obtained genotypes can then be matched to mutations observable in scRNA-seq. Recently introduced cardelino [177] is the first approach to successfully utilize the mutation mapping between the clone genotypes and the variants in scRNA-seq data. The performance of this approach, however, can be hampered by the fact that single cell transcripts contain only information on 5' part of the RNA and that the data are sparse. With such limited data, the confidence of assigning single cells to clones, and thus also of clonal genotype to gene expression phenotype mapping, is also limited.Here, we define the confidence as the concentration of the probability distribution of the cell-to-clone assignment, with high confidence corresponding to a high probability of assignment to one clone, and low confidence corresponding to a uniform probability over clones. To increase the confidence, additional available evidence should be integrated into the inference. One such evidence is a given clustering of cells, such as the grouping of cells by their similar BCR sequences in FL evolution. Combining multiple data sources has the potential to increase the resolution of tumor heterogeneity analysis [178], but is computationally challenging [179] and calls for a dedicated probabilistic model.

Here, we propose a probabilistic graphical model for integrating Clonal Architecture with genomic Clustering and Transcriptome profiling of single tUmor cellS (CACTUS). The model extends cardelino [177] and maps single cells to their clones based on comparing the allele specific transcript counts on mutated positions to given clonal genotypes, leveraging additional information about evolutionary cell clusters. As part of the model inference, CACTUS corrects the input clone genotypes and adjusts the input cell clustering using all available data. The input clusters should be defined based on additional evolutionary information, in such a way that the model can assume that cells in the same cluster tend also to belong to the same tumor clone.
We apply CACTUS to newly generated whole exome sequencing (WES), scRNA-seq and single cell BCR sequencing data of FL tumor samples from excised malignant lymph nodes of two subjects. As a result, the single cells are assigned to their clones of origin, accounting for the similarities of their BCR sequences (Fig. 5.1). We demonstrate that guided by the BCR sequence information, CACTUS assigns single cells to tumor clones in agreement with independent gene expression clustering. For both subjects, CACTUS maps cells and BCR clusters with substantially higher confidence than cardelino. These results indicate that the important challenge of tumor genotype-to-phenotype mapping can successfully be approached by probabilistic integration of multiple measurements.

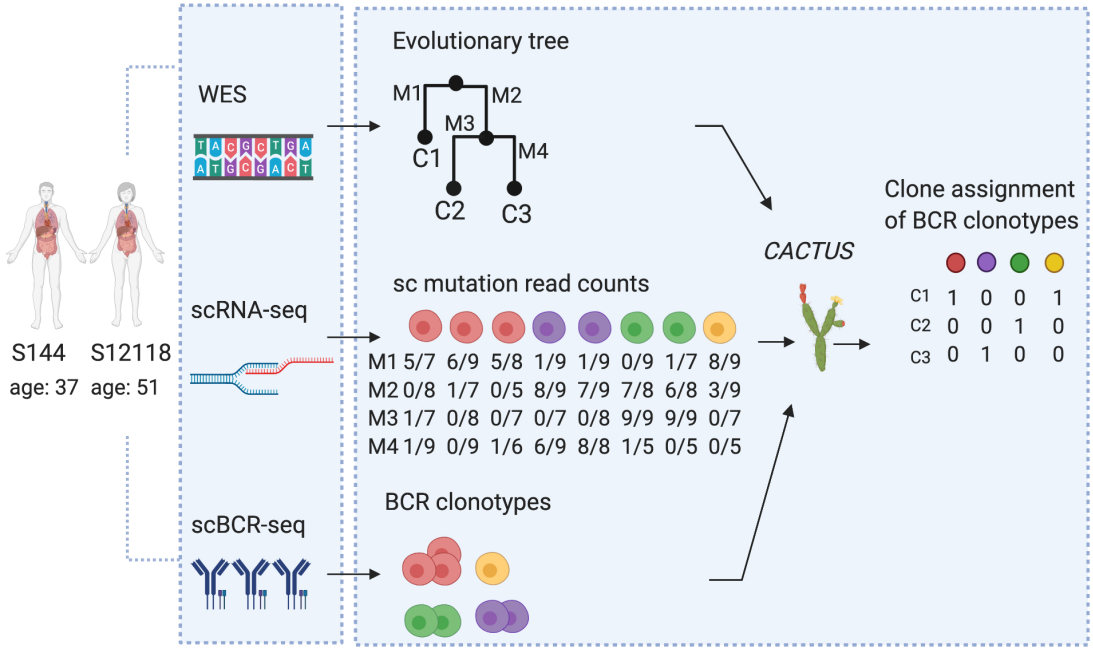


Figure 5.1: **Overview of the patient data analysis and the CACTUS model.** Whole exome sequencing and single-cell sequencing of all transcripts, as well as single-cell sequencing of BCR was performed on samples from two FL patients. Using WES, imperfect clonal evolution could be inferred and given as a prior to the model (C1, C2, ...). From scRNA-seq, allele specific transcript counts (mutated\total) were extracted at mutated positions (M1,M2, ...). Input BCR clusters were defined as clusters of cells with identical BCR heavy chain sequences. The data of input tumor clones, mutation transcript counts, and given single cell clusters (here, the BCR clusters) are combined in the CACTUS model for inference of the clonal assignment of the clusters. Both the input clone genotypes and clustering are considered potentially imperfect and are corrected during the inference using all available data.

5.1. Methods

5.1.1. Follicular Lymphoma sample preparation

Samples with histologically confirmed infiltration of follicular lymphoma were collected with approval by the institutional review board of Leiden University Medical Center according to the declaration of Helsinki and with written informed consent. Single cell suspensions were obtained by gentle mechanical disruption and mesh filtration and were cryopreserved using 10% DMSO as cryoprotectant. Remaining tissue was cultured in low-glucose DMEM to obtain stromal cell cultures for isolation of DNA of nonmalignant cells. Thawed single FL cells were purified by flow cytometry using fluorescently labeled antibodies specific for CD19 and CD10 and rested overnight followed by removal of dead cells using MACS dead cell removal kit. Cells of different patients were pooled and loaded on a 10X Genomics chip to obtain single cell cDNA libraries for an expected 1500 cells per patient. Following single cell cDNA library generation and amplification, one fraction was directly sequenced for 5' gene expression profiling. The second fraction was enriched for BCR transcripts by seminested amplification using 3' constant domain primers for all BCR genes, partially digested and sequenced. Both single cell libraries were sequenced in paired-end mode on Illumina (2x150 bp).

5.1.2. WES sequencing and mutation calling

FL single cells were purified by flow cytometry as described above to obtain bulk purified FL cells for immediate isolation of DNA. Whole exome sequencing (WES) was performed on paired FL and normal DNA at 200x and 50x coverage, respectively. Genomic DNA was isolated using the QIAamp DNA Mini kit (Qiagen). Samples were sequenced (HiSeq 4000 instrument, Illumina Inc) in paired-end mode on Illumina (2×101 bp) using TrueSeq DNA exome kit (v.6) (Illumina Inc.). Paired-end reads were aligned to the human reference genome sequence GRCh38 using BWA–MEM (V0.715-r1140) [180]. Deduplication and alignment metrics were performed using Picard tools (v2.12.1). Local realignment was performed around indels to improve SNP calling in these conflicting areas with the IndelRealigner tool. Recalibration to avoid biases was performed following the Genome Analysis Toolkit (GATK) Best Practices [127]. Single mpileup files were generated from paired bam normal/tumor using samtools mpileup (v1.6). Mutation calling and computation of somatic p-values (SPV) was performed on mpileup output files using Varscan (v2.3.9)[129] to WES data from tumor and patient-matched normal samples with a minimum coverage of 10x. Quality control metrics were assessed using FastQC (v0.11.2)[181] before and after the alignment workflow and reviewed to identify potential low-quality data files.

5.1.3. Single cell data processing

Sequencing data was processed with 10X Genomics Cell Ranger v2.1.1 with respect to GRCh38-1.2.0 genome reference to obtain UMI-corrected transcript raw gene expression count tables, BAM files and BCR all_contig.fasta files.

To generate single cell allelic transcript counts we used a custom made script to identify reads intersecting with WES-based mutated positions. For each read, to classify the allele we identified the single nucleotide overlapping the mutated base. To obtain transcript counts we used the unique molecular identifiers (UMIs) associated with the reads.

We used the vireo function from cardelino package v0.4.2 to construct clusters of cells sharing the same germline genotype. As input we provided allelic counts for the positions likely to differ between the subjects and not mutated between FL and stromal cells. For further processing we selected cells assigned to a single subject at minimum probability threshold of 0.75. Once the clusters of cells sharing the same germline genotype were identified, we assigned them to patients by comparing the cluster consensus genotype with the patient-labeled genotypes obtained from WES.

IMGT/HighV-Quest [182] was used for high-throughput BCR analysis and annotation of the BCR all_contig.fasta file [182]. IMGT/HighV-Quest output data was filtered for productive and rearranged sequences and FL cells with identical BCR heavy chains were considered unique BCR clusters within the malignant cell population and were annotated with unique identifiers. R-package 'vegan' was used to calculate Pielou's index of evenness for BCR cluster size distribution.

5.1.4. Phylogenetic analysis

For each subject, we first identified common mutations that can be found in both WES data and scRNA-seq data. Next, we used FALCON-X with default parameters for estimation of allele-specific copy numbers from WES data. As a verification, we compared the results of FALCON-X with those of GATK CNV analysis pipeline, and confirmed that the two approaches gave similar results. Finally, we run Canopy [165], providing the estimated major and minor copy number, as well as the allele-specific read counts in the tumor and matched

normal WES data as input. Taking advantage of a Bayesian framework, Canopy estimates the clonal structure of the tumor for a pre-specified number of clones. Choosing between trees with the number of clones from 2 to 4, for both subjects, the BIC criterion used by Canopy suggested trees with 4 clones as the best solution. For further analysis, for each subject, we selected the top tree returned by Canopy (see Additional file 1: Fig. S1 and Fig. S2 for the posterior likelihood and BIC plots of Canopy for subjects S144 and S12118, respectively).

5.1.5. Mapping BCR clusters to tumor clones using CACTUS

Below we introduce a probabilistic model, CACTUS, for mapping a given set of cell clusters to tumor clones based on the mutation matching between the cells in clusters and the clone genotypes (Fig. 5.2). In this analysis, the input clusters corresponded to sets of cells with identical BCR sequences. The input clustering and input clone genotypes were corrected during the inference process, taking into account all available data. Both CACTUS and cardelino are inferred using Gibbs sampling. For each subject, CACTUS was run for the top Canopy tree for a maximum of 20000 iterations of the Gibbs sampler, with 10 different starting points. For the sake of comparison, cardelino was applied with the same setup. CACTUS is a direct extension of cardelino [177], accounting for cell clustering, with the assumption that cells in the same cluster belong to the same clone. Let $i \in \{1, \ldots, N\}$ index mutation positions, which can be identified both in bulk DNA sequencing and single cell RNA seq data (see above). We assume we are given at input a set of K tumor clones, indexed by $k \in \{1, \ldots, K\}$. Each tumor clone is represented by its genotype and prevalence in the tumor population. The input clone genotypes are represented by a binary matrix $\Omega_{i,k}$ with entries equal 1 if the mutation i is present in clone k and 0 otherwise.

We are also given an independent clustering of single cells, where each cluster $q \in \{1, \ldots Q\}$ contains a number of cells and the clusters are assumed not to overlap. Let $j \in \{1, \ldots, M\}$ index cells. We assume that the input clustering is imperfect, and thus we define the true (corrected) clustering by a set of hidden categorical variables $\mathbf{T} = \{T_1, \ldots, T_M\}$, with each T_j taking values in $\{1, \ldots, Q\}$ and $T_j = q$ indicating that cell j is in cluster q. We assume a categorical distribution for T_j

$$P(T_j = q | p_{j,1}, \ldots, p_{j,Q}) = p_{j,q},$$

where $\sum_q p_{j,q} = 1$. The parameters of the categorical distribution $p_{j,q}$ are interpreted as the success probabilities for cell j to switch to cluster q. We assume these success probabilities are dependent on the input clustering of cells. Let $G_{j,q}$ denote the distance of the cell j to cluster q, obtained from the input clustering. Based on $G_{j,q}$, the probability $p_{j,q}$ is defined as

$$p_{j,q} = \frac{e^{-cG_{j,q}}}{\sum_{q'} e^{-cG_{j,q'}}},$$

where c is a constant determining the strength of the prior. This parameter should be defined by the user. Here, we set $c = 2$. In this application, the input clustering is defined as sets of cells with identical BCR sequences. Therefore, each input cluster is represented by the shared BCR sequence of its cells. Based on such input clustering, for each cell j and cluster q the distance $G_{j,q}$ is computed as the number of different mutations between BCR sequence of cell j and the representative BCR sequence of cluster q. Thus, the distance of q to its own cluster equals 0. For cells which did not have its BCR sequenced, we set their distance to their own cluster to 0, and their distance to all other clusters as equal to the mean of all known distances of cells to clusters.

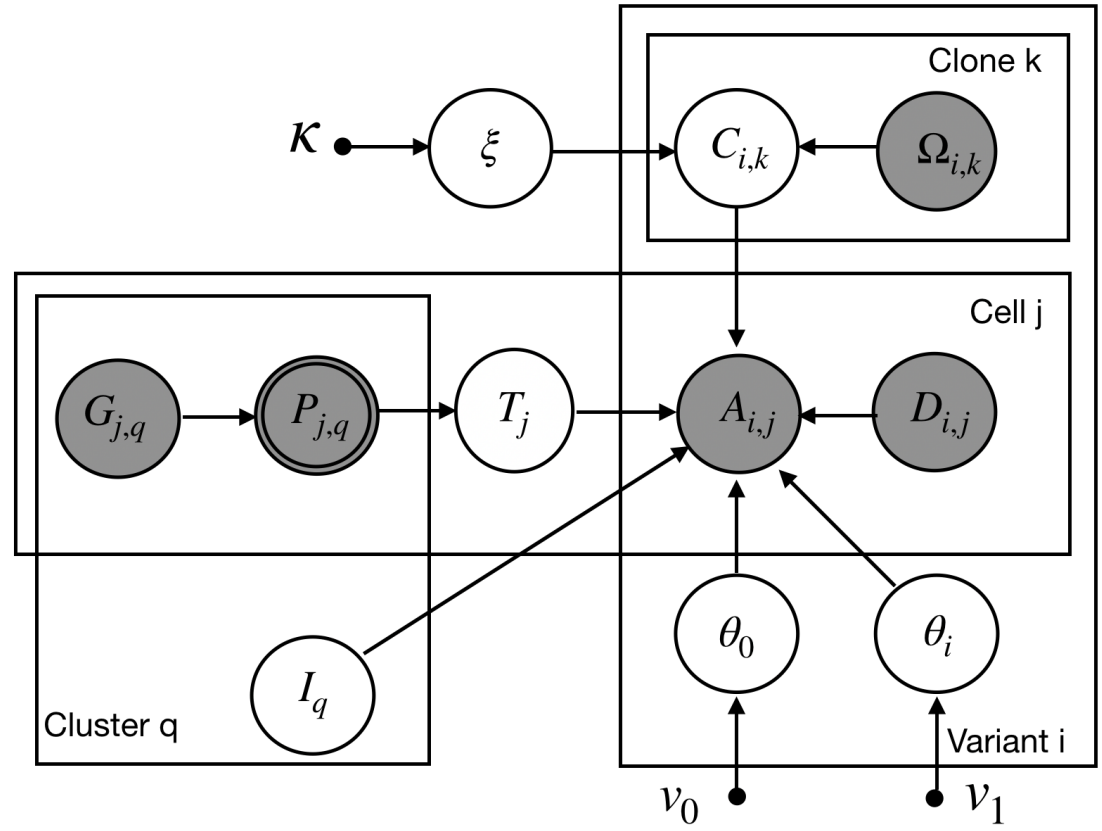


Figure 5.2: **The graphical model representation of CACTUS.** Circle nodes are labeled with random variables in the model. Arrows correspond to local conditional probability distributions of the child variables given the parent variables. Observed variables are shown as grayed nodes. Double-circled nodes are deterministically obtained from their parent variables. Small filled circles correspond to hyperparameters. $C_{i,k}$ denotes the true (corrected) genotype of clone k at variant position i. $\Omega_{i,k}$ denotes the input clone genotypes, with $\Omega_{i,k} = 1$ if the mutation i is present in clone k and 0 otherwise. $G_{j,q}$ denotes the distance of the cell j to cluster q, computed based on the input clustering of cells. $T_j = q$ indicates that cell j is in cluster q. $p_{j,q}$ is interpreted as the success probability for cell j to switch to cluster q. $A_{i,j}$ denotes the observed count of unique transcripts with alternative (mutated) nucleotide mapped to position i in cell j. $D_{i,j}$ denotes the total unique transcripts count mapped to that position in that cell. $I_q = k$ represents the assignment of cluster q to clone k. θ_i denotes the success probability of observing a transcript with the alternative nucleotide at a position i in a cell that carries this mutation, and θ_0 the success probability of observing a transcript with the alternative nucleotide in a position that is not present in the cell. ξ is the error rate for the genotypes. $\{\nu_0, \nu_1, \kappa\}$ constitutes the set of hyper-parameters in the model.

We are interested in assignment of the cell clusters to the clones. The clone assignment of each cluster q is represented in the model by a hidden variable I_q with values in $\{1, \ldots, K\}$. We assume a uniform prior for I_q and set $P(I_q = k) = \frac{1}{K}$. Alternatively, the prior could depend on the prevalences derived from the evolutionary model. The probability of cluster-to-clone assignment returned by CACTUS is computed from the Gibbs sampling iterations, as the posterior probability distribution of I_q. The single cells are assigned to each clone with the same probability as their cluster. Thus, for each cluster q and each cell j in q, the assignment probability of j to clone k equals the probability of assignment of q to k.

We assume that the input clone genotypes can contain errors with error rate ξ. The prior distribution for the error rate is parametrized by $\kappa = (\kappa_0, \kappa_1)$ and is set to $P(\xi|\kappa) = \texttt{Beta}(\xi; \kappa_0, \kappa_1)$. We define a hidden random variable $C_{i,k}$ denoting the true (corrected) genotype of clone k at variant position i, with

$$P(C_{i,k} = 1|\Omega_{i,k}, \xi) = \xi^{1-\Omega_{i,k}} \times (1-\xi)^{\Omega_{i,k}}.$$

Let matrix $\mathbf{A}$ with elements $A_{i,j}$ denote the observed count of unique transcripts with the alternative (mutated) nucleotide mapped to position i in cell j, and matrix $\mathbf{D}$ with elements $D_{i,j}$ denote the total unique transcripts count mapped to that position in that cell. Let θ_i denote the success probability of observing a transcript with the alternative nucleotide at a position i in a cell that carries this mutation, and θ_0 the success probability of observing a transcript with the alternative nucleotide in a cell that doesn't carry this mutation. The distribution of the observed read counts then becomes

$$P(A_{i,j}|D_{i,j}, I_q, C_{i,I_q}, \theta, T_j = q) = \begin{cases} \texttt{Binom}(A_{i,j}|D_{i,j}, \theta_0) & \text{if } C_{i,I_q} = 0 \\ \texttt{Binom}(A_{i,j}|D_{i,j}, \theta_i) & \text{if } C_{i,I_q} = 1. \end{cases}$$

We assume Beta priors on the θ parameters

$$\begin{aligned} P(\theta_i|v_1) &= \texttt{Beta}(\theta_i|\alpha_1, \beta_1) \\ P(\theta_0|v_0) &= \texttt{Beta}(\theta_0|\alpha_0, \beta_0), \end{aligned}$$

where $v_1 = (\alpha_1, \beta_1)$ and $v_0 = (\alpha_0, \beta_0)$. We denote $v = (v_0, v_1)$.

Let A_q be the matrix of alternative allele counts for cells contained in cluster q, at mutated positions, i.e., $A_q = (A_{i,j})_{j \in q, i=1,\ldots N}$, and let $D_q = (D_{i,j})_{j \in q, i=1,\ldots N}$. Since we assume the observed read counts at the different positions and different cells are independent, we have

$$P(A_q|D_q, I_q, \mathbf{C}, \theta, \mathbf{T}) = \prod_{j \in q} \prod_{i=1}^{N} P(A_{i,j}|D_{i,j}, I_q, C_{i,I_q}, \theta, T_j = q).$$

5.1.6. CACTUS model inference

We use Gibbs sampler, a Markov chain Monte Carlo (MCMC) algorithm for generating samples from the posterior distribution. We iteratively sample each hidden variable which is conditionally independent given the rest of the hidden variables in the model. The hidden variables in CACTUS include the cluster assignment matrix $\mathbf{I}$, the success probabilities of observing a transcript $\grave{} = (\theta_0, \theta_1, \ldots, \theta_N)$, the corrected clustering matrix $\mathbf{T}$, the corrected genotype matrix $\mathbf{C}$, and its error rate ξ. We describe the sampling steps for the full joint distribution of these hidden variables in the following.

Sampling clone assignment of clusters I_q

We sample cluster-to-clone assignment variable I_q, given the Markov Blanket of I_q in the graphical model (Fig. 5.2)

$$P(I_q = k|I_{-q}, \mathbf{A}, \mathbf{D}, \mathbf{C}, \mathbf{T}, \grave{}) \propto P(I_q = k)P(A_q|D_q, I_q = k, \mathbf{C}, \theta, \mathbf{T})$$
$$\propto \prod_{j\in q}\prod_{i=1}^{N}\{\texttt{Binom}(A_{i,j}|D_{i,j}, \theta_i)^{C_{i,k}} \times \texttt{Binom}(A_{i,j}|D_{i,j}, \theta_0)^{(1-C_{i,k})}\}. \tag{5.1}$$

Sampling success probabilities of observing a transcript `

Similarly, we sample θ from the posterior probability

$$P(\theta|\mathbf{A}, \mathbf{D}, \mathbf{C}, \mathbf{I}, \mathbf{T}, v) \propto P(\theta|v)\prod_{q=1}^{Q}\prod_{j\in q}\prod_{i=1}^{N} P(A_{i,j}|D_{i,j}, I_q, C_{i,I_q}, \theta, T_j = q)$$
$$\propto \texttt{Beta}(\theta_0|\alpha_0, \beta_0)\prod_{i=1}^{N} Beta(\theta_i|\alpha_1, \beta_1)$$
$$\times \prod_{q=1}^{Q}\prod_{j\in q}\prod_{i=1}^{N}\{\texttt{Binom}(A_{i,j}|D_{i,j}, \theta_i)^{C_{i,I_q}}\texttt{Binom}(A_{i,j}|D_{i,j}, \theta_0)^{1-C_{i,I_q}}\}$$
$$= \{\texttt{Beta}(\theta_0|\alpha_0, \beta_0)\prod_{q=1}^{Q}\prod_{j\in q}\prod_{i=1}^{N}\texttt{Binom}(A_{i,j}|D_{i,j}, \theta_0)^{(1-C_{i,I_q})}\}$$
$$\times\{\prod_{i=1}^{N}\texttt{Beta}(\theta_i|\alpha_1, \beta_1)\prod_{q=1}^{Q}\prod_{j\in q}\texttt{Binom}(A_{i,j}|D_{i,j}, \theta_i)^{C_{i,I_q}}\}. \tag{5.2}$$

Using the Beta-Binomial conjugacy, θ_0 and θ_i, for $0 < i < N$ are sampled from the Beta distribution

$$\theta_0|\mathbf{A}, \mathbf{C}, \mathbf{I}, \mathbf{T} \sim \texttt{Beta}(\alpha_0 + u_0, \beta_0 + v_0),$$
$$\theta_i|\mathbf{A}, \mathbf{C}, \mathbf{D}, \mathbf{I}, \mathbf{T} \sim \texttt{Beta}(\alpha_1 + u_i, \beta_1 + v_i), \tag{5.3}$$

where

$$u_0 = \sum_{q=1}^{Q}\sum_{j\in q}\sum_{i=1}^{N} A_{i,j}(1 - C_{i,I_q}), \qquad v_0 = \sum_{q=1}^{Q}\sum_{j\in q}\sum_{i=1}^{N}(D_{i,j} - A_{i,j})(1 - C_{i,I_q}),$$
$$u_i = \sum_{q=1}^{Q}\sum_{j\in q} A_{i,j}C_{i,I_q}, \qquad v_i = \sum_{q=1}^{Q}\sum_{j\in q}(D_{i,j} - A_{i,j})C_{i,I_q}.$$

Sampling the corrected clustering matrix T

The corrected sampling matrix $\mathbf{T}$ is sampled based on the Markov Blanket of $\mathbf{T}$ in the graphical model (Fig. 5.2),

$$P(T_j = q|\mathbf{p}, \mathbf{A}, \mathbf{C}, \mathbf{D}, \mathbf{I}, \theta) = \frac{P(T_j = q|p_{j,1}, \ldots, p_{j,Q})\prod_{i=1}^{N} P(A_{i,j}|D_{i,j}, I_q, C_{i,I_q}, \theta, T_j = q)}{\sum_{q'=1}^{Q} P(T_j = q'|p_{j1}, \ldots, p_{jQ})\prod_{i=1}^{N} P(A_{i,j}|D_{i,j}, I_q, C_{i,I_{q'}}, \theta, T_j = q')},$$

where we assume the categorical prior over T,

$$P(T_j = q|\mathbf{p}, \mathbf{A}, \mathbf{D}, \mathbf{C}, \mathbf{I}, \theta) \quad = \quad \frac{p_{j,q} \prod_{i=1}^{N} P(A_{i,j}|D_{i,j}, I_q, C_{i,I_q}, \theta, T_j = q)}{\sum_{q'=1}^{Q} p_{j,q'} \prod_{i=1}^{N} P(A_{i,j}|D_{i,j}, I_q, C_{i,I_{q'}}, \theta, T_j = q')}. \quad (5.4)$$

Sampling the corrected genotype matrix C

Similarly, the corrected genotype matrix $\mathbf{C}$ is sampled using the Markov Blanket of $\mathbf{C}$ in the graphical model

$$P(C_{i,k} = 1|C_{-(i,k)}, \mathbf{A}, \mathbf{D}, \theta, \mathbf{I}, \xi, \Omega_{i,k}, \mathbf{T}) = \frac{|\Omega_{i,k} - \xi| \prod_{q=1}^{Q} \prod_{j \in q} \texttt{Binom}(A_{i,j}|D_{i,j}, \theta_i)^{\mathbb{1}(I_q = k)}}{|\Omega_{i,k} - \xi| \prod_{q=1}^{Q} \prod_{j \in q} \texttt{Binom}(A_{i,j}|D_{i,j}, \theta_i)^{\mathbb{1}(I_q = k)} + (1 - |\Omega_{i,k} - \xi|) \prod_{q=1}^{Q} \prod_{j \in q} \texttt{Binom}(A_{i,j}|D_{i,j}, \theta_0)^{\mathbb{1}(I_q = k)}}, \quad (5.5)$$

where

$$|\Omega_{i,k} - \xi| \prod_{q=1}^{Q} \prod_{j \in q} \texttt{Binom}(A_{i,j}|D_{i,j}, \theta_i)^{\mathbb{1}(I_q = k)} = P(C_{i,k} = 1|\Omega_{i,k}, \xi) \prod_{q=1}^{Q} \prod_{j \in q} P(A_{i,j}|D_{i,j}, I_q, C_{i,I_q} = 1, \theta, T_j = q)$$

and

$$(1 - |\Omega_{i,k} - \xi|) \prod_{q=1}^{Q} \prod_{j \in q} \texttt{Binom}(A_{i,j}|D_{i,j}, \theta_0)^{\mathbb{1}(I_q = k)} = P(C_{i,k} = 0|\Omega_{i,k}, \xi) \prod_{q=1}^{Q} \prod_{j \in q} P(A_{i,j}|D_{i,j}, I_q, C_{i,I_q} = 0, \theta, T_j = q)$$

Here, we assume Bernoulli distribution over $\mathbf{C}_{i,k}$,

$$P(C_{i,k} = 1|\Omega_{i,k}, \xi) = \xi^{1 - \Omega_{i,k}} \times (1 - \xi)^{\Omega_{i,k}}$$

Indeed, we have $P(C_{i,k} = 1|\Omega_{i,k} = 1, \xi) = 1 - \xi$ and $P(C_{i,k} = 1|\Omega_{i,k} = 0, \xi) = \xi$. Thus, we can shortly write $P(C_{i,k} = 1|\Omega_{i,k}, \xi) = |\Omega_{i,k} - \xi|$. Similarly, for $C_{i,k} = 0$, we can write $P(C_{i,k} = 0|\Omega_{i,k}, \xi) = 1 - |\Omega_{i,k} - \xi|$.

Sampling the error rate ξ

We can compute the distribution of the error rate ξ having the corrected genotype matrix C, as well as the input clone genotype matrix Ω and hyperparameters κ as follows,

$$\begin{aligned} P(\xi|C, \Omega, \kappa) &= P(\xi|\kappa) \prod_{i}^{N} \prod_{k}^{K} P(C_{i,k} = 1|\Omega_{i,k}, \xi) \\ &= \texttt{Beta}(\xi; \kappa_0, \kappa_1) \times \xi^{1 - \Omega_{i,k}} (1 - \xi)^{\Omega_{i,k}}. \end{aligned}$$

From the Beta-Bernoulli conjugacy we obtain

$$P(\xi|\mathbf{C}, \Omega, \kappa) = \texttt{Beta}\left(\kappa_0 + \sum_{i,k} \mathbb{1}(\Omega_{i,k} \neq C_{i,k}), \kappa_1 + \sum_{i,k} \mathbb{1}(\Omega_{i,k} = C_{i,k})\right). \quad (5.6)$$

Finally, the Gibbs sampling algorithm for CACTUS was derived as a straightforward modification of the algorithm used for cardelino [177]. In the algorithm, I_q is iteratively sampled using Eq. (5.1) for $q = 1, \dots Q$, θ_i for $i = 1, \dots, N$ is sampled using Eq. (5.3), T_j is sampled for $j = 1, \dots, M$ using Eq. (5.4), $C_{i,k}$ for $i = 1, \dots, N$ and $k = 1, \dots K$ is sampled using Eq. (5.5), and ξ is sampled using Eq. (5.6).

5.2. Results

5.2.1. Single cell and WES profiling of two FL patients

The analyzed tumor cell populations were collected from lymph nodes of two FL patients: a male patient (S144) at the age of 37, who was diagnosed with an IgM expressing FL stage IV and a female patient (S12118) at the age of 51, who was diagnosed with an IgG expressing FL stage IV. To detect (sub-)clonal mutations, we performed WES at 200x coverage and called mutations between FL cells and paired stromal non-hematopoietic cells. We detected 398 somatic mutations for patient S144 and 1034 somatic mutations for patient S12118 with somatic p-value (SPV) < 0.1.

Next, for pooled samples of both subjects, we performed single cell sequencing of purified FL cells for full transcriptomes and BCR enriched libraries. We used the Vireo method [183] to group single cells back to the patients based on matching of alleles expressed in the single cells with germline mutations detected by bulk WES. Deconvolution of the whole transcriptome data yielded 1524 cells of subject S144 and 874 cells of subject S12118, respectively. BCR sequencing yielded BCR heavy chain sequences for approx. 70% of cells in both patients. Both samples were dominated by a limited number of larger BCR clusters (further referred to as multiplet BCR clusters), with many BCR clusters containing only one element (singleton BCR clusters). 'Pielou evenness index' was 0.59 for S144 and 0.53 for S12118, indicating moderate intraclonal diversification [184]. For generality, cells without BCR heavy chain sequences were considered to form a separate singleton cluster (see Additional file 1: Fig. S3 for BCR cluster size distribution).

5.2.2. A probabilistic model for assigning cell clusters to evolutionary tumor clones.

CACTUS is a Bayesian method that integrates three different sources of prior knowledge: (1) a set of tumor clones with their genotypes, (2) independently obtained non-overlapping cell clusters, and (3) scRNA-seq transcripts at mutated sites, to map each cell cluster to its corresponding tumor clone (Methods). Cells of the same cluster are assumed to come from the same tumor clone. Since the clusters are non-overlapping sets of cells, the cluster assignment to clones defines also the cell assignment (each cell in a given cluster is assigned to the same clone as its cluster).

Here, the input cell clustering was defined by the BCR sequences. Cells with the same BCR sequence are expected to be more likely to come from the same tumor clone. Thus, here CACTUS takes advantage of the extra information of BCR sequences to gain power and confidence of the assignment. During model inference, both the input clone genotypes and the input cell clustering are corrected, taking into account all available data. Thus, although the input clusters are defined as sets of cells with identical BCR sequences, during model inference the cells may swap between clusters, based not only on BCR sequence similarity but also based on shared sets of mutations.

CACTUS yields the posterior probability estimate for each given cell cluster to be mapped to each given clone. This probability is defined using a beta-binomial model for the allele specific transcript counts for each mutation and cell in this cluster. The model estimates the error rate for the given imperfect genotypes of the clones and outputs corrected genotypes. Similarly, the corrected clustering of single cells is returned. The likelihood of assigning a cluster to a given clone increases with the similarity of the mutation signal observed in the cells of the corrected cluster to the corrected genotype of that clone. Overall, the three most important hidden variables in the model are: the corrected clone genotypes, the corrected

clusters, and the assignment of corrected clusters to the clones by matching to their corrected genotypes. The final assignment of the clusters (and thus also their contained single cells) is obtained by selecting the most probable tumor clone for each corrected cluster (Fig. 5.1).

For both subjects, to define the input clonal structures, we first identified a set of mutations that could be identified both in WES and scRNA-seq data. We consider the mutation to be present in scRNA-seq if at least one variant read is observed. From the identified 398 mutations with SPV < 0.1 for subject S144 and 1034 mutations for subject S12118, for further analysis we selected only these mutations, for which any transcript expression was observed in scRNA-seq. Despite the relaxed significance level of 0.1 for the somatic p-values, we consider the common mutations as reliable, since they have evidence in both data sources. Only 5 out of 95 total resulting common mutations for subject S144, and 5 out of 133 common mutations for subject S12118, had somatic p-value in the $(0.05, 0.1)$ interval (Additional file 1: Fig. S4). Numbers of the common mutations vary in different cells (Additional file 1: Fig. S5). For further analysis we considered only cells which contain at least one of the common mutations. This included 1262 out of 1524 cells in subject S144 and 799 out of 874 cells in subject S12118.

We next applied Canopy to the WES data for the common mutations, and extracted the top tree and its corresponding clones, with their genotypes. To obtain the cell-to-clone assignment, CACTUS was applied to the obtained clonal structure, with a clustering of single cells defined by identical BCR sequences and scRNA-seq transcript counts as input. To demonstrate how the addition of the BCR clustering information improves the assignment of cells to clones, we applied cardelino [177] to the same Canopy trees and the scRNA-seq transcript counts. From these data, cardelino derived cell assignment to tumor clones. The two models (CACTUS and cardelino) are similar, but CACTUS can exploit the data more fully as it additionally takes into account the cell clustering (here, by BCR sequence) information into account. In fact, for the specific case of such uninformative clusters that contain exactly one cell, CACTUS reduces to cardelino. Thus, naturally, the advantage of CACTUS should be visible for such cells that are contained in clusters of more than one cell. It is important to note that both CACTUS and cardelino correct the input clone genotypes in their own way. Thus, the final genotypes of the clones might be similar, but obtained by correcting different initial clone genotypes. Therefore, keeping original labels of the clones would introduce artificial differences between the outputs of the two methods. To make a comparison of CACTUS to cardelino feasible, we first adjust the clone labels in such a way that clones with most similar corrected genotypes between the two methods share the same label (Additional file 1: Fig. S6).

5.2.3. CACTUS solution verified by an independent gene expression analysis

To validate the returned cluster-to-clone assignment and the induced cell assignment, we performed independent analysis of transcript expression levels obtained from scRNA-seq of the same cells. Note that here we describe gene expression as independent data since the transcript counts across all sites in the gene sequences are not used by CACTUS during inference. In contrast, CACTUS uses specific counts of those reads that map to the variant sites. Gene expression information is thus not used for model inference, only the signal for existence of mutations. We investigated whether the grouping of cells into the inferred clones tends to coincide with similarity of their expression profiles visually (Fig. 5.3, 5.4). To this end, we reduced the dimensionality of expression data using UMAP [185] provided in the Seurat package [186] and colored each cell with its corresponding clone inferred using CACTUS, and for a comparison, cardelino [177].

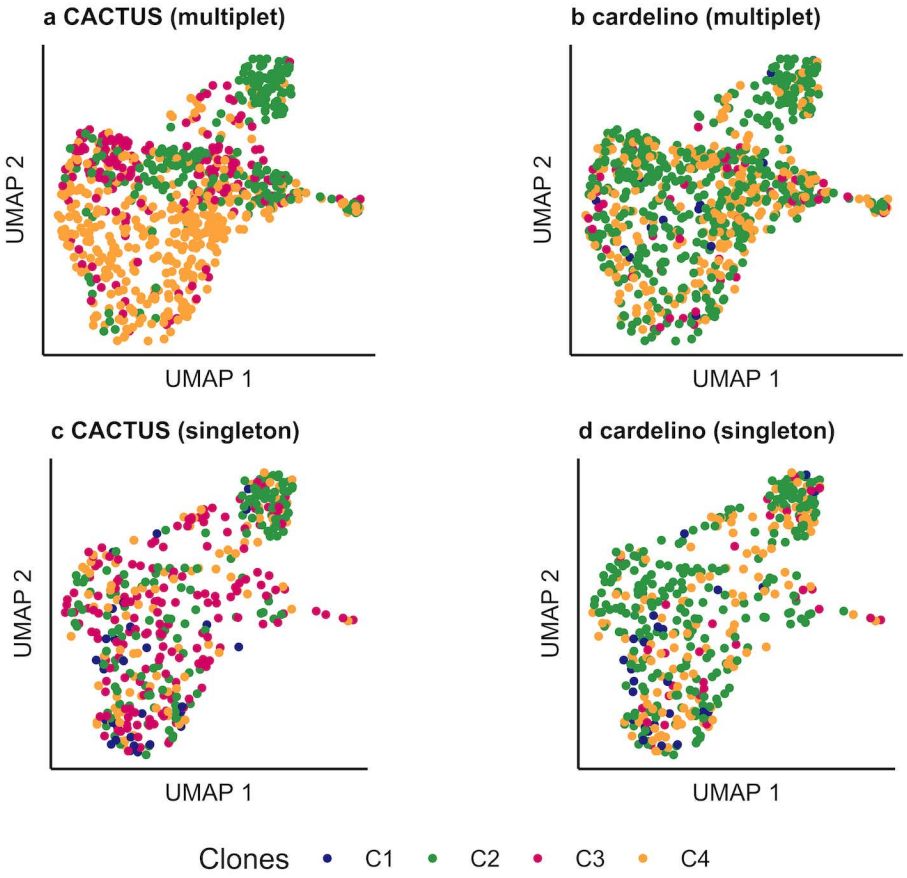


Figure 5.3: **Validation of cell-to-clone assignment with gene expression for subject S144. a, b, d, e** Transcript expression of the cells reduced to two dimensions using UMAP, shown separately for the cells in multiplet BCR clusters (**a**, **b**) and for cells belonging to singleton BCR clusters (**d,e**). Each point corresponding to a cell is colored by its clone assigned by CACTUS (**a**, **d**) and by cardelino [177] (**b**, **e**). The advantage of CACTUS in terms of agreement with gene expression is more pronounced for cells in multiplet BCR clusters.

As expected, CACTUS leverages information obtained from the multiplet BCR clusters. For cells in such BCR clusters, the results of CACTUS are more consistent with gene expression (visualized for UMAP in Fig. 5.3a and Fig. 5.4a) than the results of cardelino (Fig. 5.3b and Fig. 5.4b). For subject S144 and cells contained in the multiplet BCR clusters, CACTUS identifies clone C2 as a set of cells that is separated in gene expression space from a large cluster of cells, which is populated mostly by clone C4 and in part by clone C3. In contrast, cardelino finds clones which are mixed in the reduced gene expression space (Fig. 5.3a,b). For subject S12118, both methods associate clone C3 with one gene expression cluster and clone C4 with another, with the two gene expression clusters clearly separated in the reduced space. For CACTUS, the identified clones are slightly less intermixed with others than for cardelino (Fig. 5.4). For CACTUS, the clone assignments of cells in the singleton BCR clusters show less agreement with expression than assignments of cells in multiplet clusters (Fig. 5.3c and Fig. 5.4c). The agreement for those cells is comparably low for cardelino (Fig. 5.3d and Fig. 5.4d).

To quantify the agreement of the obtained assignment of cells to the clones with gene expression, we used a several quality measures [187]. To this end, for each cell and each subject,

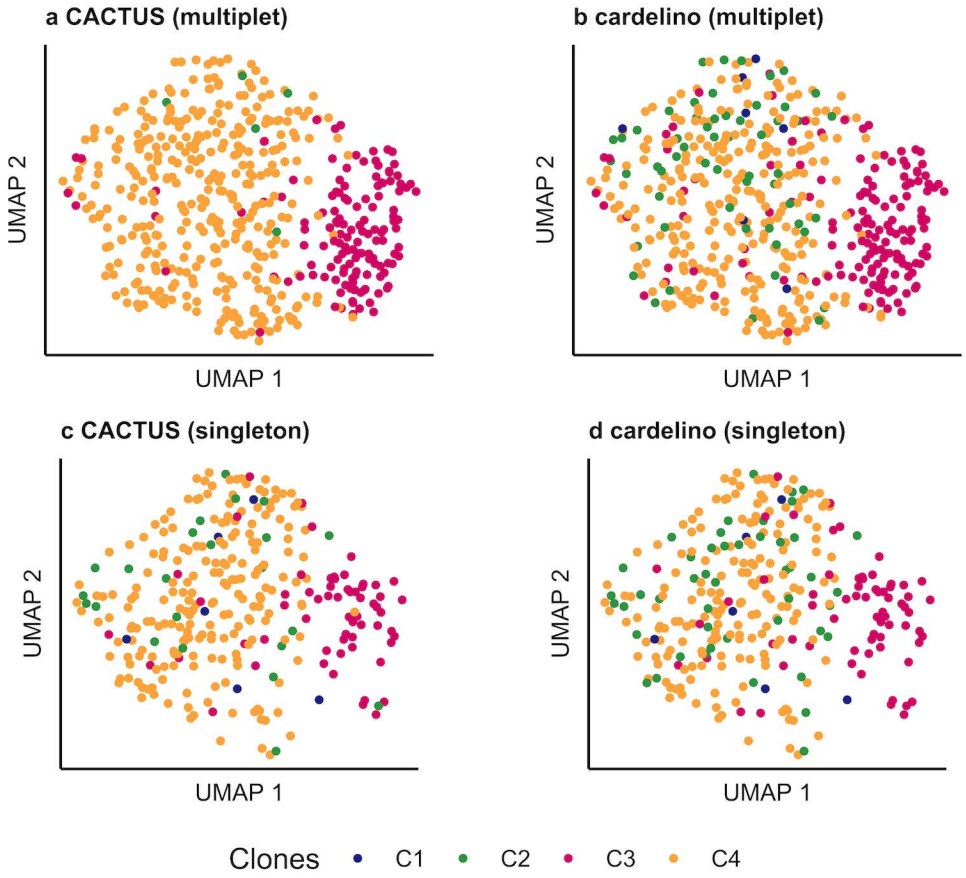


Figure 5.4: **Validation of cell-to-clone assignment with gene expression for subject S12118.** Figure panels as for subject S144 in Fig. 5.3. Also for subject S12118, assignment to clones for cells in multiplet BCR clusters using CACTUS (**a**) improves agreement with gene expression data compared to assignment of cells in singleton BCR clusters (**d**) and assignment using cardelino [177] (**b**), as quantified using connectivity measure (**c**). For singleton BCR clusters CACTUS performs comparably well as cardelino.

we first reduced the dimension of the normalised expression measurement to 25 using PCA. Next, we computed the Root mean square standard deviation (RMSSTD), connectivity, Dunn index and, Calinski–Harabasz (CH) index for the reduced gene expression vectors, grouped according to the assignment of cells to the clones [188, 154, 189, 190, 191] (Table 5.1). In this way, we measured to what extent the gene expression of the cells inside each clone is homogeneous and differs between the clones. A RMSSTD is a measure of compactness - a low value of RMSSTD indicates low variance of gene expression in each set of cells assigned to the same clone. The connectivity measure takes values between 0 and infinity and uses the k-nearest neighbors to indicate the degree of connectedness of the clusters. We used $k = 10$ for the computation, but we noted that other values of k gave similar results. If the cells assigned to the same clone would also be close in terms of Euclidean distance in the reduced 25-dimensional expression space, the connectivity would be minimized. High Dunn index values imply increased compactness of each clone and better separation between the clones, computed for the reduced expression profiles of cells assigned to the clones. The CH index is another measure for evaluating both compactness and separation simultaneously, using average between and within clone sum of squares. The higher CH score indicates

more agreement of the assignment of cells into clones with their gene expression values. For cells in the multiplet BCR clusters, these quality measures clearly indicate that CACTUS obtains better agreement between cell-to-clone assignment and gene expression than cardelino (Table 5.1). In contrast, for cells in singleton clusters, CACTUS obtains similar quality measures as cardelino.

Table 5.1: **Quantification of the agreement of the cell-to-clone assignment with gene expression profiles of the cells.**

Bolded values indicate which method (CACTUS or cardelino) obtained better agreement for the given subject and type of cluster that the cells assigned to the clones come from. High values of the Dunn Index and the Calinski–Harabasz (CH) index, as well as low values of the Root mean square standard deviation (RMSSTD) and Connectivity quantify to what extent the gene expression of the cells is homogeneous inside each clone and differs between the clones.

Subject	Type	Method	Dunn Index	RMSSTD	CH	Connectivity
S144	multiplet	CACTUS	**0.066**	**57.0**	**15.6**	**898.9**
		cardelino	0.057	77.1	3.2	1250.9
	singleton	CACTUS	**0.054**	110.9	**3.2**	839.5
		cardelino	0.052	**109.5**	1.9	**711.9**
S12118	multiplet	CACTUS	**0.098**	**79.2**	**11.9**	**169.6**
		cardelino	0.084	96.2	10.0	495.0
	singleton	CACTUS	0.085	105.4	**4.1**	**285.4**
		cardelino	**0.092**	**99.4**	3.9	396.5

We performed independent clustering of cells by their normalised expression using Seurat [186]. Then, we compared the resulting clustering of cells by expression to the grouping of cells to clones inferred by CACTUS and by cardelino using the Adjusted Rand Index (ARI; [192]). The index, with values in the [-1,1] interval, is a corrected-for-chance version of the Rand index, measuring similarity between two given clusterings. ARI is negative when the agreement is lower than expected by chance and is maximized when the compared clusterings are identical. For the subject S144 and the cells that are in the singleton BCR clusters, both clones inferred by CACTUS and by cardelino show very low similarity to expression clusters (with ARI 0.03 and 0.02, respectively). Compared to cardelino (ARI 0.01), CACTUS achieves a higher agreement with the gene expression clustering for cells contained in the multiplet BCR clusters (ARI 0.13). For subject S12118, the CACTUS clones have the same similarity to expression clusters as cardelino. For cells that are in the singleton BCR clusters, both CACTUS and cardelino yield ARI of 0.12. Finally, for the cells in the multiplet BCR clusters, the ARI for both CACTUS and cardelino is 0.21.

Overall, these results indicate that by accounting for the BCR sequence similarity, CACTUS improves the genotype-to-gene expression phenotype mapping.

5.2.4. CACTUS enhances the confidence of cell-to-clone assignment

For both subjects, the top identified evolutionary trees consisted of four clones (Fig. 5.5a, b). The number of mutations acquired along the branches of the trees ranges from 0 to 57. The genotype of each input clone is defined as the set of the mutations acquired on the path from the root of the tree to the leaf corresponding to the clone (Additional file 2: Table S1). Notably, the clone genotypes and frequencies derived by Canopy (Fig. 5.5a, b) were corrected both by CACTUS (Fig. 5.5c, g, e, i) and cardelino (Fig. 5.5d, h, f, j). CACTUS, in addition,

corrected the input BCR clustering. All results discussed below are for the corrected genotypes and corrected clusters. We investigated the confidence of assignment of cells to the tumor clones for both subjects (Fig. 5.5). The assignment of cells to the clones was directly derived from the assignment of their BCR clusters. In general, thanks to the additional information from the BCR clusters, CACTUS assigns cells to clones with a clearly higher confidence than cardelino [177]. From both methods, the probability of assigning each cell to each clone can be derived as output. For subject S144 and a majority of cells, the probability of assignment by cardelino is almost uniform across the clones (Fig. 5.5d, h). In contrast, for the subset of cells in the multiplet BCR clusters, the probability of assignment by CACTUS makes confident assignments (Fig. 5.5c). For the cells in the singleton BCR clusters, CACTUS assigns cells with similar confidence to cardelino (Fig. 5.5g).

Compared to S144, for subject S12118 the confidence of assignment is larger for both methods (Fig. 5.5). Again, CACTUS has an advantage over cardelino, especially for cells in the multiplet BCR clusters , assigning majority of those cells to one clone with high probability (Fig. 5.5e,i). In contrast, for a majority of cells, cardelino yields similar probabilities of assignment to clones C2 and C4 (Fig. 5.5f, j).

Overall, the confidence of the assignment is clearly higher for CACTUS than for cardelino, for both subjects (Table 5.2). Here, we quantified confidence as the concentration of the assignment probability distribution over the clones, averaged over the cells, using the measures of entropy and the Gini index [193, 194]. Both entropy and Gini index should be lower for larger concentration of the probability distribution (equivalently, smaller dispersion).

Table 5.2: **Quantification of the confidence of cell-to-clone assignment** Confidence is measured as the concentration of the probability distribution of assigning a cell to clones, averaged across cells. Bolded values indicate which method (CACTUS or cardelino) obtained higher confidence. Both normalized Entropy (entropy divided by the maximum possible value) and the Gini Index are supposed to have lower values for more concentrated distributions, and larger values for more dispersed ones.

Subject	Type	Method	Entropy	Gini Index
S144	multiplet	CACTUS	**0.42**	**0.46**
		cardelino	0.85	0.90
	singleton	CACTUS	**0.79**	**0.84**
		cardelino	0.87	0.90
S12118	multiplet	CACTUS	**0.04**	**0.04**
		cardelino	0.39	0.45
	singleton	CACTUS	**0.36**	**0.38**
		cardelino	0.47	0.54

5.2.5. Assignment of BCR clusters to tumor clones

Finally, we inspected the assignment of BCR clusters to clones by CACTUS. For a comparison, for each clone we computed the proportion of each multiplet BCR cluster (the fraction of cells in that BCR cluster) that were assigned to this clone using cardelino (Fig. 5.6). In the case of ties in the highest proportions across clones, we assumed the BCR cluster was assigned to the same clone as by CACTUS.

As expected by construction of the underlying probabilistic model, for both subjects, CACTUS assigns entire BCR clusters to single clones (Fig. 5.6a, c). For cardelino, the proportions of BCR clusters are more distributed across the clones (Fig. 5.6b, d). Given the

uncertainty of assignment of cells to clones by cardelino for subject S144 (Fig. 5.5), it is not surprising that for some of the BCR clusters, the clone assigned by CACTUS does not agree with the clone with the highest proportion of cells assigned by cardelino. CACTUS did not assign any BCR cluster to clone C1, while cardelino assigned cluster U to that clone. All of 11 BCR clusters assigned to clone C2 by CACTUS, were assigned to the same clone by cardelino. Out of 15 BCR clusters assigned to clone C3 by CACTUS, however, none were assigned to clone C3 also by cardelino. This large disagreement comes mainly from the fact that cardelino assigned the highest proportion of cells contained in 13 of these 15 clusters again to clone C2. Finally, out of 11 BCR clusters assigned to clone C4 by CACTUS, 4 were assigned in the highest proportion to the same clone also by cardelino.

For subject S12118, the assignment of cluster agrees between the two methods, with the only exception of cluster O. This is in accordance with the increased confidence of assignment of cells to clones by both methods for that subject (compare Fig. 5.5).

In summary, the agreement of both cell-to-clone and BCR cluster-to-clone mapping between the CACTUS and cardelino increases with the confidence of assignment. For subject S144, for which cardelino yielded low-confidence assignments, 736 out of 1262 cells in total (58%) and 22 out of 37 multiplet BCR clusters (59%) were assigned to different clones by the two methods. Here, we assume cardelino assigns a BCR cluster to the clone to which it assigned the highest proportion of cells. For subject S12118, where both methods increased confidence of assignment, only 123 cells out of 799 (15%) and only one BCR cluster out of 26 multiplet BCR clusters (4%) was assigned differently.

5.3. Discussion

Here, we propose a probabilistic model for accurate and confident mapping of single tumor cells to their evolutionary clones of origin. In this way, it allows clone-specific gene expression profiling, opening the possibility to reconstruct genotype-to-phenotype maps. The task of cell-to-clone mapping is challenged by multiple technical obstacles. First, although multiple methods exist for the inference of tumor evolution, resolving tumor clones and their genotypes is in itself a difficult computational problem and errors are expected [168]. Thus, CACTUS, uses the additional signal both in the scRNA-seq and in clustering data to correct the given genotypes of the clones. Second, the information in scRNA-seq data is only sparse, prone to errors such as dropout and uneven coverage, and biased to mutations observable in typically sequenced first 150 nt of transcripts. It is thus important to realize that the analysed tumor history is limited only to the mutations measurable in single cells, and is potentially more coarse-grained than the true clonal structure of the tumor. These limitations are purely technical, and in this respect analysis using CACTUS would benefit from full-length transcript sequencing with high depth, as well as further developments increasing the quality of scRNA-seq technology.

The key aspect of our model is the ability to borrow information across different measurements (both of DNA and RNA) of the cells in the sample. In particular, in addition to clone genotypes derived from WES, and allele specific transcript counts measured using scRNA-seq, the model leverages information given by independent clustering of single cells. Our results show that this additional evidence is crucial to overcome the challenges of the cell-to-clone assignment problem. Not any given cell clustering, however, can empower CACTUS to deliver more confident results. The assumption that cells contained in the same cluster tend to belong to the same clone is critical for model performance. In particular, such cell clustering, where the cells in the same cluster are not expected to belong to the same clone, can misguide

model inference. Apart from clustering by genomic features, which is expected to agree with the clonal structure of the tumor cell population, for example, clustering by location in the tissue could be provided as input to CACTUS. Here, we used single cell BCR heavy chain sequences to define the input clustering. As would other relevant genomic features, mutations in BCR loci bring evolutionary information. On a general level, they indicate whether a subpopulation of tumor cells sharing a BCR sequence with a low number of BCR mutations evolved relatively early, or if it has more recently evolved and carries a higher number of mutations. Similar BCR sequences indicate common evolutionary origin, as otherwise they would be disrupted by acquisition of additional mutations. Importantly, although the input clustering is defined by identical BCR sequences, cells are shifted between clusters during the model inference process, both re-distributing cells among multiplet clusters and joining singleton clusters to multiplets. This process is influenced by all available data, i.e., not only the similarity of BCR sequences, but also the variants found in scRNA-seq and in the genotypes derived from WES. Here, the quality of additional information brought in by the BCR clusters is assured by the complete and deep sequencing coverage of BCR loci in the applied scRNA-seq strategy. Errors in sequencing, however, may still occur, which further supports the need for updating the input cell clusters.

CACTUS could be extended in the future to further broaden its functionality and to account for even more additional measurements. The input clone genotypes and the number of clones are corrected, but need to be inferred *a priori* to applying the model, and the evolutionary tree structure is not utilized by the model. The possible errors in the prior tree inference, or a wrong assumption about the number of clones, can potentially hamper the model performance. To some extent this problem is avoided by the fact that CACTUS corrects the input clone genotypes during inference. Instead, CACTUS could be extended to simultaneously infer the evolutionary tree, yielding the clones and their genotypes, together with the cell assignment to the clones. Finally, other measurements could be incorporated to statistically strengthen model inference. For example, gene expression similarities between cells, here used for model validation, could be used as input, as cells with similar expression profiles are expected to come from the same clone.

The model is applied to newly generated FL patient data, for the first time shedding light on how clonal evolution in this cancer type induces clone-specific gene expression and agrees with BCR clusters. Accurate mapping of clonal structures with gene expression patterns allows detection of potential therapy-resistant clones, which is essential for effective personalized treatment. Our results demonstrate applicability of CACTUS to the complex cancer samples. The model, however, is more generally applicable and can describe somatic evolution also in other diseases or in the healthy tissue.

5.4. Conclusions

Here, we deal with the task of gene expression profiling of tumor clones by matching the genotypes of the clones to the mutations found by RNA sequencing in the single cells. As applied here, CACTUS benefits from the additional information contained in clusters of single cells sharing similar BCR sequences to assign cells to clones, to successfully deal with errors and dropouts in single cell RNA sequencing, and the difficulty of inferring the correct clonal structure. In summary, this contribution is a step forward in establishing computational tools for resolving the tumor heterogeneity and, by combining genotype with gene expression profiles, its impact on functional diversification of the tumor cell subpopulations.

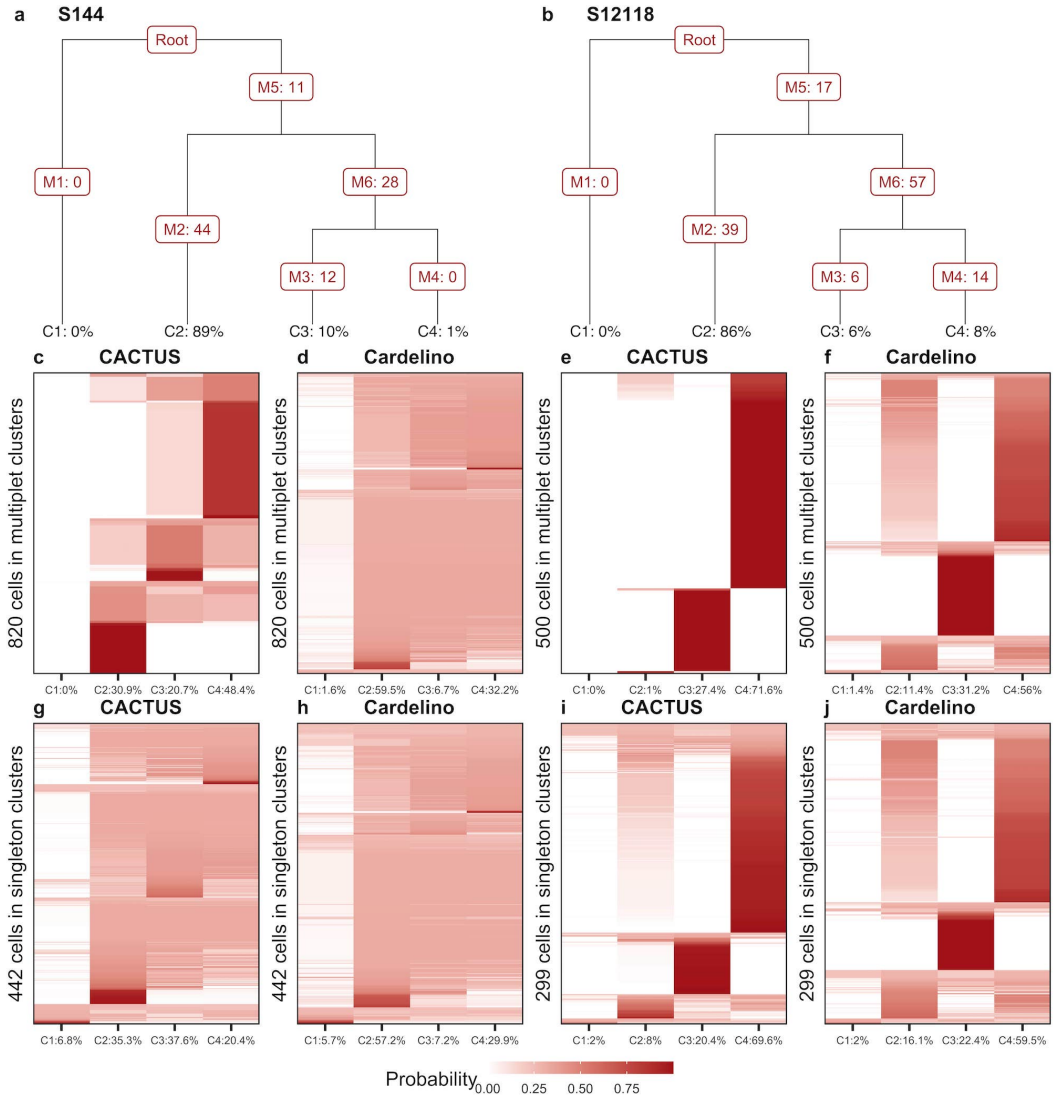


Figure 5.5: **Confidence of cell assignment to the tumor clones. a, b** Evolutionary trees inferred by Canopy [165] for subject S144 (**a**) and S12118 (**b**). Leaf labels: clone prevalences. Branch labels: numbers of acquired mutations. Canopy considers also CNVs, but they are not used for cell-to-clone mapping and hence not visualized here. Thus, the branch labels can be zero when the alterations acquired along that branch are copy number changes. Clone 1 corresponds to the base, normal clone. In tree **a**, clone 4 (C4) differs from clone 3 (C3) by the 12 SNVs acquired on the branch leading to the leaf C3. **c-j** Shades of brown indicate the probability of assignment of cells (y axis) to the clones (x axis; labeled with corrected prevalences, computed as the fraction of single cells assigned to the clones) by CACTUS (**c, g, e, i**) and cardelino [177] (**d, h, f, j**). For cells in multiplet BCR clusters (second row), CACTUS yields higher confidence of cell-to-clone assignment (**c, e**) than cardelino (**d, f**). For cells in singleton BCR clusters (third row) for subject S144 the confidence of cell-to-clone assignment by CACTUS (**g**) is similarly weak as by cardelino (**h**), while for S12118 and for CACTUS (**i**) the confidence is higher than for cardelino (**j**).

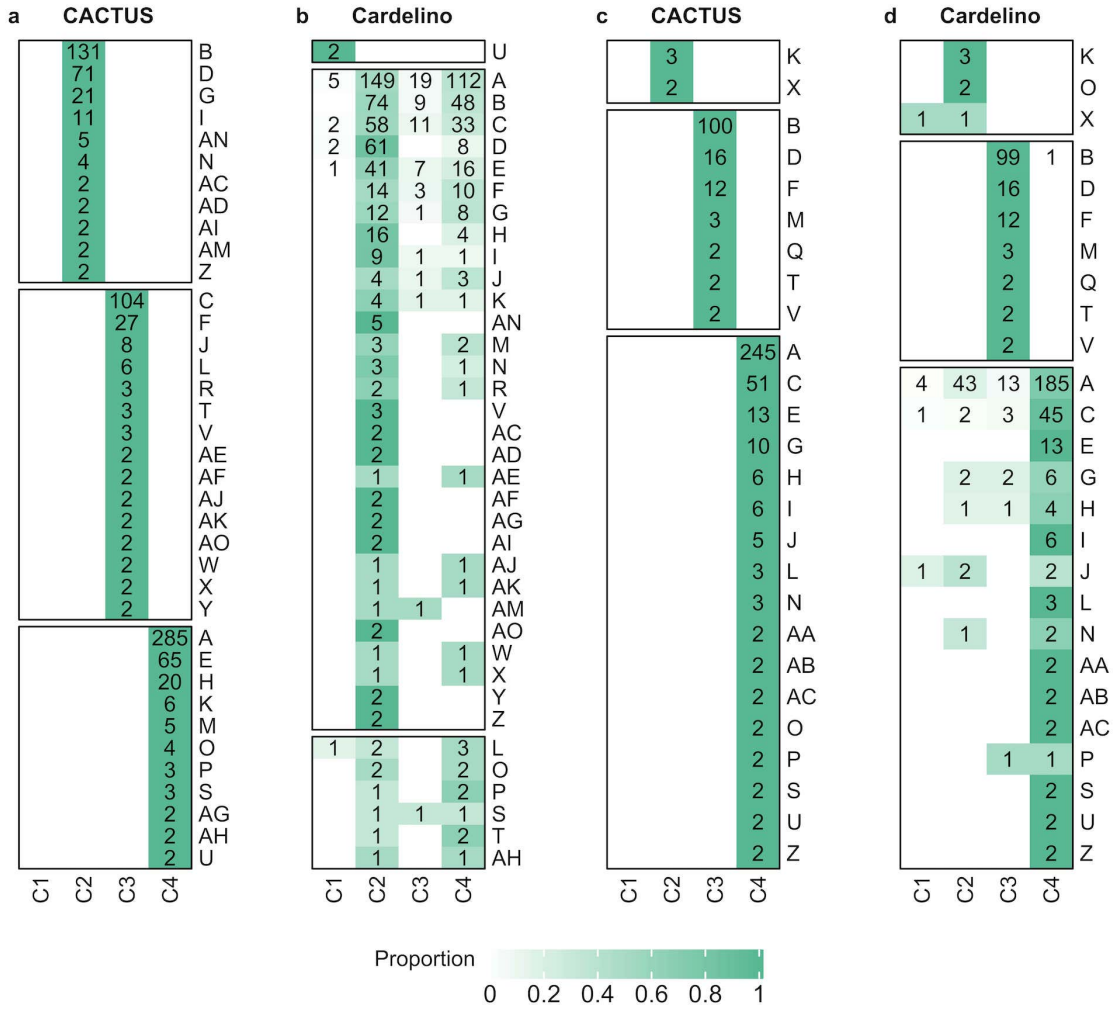


Figure 5.6: **BCR cluster assignment to tumor clones,** for both subjects: S144 (**a, b**) and S12118 (**c, d**), using CACTUS (**a, c**) and cardelino [177] (**b, d**). Heatmaps with shades of green indicate the proportion of cells in multiplet cluster (y axis) assigned to clones (x axis). Each number in a green entry indicates the nonzero number of cells of the corresponding BCR clusters assigned to the corresponding clone. Only BCR clusters of at least two cells are featured. As expected, for both subjects, CACTUS assigns entire BCR clusters to single clones (**a, c**). For cardelino, the proportions of BCR clusters are more distributed across the clones (**b, d**).

Chapter 6

Tumoroscope: a probabilistic model for mapping cancer clones in tumor tissues

Tumor evolution proceeds by the accumulation of mutations, resulting in the emergence of distinct cancer cell subpopulations, called *clones*, characterized by their genotypes. The spatial distribution of these clones may vary drastically across tumor tissue. This *genetic* and *spatial* tumor heterogeneity are the two key determinants of patient prognosis, survival, and treatment [195, 96, 196]. Characterization of the *phenotypic* heterogeneity of tumors, i.e., linking the potential differences between expression profiles of clones and their spatial distribution has up to now remained an uncharted territory.

The vast majority of studies investigate intra-tumor heterogeneity based on bulk DNA sequencing (DNA-seq) or single-cell DNA-seq (scDNA-seq) data [197, 198]. Unfortunately, bulk DNA-seq measures a mixture of millions of cells from different tumors and healthy cells and thus provides only aggregated information of variant allele frequencies. There are several approaches for clonal deconvolution of bulk DNA-seq data, reconstructing the clone genotypes, the frequencies of the clones, and their phylogenetic relationships [199, 200, 201, 165, 202, 203]. More recently, several methods for identifying clonal evolution from mutations found in scDNA-seq [204] or from combined bulk and scDNA-seq [43, 177, 205] were proposed. Despite recent technological advances [206], scDNA-seq remains much more laborious, more inaccurate, and less affordable than the highly established bulk DNA-seq [207]. Unfortunately, both bulk and scDNA-seq require tissue disaggregation and thus lose spatial information. As methods based on DNA-seq, they cannot be used to elucidate the phenotypic heterogeneity.

The localization of cancer clones was previously analyzed using multi-region single-cell or bulk DNA-seq, combined with computational clone inference for each region [208, 209, 210, 211]. This approach, however, is coarse-grained, as each of the regions is itself a bulk sample and is composed of multiple clones with an unknown position in the tissue. Unfortunately, currently there exists no experimental approach for large-scale sequencing of the DNA of single cells *in situ*. However, the recent technology of spatial transcriptomics (ST) offers spatially-resolved RNA sequencing (RNA-seq) of mini-bulks of only 1-100 cells, localized in spots of an ST array [42, 212]. Thereby ST enables an analysis of spatial gene expression patterns across the analyzed tissue. Although the resolution of ST is orders of magnitude higher than multi-region bulk sequencing, it still provides only an aggregated signal for mixtures of cells. Recently, methods addressing the localization of clonal copy number alterations from ST data were developed, but did not account for somatic point mutations, which constitute the major

factor in tumor development [213, 214]. Since ST is an RNA-seq protocol and does not have single-cell resolution, it is non-trivial to infer the point mutation genotypes of clones at the spots. Finally, single tumor cell phenotypes are widely studied using scRNA-seq, but since the DNA of these cells is not usually measured, the phenotypes are not assigned to the cancer clones. In summary, there exists no state-of-the-art approach for the study of tumor genetic, spatial and phenotypic heterogeneity in a high, close to cellular resolution.

To address this issue, we propose Tumoroscope, a probabilistic graphical model that exploits somatic point mutation information in ST reads, genotypes of clones reconstructed from bulk DNA-seq and tumor regions and cancer cell counts annotated in hematoxylin and eosin-stained (H&E) images to deconvolute the clonal composition of each localized spot in the tumor sample. In this way, we are able to localize the somatic point mutations and clones derived from DNA-seq in the tissue. On top of that, we devise a regression model for inference of gene expression profiles of the clones. After validating Tumoroscope on simulated data, we set out to answer key questions about co-localization and mutual exclusion patterns of the spatial arrangement of clones and their phenotypes in a newly generated breast, and a previously published prostate cancer dataset [44]. Our approach enables close to single-cell resolution spatial mapping and infers gene expression profiles of the clones in the tumor tissue, opening novel avenues in the study of spatial, genetic, and phenotypic heterogeneity of tumors.

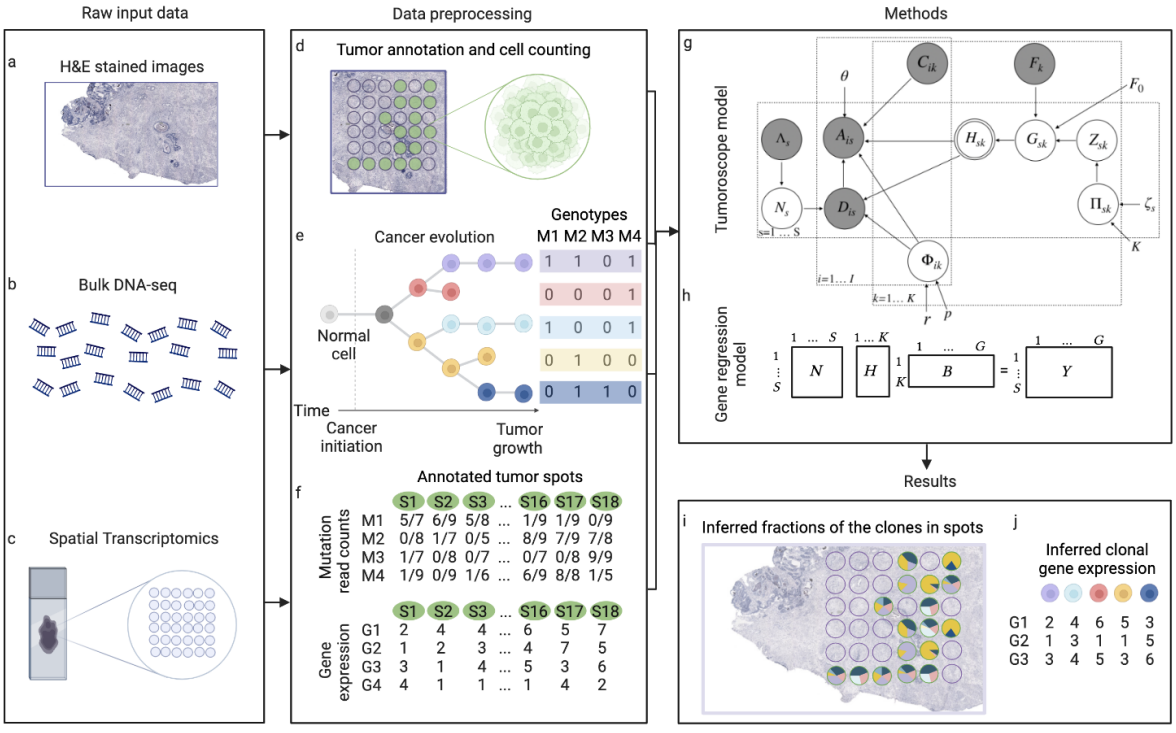


Figure 6.1: **Tumoroscope framework overview.** **a-c** Input data. **d-f** Data preprocessing. **g** Tumoroscope probabilistic model. **h** Regression model for inference of gene expression profiles of the clones. **i** Results of Tumoroscope. **j** Output of the regression model.

6.1. Results

Tumoroscope is a comprehensive probabilistic framework for mapping cancer clones across tumor tissues based on integrated signals from H&E stained images (Fig. 6.1a), spatially-resolved transcriptomics (Fig. 6.1b), and bulk DNA-seq (Fig. 6.1c). The data preprocessing pipeline starts with a two-staged analysis of the H&E-stained image of the tissue (Fig. 6.1d). Firstly, ST spots lying within regions containing cancer cells are indicated. Secondly, for each of such ST spots, we estimate the number of cells contained in that spot (using custom scripts in QuPath [215]; Methods). Next, we reconstruct cancer clones, their genotypes, and frequencies from the bulk DNA-seq data (using the existing methods Vardict [130], FalconX [142], and Canopy [165], see Methods, Fig. 6.1e). Afterwards, we analyse the data in the form of the number of alternated reads and the total number of reads for each mutation (mutation coverage), along with gene expression observed in each spot indicated as tumor (Fig. 6.1f). Notice, that the key assumption behind Tumoroscope is that each ST spot contains a hidden mixture of the clones reconstructed from the bulk DNA-seq data. Tumoroscope leverages: i) the estimated cell counts per spot provided as priors, ii) the alternated and total read counts for mutations in ST spots, and iii) the genotypes and frequencies for the clones using a probabilistic deconvolution model (Fig. 6.1g). The result of Tumoroscope is the identification of proportions of the clones in each spot (Fig. 6.1i). Additionally, for each spot, the method corrects the prior cell counts estimated from H&E images, using an inference from the ST data. Finally, we employ a regression model with gene expression data taken as independent variables and the inferred proportion of the clones in the ST spots as dependent variables (Fig. 6.1h) to infer gene expression profiles of the clonal populations (Fig. 6.1j).

6.1.1. Tumoroscope correctly estimates the proportion of clones in each spot and is robust to noise in input cell counts

In order to evaluate Tumoroscope's performance in the case when the ground truth is known, we assessed its accuracy of estimating the proportion of clones in spots using simulated data. The simulation setups varied with respect to the number of mutations present in the clones, the expected number of clones in each spot, and the coverage of mutations. Specifically, we first designed a basic setup with five clones in the evolutionary tree, 30 mutations in the genotype matrix, an average number of 13.6 mutations per clone, and an expected number of 2.5 clones per spot. Next, we created four additional setups by decreasing and increasing the average number of mutations per clone to 5.1 and 15, respectively, and the expected number of clones per spot to 1 and 4.5, respectively. Furthermore, to test the influence of the level of coverage per mutation, for each of these five different setups we additionally varied the average coverage, with settings called very low, low, medium, and high (corresponding to an average number of reads present in each spot of 18, 50, 80, and 110, respectively; Methods). We simulated 10 datasets for each of the 20 setups resulting from the five aforementioned setups and four different coverage levels (amounting to 200 different simulated datasets used for evaluation in total; see Extended Data Table 6.1 for the detailed specification of simulation setups).

To inspect the robustness of our model to noise in the counts for number of cells per spot, we considered three different levels of noise in this input, as well as two versions of Tumoroscope, differing by how this input was modeled. Specifically, the model was either given the true simulated values of cells per each spot at the input, or we introduced small and large additive noise to these counts (Methods). In the first, default model version, referred to as Tumoroscope, the provided cell counts were used as priors and the number of cells per

each spot was inferred accounting for all available data. In the second, simplified version, referred to as Tumoroscope-fixed, this input was used to fix the values of cell numbers in the spots. As both model variations were evaluated for the three levels of noise on each of the 200 simulated datasets, inference was made for 1200 synthetic datasets in total (Fig. 6.2). The performance was evaluated by calculating the Mean Average Error (MAE), that is, the average of the difference between the inferred proportions of the clone and the true values in all the spots and clones.

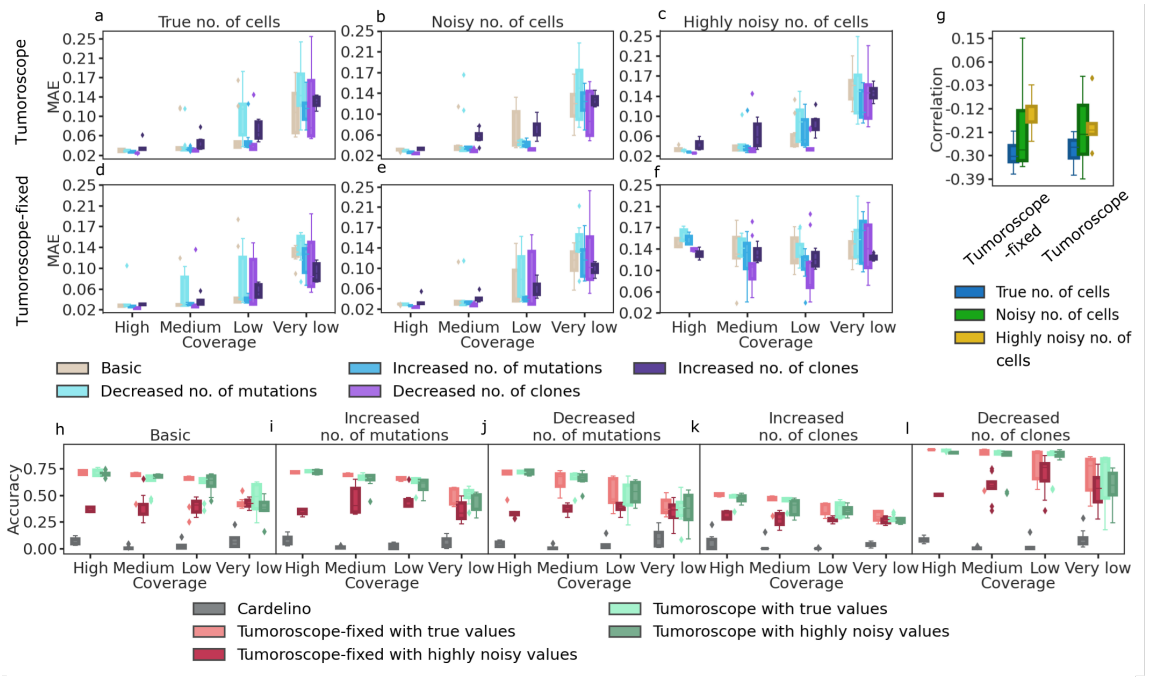


Figure 6.2: **Performance of Tumoroscope on simulated data. a-c** Mean Average Error (MAE; y-axis) as a function of mutation coverage (x-axis) in different simulation setups (colors) for Tumoroscope, for different noise levels in the cell count provided at input: no noise (**a**), medium noise (**b**) and high noise (**c**). **d-f** The same as in (**a-c**), but for Tumoroscope-fixed. **g** Correlation (y-axis) between the average mutation coverage and the average error in all the setups is negative for both model versions (x-axis), regardless of the noise in the number of cells provided at the input (colors). **h-l** Comparison of the accuracy (y-axis) of the model between cardelino (gray) and two versions of the model given true and highly noisy values for the number of cells (colors), depending on the mutation coverage (x-axis), in different simulation setups: basic (**h**), increased (**i**) and decreased (**j**) number of mutations, increased (**k**) and decreased (**l**) number of clones.

For both model versions, the error tended to increase with decreasing coverage (Fig. 6.2 a-g), indicating that better clone deconvolution can be obtained with deeper sequencing of the spots in ST data. Tumoroscope obtained low error (median MAE between 0.02 and 0.15, depending on the coverage), regardless of the level of noise in the input cell counts per spot (Fig. 6.2 a-c). Notably, in the case when the true, simulated cell counts were given as input, Tumoroscope performed equally well as Tumoroscope-fixed, despite the advantage that the latter had by fixing the counts to the true values (Fig.6.2a vs 6.2d). This advantage turned into bias when the input cell numbers became noisy, and Tumoroscope-fixed obtained a larger MAE than Tumoroscope (Fig. 6.2b,e, c, f). Similar results were obtained when higher coverages per mutation were considered (Extended Data Fig. 6.1). These results emphasize the importance of keeping the input cell count per spot as priors rather than fixing them as observed values, especially in the case of noise in this input, which is expected for real data.

Indeed, in the real data, these input cell counts per spot are estimated from H&E images using algorithms for nuclei detection, which becomes particularly difficult when the cells are densely packed and the nuclei overlap (Methods).

6.1.2. Accounting for the mixture of clones in each spot is key for model performance

To demonstrate the necessity of accounting for the mixture of clones in each spot, we compared Tumoroscope to an alternative method called cardelino [177]. Since cardelino was originally designed to assign single cells to clones based on scRNA-seq data, here, we applied cardelino providing each spot as a single cell, effectively assuming that the spot was a homogeneous readout from a single clone. For the sake of comparison, we only considered the major clone in each spot inferred by Tumoroscope and we defined the accuracy as the percentage of the agreement of the major inferred clone and the major true clone in the simulated data. Again, we evaluated both Tumoroscope and Tumoroscope-fixed. Tumoroscope obtained the worst-case median accuracy of around 0.27 for the increased number of clones and very low read count setup, and best-case accuracy of around 0.92 for the decreased number of clones and very high read count setup. With these results, Tumoroscope significantly outperformed cardelino, which obtained median accuracy between 0 and 0.09 in all simulation setups (Fig. 6.2 h–l). Similar to (Fig. 6.2 a–g), Tumoroscope's accuracy tended to decrease with decreasing coverage. Interestingly, Tumoroscope obtained the highest accuracy for the decreased number of clones setup, and the worst for the increased number of clones, indicating that the number of clones per spot is a decisive factor for method's performance. Tumoroscope-fixed obtained lower accuracy than Tumoroscope, especially when provided with highly noisy input cell counts but still outperformed cardelino by a large margin. This result strongly emphasizes the importance of accounting for the clone mixtures in spots.

6.1.3. Tumoroscope deconvolutes spatial clonal composition in a breast tumor and finds spatial patterns of cancer clones in sub-areas.

To investigate the spatial clonal structure of a real tumor sample, we applied Tumoroscope to a newly generated dataset including three breast tumor sections from one patient. As input data, for each section, we generated deep whole-exome sequencing data (WES) and spatial transcriptomics (10x Genomics) of two neighboring layers (Methods). We assayed 4885-4992 spots per sample. In the data pre-processing stage, we selected the spots that were cancerous based on the expert pathologist's annotations (Fig. 6.3e) and estimated cell counts from H&E images of the sections. We considered 608 high-confidence somatic single-nucleotide mutations (SNVs) identified from WES data that were co-observed in the annotated ST data (Methods). Next, we reconstructed the evolutionary tree of somatic mutations that were also present in ST data reads (Methods). We identified seven clones, including a base clone without somatic mutations (Fig. 6.3a,b; Extended data Fig. 6.3). Finally, given the selected 11461 cancerous spots, their estimated cell counts, total and alternated read counts at identified mutations, and the reconstructed clone genotypes, we used Tumoroscope to deconvolute the transcriptomics mutation profiles from the spots to obtain the proportions of the underlying clones.

The composition of the seven clones in the investigated breast cancerous tissue identified by Tumoroscope revealed fascinating patterns of spatial arrangement (Fig. 6.3d). Generally, there was no single clone that fully dominated a specific contiguous sub-area of tissue. However, we did observe subsets of clones that coexisted in sub-areas. For section SB1, clone 4 was present in medium proportions in all analyzed spots of both layers. Very interestingly, there

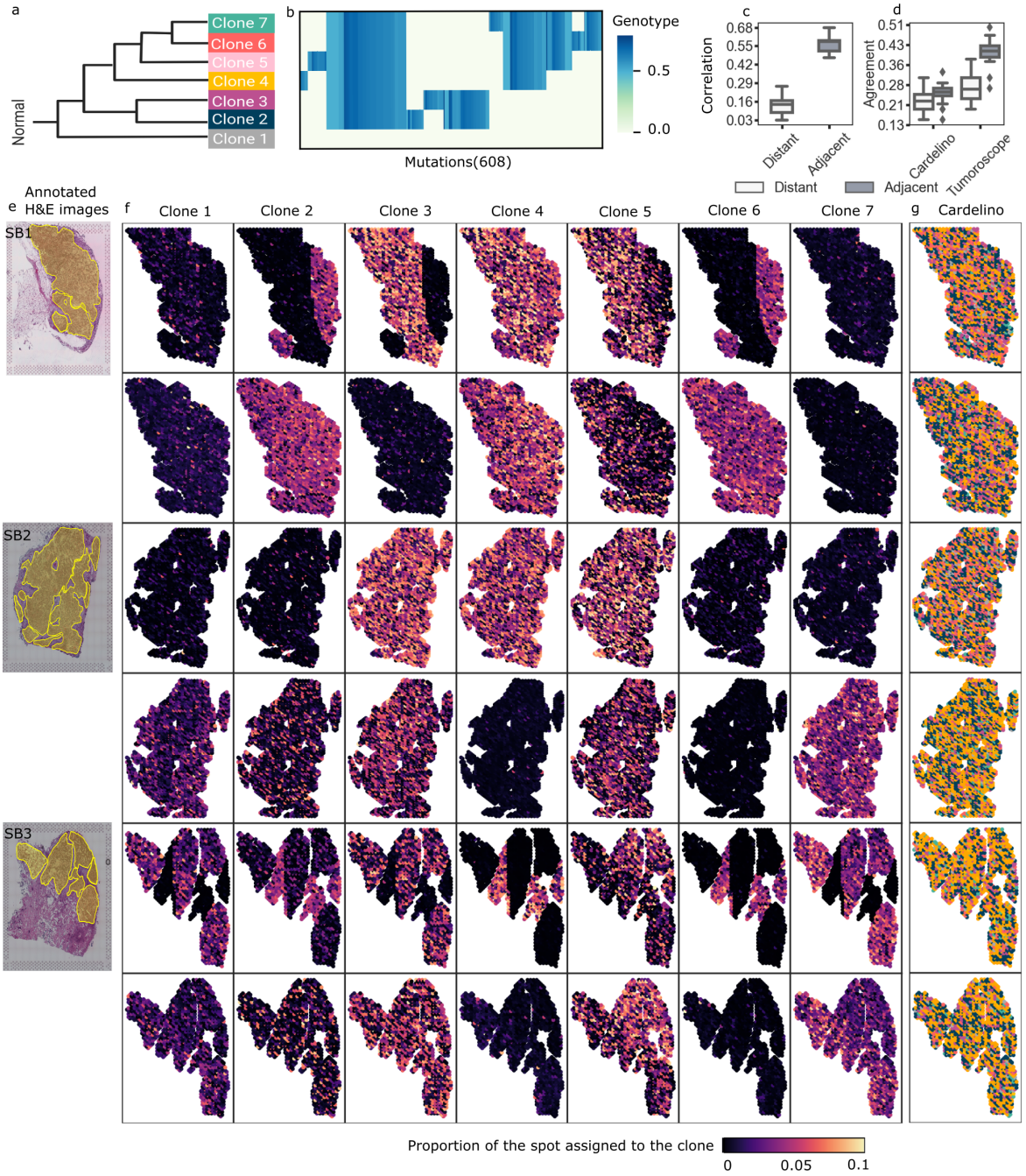


Figure 6.3: **Spatial arrangement of cancer clones found for the breast cancer dataset. a-b** Evolutionary tree and genotypes of the inferred clones. **c** Distribution of the correlation (y-axis) of the clonal composition of the spots that are distant and adjacent, computed for 100 pairs of spots sampled at random 20 times each (x-axis). **d** Distribution of the agreement of the distant and adjacent spots in cardelino and Tumoroscope, computed for the same randomly sampled pairs. For the computation of the agreement, we use the single inferred clone by cardelino and the major inferred clone by Tumoroscope. **e** Pathologist's annotation of the cancerous areas on the H&E images for sections SB1, SB2, and SB3. **f** For each section, two rows correspond to the two nearby samples and 7 columns correspond to the proportion of the spots assigned to each clone. **g** The clonal assignment of the spots by cardelino for the same samples (see Extended data Fig. 6.2 for expanded cardelino results).

was a clearly separated sub-area in the right-hand part of section SB1 first layer, where clones 2, 4, and 6 co-occurred. The rest of this layer was dominated by clones 3, 4, and 5. In the second layer of section SB1, clones 2, 4, 5, and 6 coexisted, although with larger proportions of clone 4, low proportions of clones 2 and 6, and clone 5 being present in fewer spots than other clones. Clone 7 was not present in either layer of this section. Similarly, contiguous sub-areas that were predominantly occupied by small subsets of clones could be found in both layers of sections SB2 and SB3. As expected, clone 1, which lacked somatic mutations characteristic of the remaining cancerous clones, was found in only small proportions in the analyzed spots across all sections and layers. Patterns of clonal co-occurrence and mutual exclusion could be observed across all sections and layers, indicating a systematic mechanism. For example, the pairs of clones 2 and 6, as well as 3 and 5, although evolutionarily distant and with different genotypes (Fig. 6.3a, b) were always present together in the same sub-areas, while clones 4 and 7 excluded each other.

In contrast to Tumoroscope, in the assignments of single clones to spots inferred by cardelino, there was no detectable spatial pattern of domination of clones in sub-areas, as all clones were present in all sections uniformly (Extended Data Fig. 6.4 and 6.2). Again, this underlined the importance of spot deconvolution.

To validate the decomposition results in the absence of ground truth, we exploited that it is natural to expect the similarities in the clonal composition of the adjacent spots to be high, due to the growth process of the tumor in space. We found that the median correlation of clone proportions inferred by Tumoroscope between adjacent spots was significantly higher than the median correlation between distant spots (computed between 100 pairs of spots each, sampled at random 20 times; Fig. 6.3c). Since Tumoroscope treated each spot as independent and did not enforce any spatial similarities by design, this result strongly supports the correctness of the deconvolution of ST spots using Tumoroscope.

Importantly, we compared Tumoroscope's and cardelino's performances. Since cardelino was originally designed to analyze scRNA-seq data, when applied to ST data, it assigned only one clone to each spot. Thus, to enable the comparison, for every spot of interest, we determined the major clone (characterized by the highest proportion) indicated by Tumoroscope and computed the agreement for each out of 20 randomly sampled sets of adjacent and distant pairs of spots considered previously. The median fraction of adjacent pairs of spots with clonal assignment in agreement was much higher for Tumoroscope (0.41) than for cardelino (0.25). Moreover, the difference between the agreement for the distant and adjacent pairs was larger for Tumoroscope (distance between medians 0.14; one-sided Wilcoxon p-value 1.9e-06) than for cardelino (distance between medians 0.03; one-sided Wilcoxon p-value 0.063).

6.1.4. Tumoroscope assigns the ST spots to clones in a prostate tumor

Next, we applied Tumoroscope to three prostate tumor sections from one patient, for which deep WES and ST data (custom arrays) of neighboring layers were generated [44]. As before, we selected the spots that were cancerous based on the regions that were annotated as tumor areas by an expert pathologist (obtaining 968-1001 spots per sample) and counted cells in spots from H&E images (obtaining 1-188 cells per spot; Fig. 6.4c; Methods). We then called the somatic mutations from WES data and identified 282 high-confidence somatic SNVs that co-occured in the ST data. Next, we reconstructed the evolutionary tree using Canopy [165] for that tumor from the WES data, identifying four clones, including a base clone without somatic mutations (Fig. 6.4a,b; Extended Data Fig. 6.4). Finally, we used Tumoroscope to deconvolute the transcriptomic signal from 294 spots in the ST data to reveal the proportions of the underlying clones.

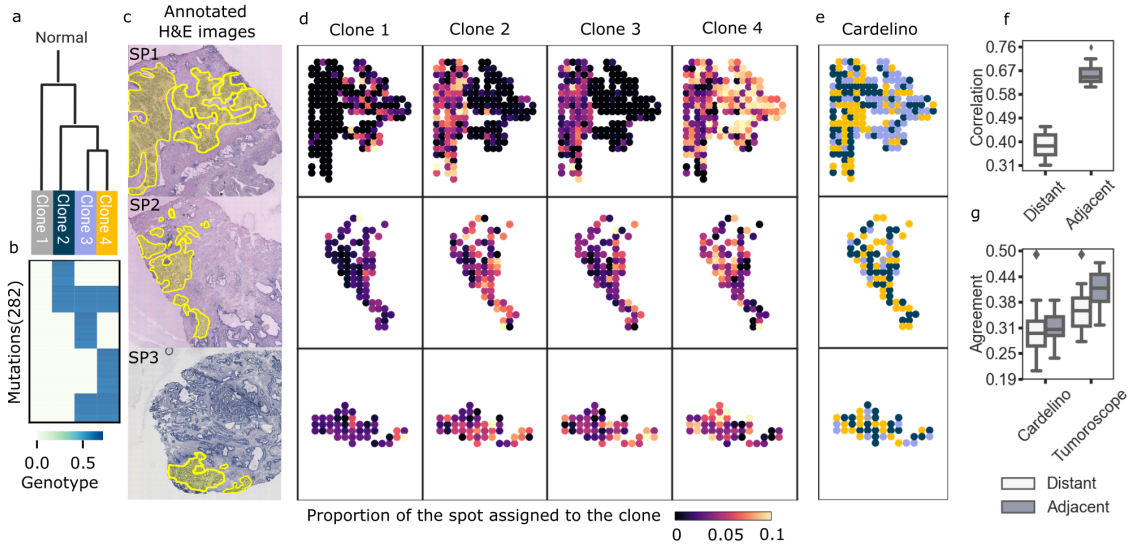


Figure 6.4: **Results obtained for the prostate cancer dataset. a-b** Evolutionary tree and genotype of the clones. **c** Pathologist's annotation of the cancerous areas on the H&E images for sections SP1, SP2, and SP3. **d** For each section (rows), 4 columns correspond to the proportion of the spots assigned to each clone. **e** The clonal assignment by cardelino (see Extended data Fig. 6.5 for expanded cardelino results). **f** Distribution of the correlation of the clonal composition of the spots that are distant and adjacent, computed for 100 pairs of spots sampled at random 20 times each. **g** Distribution of the agreement (y-axis) of the distant and adjacent spots for cardelino and Tumoroscope, computed for the same randomly sampled pairs. For the computation of the agreement, we use the single inferred clone by cardelino and the major inferred clone by Tumoroscope.

Similar to the results obtained for breast cancer, for prostate cancer, we observed a pattern of sub-areas with marked presence of subsets of clones (Fig. 6.4d). Interestingly, section SP1 was divided into two sub-areas, with the left-hand sub-area containing all cancer clones 2, 3 and 4, while the right-hand sub-area was predominantly occupied by clone 4 with a small admixture of normal cells (clone 1). Sections SP2 and SP3 were smaller than SP1, but also showed distinct sub-areas with different clonal compositions.

For comparison, we again applied cardelino, by considering each spot in the ST data as a single cell measured using scRNA-seq (Fig. 6.4e). Interestingly, similarly to Tumoroscope, for section SP1 cardelino also divided the tissue into two different subareas, confirming their distinct clonal composition. However, the clones assigned by cardelino did not agree with the clones identified as taking the most proportion of the same spots by Tumoroscope. For example, for the right-hand sub-area of section SP1, cardelino mostly assigned spots to clone 3, and not 4.

We further verified whether Tumoroscope inferred more similar clonal profiles for adjacent spots than for distant spots. As expected, the correlations of the inferred clone proportions between of adjacent spots (median 0.65) were significantly higher than the correlations between distant spots (median 0.38; computed for 100 randomly selected pairs each and sampled 20 times; Fig. 6.4f), validating the results of Tumoroscope.

Furthermore, we compared the percentage of the agreement of the major clone in each spot in the adjacent and distant pairs of spots found using Tumoroscope, with the agreement of the clones in the same pairs of spots assigned by cardelino (Fig. 6.4g). With a median of 0.41, the agreement for adjacent spots was significantly higher for Tumoroscope than for cardelino (median 0.31). Furthermore, the difference between the agreement of the adjacent

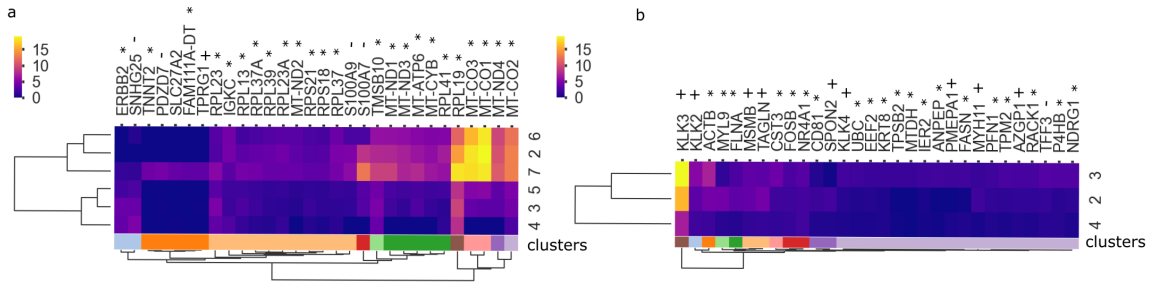


Figure 6.5: **Genes are expressed differently in various cancer clones.** The expression of the 30 genes that were inferred by the regression model as the most active in at least one clone, clustered in rows and columns, for breast (**a**) and prostate cancer (**b**) tissues. * cancer gene found in all cancer tissues (not cancer type specific) according to the HPA database [216]; + cancer gene with nTPM (normalized gene expression value) in the desired cancer type (either breast in **a** or prostate in **b**) at least four times higher than in other cancer tissues, according to [216]; - not detected in cancer tissues, or nTPM at least four times higher in another cancer tissue than the desired one, according to [216].

and distant spots was significant for Tumoroscope (difference between medians 0.05; one-sided Wilcoxon p-value 0.004) and was notably higher than for cardelino (0.01; one-sided Wilcoxon p-value 0.556).

6.1.5. Similarity in gene expression profiles coincides with spatial co-occurrence of clones

Next, we applied the regression model (Methods) to deconvolve the expression values of genes in each clone from the aggregated gene expression values in spots. The regression model assumed that the expression of each gene in each spot is given by a mixture of expression values of that gene coming from clones present in that spot, weighted by their proportion inferred by Tumoroscope, and scaled by the inferred cell number in that spot. Both for breast and prostate cancer data, we ranked the genes by their maximum inferred expression across the clones and selected the first 30 genes at the top of the ranked list. We found that out of 30 genes selected for prostate cancer, 9 of them (*KLK2*, *KLK3*, *MSMB*, *TAGLN*, *SPON2*, *KLK4*, *PMEPA1*, *MYH11*, *AZGP1*) are known to be enriched in prostate cancer tissues (i.e, have elevated expression specifically in the prostate cancer tissue, according to Human Protein Atlas; HPA; [216]), and 19 are known to be upregulated in cancer (i.e, expressed in several cancer types according to HPA). For breast cancer, we found the *TPRG1* gene, known to be enriched in breast cancer tissues, and 25 genes known to be upregulated in cancer. The deconvolved expression profiles with respect to clones varied between the genes (Fig. 6.5). Using hierarchical clustering of the genes by their expression values, we found 10 clusters for breast, and 10 clusters for prostate cancer, respectively (Fig. 6.5a,b). Interestingly, we found clusters with genes expressed highly only in specific subsets of clones. For instance, for breast cancer data, we found that the genes *MT-CO1* and *MT-CO3*, associated with promoting cancer phenotype [217, 218] and gene *RPL19* associated with poor cancer patient survival [219], were highly expressed exclusively in clones 2, 6 and 7. Furthermore, in the results obtained for prostate cancer data, we found that the gene *KLK3*, known as prostate-specific antigen and the most frequently clinically used prostate cancer biomarker [220], was active in clones 2 and 3. These results suggest that each cancer cell clone might have its specific function in tumor progression and development.

Finally, we performed clustering of the clones with respect to their gene expression profiles

(Fig. 6.5). Intriguingly, for both prostate and breast cancer, clones with similar inferred phenotypes, which were clustered together by their expression profiles, also coexisted in space across the tissue (compare with Fig. 6.3 and 6.4). For breast cancer, since the correlation between the fraction of the clones 2 and 6 in the spots was relatively high (Pearson correlation $r = 0.64$; see Extended data Fig. 6.7), it was expected to find them to be similar in terms of their gene expression profiles, per construction of the regression model. In contrast, while the fractions of clones 3 and 5 in the same spots were not correlated (Pearson correlation $r = -0.28$; Extended data Fig. 6.7), they were still co-localized in tissue space as they co-occurred in adjacent spots (Fig. 6.3; average correlation of fractions in the adjacent spots $r = 0.16$; Extended data Fig. 6.8). In this case, similar expression profiles for these clones were not expected per construction of the regression model, but still, these clones were inferred as the second most similar. For the prostate cancer sample, the two clones 2 and 3 which were co-localized in the tissue (Fig. 6.4), were also the most similar in terms of their gene expression profiles. Both for breast and prostate cancer, the pairs of correlated clones were not close in terms of their mutations (Fig. 6.3b) and thus placed in distinct sub-trees of the evolutionary tree (Fig. 6.3a). This indicates that even distant clones may have similar phenotypes and play analogous roles in tumor progression.

6.2. Discussion

Tumoroscope is the first approach for mapping cancer clones based on point mutations in tissue space and resolving their expression profiles in close to single-cell resolution. This resolution amounts to the diameter of the deconvoluted spots, ranging from 100 µm (as for the prostate cancer dataset [44]) to 55 µm (as for the breast cancer dataset), depending on the ST technology. Effectively, this means the model is able to assign clone proportions for spatially resolved mini bulks of the order of 1-40 cells (1-10 cells for breast cancer dataset [212] and 10-40 for prostate cancer dataset [42]). Tumoroscope achieves this result by innovative integration of data from technologies that were not originally developed for this task: H&E, WES, and ST. The key signal exploited by Tumoroscope to identify the clonal composition of ST spots is the matching between mutations present in the genotypes of the clones to the mutations found in the RNA sequencing of the spots. On top of that, the method estimates additional variables, such as the number of cells in each spot and the average expression of each variant site per single cell. Finally, with the proportion of the spots coming from specific clones, alongside the gene expression observed in spots in hand, we solve the problem of clone-specific gene expression deconvolution.

Our comprehensive simulation study demonstrates Tumoroscope's robustness to noise in the estimation of the number of cells in ST spots. The results clearly indicate that the deconvolution task becomes easier with increasing coverage of mutations in ST spots and with a decreasing number of coexisting clones in each spot. In application to breast and prostate cancer data, Tumoroscope reveals spatial patterns of clonal arrangement, indicating a well-mixed coexistence of small subsets of all clones in subareas of the tumor tissue.

Applying our regression model to infer gene expression levels in the different clones allows us to identify the distinct phenotypes of the clones, effectively assigning spatial resolution to the function of the different tumor subpopulations, and thus profiling the functional heterogeneity of tumors. Moreover, our findings in both analyzed cancer types indicate that it is the phenotypic, and not genotypic similarity, which could drive the spatial co-occurrence of clones. However, this result should be further validated in additional patient samples and using independent data.

To our knowledge, there exists no competing technology that could be applied to resolve the spatial clonal heterogeneity of tumors in a comparable resolution to Tumoroscope. Spatial capturing of DNA sequences is still at the very early stage of development [221]. The very low resolution obtained with current spatial DNA sequencing technology requires merging beads located nearby in the array and, thus, provides spatial mini-bulk data, akin to ST spots. Additionally, ignoring the evolutionary origin of distinct clones and clustering beads with no information about variant allele frequency, as performed in [221], oversimplifies the complex problem of spatial clonal deconvolution. Considering the difficulties intrinsic to spatial DNA-seq data, close to single-cell resolution ST data proves to be a highly attractive alternative for the spatial inference of clonal evolution. The recently developed method STARCH [213] combines RNA-sequencing of ST spots and DNA-sequencing from neighboring tissues in the same tumor sample to infer the spatial arrangement of clones based on their copy number profiles. Besides, Erickson *et al.* [214] developed a method to infer genome-wide copy number variations (CNVs) from spatially resolved mRNA profiles in situ that reveal distinct CNV based clonal patterns within tumors. In contrast to Tumoroscope, however, both these methods ignore the impact of point mutations in tumor heterogeneity. Moreover, they do not directly address the problem of deconvoluting the mixture of cancer clones per spot.

The quality of the obtained results could be further improved with better technology. For example, replacing WES with scDNA-seq data would allow more accurate inference of cancer clones, their evolutionary relationships, and genotypes using dedicated computational approaches [222, 223]. As the ST technology improves, smaller spots are expected, limiting the number of clones per spot, which makes the deconvolution problem easier. Finally, currently, only the first 300 bp of gene sequences are sequenced in the process of ST data generation. For our approach, ideally, whole gene bodies should be sequenced so that all mutations detectable from WES could also be observed in ST and matched for more accurate deconvolution of spots into clones. Such a sequencing was recently shown to be possible [224] but was not available for the data that we analyzed and is not in the standard ST protocols.

Despite these technological limitations, already now Tumoroscope offers a major breakthrough in the integrated analysis of spatial, genomic and phenotypic tumor heterogeneity. The model could be applied in further studies profiling adjacent tumor samples to provide 3D maps of clones. With our ability to compute gene expression profiles of the clones, we could make it possible to predict the most proliferating areas and, thereby, the most probable expansion sites of the 3D structure of the tumor. Furthermore, studies combining H&E, WES, and ST for large cohorts of patients, could explore the dependencies between patient clinical features and the spatial patterns of clones found using Tumoroscope. Combined with cell-type deconvolution approaches for ST data in the tissue surrounding the tumors [225, 226, 227], our framework has the potential to bring unprecedented insights into the interactions of specific cancer clones, their phenotypes, and the surrounding microenvironment. In summary, Tumoroscope opens up a new avenue in cancer research with broad applications for a basic understanding of the disease and its clinical applications.

6.3. Methods

6.3.1. Breast tumor samples

The breast tumor study was approved by the Swedish Ethical Review authority (no. 2016/957-31 with amendments 2017-742-32, 2020-00323 and 2021-00795). Breast tumor tissues were obtained by Dr. Johan Hartman (Institute of Oncology and Pathology, Karolinska Institute). The samples were collected from tumor material removed from a patient with untreated

invasive ductal carcinoma during breast cancer surgery. Histological evaluation of the patient's tumor was performed by pathologists for diagnostic purposes and defined as HER2 (+3), ER (30%), and Ki67 (79%). For this tumor sample, different regions ($n = 5$) were selected by the pathologists. From each region, tissue was isolated for immediate embedding in OCT for gene expression analysis with spatial transcriptomics. Samples for spatial transcriptomics were immediately frozen and stored at −80℃ until further analysis.

6.3.2. Preparation and sequencing of Spatial Gene Expression Libraries for the breast tumor samples

Sections of fresh-frozen breast tumor tissue were cut at 10 µm thickness and mounted onto slides from the Visium Spatial Gene Expression Slide & Reagent kit (10X Genomics). Sequencing libraries were prepared following the manufacturer's protocol (Document number CG000239 Rev A, 10x Genomics). Prior to imaging, coverslips were mounted on the slides according to the protocol's optional step Coverslip Application & Removal. Tissue images were taken at 20x magnification using Metafer Slide Scanning platform (MetaSystems) and raw images stitched with VSlide software (MetaSystems). Adaptations of the protocol were made in that the Hematoxylin staining time was reduced to 4 minutes and tissue permeabilization was performed for 12 minutes. Final libraries were sequenced on NextSeq2000 (Illumina) or NovaSeq6000 (Illumina).

6.3.3. Data processing of spatial gene expression libraries for the breast tumor samples

Following demultiplexing of bcl files, read 2 fastq files were trimmed using Cutadapt [228] to remove full-length or truncated template switch oligo (TSO) sequences from the 5' end (beginning of read 2) and polyA homopolymers from the 3' end (end of read 2). The TSO sequence (AAGCAGTGGTATCAACGCAGAGTACATGGG) was used as a non-internal 5' adapter with a minimum overlap of 5, meaning that partial matches (up to 5 base pairs) or intact TSO sequences were removed from the 5' end. The error tolerance was set to 0.1 for the TSO trimming to allow for a maximum of 3 errors. For the 3' end homopolymer trimming, a sequence of 10 As was used as a regular 3' adapter to remove potential polyA tail products regardless of its position in the read, also with a minimum overlap of 5 base pairs. The trimmed data were processed with the spaceranger pipeline (10X Genomics), version 1.0.0 (BC) and mapped to the GRCH38 v93 genome assembly.

6.3.4. Prostate cancer sample

The prostate cancer dataset was generated and published by Berglund *et al.* [44]. This dataset consists of twelve sections, with H&E images, bulk DNA-seq and spatial transcriptomics provided for each section. The data were generated and processed using protocols as described in [44].

6.3.5. Identifying the spots that contain tumor cells

To select the spots that contain tumor cells, we took advantage of H&E staining images of the analyzed tissues. For both breast and prostate cancer, regions containing cancer cells were annotated by an expert pathologist Dr. Łukasz Koperski using QuPath [215]. We further selected spots whose area overlapped with the pathologist's annotated regions, using a custom script in QuPath [215].

6.3.6. Counting cells in spots

We developed a custom script in QuPath [215] to count cells in each ST spot visible in the H&E images [215]. The script takes as input coordinates and diameters of spots to define target areas. Then, we employ QuPath's inbuilt cell counting algorithm for detecting and counting nuclei. In order to adjust parameters of the algorithm, we examined random spots by manually counting cells to verify the accuracy of the results.

6.3.7. Spatial transcriptomics data preprocessing

For prostate cancer sample, the ST data bam files were provided by Berglund *et al.* [44]. For breast cancer sample, to create the genome index, we used the STAR program [229] with the GRCh38 reference genome as input. Next, we applied the ST Pipeline [230], providing the genome index, FASTQ files, barcodes and array coordinates as input. We obtained the gene expression matrix as counts of reads for each gene, which the ST Pipeline produces by default. In addition, we modified the default settings, to obtain BAM files with the mapped reads.

6.3.8. Bulk DNA-seq and somatic mutation calling

We identified somatic mutations that appeared in at least one of the bulk DNA-seq sections, by calling the mutations using Vardict [130] for each section with a p-value threshold equal to 0.1. Then we used their union over sections as the set of mutations called in bulk DNA-seq data. This procedure was performed in the same way for the prostate and the breast dataset.

6.3.9. Selection of somatic mutations that are detected both in bulk DNA-seq and ST data

Next, we identified the bulk DNA-seq mutations that were also present in ST data. For calculating the total and alternated reads over the mutations in ST data, we located the selected bulk DNA-seq mutations in the ST bam files and counted the corresponding mapped reads with our script. The reads with a different nucleotide as compared to the reference genome were called the alternated reads.

Finally, we selected the mutations for which there existed at least one alternated read in at least one section. The alternated and total read counts in bulk DNA-seq data for the selected mutations were given as input for phylogenetic inference, while the alternated and total read counts in ST data for the same mutations were given as input to Tumoroscope. The median read coverage for the selected variant sites in bulk DNA-seq data for breast and prostate cancer were 214.5 and 134.75, respectively.

6.3.10. Phylogenetic tree analysis

To identify the phylogenetic tree and infer the genotype and prevalence of each clone in the tree we used a statistical method called Canopy [165]. The input to Canopy are variant allele frequencies of somatic single nucleotide alterations (SNAs), along with allele-specific coverage ratios between the tumor and matched normal sample for somatic copy number alterations (CNAs). We used FalconX for producing the allele-specific coverage ratio between tumor and normal sample [142]. We used multi-sample feature of Canopy to infer the clonal evolution across the sections for both prostate and breast datasets.

6.3.11. Mapping fractions of cells in ST spots to cancer clones using Tumoroscope

Tumoroscope is a probabilistic graphical model for estimating proportions of cancer clones in ST spots given alternated and total read counts over the analysed somatic mutations, genotypes and frequencies of the clones, and estimated cell counts per each spot (Fig. 6.1f). Let $i \in \{1, \ldots, M\}$ index the selected mutation positions, identified both in bulk DNA sequencing and ST data. We are given a set of K cancer clones, indexed by $k \in \{1, \ldots, K\}$ as input, which has been derived from bulk DNA sequencing data. The genotypes of the input clones are represented as a matrix C with entries between 0 and 1 corresponding to the zygosity. $C_{i,k}$ equals 0 if there is no mutation on position i in clone k, equals 1 in case all alleles of that position carry the mutation, and equals 0.5 when the half of the alleles of that position carry the mutation. Note that there can be multiple alleles for position i. In general, the zygosity is defined as the ratio of the number of mutated alleles to the total number of alleles and we estimate it by the ratio of the major allele frequency to the total read count. The prevalence of the clones in the bulk DNA sequencing is represented by the vector $F = (F_1, ..., F_K)$, with values summing up to one. Let $s \in \{1, \ldots, S\}$ index the spots. We use a feature allocation model to account for the presence of clones in spots [231]. Specifically, we define $Z_{s,k} \in \{0, 1\}$ as an indicator of the presence of clone k in spot s. We assume a Bernoulli distribution over $Z_{s,k}$ and a Beta prior over its parameter Π with hyper-parameter ζ_s:

$$\mathbb{P}(Z_{s,k} \mid \Pi_{s,k}) \sim Bern(\Pi_{s,k}),$$

$$\mathbb{P}(\Pi_{s,k} \mid \zeta_s, K) \sim Beta\left(\frac{\zeta_s}{K}, 1\right).$$

Let $\mathbf{1} = \{1\}^K$ denotes a K-dimensional vector with all elements equal to 1. Bearing in mind the assumption about Beta prior over $\Pi_{s,k}$, we calculate the expected number of nonzero entries in each spot $\mathbb{E}[Z_{.,k}^T \mathbf{1}]$ using the formula for the mean of the Beta distribution as [232, 233]

$$\mathbb{E}[Z_{.,k}^T \mathbf{1}] = \sum_{k=1}^{K} \mathbb{E}[Z_{s,k}] = K\mathbb{E}[Z_{s,k}] = K\frac{\frac{\zeta_s}{K}}{\frac{\zeta_s}{K} + 1} = \frac{K\zeta_s}{\zeta_s + K}.$$

Given this formula and the number of the clones K, we are able to control the expected number of clones in each spot by tuning shape parameter of the beta distribution, $\frac{\zeta_s}{K}$.

Our main goal is to estimate the proportions of clones in the spots, which are represented by the variable H, a matrix with S rows and K columns. The value of an element $H_{s,k}$ is the fraction of spot s coming from clone k. We consider a Dirichlet distribution over $H_{s,.} = (H_{s,1}, \ldots, H_{s,K})$,

$$\mathbb{P}(H_{s,1}, \ldots, H_{s,K} \mid F', F_0, Z_{s,.}) \sim \text{Dirichlet}({F'_1}^{Z_{s,1}} {F_0}^{1-Z_{s,1}}, \ldots, {F'_K}^{Z_{s,K}} {F_0}^{1-Z_{s,K}}).$$

Here, F_0 corresponds to a "pseudo-frequency", and results in non-zero proportions for all clones for each spot. We set F_0 to a small number, effectively assigning small proportions to clones which are not present in the spot. The $F' = (F'_1, ..., F'_K)$ are obtained as discretized frequencies F. Specifically, we discretize the values of F by dividing the range from 0 to 1 into 20 equal-sized bins and then round up the values to the upper-bounds of the bins and scale them by multiplicative factor l

$$F'_k = l \times \frac{\lceil 20 \times F_k \rceil}{20},$$

where we used $l = 100$, but it can be specified by the user.

To sample H, we take advantage of the relation between Dirichlet and Gamma distribution [234] and draw K independent random samples $(G_{s,1}, \ldots, G_{s,K})$ from K Gamma distributions,

$$\mathbb{P}(G_{s,k} \mid F'_k, F_0, Z_{s,k}) \sim \text{Gamma}({F'_k}^{Z_{s,k}} {F_0}^{1-Z_{s,k}}, 1),$$

and then we calculate the proportions H:

$$H_{s,k} = \frac{G_{s,k}}{\sum_{l=1}^{K} G_{s,k}}.$$

The total read count at position i in spot s is represented by observed variable $D_{i,s}$. We assume a Poisson distribution over $D_{i,s}$,

$$\mathbb{P}(D_{i,s} \mid H_{s,.}, \Phi_{i,.}, N_s) \sim \text{Pois}\left(N_s \sum_k H_{s,k}\Phi_{i,k}\right),$$

where $\Phi_{i,k}$ is the average coverage for the position i across the cells from clone k, and N_s is the number of cells in spot s. The variables N_s can be fixed to *a priori* known values.

However, in most practical applications, the number of cells per spot is not known. This gives a compelling reason to estimate them as a part of model inference. We assume a Poisson distribution over N_s,

$$\mathbb{P}(N_s \mid \Lambda_s) \sim \text{Pois}\left(\Lambda_s\right),$$

where Λ_s is the expected number of cells in spot s. Also, we assume a Gamma distribution over $\Phi_{i,k}$,

$$\mathbb{P}(\Phi_{i,k} \mid r, p) \sim \text{Gamma}(r, p),$$

where r and p are the shape and rate hyperparameters, respectively.

$A_{i,s}$ represents the number of alternated reads for position i in spot s. We assume a Binomial distribution over $A_{i,s}$,

$$\mathbb{P}(A_{i,s} \mid D_{i,s}, H_{s,.}\Phi_{i,.}C_{i,.}) \sim \text{Binom}\left(D_{i,s}, \frac{\sum_{k=1}^{K} H_{s,k}\Phi_{i,k}C_{i,k}}{\sum_{k=0}^{K} H_{s,k}\Phi_{i,k}}\right).$$

Where the success probability of Binomial distribution is the probability of observing $A_{i,s}$ alternated reads out of $D_{i,s}$ reads in total. Given the variables N_s,$H_{s,.}$ and $\Phi_{i,.}$,we calculate the expected number of alternated reads and the total reads in spot s using $N_s \sum_{k=1}^{K} H_{s,k}\Phi_{i,k}C_{i,k}$ and $N_s \sum_{k=1}^{K} H_{s,k}\Phi_{i,k}$, respectively. Therefore, we calculate the success probability by calculating the fraction of the expected number of alternated reads and the total reads.

6.3.12. Metropolis-Hasting inside Gibbs Sampling

In the Gibbs sampling, we iteratively generate samples from each hidden variable's conditional distribution, given the remaining variables, in order to estimate the posterior distribution of the hidden variables. Each hidden variable given the variables in its Markov Blanket is conditionally independent of all variables outside its Markov Blanket in the graphical model [235]. A variable's Markov Blanket includes its parents, children, and children's parents. If the conditional distribution does not have a closed analytical form, we use a Metropolis-Hasting step inside the Gibbs sampler. In the following, we describe the sampling steps for each hidden variable.

The variables with the closed-form sampling distribution

$\Pi_{s,k}$ and $Z_{s,k}$ are the only variables with analytical sampling distributions.

Sampling $\Pi_{s,k}$

For sampling $\Pi_{s,k}$ we take advantage of the conjugacy of Beta and Bernoulli distributions:

$$\mathbb{P}(\Pi_{s,k} \mid \zeta_s, K, Z_{s,k}) \propto \mathbb{P}(\Pi_{s,k} \mid \zeta_s, K)\mathbb{P}(Z_{s,k} \mid \Pi_{s,k})$$

$$= \text{Beta}\left(\Pi_{s,k} \mid \frac{\zeta_s}{K}, 1\right) \text{Bern}(Z_{s,k} \mid \Pi_{s,k}) = \text{Beta}(\Pi_{s,k} \mid \frac{\zeta_s}{K} + Z_{s,k}, 2 - Z_{s,k}).$$

Sampling $Z_{s,k}$

For sampling $Z_{s,k}$ we utilize the fact that this variable only accepts binary values. Therefore, we sample 0 or 1, proportional to their corresponding calculated probabilities.

$$\mathbb{P}(Z_{s,k} \mid \Pi_k, G_{s,k}, F_k, F_0, l) \propto \mathbb{P}(Z_{s,k} \mid \Pi_k)\mathbb{P}(G_{s,k} \mid F_k, F_0, l) = \text{Bern}(Z_{s,k} \mid \Pi_k)\text{Gamma}(G_{s,k} \mid F_k^{Z_{s,k}} F_0^{1-Z_{s,k}}, 1).$$

Metropolis-Hasting adaptive steps inside Gibbs sampler

In our model, there is no closed analytical form of conditional distribution for variables $\Phi_{i,k}$, $G_{s,k}$ and N_s. Therefore, we take advantage of Metropolis-Hasting inside Gibbs sampler. We compute the acceptance ratio A as the following

$$A = \frac{f(x_c)Q(x_n \mid x_c)}{f(x_n)Q(x_c \mid x_n)}.$$

Where $f(x)$ is a function that is proportional to the desired density function $P(x)$ and Q is the proposal distribution. Bearing in mind the non-negativity of the variables of our interest, we choose a Truncated Normal distribution for Q with the mean value of the current sample x_c and variances σ_Φ, σ_G and σ_N corresponding to each variable. The variance of the Truncated Normal distribution determines the proximity of the new sample from the current one, which is interpreted as the step size. The choice of the step size has a major impact on the acceptance rate of the Metropolis Hasting. We tune the σ_Φ, σ_G and σ_N every b steps starting with an arbitrary value based on the feedback from the acceptance rate. Firstly, we choose an optimal acceptance rate R_o for each variable. Secondly, we modify the variance by δ percent of the current variance and δ is calculated by the difference of the optimal and current acceptance rate R_c. Ultimately, during the sampling steps, we learn the optimal variance value for each variable.

$$\delta_t = R_o - R_c$$

$$\sigma_{t+1} = \sigma_t(1 + \delta + t)$$

In the following, we describe the conditional distribution for each variable.

Conditional distribution for $\Phi_{i,k}$

$$\mathbb{P}(\Phi_{ik} \mid r, p, A_{i,.}, H_{.,k}, C_{i,.}, D_{i,.}, N_.) \propto \mathbb{P}(\Phi_{ik} \mid r, p) \prod_s \mathbb{P}(A_{i,s} \mid H_{s,k}, \Phi_{i,k}, C_{i,k}) \prod_s \mathbb{P}(D_{i,s} \mid \Phi_{i,k}, H_{s,k}, N_s)$$

$$= \text{Gamma}(r, p) \prod_s \text{Binom}\left(A_{i,s} \mid D_{i,s}, \frac{\sum_{k=1}^K H_{s,k}\Phi_{i,k}C_{i,k}}{\sum_{k=0}^K H_{s,k}\Phi_{i,k}}\right) \prod_s \text{Pois}\left(D_{i,s} \mid N_s \sum_k H_{s,k}\Phi_{i,k}\right).$$

Conditional distribution for $G_{s,k}$

$$\mathbb{P}(G_{s,k} \mid F'_k, F_0, Z_{s,k}, A_{.,s}, D_{.,s}, \Phi_{.,k}, C_{.,k}, N_s)$$

$$\propto \mathbb{P}(G_{s,k} \mid F_k, F_0, Z_{s,k}) \prod_i \mathbb{P}(A_{i,s} \mid H_{s,k}, \Phi_{i,k}, C_{i,k}) \prod_i \mathbb{P}(D_{i,s} \mid \Phi_{i,k}, H_{s,k}, N_s)$$

$$= \text{Gamma}({F'_k}^{Z_{s,k}} {F_0}^{1-Z_{s,k}}, 1) \prod_i \text{Binom}\left(A_{i,s} \mid D_{i,s}, \frac{\sum_{k=1}^{K} H_{s,k}\Phi_{i,k}C_{i,k}}{\sum_{k=0}^{K} H_{s,k}\Phi_{i,k}}\right) \prod_i \text{Pois}\left(D_{i,s} \mid N_s \sum_k H_{s,k}\Phi_{i,k}\right).$$

Sampling N_s

$$\mathbb{P}(N_s \mid \Lambda_s, D_{.,s}, \Phi, H_{s,.}) \propto \mathbb{P}(N_s \mid \Lambda_s) \prod_i \mathbb{P}(D_{i,s} \mid \Phi_{i,.}, H_{s,.}, N_s)$$

$$= \text{Pois}\left(N_s \mid \Lambda_s\right) \prod_i \text{Pois}\left(D_{i,s} \mid N_s \sum_k H_{s,k}\Phi_{i,k}\right).$$

Parameter setting for different simulation setups

First, we calculate the parameter of the Beta distribution over variable $\Pi_{s,k}$ based on the assumed expected value of the number of clones:

$$\frac{\zeta_s}{k} = \frac{\mathbb{E}[Z_{.,k}^T \mathbf{1}]}{K - \mathbb{E}[Z_{.,k}^T \mathbf{1}]}.$$

Considering expected values of 1, 2.5, and 4.5 for the number of clones found in each spot, we obtain 0.25, 1, and 9 and use these values for the Beta distribution parameter.

Second, we exploit $\Phi_{i,k}$ that represents the expected number of reads for mutation i in each cell for generating different read coverage for total and alternated read counts. We set $p = 1$. With this, we control the expected value of $\Phi_{i,k}$ using parameter r.

$$\mathbb{P}(\Phi_{i,k} \mid r, p) \sim Gamma(r, p),$$

$$\mathbb{E}[\Phi_{i,k}] = \frac{r}{p^2}.$$

For the very low, low, medium and high number of reads, we consider $r = 0.02$, $r = 0.07$, $r = 0.09$ and $r = 0.19$, respectively, leading to the 18, 50, 80, and 110 average total reads for each spot.

Last, we generate three datasets for the number of cells with different level of noise to compare our two models having number of cells as observed and hidden variable. We add the noise value ϵ to the true values.

$$N_s = N_s + \epsilon.$$

We consider $\epsilon = 0$, $\epsilon \sim$ Pois(1) and $\epsilon \sim$ Pois(10) for generating without noise, noisy and highly noisy number of cells.

Parameter estimation obtained for the real data

For the higher accuracy of the graphical model reflecting the real data, we estimate the input parameters of the model based on the characteristics of the data. The first parameter is λ_s, the expected number of the cells in spot s, which affects the estimation of the number of cells and, ultimately, the number of reads we are expecting, which is a crucial element for estimating the fraction of the clones. Therefore, we estimate the number of cells using the H&E images and a customized script in QuPath and use them as the mean parameter for the Poisson distribution over N(described above) [215]. Next parameters are r and p, the shape and rate in the Gamma distribution over variable Φ. We use mixed type log-moment estimators for calculating r and p [236].

$$\hat{r} = \frac{I \sum_{i=1}^{I} x_i}{I \sum_{i=1}^{I} x_i \ln(x_i) - \sum_{i=1}^{I} \ln(x_i) \sum_{i=1}^{I} x_i}.$$

$$\hat{p} = \frac{I^2}{I \sum_{i=1}^{I} x_i \ln(x_i) - \sum_{i=1}^{I} \ln(x_i) \sum_{i=1}^{I} x_i}.$$

Where x_i with $i \in \{1, \ldots, I\}$ are the sample from Gamma distribution. We generate these samples using the total number of reads D. We calculate the average number of reads from every cell dividing the reads from the spots to the number of estimated cells as input which gives us I samples, equal to the number of mutations.

$$x_i = \frac{1}{S} \sum_{s} \frac{D_{i,s}}{n_s}.$$

Clonal composition resemblance in adjacent spots

The evolutionary process imposes the similarity of the clonal composition in the adjacent spots. Therefore, we expect to have a higher correlation between the clonal composition of the adjacent spots as compared to distant spots. To make this comparison, we randomly generate N pair of adjacent spots $([(X_1, Y_1), (X_1^{'}, Y_1^{'})] \ldots [(X_N, Y_N), (X_N^{'}, Y_N^{'})])$ with X and Y corresponding to their coordinates. These adjacent pairs satisfy two constraints of $X_j - X_j^{'} \leq 1$ and $Y_j - Y_j^{'} \leq 1$ indexed by $j \in \{1, \ldots, N\}$. We also generate N pair of distant spots with the two constraints of $X_j - X_j^{'} > 1$ and $Y_j - Y_j^{'} > 1$. We define $[V_{k,j}, V_{k,j}^{'}]$ as the fraction of clone k in spots corresponding to the j^{th} pair in the adjacent spots. Then we calculate the Pearson correlation for the vector $[(V_{k,1}, V_{k1}^{'}), \ldots, (V_{k,N}, V_{k,N}^{'})]$. The procedure is repeated for all the clones and distant spots for the sake of comparison.

6.3.13. Clonal assignment of the spots using cardelino

Cardelino [177] is a statistical method originally developed for inferring the clone of origin of individual cells using single-cell RNA-seq (scRNA-seq). It integrates information from imperfect clonal trees inferred from whole-exome sequencing data and sparse variant alleles expressed in scRNA-seq data. However, here, we applied it on spatial transcriptomics instead of scRNA-seq to validate the assumption of mixture of clones in each ST spot instead of assuming homogenous spots contaning only one clone. We used *clone_id* function with "sampling" inference mode, minimum iteration of 100000 and maximum iteration of 250000. We used 3 parallel chains for prostate cancer data and 1 chain for breast cancer data due to the high RAM demand of the cardelino.

Property	Values
number of the clones in the evolutionary tree	5
number of mutations in the genotype	30
average clone per spot	1, 2.5, 4.5
average number of mutations per clone	5.1, 13.6,15
average number of reads present in each spot	18, 50, 80, 110
average noise introduced to the number of cells	0,1,10

Extended data Table 6.1: The setups used for simulation.

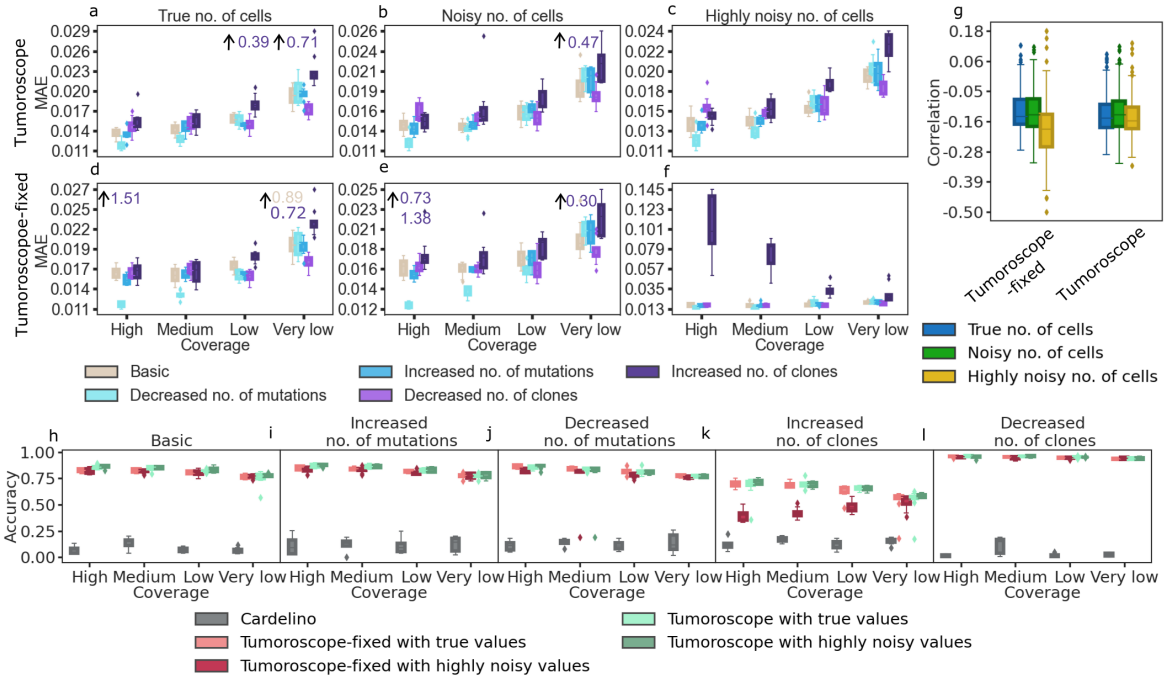
6.3.14. Estimating gene expression of the clones

Having the proportions of the clones in each spot inferred using Tumoroscope and gene expression data from spatial transcriptomics, we estimate average clonal gene expression using a regression model. Let $g \in \{1, \ldots, G\}$ index genes and Y be a matrix with S rows and G columns, where $Y_{s,g}$ is the measured gene expression of gene g in spot s. We are interested in estimating $B_{k,g}$ - average gene expression of gene g in one cell of clone k. We use H and N variables inferred by Tumoroscope, and we rewrite N as an $S \times S$ diagonal matrix N', where $N'_{s,s}$ is the number of cells in spot s and other elements of the matrix are equal to zero. We describe the relationship between the variables with an overdetermined system of equations $N'HB = Y$. Then we try to find the optimal solution of this equation using linear regression with a lower bound of $B_{k,g} \geq 0$ and no intercept. For this purpose, we apply a python function *scipy.optimize.lsq_linear* to the data.

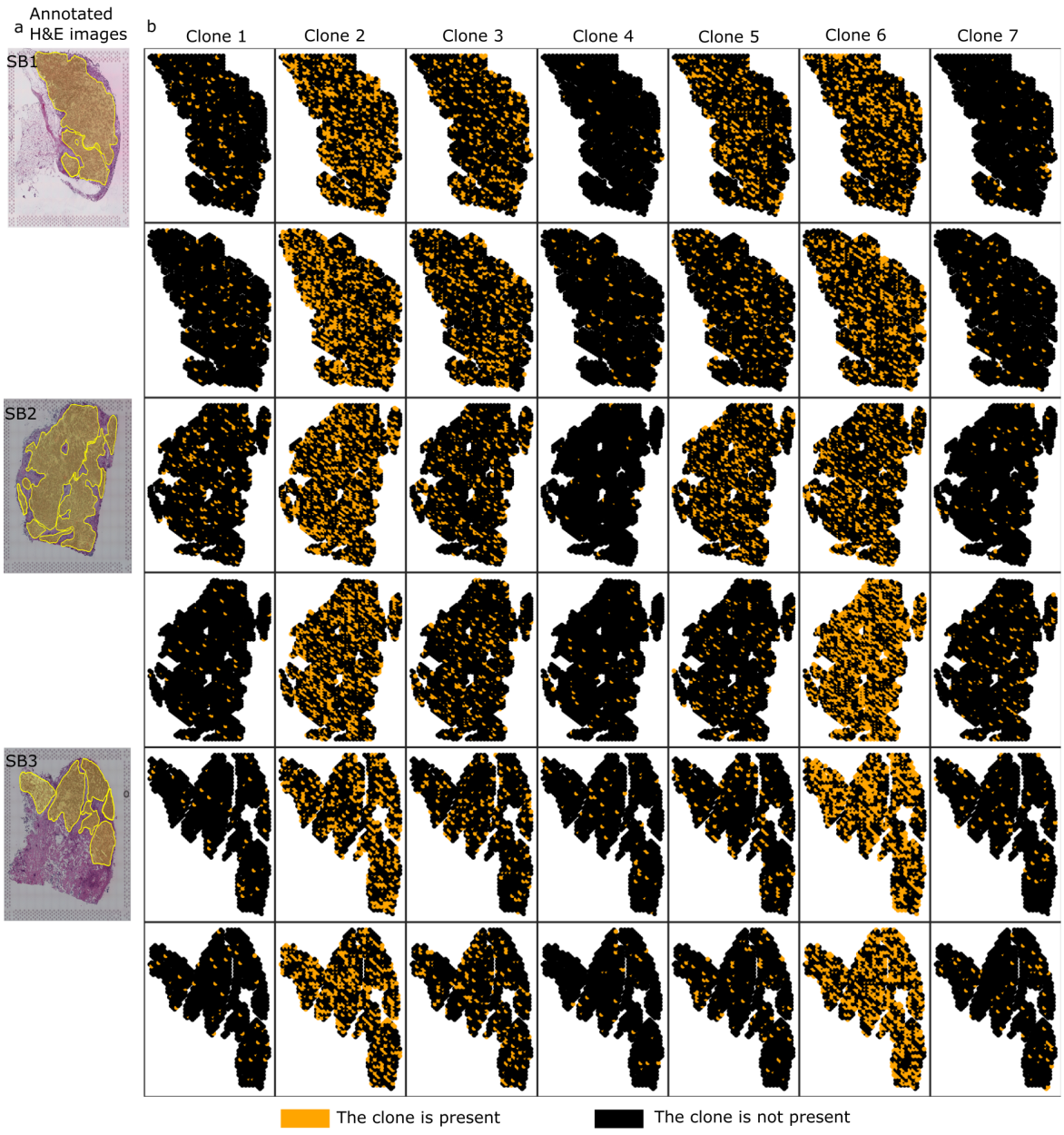
6.4. Data and Code availability

Tumoroscope can be obtained as an installable Python package, via 'pip install tumoroscope', and is available under the GNU General Public License v3.0. Tumoroscope implementation, package updates, and datasets supporting the conclusions of this article will be maintained at `https://github.com/szczurek-lab/Tumoroscope`.

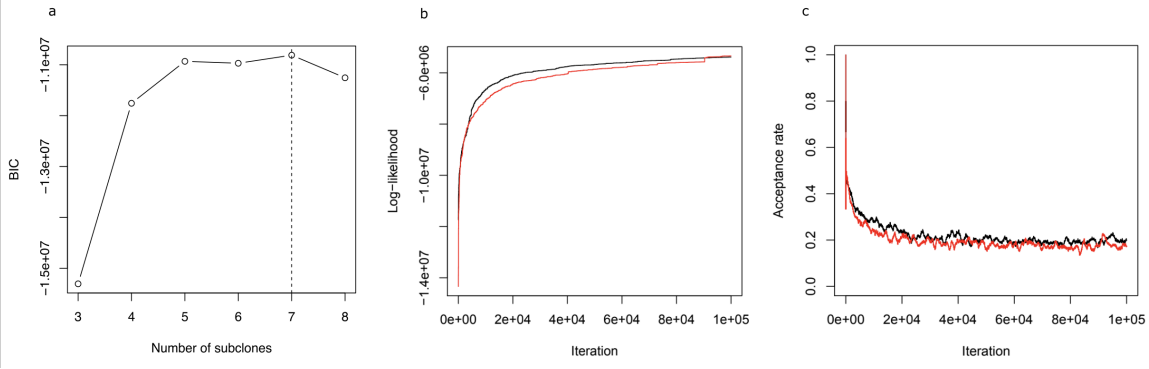
6.5. Extended data



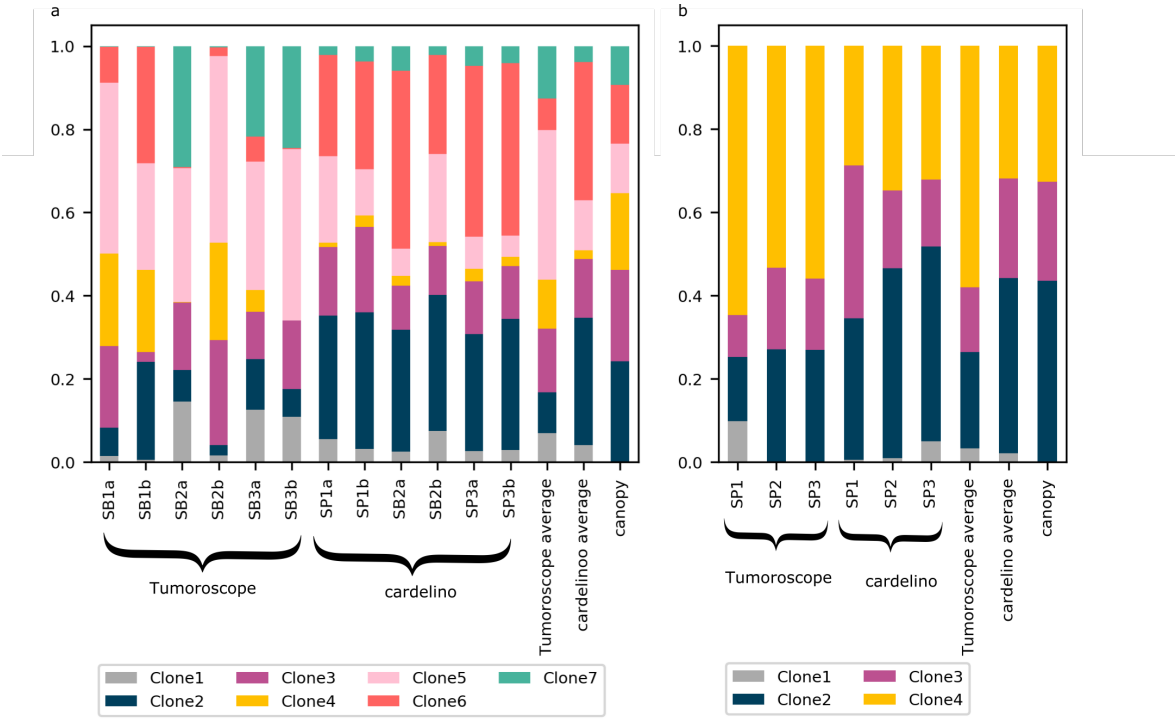
Extended data Figure 6.1: **Performance of Tumoroscope on simulated data with a high number of reads.** The axess and legends are the same as in Fig. 6.2 in the main text. Here, high, medium, low, and very low are corresponding to the average number of reads present in each spot of 297, 734, 1488, and 2246, respectively. Overall, compared to Fig. 6.2, having a higher number of reads increased the performance (strongly decreased MAE) for the estimation of the fraction of the clones in spots. **f** In the case when Tumoroscope-fixed is given a fixed number of cells that is highly noisy, increasing the number of clones in spots entangles the deconvolution problem. Consequently, for Tumoroscope-fixed, the highly noisy input confounds the model the most when the read counts are high and the model cannot assign the right clones to the spots, resulting in the largest MAE.



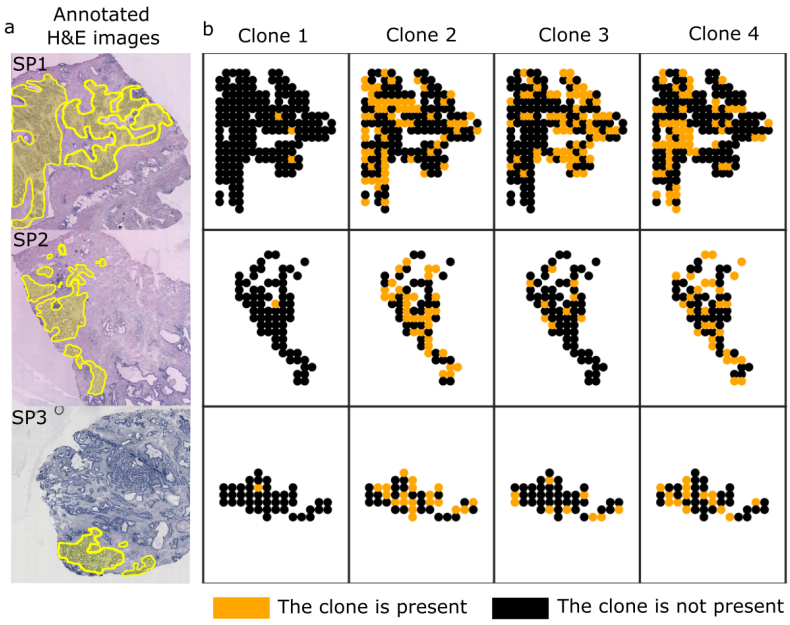
Extended data Figure 6.2: **Spatial arrangement of cancer clones found by the cardelino for the breast cancer dataset. a** Pathologist's annotation of the cancerous areas on the H&E images for sections SB1, SB2, and SB3. **b** For each section, two rows correspond to the two nearby samples and seven columns correspond to the presence of the clone in the spots inferred by cardelino.



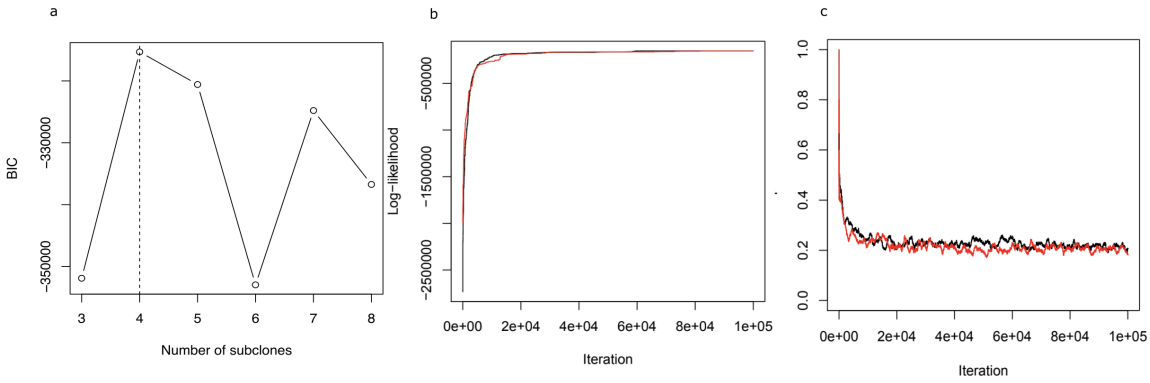
Extended data Figure 6.3: **Analysis of the Canopy tree inference for the breast cancer dataset. a** Bayesian Information Criterion (BIC; y-axis) of the Canopy model for different numbers of clones in the tree (x-axis). We selected the tree with seven clones, for which the BIC was the largest (indicated with the dotted vertical line). **b** Log-Likelihood of two MCMC chains of Canopy (y-axis) across MCMC iterations (x-axis), showing the convergence of the MCMC procedure. **c** Acceptance rate (y-axis) across iterations (x-axis). The acceptance rate converges to around the desired value of around 0.2.



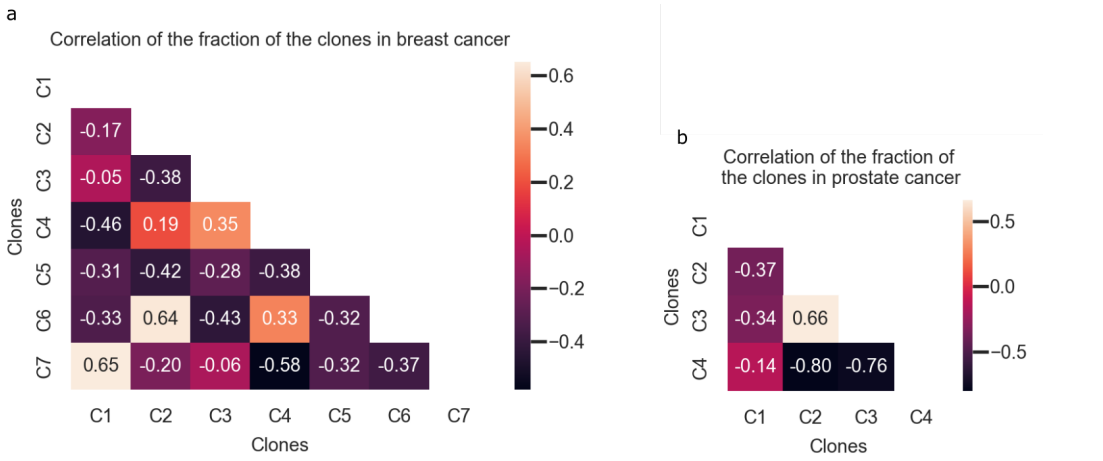
Extended data Figure 6.4: **Proportion of each clone in each section. a** Proportions of inferred clones by Tumoroscope and cardelino for the breast cancer dataset. The proportions were computed by summing the inferred fractions across spots for each ST section. Averages over sections and clone frequences inferred by Canopy from bulk DNAseq data are also shown. **b** Proportions of inferred clones by Tumoroscope and cardelino for the prostate cancer dataset.



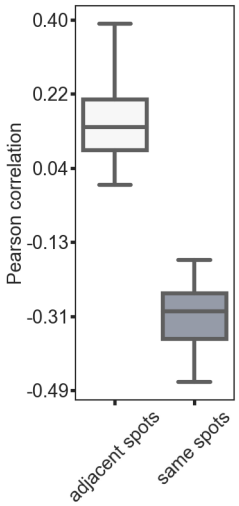
Extended data Figure 6.5: **Results obtained by cardelino for the prostate cancer dataset. a** Pathologist's annotation of the cancerous areas on the H&E images for sections SP1, SP2, and SP3. **b** For each section (rows), there are four columns corresponding to the presence of the clones in the spots.



Extended data Figure 6.6: **Analysis of the Canopy tree inference for prostate cancer dataset. a** Bayesian Information Criterion (BIC; y-axis) of the Canopy model for different numbers of clones in the tree (x-axis). We selected the tree with four clones, for which the BIC was the largest (indicated with the dotted vertical line). **b** Log-Likelihood of two MCMC chains of Canopy (y-axis) across MCMC iterations (x-axis), showing the convergence of the MCMC procedure. **c** Acceptance rate (y-axis) across iterations (x-axis). The acceptance rate converges to around the desired value of around 0.2.



Extended data Figure 6.7: Pairwise Pearson correlations of the proportions of all the spots taken by the clones for breast cancer (**a**) and for prostate cancer (**b**) data.



Extended data Figure 6.8: **Distribution of the Pearson correlation between the proportions of the spots taken by the clones 3 and 5**. The correlation between proportions of clone 5 and 3 in adjacent spots was computed for 20 different sets of randomly sampled pairs of 100 adjacent spots. The correlation between proportions of clone 5 and 3 in the same spots was computed for 20 sets of 100 randomly sampled spots.

Chapter 7

ClonalGE: Clonal gene expression analysis from spatial transcriptomics data

Tumor cells evolve through the acquisition of genetic alterations, leading to the formation of subpopulations of cells known as clones. Targeted cancer therapies often fail to eradicate all of the clones, thereby exerting selective evolutionary pressure on the remaining cells [237]. Understanding the heterogeneity of a tumor, particularly in terms of the variation in phenotype and localization of the different clones, is critical for effectively targeting and eliminating the cancerous cells while minimizing the risk of recurrence [238].

Previous studies of cancer heterogeneity mainly focused on solely resolving the genotype of the clones. These studies infer the evolutionary history of somatic alterations in the tumor from bulk DNA sequencing data [199, 200, 201, 165, 202, 203], or multi-region bulk DNA sequencing [239], or single cell DNA sequencing [240, 222]. While these methods can indicate the number of the clones, their frequencies and genotypes in the sample, they cannot estimate their gene expression profiles, nor identify genes that are deferentially expressed between the clones, nor localize in high resolution where the clones are in the tumor tissue.

The association between gene expression and clones was studied by [241] using semi-supervised identification of clones based on single cell RNA sequencing (scRNA-seq) data. This involved analyzing the gene expression and activated pathways of each clone. An alternative approach, called CONICS, predicted gene expression of copy number variant (CNV)-based clones using previously learned global correlation of CNV and gene expression from multiple scRNA-seq datasets and corresponding whole exome sequencing (WES) data [242]. [243] proposed clonealign, an approach that assigned gene expression states to cancer clones using RNA-seq and DNA-seq for single cells independently sampled from a heterogeneous tumor population. Other approaches, such as Cardelino [177] and CACTUS [43], combined WES and scRNA-seq data to first infer cancer clones and their genotypes from WES and then map single cells to the clones based on the shared single nucleotide variants (SNVs). By aggregating RNA signal from the single cells into clones, these approaches identified clone-specific phenotypes. These methods, however, were unable to identify the localization of the clones in the tissue and required availability of single cell RNA-sequencing data.

Recent spatial transcriptomics (ST) technology measures gene expression across localized spots in the tissue [42]. This technology utilizes the RNA sequencing protocol and generates a mixture of gene expression from the cells present in a given spot. Spatial transcriptomics can be seen as an alternative source of gene expression measurement for identifying clonal

expression profiles simultaneously with their localization. Several methods emerged that combine signal in WES and ST data to identify cancer clones and their mapping in the tissue based on CNVs [214, 213, 244]. These methods, however, ignore important evolutionary events caused by SNVs, which play a crucial role in tumor progression and treatment.

To address this issue, recently, we proposed a model called Tumoroscope [245], which combines WES and ST data alongside H&E images to infer and localize clones based on SNVs. The main functionality of Tumoroscope is inference of the proportions of each ST spot occupied by each clone. Estimation of clonal gene expression profiles in the Tumoroscope framework is done as a post-processing step, using linear regression with weights set to the inferred proportions [245]. Therefore, the accuracy of the gene expression profile estimation using Tumoroscope is limited. Moreover, the framework does not include the ability to perform differential gene expression analysis. This is due to the fact that Tumoroscope followed by linear regression does not estimate the full distribution of clone-specific gene expression profiles, but rather provides its point estimates, making it impossible to perform differential gene expression analyses. Therefore, an approach that integrates WES and ST data in a single model, focusing on gene expression analysis of the clones, identification of their gene expression distributions, differentially expressed genes, and explaining the ST measurements as compositions of gene expression stemming from the localized clones is lacking.

In this study, we present ClonalGE, a novel statistical graphical model that infers clone-specific gene expression and the composition of clones present in localized spots within a tissue sample. In addition to utilizing WES data, H&E images, as well as read counts for somatic SNVs shared between WES and ST, ClonalGE leverages expression measurements of genes across the tissue sample to improve the accuracy of mapping clones to specific regions. Furthermore, ClonalGE enables the generation of distributions of clone-specific gene expression profiles, allowing for subsequent differential expression analysis between pairs of clones. Our model contributes to the increased understanding of tumor heterogeneity, not only on genetic but also on the phenotypic level.

7.1. Methods

7.1.1. A novel framework for inference of gene expression profiles specific for cancer clones, together with spatial mapping of the clones in tumor tissue and estimation of differential gene expression between clones

We propose a novel probabilistic framework for inference of gene expression profiles of cancer clones, alongside the spatial mapping of the clones across the tissue and estimation of differential gene expression, using H&E stained images, spatially-resolved transcriptomics, and bulk DNA-seq. This framework takes as input pre-processed data, including: regions of the tissue annotated as cancer and the number of cells in the annotated ST spots (based on H&E images; Fig. 7.1a), genotypes and frequencies of the inferred clones (based on bulk DNA-seq data; Fig. 7.1b), the total and alternated read counts over selected mutations (based on ST reads; Fig. 7.1c), and the gene expression profile of the spots (based on ST reads; Fig. 7.1d). The core of the framework is a probabilistic graphical model called ClonalGE. ClonalGE combines inference of clonal gene expression and composition of ST spots. The model is a profound extension of our previous model, Tumoroscope [245], by incorporating additional variables for the gene expression in the spots. The assumptions behind this extension are that each clone has its specific average gene expression profile; and that the expression of the genes observed at the spots is a mixture of the gene expressions in the clones that are present in those spots. For inference, Tumoroscope is run first to provide initialisation of all

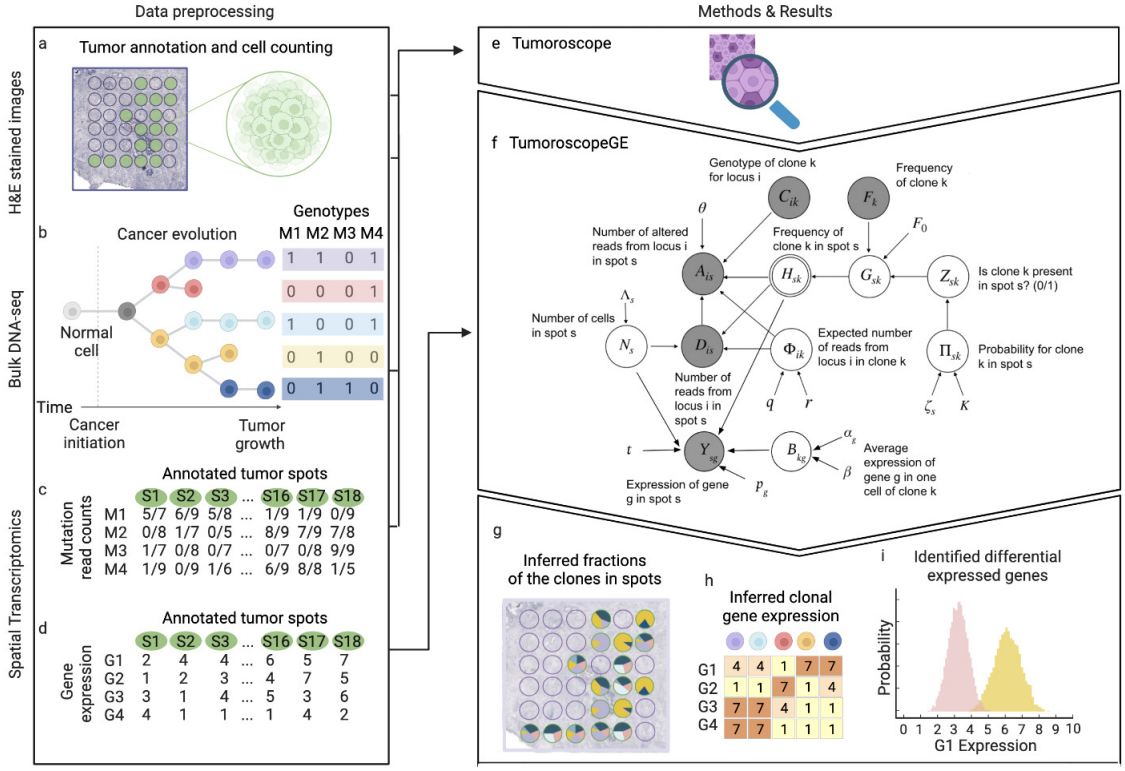


Figure 7.1: **ClonalGE framework overview. a-d** Data preprocessing. **e** Initialization of hidden variables by running Tumoroscope. **f** ClonalGE probabilistic model. **g-i** Outputs: **g** Inferred fraction of the clones in the spots. **h** Inferred clonal gene expression. **i** Identified differentially expressed genes.

the hidden variables to ClonalGE (Fig. 7.1e). Next, we use Metropolis Hastings within Gibbs for estimation of ClonalGE (Fig. 7.1f). The model returns the fraction of each annotated ST spot occupied by each clone (Fig. 7.1g) together with the gene expression profile of each clone (Fig. 7.1h). Finally, we use the model-inferred distributions of gene expression for each clone to perform differential gene expression analysis between clones (Fig. 7.1i).

7.1.2. Prostate cancer sample

We used a published prostate cancer dataset including bulk DNA-seq, H&E images, and spatial transcriptomics for twelve sections regarding one patient. Besides, we have the bulk DNA-seq data from the patient's blood to serve as a control for detecting the somatic mutations. The generating and pre-processing steps alongside with the ST data bam files and gene expression matrix are available in [44] and [245].

7.1.3. Spots that contain tumor cells and cell counts in spots

For each section, we retrieved the same spots that were identified by [245] as cancerous. For cancerous spot identification, the regions containing cancer cells were annotated by an expert pathologist taking advantage of the H&E image of each section. Afterward, using a custom script in QuPath [215], the spots inside the annotated regions were identified.

Similarly, for each section, we used the cell counts that were obtained by [245] based on $H\&E$ images using a custom script in QuPath [215].

7.1.4. Bulk DNA-seq and somatic mutation calling

We used the same set of somatic mutations as analyzed by [245]. This set contained mutations that were called using Vardict [130] with a p-value threshold equal to 0.1 in at least one of the bulk DNA-seq sections (i.e., the set of mutations is the union over mutations found in individual sections).

7.1.5. Selection of somatic mutations that are detected both in bulk DNA-seq and ST data

We used the somatic mutations and their corresponding total and alternated reads as analyzed by the Tumoroscope model previously and provided by [245]. Specifically, calculation of the the total and alternated reads over the mutations in ST data was done by a script provided by [245], which finds the selected bulk DNA-seq mutations in the ST bam files and counts the corresponding mapped reads. Finally, we selected the mutations for which at least one alternated read in at least one section were found. We give alternated and total read counts in bulk DNA-seq data as input to the method of phylogenetic tree inference. On the other hand, we give the alternated and total read counts in ST data for the same mutations as input to ClonalGE.

7.1.6. Phylogenetic tree analysis

We used the genotypes and frequencies of the clones provided by [245] that were obtained using a statistical method for inference of the phylogenetic tree called Canopy [165]. Canopy, as the inputs take 1) variant allele frequencies (VAF) of somatic single nucleotide alterations (SNAs), obtained by Vardict [130] and 2) ratios of the coverage between the tumor and normal sample for somatic copy number alterations (CNAs), obtained by FalconX [142]. Furthermore, the multi-sample feature of the Canopy was used to infer the clonal evolution across the sections.

7.1.7. ClonalGE

ClonalGE is a probabilistic graphical model for estimating the gene expression profile of the clones alongside with the fraction of the clones in the ST spots across the tumor sample.

Similarly as in the previous Tumoroscope model, let $k \in \{1, \ldots, K\}$ index the clones derived from DNA-sequencing data and $i \in \{1, \ldots, I\}$ index the loci with somatic point mutations observed both in ST and DNA-sequencing data. F_k indicates the frequency of occurrence of clone k in the whole sample and C_{ik} codes the inferred genotype, taking values between 0 and 1 that is defined by the ratio of the number of alleles of clone k carrying a mutation on locus i to the total number of alleles. Both the frequencies F_k and genotypes C_{ik} are specifications of the clones and are given as observed variables to the model.

Let $s \in \{1, \ldots, S\}$ index the spots in the ST data. Z_{sk} is a hidden binary variable, indicating the presence of clone k in spot s. It follows a Bernoulli distribution with success parameter Π_{sk}, following a Beta prior distribution.

$$\mathbb{P}(Z_{sk}|\Pi_{sk}) \sim \text{Bern}(\Pi_{sk})$$

$$\mathbb{P}(\Pi_{sk}|\zeta_s, K) \sim \text{Beta}\left(\frac{\zeta_s}{K}, 1\right)$$

The expected number of clones in a spot can be controlled with the ζ_s hyperparameter.

Depending on the indicator variables Z_{sk} and the observed clone genotypes and frequencies, we expect higher or lower values of unnormalized abundances of the clones in the spots.

These values for each spot s and clone k are described in the model using Gamma-distributed hidden variables G_{sk}.

$$\mathbb{P}(G_{sk}|F_k, F_0, Z_{sk}) \sim \text{Gamma}(f(F_k)^{Z_{sk}} {F_0}^{1-Z_{sk}}, 1)$$

The first parameter of the gamma distribution is calculated using a discretized value of F_k and F_0. The discretization function f is defined by $f(x) = \lceil \frac{x}{0.05} \rceil \cdot 100$. The hyperparameter F_0 is used as a pseudo-count for clones not assigned to the spot, indicated by $Z_{sk} = 0$.

H_{sk} is a deterministic variable that can be interpreted as a fraction of spot s assigned to clone k. It follows a Dirichlet distribution and is calculated based on the unnormalized abundances G_{sk} as

$$H_{sk} = \frac{G_{sk}}{\sum_{l=1}^{K} G_{sk}}$$

Note that H_{sk}'s are one of the two central sets of random variables in the model - they answer the question of the proportions of the clones in the spots.

Φ_{ik} is a latent variable that is interpreted as an average read count at position i per one cell of clone k. It follows a Gamma distribution with hyperparameters r and q:

$$\mathbb{P}(\Phi_{ik}|r, q) \sim \text{Gamma}(r, q).$$

The number of cells in spot s, represented by a variable N_s is modeled with a Poisson distribution:

$$\mathbb{P}(N_s|\Lambda_s) \sim \text{Poiss}(\Lambda_s),$$

where Λ_s is the provided at input, estimated from H&E images, cell count at spot s.

D_{is} represents counts of reads over locus i coming from spot s and A_{is} indicates how many of them are mutated (altered). D_{is} and A_{is} are observed variables and are given to the model and assumed to follow a Poisson distribution and a Binomial distribution respectively:

$$\mathbb{P}(D_{is}|H_{s.}, \Phi_{i.}, N_s) \sim \text{Pois}\left(N_s \sum_k H_{sk}\Phi_{ik}\right)$$

$$\mathbb{P}(A_{is}|D_{is}, \Phi_{i.}, H_{s.}, C_{i.}) \sim \text{Binom}\left(D_{is}, \frac{\sum_k \Phi_{ik}C_{ik}H_{sk}}{\sum_k \Phi_{ik}H_{sk}}\right).$$

All the above definitions concern also the previous model Tumoroscope by [245]. We now explain the model extension to account for gene expression values in the spot, allowing to estimate gene expression profiles of clones as an integral part of the model.

Let $g \in \{1, \ldots, G\}$ index the genes observed in the ST data. Our main goal is to estimate the average gene expression of gene g per one cell in clone k, represented by a hidden random variable B_{kg}, which is assumed to follow a Gamma distribution with hyperparameters α_g and β

$$\mathbb{P}(B_{kg}|\alpha_g, \beta) \sim \text{Gamma}(\alpha_g, \beta)$$

An observed variable Y_{sg} signifies the expression of gene g in spot s which is the main measurement in ST data and is given to the model at input. We assume a Negative Binomial distribution over Y_{sg} with expected value of $N_s \sum_k H_{sk}B_{kg}$, which is the clonal gene expression multiplied by the number of cells of each clone in spot s. To overcome the batch effect observed when integrating the data from different sections, we added a scaling parameter t_j to the model (with j indexing the sections and t_j assumed to be common for all spots in section j). Then the expected value for Y_{sg} is $t_j \cdot N_s \sum_k H_{sk}B_{kg}$ given that spot s is in section j. Considering

r_{sg} and p_g as the parameters of the Negative Binomial distribution over Y_{sg}, the expected value of Y_{sg} is $\mathbb{E}[Y_{sg}] = \frac{(1-p_g)r_{sg}}{p_g}$ and the variance is equal to $Var(Y_{sg}) = \frac{(1-p_g)r_{sg}}{p_g^2}$. Using the formula for $\mathbb{E}[Y_{sg}]$ we substitute r_{sg} by

$$r_{sg} = \frac{\mathbb{E}[Y_{sg}]p_g}{1-p_g},$$

which results in a useful reparametrization:

$$Y_{sg} \sim \text{NB}\left(\frac{\mathbb{E}[Y_{sg}]p_g}{1-p_g}, p_g\right). \tag{7.1}$$

Using this parametrization, we can rewrite the distribution of Y_{sg}:

$$\mathbb{P}(Y_{sg}|t_j, B_{.g}, p_g, H_{s.}, N_s) \;\sim\; \text{NB}\left(t_j N_s \sum_k H_{sk}B_{kg}\frac{p_g}{1-p_g}, p_g\right).$$

where spot s is in section j.

7.1.8. Metropolis-Hasting inside Gibbs sampling

For estimating the posterior distribution of the hidden variables in our probabilistic model, we iteratively generate samples from each hidden variable's conditional distribution, given the remaining variables. We use a Metropolis-Hasting step inside the Gibbs sampler in case a closed-form distribution for sampling a random variable does not exist.

The sampling procedure for variables $\Phi_{i,k}$, $Z_{s,k}$ and $\Pi_{s,k}$ remain the same as is described in [245]. We use Metropolis-Hastings inside Gibbs sampler for variables $\Phi_{i,k}$, $G_{s,k}$, N_s and $B_{k,g}$ since there exists no closed form conditional distribution for them. We choose a Truncated Normal distribution for the proposal distribution due to the non-negativity of the variables of our interest. x_c represents the mean value and σ_G and σ_N represent the variances of the proposal distribution corresponding to each variable. The proximity of the new sample from the current one, which is interpreted as the step size, is determined by the variance of the proposal distribution. During the sampling steps, we learn the optimal variance value for each variable as described by [245].

Conditional distribution for G_{sk}:

$$\begin{aligned}
&\mathbb{P}(G_{sk}|F_k, F_0, Z_{sk}, A_{is}, D_{is}) \\
&\qquad \propto \mathbb{P}(G_{sk}|F_k, F_0, Z_{sk}) \prod_i \mathbb{P}(D_{is}|\Phi_{i.}, H_{s.}, N_s) \\
&\qquad\quad \times \prod_i \mathbb{P}(A_{is}|D_{is}, \Phi_{i.}, H_{s.}, C_{i.}) \\
&\qquad\quad \times \prod_g \mathbb{P}(Y_{sg}|t, B_{.g}, p_g, H_{s.}, N_s) \\
&\quad = Gamma(f(F_k)^{Z_{sk}} F_0^{1-Z_{sk}}, 1) \prod_i Pois(N_s \sum_k H_{sk}\Phi_{ik}) \\
&\qquad\quad \times \prod_i Binom(D_{is}, \frac{\sum_k \Phi_{ik}C_{ik}H_{sk}}{\sum_k \Phi_{ik}H_{sk}}) \\
&\qquad\qquad\quad \times \prod_g NB(tN_s \sum_k H_{sk}B_{kg}\frac{p_g}{1-p_g}, p_g).
\end{aligned}$$

where with italic names we write the density functions of the respective probability distributions. Conditional distribution for N_s:

$$\begin{aligned}\mathbb{P}(N_s|\Lambda_s, D_{is}, Y_{sg}) &\propto \mathbb{P}(N_s|\Lambda_s) \\ &\times \prod_i \mathbb{P}(D_{is}|\Phi_{i.}, H_{s.}, N_s) \prod_g \mathbb{P}(Y_{sg}|t, B_{.g}, p_g, H_{s.}, N_s) \\ &= Pois(\Lambda_s) \prod_i Pois(N_s \sum_k H_{sk}\Phi_{ik}) \\ &\times \prod_g NB(tN_s \sum_k H_{sk}B_{kg}\frac{p_g}{1-p_g}, p_g).\end{aligned}$$

Conditional distribution for B_{kg}:

$$\begin{aligned}\mathbb{P}(B_{kg}|\alpha_g, \beta, Y_{sg}) &\propto \mathbb{P}(B_{kg}|\alpha_g, \beta) \prod_s \mathbb{P}(Y_{sg}|t, B_{.g}, p_g, H_{s.}, N_s) \\ &= Gamma(\alpha_g, \beta) \prod_s NB(tN_s \sum_k H_{sk}B_{kg}\frac{p_g}{1-p_g}, p_g).\end{aligned}$$

7.1.9. Hyper-parameter estimation and initialization of random variable values prior to MCMC

We estimate the hyper-parameters of the model based on the real data. First, the value of Λ_s is set to the estimated number of cells in spot s, calculated based on the H&E images. Second, having a Negative Binomial distribution over Y_{sg} with the success probability p_g, we have

$$p_g = \frac{\mathbb{E}[Y_{sg}]}{Var(Y_{sg})}.$$

Thus, we approximate p_g by estimating $\mathbb{E}[Y_{sg}]$ and $Var(Y_{sg})$ using arithmetic mean and the variance of the given gene expression data for gene g over the spots.

For estimating the gene expression scaling factor, t_j, we calculate the mean value of Y_{sg} for spots s that are present in section j. Let $m_1, \ldots, m_J$ be the average gene expression for spots in the J analyzed sections and let $m = \frac{1}{J}(m_1 + \ldots + m_J)$. Then for section j the estimated scaling factor is $\hat{t}_j = \frac{m_j}{m}$.

For estimating the hyperparameters α_g and β of the Gamma distribution of the B_{kg} variables we use the formulas for expected value, $\mathbb{E}[B_{kg}] = \alpha_g\beta$, and the variance, $Var(B_{kg}) = \alpha_g\beta^2$ of the Gamma distribution. Therefore, one could obtain β by dividing the two values.

$$\beta = \frac{Var(B_{kg})}{\mathbb{E}[B_{kg}]}.$$

Since we do not have B values as they are latent variables, we approximate them by the average expression of the gene per cell over clones:

$$B_{kg} \approx \frac{\sum_{k=1}^{K} B_{kg}}{K} \approx \frac{1}{S \cdot J} \sum_j \sum_{s \text{ in section } j} \left(\frac{Y_{sg}}{t_j N_s}\right).$$

Therefore, we estimate $\mathbb{E}[B_{kg}]$ and $Var(B_{kg})$ by calculating the mean and variance of these average values across genes. By considering the average over clones, we are underestimating the variance and accordingly the values of β. Therefore, we multiply the estimated variance

by a correction parameter ω_B. The value of ω_B was estimated based on simulations and fixed to $\omega_B = 2$.

Next, based on the expected value of the gamma distribution over B_{kg}, $\mathbb{E}[B_{kg}] = \alpha_g \beta$, we approximate α_g using the estimated $\hat{\beta}$ and the already calculated average expression of the genes over the clones.

We estimate the hyper-parameters of gamma distribution over Φ, r and q, similarly to the α_g and β, using the approximation of the expected value and variance of Φ.

$$\hat{E}(\Phi) = \frac{\sum_i \sum_s \frac{D_{is}}{N_s}}{I \times S}$$

$$\hat{Var}(\Phi) = \frac{\sum_i \sum_s (\frac{D_{is}}{N_s} - E(\Phi))}{I \times S}.$$

Then, we estimate the hyper-parameters with

$$\hat{r} = \omega_\phi \frac{\hat{Var}(\Phi)}{\hat{E}(\Phi)} \quad \text{and} \quad \hat{q} = \frac{\hat{E}(\Phi)}{\hat{r}}.$$

Here, we use a constant, ω_ϕ, for compensation of the underestimation of the r. This correction constant was again estimated based on simulations and fixed to $\omega_\phi = 2$.

To find initial values of hidden variables prior to MCMC, we first run Tumoroscope for at least 7000 iterations stopping after reaching the Geweke's convergence criterion, with the upper bound of 15000 iterations. Afterward, we use the inferred variables to initialize the corresponding hidden variables in ClonalGE. We further use the regression model proposed in [245] to initialize the B matrix.

7.1.10. Examining convergence using Geweke's diagnostic

We examine the convergence of the H variable matrix using Geweke's convergence diagnostic [246]. We consider different sizes of the burn-in period and examine them by calculating the percentage of the convergence after each batch (we use batches of 1000 iterations). We finish the sampling process if at least 99.5% of the chains have already converged and the sampling has reached a selected minimal number of iterations.

7.1.11. Inferring the variables based on the samples

For the real data, we compute final values of the inferred variables using multiple chains. Specifically, we run the model 10 times, each time with 10 chains. Then, from each run we select the chain with the highest likelihood, and all chains that obtained similar results. We consider two chains similar if the Pearson correlation between the inferred H_{sk} values is higher than 0.9. If the chain with the highest likelihood does not have any other similar chains, we omit it. Afterward, we take an average out of all the chains that are similar to the one with the highest likelihood out of the remaining chains.

7.1.12. Differential gene expression analysis between clones

We obtain distributions of the inferred gene expression per each clone k and gene g by selecting every 1000-th B_{kg} sample after a burn-in period of 5000 iterations, from the chains that agreed on the inferred H_{sk} values. Then we apply the two-sided Wilcoxon rank-sum test [247] to compare the distributions between different clones. We calculate this statistic for each pair of clones, excluding clone 1, and each gene from a given set of n genes. We use the Bonferroni correction to account for the multiple comparisons (for 3 clones it is $3 \times n$ comparisons).

7.1.13. Computing the expected gene expression values across spots based on the inferred gene expression profiles of the clones

Having the inferred clonal expression B_{kg}, we compute the gene expression profile across spots in the tissue section j as:

$$\hat{Y}_{sg} = \sum_k t_j N_s H_{sk} B_{kg} = t_j N_s \sum_k H_{sk} B_{kg}.$$

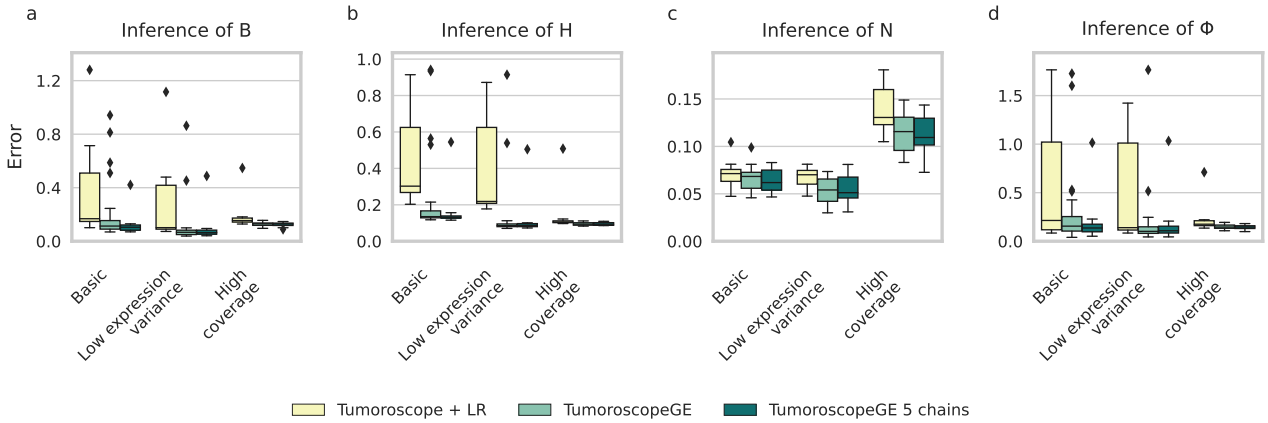


Figure 7.2: **P**erformance of ClonalGE on simulated data. **a** Error (computed as the Mean Absolute Error divided by the mean of the true values; y-axis) of inferred clonal gene expression profile in different simulation setups (y-axis) for different methods (colors). **b** The same as in (**a**), but for the inferred fraction of the clones in the spots. **c** The same as in (**a**), but for the corrected number of cells in the spots. **d** The same as in (**a**), but for the inferred number of total reads per cell in each clone.

7.2. Results

7.2.1. ClonalGE correctly estimated the clonal gene expression in the spots

We first assessed the performance of the ClonalGE on simulated data, where the ground truth was known, creating three different simulation setups. First, we generated a *basic* setup, with five clones randomly mixed in 300 spots. The total number of different mutations was 200 and the number of analyzed genes was 50. The expected value for the number of cells and clones in each spot were 45 and 2.5 respectively. The sum of simulated read counts per each spot was in the (36, 44) interval, and the expected expression variance was 1715, reflecting the levels observed in real data. Next, to test the influence of the coverage per mutation and the different expression levels, we designed two additional setups (*low expression variance* and *high coverage*), lowering the expected value of the expression variance and increasing the average number of reads, respectively (see Table 7.1). For each setup, we repeated the simulation 20 times.

For inference, we first ran the sampling process of ClonalGE for at least 10000 iterations, saving every 5th sample and stopping after reaching convergence with an upper bound of 30000 iterations. The initial parameters were set using a pre-run of Tumoroscope (see Methods). We discarded the burn-in samples and used the remaining for inference of the latent variables. We used known values of the hyper-parameters, except for r and q, which were estimated (see Methods).

setup property	basic	low expression variance	high coverage
interval of the average read counts	$(36, 44)$	$(36, 44)$	$(180, 221)$
expected value of expression variance	1715	315	1715

Table 7.1: Parameters for generating three different simulation setups.

We also evaluated another running procedure of ClonalGE that aimed at decreasing the chances of getting stuck in local optima. To this end, we ran five MCMC chains for ClonalGE and selected the chain with the highest likelihood. This procedure is referred to as ClonalGE 5 chains.

For comparison, we ran Tumoroscope with its parameter estimation procedure for the same lower and upper bound numbers of iterations as ClonalGE. To estimate gene expression per clone after running Tumoroscope, similarly to [245], we used linear regression (LR) with weights fixed to the fractions of clones in spots inferred by Tumoroscope. Hence, this entire procedure is referred to as Tumorocope + LR.

Overall, both TumroscopeGE and ClonalGE 5 chains obtained significantly better results compared to Tumoroscope + LR in all simulation setups (Fig. 7.2). In particular, ClonalGE obtained much more accurate estimates of gene expression values in clones (Fig. 7.2a). This result confirms that the joint modeling of the gene expression values with the mutation read counts as performed in ClonalGE gives increased statistical strength and yields higher performance than the double step procedure employed by combining Tumoroscope with linear regression. This advantage may also come from the appropriate model of the Negative Binomial distribution for the expression data per gene and spot employed by TumoroscpeGE, in contrast to linear regression.

Interestingly, ClonalGE also outperforms Tumoroscope in the functionality that is shared between these two models: inference of the fraction of clones in spots (Fig. 7.2b), numbers of cells in spots (Fig. 7.2c), and read coverages per mutation per cell (Fig. 7.2d). This is likely due to the fact that ClonalGE, in contrast to Tumoroscope, benefits from the additional signal in gene expression measurements across spots. The most significant improvement over Tumoroscope is visible for the *basic* and *low expression variance* setups. In both these setups, the coverage is low. These demanding setups illustrate the sensitivity of Tumoroscope to coverage, which is not the case for ClonalGE.

As expected, ClonalGE with five chains achieved more robust results with lower outliers compared to both ClonalGE and Tumoroscope + LR. The mean and median of the error in all the setups and variables decreased or stayed equal when we used five chains instead of one. These results emphasize the importance of running more chains to decrease the probability of the model getting trapped in the local minimum.

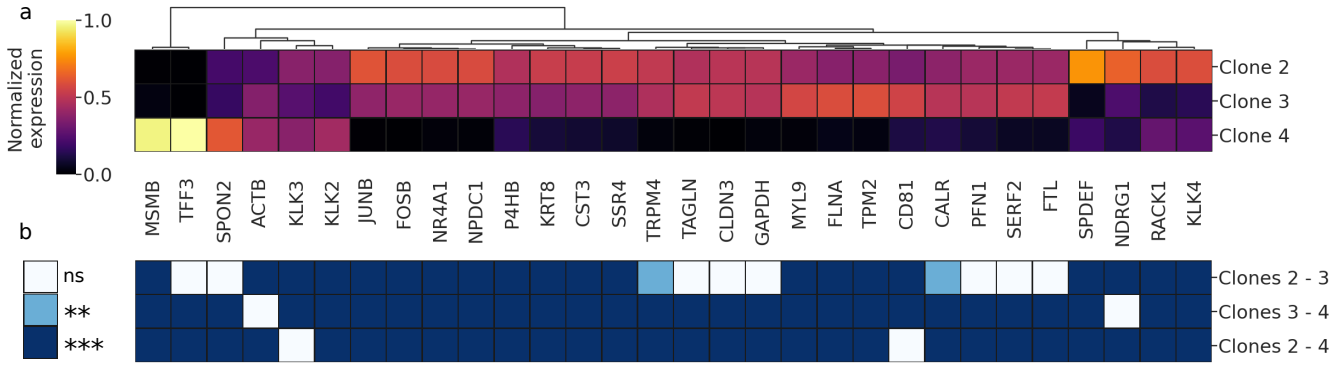


Figure 7.3: **G**enes are expressed differently in various cancer clones. **a** The expression of the 30 genes that were inferred by ClonalGE as the most active in at least one clone, clustered in rows and columns, for prostate cancer tissues. **b** The p-value of the differential gene expression analysis ns: statistically non-significant; **: p-value between 0.001 and 0.01; ***: p-value ≤ 0.001

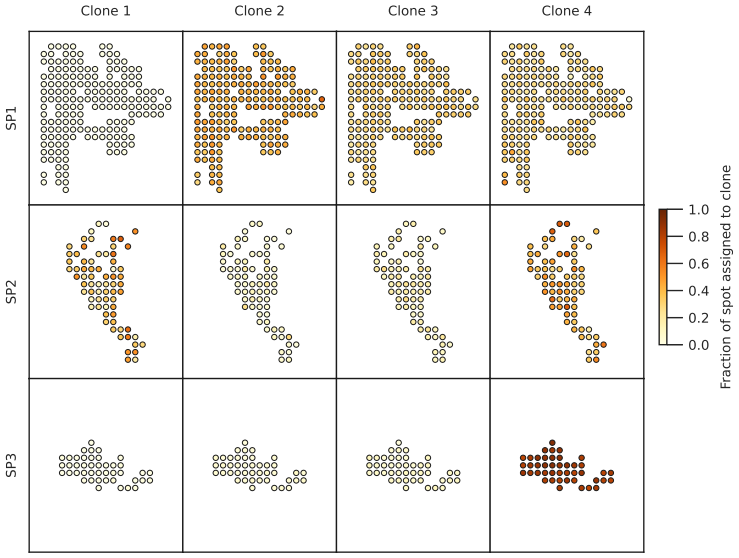


Figure 7.4: The proportion of the spots assigned by ClonalGE to each clone (columns) for each section (rows) of the prostate sample.

7.2.2. ClonalGE reconstructs the gene expression profile of the clones on prostate cancer data

To showcase the performance of ClonalGE on real data, we applied it to prostate cancer dataset generated by [44]. To this end, we selected 3 sections: SP1, SP2, and SP3 out of 12 sections, based on a pathologist's annotation of cancer regions. These sections were sampled from one patient, for which deep DNA-seq and ST data (custom arrays) of neighboring layers were generated [44]. From these sections, only the spots overlapping with cancer regions were selected, namely 198, 70, and 43 spots from sections SP1, SP2, and SP3 respectively, out of 968-1001 spots per section. Furthermore, we excluded 17 spots in section SP3 that lacked gene expression data. In the remaining 294 spots, the sequenced transcripts were mapped to 12873 different genes. We selected the data for the 1000 most variable genes to 1) limit the computation time and 2) concentrate on the genes that carry the most information for

more accurate inference. We used the mutations present in both bulk DNA-seq and ST data together with the evolutionary tree reconstructed from DNA-seq data by [245]. The tree included a normal clone without somatic mutations and three other clones. Finally, we applied ClonalGE to deconvolute the transcriptomic signal from 294 spots in the ST data in order to achieve the proportions of the underlying clones and the expression profile of the clones.

For running the model on the prostate cancer dataset, we used the same procedure and number of iterations as described for the simulated data. We repeated the sampling and inference 20 times, with different random seeds, and selected only the run with the highest likelihood. For the real data, the true values of the hyperparameters are unknown, so we estimated the hyperparameters (see Methods).

Finally, after inferring the gene expression profile of the clones, we ranked the genes based on the maximum expression variance of the genes across the clones in the inferred clone-specific gene expression profiles. Afterwards, we selected top 30 genes and plotted the normalized expression across the clones (Fig. 7.3a). According to the Human Protein Atlas [216], top 9 out of 30 ranked genes (*MSMB*, *SPON2*, *KLK3*, *KLK2*, *NPDC1*, *TRPM4*, *TAGLN*, *CLDN3*, *SPDEF*, *KLK4*) were found to have elevated expression in all prostate cancer tissues and can be referred to as "prostate enhanced genes". Among these prostate enhanced genes, there are 4 genes (*NPDC1*, *TRPM4*, *CLDN3*, *SPDEF*) that were not observed in the top ranked genes (with the highest inferred expression) found previously by [245] for the same dataset using Tumoroscope.

Furthermore, we performed Wilcoxon signed-rank test on each pair of clones in order to conduct a differential gene expression analysis (Methods; Fig. 7.3b). The results indicate that the majority of the top-ranked genes exhibited different expression levels across the clones with high p-value. Notably, the greatest degree of similarity in gene expression was observed for clones 2 and 3, where 8 genes were found to have non-significant p-values. This similarity is also reflected in normalized gene expression profiles of clones 2 and 3 (Fig. 7.3a).

Importantly, this pair of clones (2 and 3) were found to be localized in the same area across all three sections, suggesting a potential correlation between spatial localization and gene expression similarity (Fig. 7.4). Interestingly, the distribution of the clones across the analyzed tissue sections inferred by ClonalGE is different than the one inferred by Tumoroscope (compare Fig.4 in [245]), which can be attributed to the adoption of the extra information of gene expression in the inference by ClonalGE. Still, both analyses indicated similarity of clones 2 and 3 on the level of inferred gene expression and on the level of their spatial proximity in the tissue.

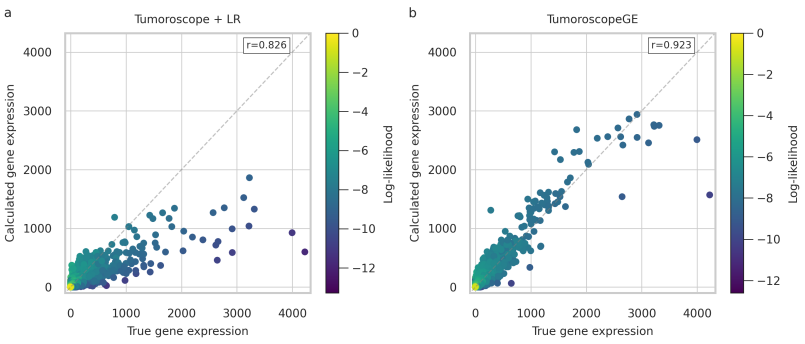


Figure 7.5: Comparison of the calculated (y-axis) and the true value (x-axis) of gene expression in the spots for Tumoroscope followed by linear regression (Tumoroscope + LR; **a**) and TumorscopeGE (**b**). The spots are colored based on the log-likelihood of the correspondin model.

7.2.3. ClonalGE reconstructs the gene expression profile of the tissue more accurately than Tumoroscope

Finally, we evaluated the accuracy of the gene expression reconstruction using ClonalGE. To this end, we compared the true gene expression values of the spots measured in the prostate tissue with the values expected for the same genes at the same spots based on the model estimation of clone-specific gene expression profiles (the real $\hat{Y}_{sg}$ values with the $\hat{Y}_{sg}$ values, see Methods for their calculation). For an accurate model, the real values should perfectly agree with the ones expected from the model.

For a comparison, we applied the entire procedure of Tumoroscope + LR on the same data, with the same number of iterations as for ClonalGE, and also obtained the expected gene expression values at the spots for Tumorscope + LR and compared them to the real ones.

We observed a much better agreement between the true gene expression values across spots and the expected values for ClonalGE than for Tumoroscope + LR (Pearson correlation $r = 0.923$ and $r = 0.826$ for ClonalGE and Tumoroscope+LR respectively; see Fig. 7.5a,b). Both methods show high agreement for small expression values. However, in contrast to TumroscopeGE, Tumoroscope + LR tends to underestimate gene expression values (Fig. 7.5a). Interestingly, the worst agreement is obtained for such genes which also have the lowest likelihood. This may be due to the fact that Tumoroscope + LR does not account for the Negative Binomial distribution of gene expression data across spots. With only a few exceptions, expected gene expression values based on ClonalGE estimates correctly recover the true measured values, even for large expression (Fig. 7.5b).

7.3. Discussion

This study demonstrates that using ST data together with the WES data and H&E images, we can robustly identify the tumor clones, their specific gene expression profiles and map the spatial distribution of these clones within a tissue. Moreover, the inferred gene expression profile distribution of the clones enables us to perform differential expression analysis based only on a single biological sample from the tumor.

The core of the ClonalGE model is a probabilistic graphical model that extends a previous model, Tumoroscope, by incorporating additional variables for gene expression in the spots. The model assumes that each clone has a specific average gene expression profile and that the expression of genes observed at the spots is a mixture of the gene expressions in the clones present in those spots. Overall, this framework provides a comprehensive approach for inferring and analyzing gene expression profiles of cancer clones within a tissue.

The ClonalGE model was determined to be superior to the framework where the predecessor Tumoroscope is followed by linear regression (Tumoroscope + LR) in all simulated scenarios. ClonalGE was able to provide more precise estimates of gene expression levels in clones by incorporating information on mutation read counts. Additionally, by utilizing this extra data, the ClonalGE model was found to have a better performance in determining the fraction of clones in spots, the number of cells in spots, and read coverages per mutation per cell. The study also found that running multiple chains improves the robustness of the results by reducing the likelihood of the model getting stuck in local optima.

Previous studies did not decisively determine whether the genotypes of the clones determine their gene expression. On the one hand, since the population of tumor cells co-evolves in a relatively restricted area in the body, the subclones are not expected to show major differences in their expression profiles. In accordance, with decomposing gene expression pro-

grams purely from scRNA-seq data, a previous study by [248] found a little overlap of cell subpopulations expressing different functional programs with their genotypes. This approach, however, did not incorporate the knowledge of the clonal genotypes and mutations present in a single cell in the procedure for identifying the subpopulations.

On the other hand, it is naturally expected that the occurrence of mutations may lead to changes in gene expression, some of which may lead to phenotypic changes in the subpopulation of cells that carry those mutations. Using ClonalGE on prostate cancer data, the model was able to detect genes commonly found mutated in prostate cancer, and simultaneously show specific expression in certain clones. This shows that there exist some key genes with variable activation across the clones. Moreover, our study revealed a possible link between the location of the clones and their gene expression profiles. These results suggest that ClonalGE has the potential to be a valuable tool for further understanding the genetic characteristics of cancer and developing targeted treatments.

We hypothesize that not all, but a subset of genes may show clone-specific gene expression. In our analysis, we estimated gene expression profiles of 1000 genes that had the most variable expression profiles across ST spots, presuming that these have the most potential to be clone-specific. Including non-specific genes in the analysis would result in the incorporation of noise and could negatively affect the results. As an idea for future extension of the model, we could incorporate an additional hidden variable indicating whether the gene is clone specific or not. This information could then be inferred from the data.

Already in its current form, the model yields important insights into clonal phenotypes, marker genes, and their localization. Taken together, ClonalGE is a step forward in the clonal expression analysis across the tumor tissue.

Chapter 8

Conclusion

In this thesis, I proposed and implemented novel computational techniques for studying the presence of genetically distinct sub-populations of cells with different phenotypic behavior across a tumor, which defines the intra-tumor heterogeneity. For this, we integrated different datasets to observe different aspects of the tumor and increase the reliability of the outcomes of our analyses. Each of the three described projects addressed the research problems outlined in the Introduction in a specific manner.

In my first project, we proposed CACTUS, a Bayesian framework that integrates bulk DNA-seq, scRNA-seq, and BCR sequencing data to map the individual tumor cells to their clone of origin. For inferring the existing clones in the tumor, we utilized the aggregated reads over somatic mutations in the bulk DNA-seq. For deconvolving the reads coming from different clones, from the bulk data and for learning the clone genotypes, we used and existing Bayesian model, canopy [165]. Besides, knowing the existence of errors in the learned genotypes, we integrated error correction in CACTUS by explicit modeling of the errors. Moreover, for deconvolving the cells coming from different BCR clusters, we employed the categorical distribution in the CACTUS model. We also utilized the alternated and total reads over somatic mutations observed in the single cells to map them to the indicated clones. These read counts could be affected by the state of the cell, its gene expression and the clone it came from. We modeled the number of alternated reads relative to the total reads using the Binomial distribution. We calculated the probability of success in the Binomial distribution as the fraction of alternate to total reads, so that we can reduce the effects of bias and error.

In the end, due to the lack of ground truth, we employed external validation using gene expression data in the reduced dimension as an independent source of information. We verified our assignment of the labels (clones) to the cells by showing the same clustering in the independent feature space. Additionally, we used the probability of the assignment in the graphical model as the model's confidence and compared it to the confidence of cardelino in the same task.

The first project could be extended in a number of ways. Having the genotype of the cells and their gene expression could make it possible to look into the behavior of the cells inside and between the clones in the future work. This could enhance the cancer treatment and prevention of resistance by studying and predicting if a clone has the potential to resist.

Moreover, we used auxiliary clustering of cells to improve the assignment of the cells to the clones. This clustering was defined based on identical BCR sequences. Although we consider the sequence error and make it possible to move the cells between the clusters, we did not include the situations in which two different sequences can cluster together based on biological similarities. Embedding such a clustering inside the model could improve the clusters and the

cells' assignment to the clones. Moreover, the coverage and quality of scRNA-seq data are improving every year, and using a dataset with more coverage and fewer errors could increase the certainty of the results. Finally, using the scDNA-seq instead of scRNA-seq data could make it possible to infer the clones from the cells directly and increase the number of observed mutations instead of being limited only to the first part of the reads in scRNA-seq.

In the second project, we focused on the localization of the clones across the tumor and studying the difference of their phenotype. We integrated three different data sources: H&E images, WES, and reads over mutations in ST data. Although in this project we did not have the resolution of individual cells anymore, we gained valuable spatial perspective of the clones. Same as in the first project, we firstly used the aggregated reads in the bulk DNA-seq data and deconvolved them into the clones and their genotypes using canopy [165]. We accounted for potential errors in counting cells using $H\&E$ images in each ST spot and introduced a hidden random variable corresponding to the true number. Afterwards, for the decomposition of reads coming from different cells in a spot, we took advantage of the Dirichlet distribution and for modeling the allocation of the clones (features) to the spots (samples), we took advantage of the Beta-Bernoulli Process (BBP). Tumoroscope also used Binomial distribution to model the success parameter as the fraction of the alternative and total read counts, which is decreasing the effect of bias.

Due to the lack of ground truth, we carried out extensive simulation validation alongside external validation using the independent gene expression data. Our results suggest that the co-localization of the clones could have correlation with their phenotype. This is the point that three different heterogeneities (genotypic, spatial and phenotypic) cross each other and our studies gave insights their relation in the tumor tissue. This inferred information could be helpful for prediction of the possible location of metastasis or the possible resistant clone to the treatment.

Using ST data with an average of 50 cells per spot had limitations in terms of accuracy and precision. To address this issue, we believed that using higher resolution ST data would give us more accurate assignment of them to the clones and finally more accurate estimation of the clone-specific gene expression profile. Additionally, we used regression methods to estimate the gene expression profile of the clones which means assuming their distribution as Normal. However, this strong assumption could introduce extra error especially when calculating clone-specific gene expression profiles separately as post-processing. This could be improved by incorporating the inference of the clone-specific gene expression profiles inside the model, alongside with the inference of the fraction of the clones in the spots. We pursued this idea in my third project.

In the third project, we explored the phenotypic differences of the different clones alongside with their localization. In addition to the H&E images, WES, and ST reads, we included the observed gene expression across the ST spots, which is the normal output of the ST data. We identified the clone-specific gene expression across the tissue robustly and confidently, which pushed the limits of characterization of the clones in the tissue.

In this project, in addition to the problems that we faced in the second project, we had the issue of gene expression deconvolution to solve. We modeled the observed gene expression data as the mixture of Negative Binomial distribution. We validated the model with measuring the accuracy of the reconstruction of the observed data (aggregated gene expression in the spots) using the inferred hidden values (deconvolved clone-specific gene expression) in the probabilistic graphical model.

The effect of genotypes on gene expression in clones has not been clearly established in previous studies. However, some studies have shown that mutations can lead to changes in gene expression and phenotypic changes in cells carrying those mutations [249, 243]. In

my second and third projects, we found a potential link between clonal location and gene expression profiles. Our hypothesis was that only a subset of genes may exhibit clone-specific expression, and we estimated gene expression of 1000 genes with the most variable expression profiles to focus on those most likely to be clone-specific. As a future improvement, the model could incorporate information on whether a gene is clone-specific or not.

Finally, in all the projects, we used an existing Bayesian model for inferring the genotypes of the clones, which introduce extra uncertainty to the model. This can be improved by including the clonal inference in the model itself. Moreover, in all the projects, we used MH inside Gibbs sampling. This method could be replaced by the variational inference, which approximates a complex probability distribution by optimizing a simpler one. Variational inference tends to be faster and more scalable than MCMC methods, but it may introduce some bias and underestimate uncertainty.

In conclusion, while our models provide a highly promising approach to explore the genetic, phenotypic and spatial heterogenity, there are still areas for improvement and further research. Continued efforts in refining and improving the model can lead to more accurate identification of clonal structures, ultimately contributing to improved cancer treatment and prevention strategies. Overall, our work represents an important step forward in the field of cancer genomics and demonstrates the potential for machine learning models to contribute to advances in cancer research and treatment. As more and more data becomes available, we will continue to refine our models and improve our understanding of cancer genomics, ultimately paving the way for more personalized and effective cancer therapies.

Bibliography

[1] National Cancer Institute, Cancer Statistics. `https://www.cancer.gov/about-cancer/understanding/what-is-cancer`. Accessed October 2021.

[2] The genetics of cancer. `https://www.cancer.gov/about-cancer/causes-prevention/genetics`. Accessed August 2022.

[3] Cancer Research UK, Types of cancer. https://www.cancerresearchuk.org/what-is-cancer/how-cancer-starts/types-of-cancer. Accessed July 2020.

[4] Matthew W Fittall and Peter Van Loo. Translating insights into tumor evolution to clinical practice: promises and challenges. *Genome medicine*, 11(1):20, 2019.

[5] Eiji Furuta, Hiroshi Okuda, Aya Kobayashi, and Kounosuke Watabe. Metabolic genes in cancer: their roles in tumor progression and clinical implications. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1805(2):141–152, 2010.

[6] J Michael Bishop. The molecular genetics of cancer. *Science*, 235(4786):305–311, 1987.

[7] Matthew G Vander Heiden. Targeting cancer metabolism: a therapeutic window opens. *Nature reviews Drug discovery*, 10(9):671–684, 2011.

[8] David M Roy, Logan A Walsh, and Timothy A Chan. Driver mutations of cancer epigenomes. *Protein & cell*, 5(4):265–296, 2014.

[9] William Pao and Nicolas Girard. New driver mutations in non-small-cell lung cancer. *The lancet oncology*, 12(2):175–180, 2011.

[10] Matthew H Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C Wendl, Jaegil Kim, Brendan Reardon, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385, 2018.

[11] Ji-Hyun Lee, Xing-Ming Zhao, Ina Yoon, Jin Young Lee, Nam Hoon Kwon, Yin-Ying Wang, Kyung-Min Lee, Min-Joo Lee, Jisun Kim, Hyeong-Gon Moon, et al. Integrative analysis of mutational and transcriptional profiles reveals driver mutations of metastatic breast cancers. *Cell discovery*, 2(1):1–14, 2016.

[12] Lucy R Yates, Stian Knappskog, David Wedge, James HR Farmery, Santiago Gonzalez, Inigo Martincorena, Ludmil B Alexandrov, Peter Van Loo, Hans Kristian Haugland, Peer Kaare Lilleng, et al. Genomic evolution of breast cancer metastasis and relapse. *Cancer cell*, 32(2):169–184, 2017.

[13] Michael Fraser, Julie Livingstone, Jeffrey L Wrana, Antonio Finelli, Housheng Hansen He, Theodorus van der Kwast, Alexandre R Zlotta, Robert G Bristow, and Paul C Boutros. Somatic driver mutation prevalence in 1844 prostate cancers identifies znrf3 loss as a predictor of metastatic relapse. *Nature communications*, 12(1):6248, 2021.

[14] Phenotype. `https://www.genome.gov/genetics-glossary/Phenotype`. Accessed: 2023-03-19.

[15] Martin Ackermann. A functional perspective on phenotypic heterogeneity in microorganisms. *Nature Reviews Microbiology*, 13(8):497–508, 2015.

[16] Dana Pe'er, Seishi Ogawa, Ofer Elhanani, Leeat Keren, Trudy G Oliver, and David Wedge. Tumor heterogeneity. *Cancer cell*, 39(8):1015–1017, 2021.

[17] Yinyin Yuan. Spatial heterogeneity in the tumor microenvironment. *Cold Spring Harbor perspectives in medicine*, 6(8):a026583, 2016.

[18] Kimberly H Allison and George W Sledge. Heterogeneity and cancer. *Oncology*, 28(9):772–772, 2014.

[19] J Gray Camp, Randall Platt, and Barbara Treutlein. Mapping human cell phenotypes to genotypes with single-cell genomics. *Science*, 365(6460):1401–1405, 2019.

[20] Crispin T Hiley and Charles Swanton. Spatial and temporal cancer evolution: causes and consequences of tumour diversity. *Clinical medicine*, 14(Suppl 6):s33–s37, 2014.

[21] Bo Tang, Steven Kay, and Haibo He. Toward optimal feature selection in naive bayes for text categorization. *IEEE transactions on knowledge and data engineering*, 28(9):2508–2521, 2016.

[22] TAHSEEN AHMED JILANI and SYED ALI RAZA NAQVI. A review of probabilistic graph models for feature selection with applications in economic and financial time series forecasting. *VFAST Transactions on Software Engineering*, 3(1):20–27, 2014.

[23] C Shane Reese, Alyson G Wilson, Jiqiang Guo, Michael S Hamada, and Valen E Johnson. A bayesian model for integrating multiple sources of lifetime information in system-reliability assessments. *Journal of quality technology*, 43(2):127–141, 2011.

[24] Bo Zhao, Benjamin IP Rubinstein, Jim Gemmell, and Jiawei Han. A bayesian approach to discovering truth from conflicting sources for data integration. *arXiv preprint arXiv:1203.0058*, 2012.

[25] Tianzhou Ma, Faming Liang, Steffi Oesterreich, and George C Tseng. A joint bayesian model for integrating microarray and rna sequencing transcriptomic data. *Journal of Computational Biology*, 24(7):647–662, 2017.

[26] Peter ZG Qian and CF Jeff Wu. Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics*, 50(2):192–204, 2008.

[27] Umamahesh Srinivas, Yi Chen, Vishal Monga, Nasser M Nasrabadi, and Trac D Tran. Exploiting sparsity in hyperspectral image classification via graphical models. *IEEE Geoscience and Remote Sensing Letters*, 10(3):505–509, 2012.

[28] Charalampos Rotsos, Jurgen Van Gael, Andrew W Moore, and Zoubin Ghahramani. Probabilistic graphical models for semi-supervised traffic classification. In *Proceedings of the 6th International wireless communications and mobile computing conference*, pages 752–757, 2010.

[29] Boubaker Smii. Markov random fields model and applications to image processing. *AIMS Math*, 7:4459–4471, 2022.

[30] Alireza Farasat, Alexander Nikolaev, Sargur N Srihari, and Rachael Hageman Blair. Probabilistic graphical models in modern social network analysis. *Social Network Analysis and Mining*, 5:1–18, 2015.

[31] Pedro Larrañaga and Serafín Moral. Probabilistic graphical models in artificial intelligence. *Applied soft computing*, 11(2):1511–1528, 2011.

[32] Graphical model. `https://en.wikipedia.org/wiki/Graphical_model`. Accessed January 2023.

[33] Probabilistic graphical models. `https://www.adelaide.edu.au/aiml/our-research/machine-learning/probabilistic-graphical-models`. Accessed November 2019.

[34] Li Hongmei, Hao Wenning, Gan Wenyan, and Chen Gang. Survey of probabilistic graphical models. In *2013 10th Web Information System and Application Conference*, pages 275–280. IEEE, 2013.

[35] Franz Pernkopf, Robert Peharz, and Sebastian Tschiatschek. Introduction to probabilistic graphical models. In *Academic Press Library in Signal Processing*, volume 1, pages 989–1064. Elsevier, 2014.

[36] Concha Bielza and Pedro Larranaga. Discrete bayesian network classifiers: A survey. *ACM Computing Surveys (CSUR)*, 47(1):1–43, 2014.

[37] Hao Wang and Dit-Yan Yeung. A survey on bayesian deep learning. *ACM computing surveys (csur)*, 53(5):1–37, 2020.

[38] Robert Ekblom and Jochen BW Wolf. A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary applications*, 7(9):1026–1042, 2014.

[39] Shaked Slovin, Annamaria Carissimo, Francesco Panariello, Antonio Grimaldi, Valentina Bouché, Gennaro Gambardella, and Davide Cacchiarelli. Single-cell rna sequencing analysis: a step-by-step overview. *RNA Bioinformatics*, pages 343–365, 2021.

[40] Gur Yaari and Steven H Kleinstein. Practical guidelines for b-cell receptor repertoire sequencing analysis. *Genome medicine*, 7:1–14, 2015.

[41] Andrew H Fischer, Kenneth A Jacobson, Jack Rose, and Rolf Zeller. Hematoxylin and eosin staining of tissue and cell sections. *Cold spring harbor protocols*, 2008(5):pdb–prot4986, 2008.

[42] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[43] Shadi Darvish Shafighi, Szymon M Kiełbasa, Julieta Sepúlveda-Yáñez, Ramin Monajemi, Davy Cats, Hailiang Mei, Roberta Menafra, Susan Kloet, Hendrik Veelken, Cornelis AM van Bergen, et al. Cactus: integrating clonal architecture with genomic clustering and transcriptome profiling of single tumor cells. *Genome medicine*, 13(1):1–16, 2021.

[44] Emelie Berglund, Jonas Maaskola, Niklas Schultz, Stefanie Friedrich, Maja Marklund, Joseph Bergenstråhle, Firas Tarish, Anna Tanoglidi, Sanja Vickovic, Ludvig Larsson, et al. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nature communications*, 9(1):2419, 2018.

[45] Shadi Darvish Shafighi, Agnieszka Geras, Barbara Jurzysta, Alireza Sahaf Naeini, Igor Filipiuk, Lukasz RÄ… czkowski, Hosein Toosi, Lukasz Koperski, Kim Thrane, Camilla Engblom, et al. Tumoroscope: a probabilistic model for mapping cancer clones in tumor tissues. *bioRxiv*, pages 2022–09, 2022.

[46] Yi Li and Xiaohui Xie. A mixture model for expression deconvolution from rna-seq in heterogeneous tissues. In *BMC bioinformatics*, volume 14, pages 1–11. Springer, 2013.

[47] MA Ausdemore and C Neumann. Deconvolution of dust mixtures. *Forensic science international*, 308:110144, 2020.

[48] Hector Zenil, Narsis A Kiani, Allan A Zea, and Jesper Tegnér. Causal deconvolution by algorithmic generative models. *Nature Machine Intelligence*, 1(1):58–66, 2019.

[49] Oscar Hernan Madrid Padilla, Nicholas G Polson, and James G Scott. A deconvolution path for mixtures. *arXiv preprint arXiv:1511.06750*, 2015.

[50] Li Dong, Avinash Kollipara, Toni Darville, Fei Zou, and Xiaojing Zheng. Semi-cam: a semi-supervised deconvolution method for bulk transcriptomic data with partial marker gene information. *Scientific reports*, 10(1):1–12, 2020.

[51] Marc Elosua-Bayes, Paula Nieto, Elisabetta Mereu, Ivo Gut, and Holger Heyn. Spotlight: seeded nmf regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Research*, 49(9):e50–e50, 2021.

[52] Jason A O'Rawe, Scott Ferson, and Gholson J Lyon. Accounting for uncertainty in dna sequencing data. *Trends in Genetics*, 31(2):61–66, 2015.

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

[54] Yuxin Chen, Yongsheng Chen, Chunmei Shi, Zhibo Huang, Yong Zhang, Shengkang Li, Yan Li, Jia Ye, Chang Yu, Zhuo Li, et al. Soapnuke: a mapreduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience*, 7(1):gix120, 2018.

[55] Yan Guo, Fei Ye, Quanghu Sheng, Travis Clark, and David C Samuels. Three-stage quality control strategies for dna re-sequencing data. *Briefings in bioinformatics*, 15(6):879–889, 2014.

[56] Xiao Yang, Sriram P Chockalingam, and Srinivas Aluru. A survey of error-correction methods for next-generation sequencing. *Briefings in bioinformatics*, 14(1):56–66, 2013.

[57] Fernando Marmolejo-Ramos, Denis Cousineau, Luis Benites, and Rocío Maehara. On the efficacy of procedures to normalize ex-gaussian distributions. *Frontiers in psychology*, 5:1548, 2015.

[58] Sarang Vasantrao Khond. Effect of data normalization on accuracy and error of fault classification for an electrical distribution system. *Smart Science*, 8(3):117–124, 2020.

[59] Harish Bhaskar, David C Hoyle, and Sameer Singh. Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Computers in biology and medicine*, 36(10):1104–1125, 2006.

[60] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[61] Liming Wang and Xiaodong Wang. Hierarchical dirichlet process model for gene expression clustering. *EURASIP Journal on Bioinformatics and Systems Biology*, 2013:1–14, 2013.

[62] Zoubin Ghahramani and Thomas Griffiths. Infinite latent feature models and the indian buffet process. *Advances in neural information processing systems*, 18, 2005.

[63] Thomas L Griffiths and Zoubin Ghahramani. The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12(4), 2011.

[64] Anjin Guo, Yi Zhong, Wenyi Zhang, and Martin Haenggi. The gauss–poisson process for wireless networks and the benefits of cooperation. *IEEE Transactions on Communications*, 64(5):1916–1929, 2016.

[65] Hui Li, Xuejun Liao, and Lawrence Carin. Nonparametric bayesian feature selection for multi-task learning. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2236–2239. IEEE, 2011.

[66] A Kimball Romney, Margaret Kieffer, and Robert E Klein. A normalization procedure for correcting biased response data. *Social science research*, 2(4):307–320, 1973.

[67] Ross M Stolzenberg and Daniel A Relles. Tools for intuition about sample selection bias and its correction. *American sociological review*, pages 494–507, 1997.

[68] Ian R Dohoo. Bias—is it a problem, and what should we do? *Preventive Veterinary Medicine*, 113(3):331–337, 2014.

[69] Henry Han. Diagnostic biases in translational bioinformatics. *BMC Medical Genomics*, 8:1–17, 2015.

[70] Alaa Tharwat. Classification assessment methods. *Applied computing and informatics*, 17(1):168–192, 2021.

[71] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *2010 IEEE international conference on data mining*, pages 911–916. IEEE, 2010.

[72] David A Langan, James W Modestino, and Jun Zhang. Cluster validation for unsupervised stochastic model-based image segmentation. *IEEE Transactions on Image Processing*, 7(2):180–195, 1998.

[73] National Cancer Institute. Nci dictionary of cancer terms, 2017.

[74] Cell cycle. https://www.khanacademy.org/science/ap-biology/cell-communication-and-cell-cycle#cell-cycle.

[75] Cancer. https://en.wikipedia.org/wiki/Cancer. Accessed February 2023.

[76] Global cancer observatory. https://gco.iarc.fr/.

[77] Wikipedia contributors. Dna — Wikipedia, the free encyclopedia, 2023. [Online; accessed 5-January-2023].

[78] Human genetic variation. https://en.wikipedia.org/wiki/Human_genetic_variation. Accessed January 2023.

[79] Emmanouil P Pappou and Nita Ahuja. The role of oncogenes in gastrointestinal cancer. *Gastrointestinal Cancer Research: GCR*, page S2, 2010.

[80] Eva YHP Lee and William J Muller. Oncogenes and tumor suppressor genes. *Cold Spring Harbor perspectives in biology*, 2(10):a003236, 2010.

[81] Paul Yaswen, Karen L MacKenzie, W Nicol Keith, Patricia Hentosh, Francis Rodier, Jiyue Zhu, Gary L Firestone, Ander Matheu, Amancio Carnero, Alan Bilsland, et al. Therapeutic targeting of replicative immortality. In *Seminars in cancer biology*, volume 35, pages S104–S128. Elsevier, 2015.

[82] Yixin Yao and Wei Dai. Genomic instability and cancer. *Journal of carcinogenesis & mutagenesis*, 5, 2014.

[83] Rita Nahta, Fahd Al-Mulla, Rabeah Al-Temaimi, Amedeo Amedei, Rafaela Andrade-Vieira, Sarah N Bay, Dustin G Brown, Gloria M Calaf, Robert C Castellino, Karine A Cohen-Solal, et al. Mechanisms of environmental chemicals that enable the cancer hallmark of evasion of growth suppression. *Carcinogenesis*, 36(Suppl_1):S2–S18, 2015.

[84] Ramzi M Mohammad, Irfana Muqbil, Leroy Lowe, Clement Yedjou, Hsue-Yin Hsu, Liang-Tzung Lin, Markus David Siegelin, Carmela Fimognari, Nagi B Kumar, Q Ping Dou, et al. Broad targeting of resistance to apoptosis in cancer. In *Seminars in cancer biology*, volume 35, pages S78–S103. Elsevier, 2015.

[85] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[86] Tao Yu, Yanfen Wang, Yu Fan, Na Fang, Tongshan Wang, Tongpeng Xu, and Yongqian Shu. Circrnas in cancer metabolism: a review. *Journal of hematology & oncology*, 12:1–10, 2019.

[87] Rui Silva, Irene Gullo, and Fátima Carneiro. The pd-1: Pd-l1 immune inhibitory checkpoint in helicobacter pylori infection and gastric cancer: A comprehensive review and future perspectives. *Porto Biomedical Journal*, 1(1):4–11, 2016.

[88] Melvyn T Chow, Andreas Möller, and Mark J Smyth. Inflammation and immune surveillance in cancer. In *Seminars in cancer biology*, volume 22, pages 23–32. Elsevier, 2012.

[89] Florian R Greten and Sergei I Grivennikov. Inflammation and cancer: triggers, mechanisms, and consequences. *Immunity*, 51(1):27–41, 2019.

[90] Gemilang Khusnurrokhman and Farah Fatma Wati. Tumor-promoting inflammation in lung cancer: a literature review. *Annals of Medicine and Surgery*, page 104022, 2022.

[91] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

[92] Subramanian Venkatesan and Charles Swanton. Tumor evolutionary principles: how intratumor heterogeneity influences cancer treatment and outcome. *American Society of Clinical Oncology Educational Book*, 36:e141–e149, 2016.

[93] Yassen Assenov, David Brocks, and Clarissa Gerhäuser. Intratumor heterogeneity in epigenetic patterns. In *Seminars in cancer biology*, volume 51, pages 12–21. Elsevier, 2018.

[94] Xinyi Cindy Zhang, Chang Xu, Ryan M Mitchell, Bo Zhang, Derek Zhao, Yao Li, Xin Huang, Wenhong Fan, Hongwei Wang, Luisa Angelica Lerma, et al. Tumor evolution and intratumor heterogeneity of an oropharyngeal squamous cell carcinoma revealed by whole-genome sequencing. *Neoplasia*, 15(12):1371–IN7, 2013.

[95] Anne C Rios. Resolving the spatial heterogeneity of cancer in 3d. *Nature Reviews Cancer*, 22(10):548–549, 2022.

[96] Ibiayi Dagogo-Jack and Alice T Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature reviews clinical oncology*, 15(2):81–94, 2018.

[97] Katsuyoshi Takata, Tomoko Miyata-Takata, Yasuharu Sato, and Tadashi Yoshino. Pathology of follicular lymphoma. *Journal of Clinical and Experimental Hematopathology*, 54(1):3–9, 2014.

[98] Yuri Tolkach and Glen Kristiansen. The heterogeneity of prostate cancer: a practical approach. *Pathobiology*, 85(1-2):108–116, 2018.

[99] Gulisa Turashvili and Edi Brogi. Tumor heterogeneity in breast cancer. *Frontiers in medicine*, 4:227, 2017.

[100] Patricia Banks, Wen Xu, Declan Murphy, Paul James, and Shahneen Sandhu. Relevance of dna damage repair in the management of prostate cancer. *Current problems in cancer*, 41(4):287–301, 2017.

[101] Sander Frank, Peter Nelson, and Valeri Vasioukhin. Recent advances in prostate cancer research: Large-scale genomic analyses reveal novel driver mutations and dna repair defects. *F1000Research*, 7, 2018.

[102] Joaquin Mateo, Gunther Boysen, Christopher E Barbieri, Helen E Bryant, Elena Castro, Pete S Nelson, David Olmos, Colin C Pritchard, Mark A Rubin, and Johann S de Bono. Dna repair in prostate cancer: biology and clinical implications. *European urology*, 71(3):417–425, 2017.

[103] Sidrah Shah, Rachelle Rachmat, Synthia Enyioma, Aruni Ghose, Antonios Revythis, and Stergios Boussios. Brca mutations in prostate cancer: assessment, implications and treatment considerations. *International Journal of Molecular Sciences*, 22(23):12628, 2021.

[104] Fernando Santos de Azevedo, Lanúscia Morais de Santana Sá, Uirá Maíra de Resende, Augusto Ribeiro Gabriel, and Elisângela de Paula Silveira Lacerda. Prostate cancer and dna genes repair: What should an oncologist know?–a narrative review. *Open Access Indonesian Journal of Medical Reviews*, 3(1):331–341, 2023.

[105] DW Smithers. Family histories of 459 patients with cancer of the breast. *British Journal of Cancer*, 2(2):163, 1948.

[106] Barbara S Hulka and Azadeh T Stark. Breast cancer: cause and prevention. *The Lancet*, 346(8979):883–887, 1995.

[107] Yoshio Miki, Jeff Swensen, Donna Shattuck-Eidens, P Andrew Futreal, Keith Harshman, Sean Tavtigian, Qingyun Liu, Charles Cochran, L Michelle Bennett, Wei Ding, et al. A strong candidate for the breast and ovarian cancer susceptibility gene brca1. *Science*, 266(5182):66–71, 1994.

[108] Hanne Meijers-Heijboer, Ans Van den Ouweland, Jan Klijn, Marijke Wasielewski, Anja de Snoo, Rogier Oldenburg, Antoinette Hollestelle, Mark Houben, Ellen Crepin, Monique van Veghel-Plandsoen, et al. Low-penetrance susceptibility to breast cancer due to chek2* 1100delc in noncarriers of brca1 or brca2 mutations. *Nature genetics*, 31(1), 2002.

[109] Anglian Breast Cancer Study Group. Prevalence and penetrance of brca1 and brca2 mutations in a population-based series of breast cancer cases. *British Journal of Cancer*, 83(10):1301, 2000.

[110] Philippe Bertheau, Marc Espié, Elisabeth Turpin, Jacqueline Lehmann, Louis-Francois Plassa, Mariana Varna, Anne Janin, et al. Tp53 status and response to chemotherapy in breast cancer. *Pathobiology*, 75(2):132–139, 2008.

[111] Elizabeth A Mittendorf, Anne V Philips, Funda Meric-Bernstam, Na Qiao, Yun Wu, Susan Harrington, Xiaoping Su, Ying Wang, Ana M Gonzalez-Angulo, Argun Akcakanat, et al. Pd-l1 expression in triple-negative breast cancer. *Cancer immunology research*, 2(4):361–370, 2014.

[112] Ian G Campbell, Sarah E Russell, David YH Choong, Karen G Montgomery, Marianne L Ciavarella, Christine SF Hooi, Briony E Cristiano, Richard B Pearson, and Wayne A Phillips. Mutation of the pik3ca gene in ovarian and breast cancer. *Cancer research*, 64(21):7678–7681, 2004.

[113] Kurtis E Bachman, Pedram Argani, Yardena Samuels, Natalie Silliman, Janine Ptak, Steve Szabo, Hiroyuki Konishi, Bedri Karakas, Brian G Blair, Clarence Lin, et al. The pik3ca gene is mutated with high frequency in human breast cancers. *Cancer biology & therapy*, 3(8):772–775, 2004.

[114] Emad A Rakha, Gary M Tse, and Cecily M Quinn. An update on the pathological classification of breast cancer. *Histopathology*, 82(1):5–16, 2023.

[115] Antonino Carbone, Sandrine Roulland, Annunziata Gloghini, Anas Younes, Gottfried von Keudell, Armando López-Guillermo, and Jude Fitzgibbon. Follicular lymphoma. *Nature Reviews Disease Primers*, 5(1):83, 2019.

[116] Exome sequencing. `https://www.illumina.com/techniques/sequencing/dna-sequencing/targeted-resequencing/exome-sequencing.html`.

[117] Bahareh Rabbani, Mustafa Tekin, and Nejat Mahdieh. The promise of whole-exome sequencing in medical genetics. *Journal of human genetics*, 59(1):5–15, 2014.

[118] Natalie I Vokes and Jianjun Zhang. The role of whole exome sequencing in distinguishing primary and secondary lung cancers. *Lung Cancer: Targets and Therapy*, pages 139–149, 2021.

[119] Yongmei Zhao, Li Tai Fang, Tsai-wei Shen, Sulbha Choudhari, Keyur Talsania, Xiongfong Chen, Jyoti Shetty, Yuliya Kriga, Bao Tran, Bin Zhu, et al. Whole genome and exome sequencing reference datasets from a multi-center and cross-platform benchmark study. *Scientific data*, 8(1):296, 2021.

[120] Ludvig Larsson, Jonas Frisén, and Joakim Lundeberg. Spatially resolved transcriptomics adds a new dimension to genomics. *Nature methods*, 18(1):15–18, 2021.

[121] Ada T Feldman and Delia Wolfe. Tissue processing and hematoxylin and eosin staining. *Histopathology: Methods and Protocols*, pages 31–43, 2014.

[122] Alex Skovsbo Jørgensen, Anders Munk Rasmussen, Niels Kristian Mäkinen Andersen, Simon Kragh Andersen, Jonas Emborg, Rasmus Røge, and Lasse Riis Østergaard. Using cell nuclei features to detect colon cancer tissue in hematoxylin and eosin stained slides. *Cytometry Part A*, 91(8):785–793, 2017.

[123] Alper Aksac, Douglas J Demetrick, Tansel Ozyer, and Reda Alhajj. Brecahad: a dataset for breast cancer histopathological annotation and diagnosis. *BMC research notes*, 12(1):1–3, 2019.

[124] Daniel Hebenstreit. Methods, challenges and potentials of single cell rna-seq. *Biology*, 1(3):658–667, 2012.

[125] Single cell sequencing. `https://en.wikipedia.org/wiki/Single_cell_sequencing`. Accessed February 2023.

[126] Thale Kristin Olsen and Ninib Baryawno. Introduction to single-cell rna sequencing. *Current protocols in molecular biology*, 122(1):e57, 2018.

[127] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[128] Daniel C Koboldt, Ken Chen, Todd Wylie, David E Larson, Michael D McLellan, Elaine R Mardis, George M Weinstock, Richard K Wilson, and Li Ding. Varscan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–2285, 2009.

[129] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576, 2012.

[130] Zhongwu Lai, Aleksandra Markovets, Miika Ahdesmaki, Brad Chapman, Oliver Hofmann, Robert McEwen, Justin Johnson, Brian Dougherty, J Carl Barrett, and

Jonathan R Dry. Vardict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic acids research*, 44(11):e108–e108, 2016.

[131] Sangtae Kim, Konrad Scheffler, Aaron L Halpern, Mitchell A Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, Yeonbin Kim, Doruk Beyter, Peter Krusche, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nature methods*, 15(8):591–594, 2018.

[132] Vladislav Lysenkov. Introducing deep learning-based methods into the variant calling analysis pipeline. *Science*, 6:7789, 2019.

[133] Barry S Taylor, Jordi Barretina, Nicholas D Socci, Penelope DeCarolis, Marc Ladanyi, Matthew Meyerson, Samuel Singer, and Chris Sander. Functional copy-number alterations in cancer. *PloS one*, 3(9):e3179, 2008.

[134] Ruby YunJu Huang, Geng Bo Chen, Noriomi Matsumura, Hung-Cheng Lai, Seiichi Mori, Jingjing Li, Meng Kang Wong, Ikuo Konishi, Jean-Paul Thiery, and Liang Goh. Histotype-specific copy-number alterations in ovarian cancer. *BMC medical genomics*, 5:1–13, 2012.

[135] Yi Zhang, John WM Martens, Jack X Yu, John Jiang, Anieta M Sieuwerts, Marcel Smid, Jan GM Klijn, Yixin Wang, and John A Foekens. Copy number alterations that predict metastatic capability of human breast cancer. *Cancer research*, 69(9):3795–3801, 2009.

[136] Adam Shlien and David Malkin. Copy number variations and cancer. *Genome medicine*, 1(6):1–9, 2009.

[137] Mehdi Zarrei, Jeffrey R MacDonald, Daniele Merico, and Stephen W Scherer. A copy number variation map of the human genome. *Nature reviews genetics*, 16(3):172–183, 2015.

[138] Jennifer L Freeman, George H Perry, Lars Feuk, Richard Redon, Steven A McCarroll, David M Altshuler, Hiroyuki Aburatani, Keith W Jones, Chris Tyler-Smith, Matthew E Hurles, et al. Copy number variation: new insights in genome diversity. *Genome research*, 16(8):949–961, 2006.

[139] Luísa Esteves, Francisco Caramelo, Ilda Patrícia Ribeiro, Isabel M Carreira, and Joana Barbosa de Melo. Probability distribution of copy number alterations along the genome: an algorithm to distinguish different tumour profiles. *Scientific Reports*, 10(1):14868, 2020.

[140] Xian F Mallory, Mohammadamin Edrisi, Nicholas Navin, and Luay Nakhleh. Methods for copy number aberration detection from single-cell dna-sequencing data. *Genome biology*, 21(1):1–22, 2020.

[141] Kaushalya C Amarasinghe, Jason Li, Sally M Hunter, Georgina L Ryland, Prue A Cowin, Ian G Campbell, and Saman K Halgamuge. Inferring copy number and genotype in tumour exome data. *BMC genomics*, 15(1):1–12, 2014.

[142] Hao Chen, Yuchao Jiang, Kara N Maxwell, Katherine L Nathanson, and Nancy Zhang. Allele-specific copy number estimation by whole exome sequencing. *The annals of applied statistics*, 11(2):1169, 2017.

[143] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[144] Markov blanket. `https://en.wikipedia.org/wiki/Markov_blanket`. Accessed April 2021.

[145] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

[146] Posterior probability. `https://en.wikipedia.org/wiki/Posterior_probability`. Accessed March 2023.

[147] Nathalie Peyrard, M-J Cros, Simon de Givry, Alain Franc, Stephane Robin, Regis Sabbadin, Thomas Schiex, and Matthieu Vignes. Exact or approximate inference in graphical models: why the choice is dictated by the treewidth, and how variable elimination can be exploited. *Australian & New Zealand Journal of Statistics*, 61(2):89–133, 2019.

[148] Paul A Gagniuc. *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons, 2017.

[149] Markov chain. `https://en.wikipedia.org/wiki/Markov_chain`. Accessed February 2023.

[150] Christian Robert, George Casella, Christian P Robert, and George Casella. Metropolis–hastings algorithms. *Introducing Monte Carlo Methods with R*, pages 167–197, 2010.

[151] Ferenc Kovács, Csaba Legány, and Attila Babos. Cluster validity measurement techniques. In *6th International symposium of hungarian researchers on computational intelligence*, volume 35. Citeseer, 2005.

[152] Dunnindex. `https://en.wikipedia.org/wiki/Dunn_index`. Accessed November 2022.

[153] Sriparna Saha and Sanghamitra Bandyopadhyay. A validity index based on connectivity. In *2009 Seventh International Conference on Advances in Pattern Recognition*, pages 91–94. IEEE, 2009.

[154] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Clustering algorithms and validity measures. In *Proceedings Thirteenth International Conference on Scientific and Statistical Database Management. SSDBM 2001*, pages 3–22. IEEE, 2001.

[155] Eréndira Rendón, Itzel Abundez, Alejandra Arizmendi, and Elvia M Quiroz. Internal versus external cluster validation indexes. *International Journal of computers and communications*, 5(1):27–34, 2011.

[156] Wikipedia contributors. Gini coefficient. [Online; accessed April 2023].

[157] Wikipedia contributors. Mean absolute error. [Online; accessed March 2023].

[158] Song Yi, Shengda Lin, Yongsheng Li, Wei Zhao, Gordon B Mills, and Nidhi Sahni. Functional variomics and network perturbation: connecting genotype to phenotype in cancer. *Nature Reviews Genetics*, 18(7):395, 2017.

[159] Samra Turajlic, Andrea Sottoriva, Trevor Graham, and Charles Swanton. Resolving genetic heterogeneity in cancer. *Nature Reviews Genetics*, 20(7):404–416, 2019.

[160] Robert Kridel, Laurie H Sehn, and Randy D Gascoyne. Pathogenesis of follicular lymphoma. *The Journal of clinical investigation*, 122(10):3424–3431, 2012.

[161] Laura Pasqualucci. Molecular pathogenesis of germinal center-derived b cell lymphomas. *Immunological reviews*, 288(1):240–261, 2019.

[162] Florian Scherer, Marcelo A Navarrete, Cristina Bertinetti-Lapatki, Joachim Boehm, Annette Schmitt-Graeff, and Hendrik Veelken. Isotype-switched follicular lymphoma displays dissociation between activation-induced cytidine deaminase expression and somatic hypermutation. *Leukemia & lymphoma*, 57(1):151–160, 2016.

[163] Florian Scherer, Marlon van der Burgt, Szymon M Kiełbasa, Cristina Bertinetti-Lapatki, von Minden M Dühren, Kristina Mikesch, Katja Zirlik, Liesbeth de Wreede, Hendrik Veelken, and Marcelo A Navarrete. Selection patterns of b-cell receptors and the natural history of follicular lymphoma. *British journal of haematology*, 175(5):972, 2016.

[164] Dunja Schneider, Marcus Dühren-von Minden, Alabbas Alkhatib, Corinna Setz, Cornelis AM van Bergen, Marco Benkißer-Petersen, Isabel Wilhelm, Sarah Villringer, Sergey Krysov, Graham Packham, et al. Lectins from opportunistic bacteria interact with acquired variable-region glycans of surface immunoglobulin in follicular lymphoma. *Blood, The Journal of the American Society of Hematology*, 125(21):3287–3296, 2015.

[165] 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. *Proceedings of the National Academy of Sciences*, 113(37):E5528–E5537, 2016.

[166] 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. *Genome Biology*, 16(1):1–20, 2015.

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

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

[169] Edith M Ross and Florian Markowetz. Onconem: inferring tumor evolution from single-cell sequencing data. *Genome biology*, 17(1):1–14, 2016.

[170] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome biology*, 17(1):86, 2016.

[171] Charles Gawad, Winston Koh, and Stephen R Quake. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proceedings of the National Academy of Sciences*, 111(50):17947–17952, 2014.

[172] Jack Kuipers, Katharina Jahn, and Niko Beerenwinkel. Advances in understanding tumour evolution through single-cell sequencing. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1867(2):127–138, 2017.

[173] Sören Müller, Siyuan John Liu, Elizabeth Di Lullo, Martina Malatesta, Alex A Pollen, Tomasz J Nowakowski, Gary Kohanbash, Manish Aghi, Arnold R Kriegstein, Daniel A Lim, et al. Single-cell sequencing maps gene expression to mutational phylogenies in pdgf-and egf-driven gliomas. *Molecular systems biology*, 12(11), 2016.

[174] Itay Tirosh, Andrew S Venteicher, Christine Hebert, Leah E Escalante, Anoop P Patel, Keren Yizhak, Jonathan M Fisher, Christopher Rodman, Christopher Mount, Mariella G Filbin, et al. Single-cell rna-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*, 539(7628):309–313, 2016.

[175] Jean Fan, Hae-Ock Lee, Soohyun Lee, Da-eun Ryu, Semin Lee, Catherine Xue, Seok Jin Kim, Kihyun Kim, Nikolaos Barkas, Peter J Park, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell rna-seq data. *Genome research*, 28(8):1217–1227, 2018.

[176] Olivier Poirion, Xun Zhu, Travers Ching, and Lana X Garmire. Using single nucleotide variations in single-cell rna-seq to identify subpopulations and genotype-phenotype linkage. *Nature communications*, 9(1):1–13, 2018.

[177] Davis J McCarthy, Raghd Rostom, Yuanhua Huang, Daniel J Kunz, Petr Danecek, Marc Jan Bonder, Tzachi Hagai, Ruqian Lyu, Wenyi Wang, Daniel J Gaffney, et al. Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes. *Nature Methods*, 17(4):414–421, 2020.

[178] Michael A Ortega, Olivier Poirion, Xun Zhu, Sijia Huang, Thomas K Wolfgruber, Robert Sebra, and Lana X Garmire. Using single-cell multiple omics approaches to resolve tumor heterogeneity. *Clinical and translational medicine*, 6(1):46, 2017.

[179] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

[180] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

[181] Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

[182] Marie-Paule Lefranc, Veronique Giudicelli, Chantal Ginestoux, Joumana Jabado-Michaloud, Geraldine Folch, Fatena Bellahcene, Yan Wu, Elodie Gemrot, Xavier Brochet, Jero^me Lane, et al. Imgt®, the international immunogenetics information system®. *Nucleic acids research*, 37(suppl_1):D1006–D1012, 2009.

[183] Yuanhua Huang, Davis J McCarthy, and Oliver Stegle. Vireo: Bayesian demultiplexing of pooled single-cell rna-seq data without genotype reference. *Genome Biology*, 20(1):273, 2019.

[184] Evelyn C Pielou. The measurement of diversity in different types of biological collections. *Journal of theoretical biology*, 13:131–144, 1966.

[185] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[186] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[187] Julia Handl, Joshua Knowles, and Douglas B Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005.

[188] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[189] Joseph C Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104, 1974.

[190] Julia Handl and Joshua Knowles. Exploiting the trade-off—the benefits of multiple objectives in data clustering. In *International Conference on Evolutionary Multi-Criterion Optimization*, pages 547–560. Springer, 2005.

[191] Marwan Hassani and Thomas Seidl. Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. *Vietnam Journal of Computer Science*, 4(3):171–183, 2017.

[192] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[193] Anthony B Atkinson and François Bourguignon. *Handbook of income distribution*, volume 2. Elsevier, 2014.

[194] Claude E Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948.

[195] Neil Vasan, José Baselga, and David M Hyman. A view on drug resistance in cancer. *Nature*, 575(7782):299–309, 2019.

[196] Xiao-xiao Sun and Qiang Yu. Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta pharmacologica sinica*, 36(10):1219–1227, 2015.

[197] Nicholas McGranahan and Charles Swanton. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, 168(4):613–628, 2017.

[198] Zhenhua Yu, Fang Du, and Lijuan Song. Scclone: Accurate clustering of tumor single-cell dna sequencing data. *Frontiers in genetics*, page 26, 2022.

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

[200] Niko Beerenwinkel, Chris D. Greenman, and Jens Lagergren. Computational Cancer Biology: An Evolutionary Perspective. *PLOS computational biology*, 12(2):e1004717, February 2016. Number: 2.

[201] F. Vandin. Computational Methods for Characterizing Cancer Mutational Heterogeneity. *Frontiers in Genetics*, 8:83, 2017.

[202] A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, and J. Biele, *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398, April 2014. Number: 4.

[203] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology*, 16(1):35, 2015. Number: 1.

[204] Jack Kuipers, Katharina Jahn, and Niko Beerenwinkel. Advances in understanding tumour evolution through single-cell sequencing. *Biochimica et biophysica acta*, 1867(2):127–138, April 2017. Number: 2.

[205] Salem Malikic, Katharina Jahn, Jack Kuipers, S. Cenk Sahinalp, and Niko Beerenwinkel. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature communications*, 10(1):2750, December 2019. Number: 1.

[206] Maurizio Pellegrino, Adam Sciambi, Sebastian Treusch, Robert Durruthy-Durruthy, Kaustubh Gokhale, Jose Jacob, Tina X. Chen, Jennifer A. Geis, William Oldham, Jairo Matthews, Hagop Kantarjian, P. Andrew Futreal, Keyur Patel, Keith W. Jones, Koichi Takahashi, and Dennis J. Eastburn. High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. *Genome research*, 28(9):1345–1352, September 2018. Number: 9.

[207] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J. McCarthy, Stephanie C. Hicks, Mark D. Robinson, Catalina A. Vallejos, Kieran R. Campbell, Niko Beerenwinkel, Ahmed Mahfouz, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto Attolini, Samuel Aparicio, Jasmijn Baaijens, Marleen Balvert, Buys de Barbanson, Antonio Cappuccio, Giacomo Corleone, Bas E. Dutilh, Maria Florescu, Victor Guryev, Rens Holmer, Katharina Jahn, Thamar Jessurun Lobo, Emma M. Keizer, Indu Khatri, Szymon M. Kielbasa, Jan O. Korbel, Alexey M. Kozlov, Tzu-Hao Kuo, Boudewijn P. F. Lelieveldt, Ion I. Mandoiu, John C. Marioni, Tobias Marschall, Felix Mölder, Amir Niknejad, Lukasz Raczkowski, Marcel Reinders, Jeroen de Ridder, Antoine-Emmanuel Saliba, Antonios Somarakis, Oliver Stegle, Fabian J. Theis, Huan Yang, Alex Zelikovsky, Alice C. McHardy, Benjamin J. Raphael, Sohrab P. Shah, and Alexander Schönhuth. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):31, 2020.

[208] Charles Swanton. Intratumor Heterogeneity: Evolution through Space and Time. *Cancer research*, 72(19):4875–4882, October 2012. Number: 19.

[209] Nicholas McGranahan and Charles Swanton. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*, 168(4):613–628, February 2017. Number: 4.

[210] Marco Gerlinger, Andrew J. Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, Ignacio Varela, Benjamin Phillimore, Sharmin Begum, Neil Q. McDonald, Adam Butler, David Jones, Keiran Raine, Calli Latimer, Claudio R. Santos, Mahrokh Nohadani, Aron C. Eklund, Bradley Spencer-Dene, Graham Clark, Lisa Pickering, Gordon Stamp, Martin Gore, Zoltan Szallasi, Julian Downward, P. Andrew Futreal, and Charles Swanton. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England journal of medicine*, 366(10):883–892, March 2012. Number: 10.

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

[212] 10x genomics website. `https://kb.10xgenomics.com/`.

[213] Rebecca Elyanow, Ron Zeira, Max Land, and Benjamin J Raphael. STARCH: Copy number and clone inference from spatial transcriptomics data. *Physical Biology*, 18(3):035001, 2021.

[214] Andrew Erickson, Mengxiao He, Emelie Berglund, Maja Marklund, Reza Mirzazadeh, Niklas Schultz, Linda Kvastad, Alma Andersson, Ludvig Bergenstråhle, Joseph Bergenstråhle, et al. Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature*, 608(7922):360, 2022.

[215] Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al. QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1):1–7, 2017.

[216] Fredrik Pontén, Karin Jirström, and Matthias Uhlen. The human protein atlas—a tool for pathology. *The journal of pathology: A journal of the pathological society of Great Britain and Ireland*, 216(4):387–393, 2008.

[217] Giovanna C Cavalcante, Ândrea Ribeiro-dos Santos, and Gilderlanio S de Araújo. Mitochondria in tumour progression: a network of mtdna variants in different types of cancer. *BMC genomic data*, 23(1):1–10, 2022.

[218] Rajnish Kumar Singh, Sunil Kumar Saini, Gopinath Prakasam, Ponnuusamy Kalairasan, and Rameshwar NK Bamezai. Role of ectopically expressed mtdna encoded cytochrome c oxidase subunit i (mt-coi) in tumorigenesis. *Mitochondrion*, 49:56–65, 2019.

[219] Richard Y Ebright, Sooncheol Lee, Ben S Wittner, Kira L Niederhoffer, Benjamin T Nicholson, Aditya Bardia, Samuel Truesdell, Devon F Wiley, Benjamin Wesley, Selena Li, et al. Deregulation of ribosomal protein expression and translation promotes breast cancer metastasis. *Science*, 367(6485):1468–1473, 2020.

[220] Patrick Ruch, Douglas Teodoro, UniProt Consortium, et al. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 2021.

[221] Tongtong Zhao, Zachary D Chiang, Julia W Morriss, Lindsay M LaFave, Evan M Murray, Isabella Del Priore, Kevin Meli, Caleb A Lareau, Naeem M Nadaf, Jilong Li, et al. Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature*, 601(7891):85–91, 2022.

[222] Senbai Kang, Nico Borgsmüller, Monica Valecha, Jack Kuipers, Joao Alves, Sonia Prado-López, Débora Chantada, Niko Beerenwinkel, David Posada, and Ewa Szczurek. Sieve: joint inference of single-nucleotide variants and cell phylogeny from single-cell dna sequencing data. *BioRxiv*, 2022.

[223] Alexey Kozlov, Joao M Alves, Alexandros Stamatakis, and David Posada. Cellphy: accurate and fast probabilistic inference of single-cell phylogenies from scdna-seq data. *Genome biology*, 23(1):1–30, 2022.

[224] Kevin Lebrigand, Joseph Bergenstråhle, Kim Thrane, Annelie Mollbrink, Konstantinos Meletis, Pascal Barbry, Rainer Waldmann, and Joakim Lundeberg. The spatial landscape of gene expression isoforms in tissue sections. *BioRxiv*, pages 2020–08, 2022.

[225] Agnieszka Geras, Shadi Darvish Shafighi, Kacper Domżał, Igor Filipiuk, Łukasz Rączkowski, Hosein Toosi, Leszek Kaczmarek, Łukasz Koperski, Jens Lagergren, Dominka Nowis, et al. Celloscope: a probabilistic model for marker-gene-driven cell type deconvolution in spatial transcriptomics data. *BioRxiv*, 2022.

[226] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R Vanderburg, Åsa Segerstolpe, Meng Zhang, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nature methods*, 18(11):1352–1362, 2021.

[227] Bin Li, Wen Zhang, Chuang Guo, Hao Xu, Longfei Li, Minghao Fang, Yinlei Hu, Xinye Zhang, Xinfeng Yao, Meifang Tang, et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature methods*, pages 1–9, 2022.

[228] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

[229] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[230] José Fernández Navarro, Joel Sjöstrand, Fredrik Salmén, Joakim Lundeberg, and Patrik L Ståhl. St pipeline: an automated pipeline for spatial mapping of unique transcripts. *Bioinformatics*, 2017.

[231] Lancelot F James. Bayesian poisson calculus for latent feature modeling via generalized indian buffet process priors. *The annals of statistics*, 45(5):2016–2045, 2017.

[232] J Bernardo, M Bayarri, J Berger, A Dawid, D Heckerman, A Smith, and M West. Bayesian nonparametric latent feature models. *Bayesian statistics*, 8:1–25, 2007.

[233] Zoubin Ghahramani and Thomas L Griffiths. Infinite latent feature models and the indian buffet process. In *Advances in neural information processing systems*, pages 475–482, 2006.

[234] Wikipedia contributors. Dirichlet distribution — Wikipedia, the free encyclopedia, 2019. [Online; accessed 18-October-2019].

[235] Christopher M Bishop. Model-based machine learning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20120222, 2013.

[236] Wikipedia contributors. Gamma distribution — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Gamma_distribution&oldid=1067698046`, 2022. [Online; accessed 2-February-2022].

[237] Mel Greaves and Carlo C Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 2012.

[238] Corbin E Meacham and Sean J Morrison. Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328–337, 2013.

[239] Elza C De Bruin, Nicholas McGranahan, Richard Mitter, Max Salm, David C Wedge, Lucy Yates, Mariam Jamal-Hanjani, Seema Shafi, Nirupa Murugaesu, Andrew J Rowan, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*, 346(6206):251–256, 2014.

[240] Magda Markowska, Tomasz Cąkała, Błażej Miasojedow, Bogac Aybey, Dilafruz Juraeva, Johanna Mazur, Edith Ross, Eike Staub, and Ewa Szczurek. Conet: Copy number event tree model of evolutionary tumor history for single-cell data. *Genome Biology*, 23(1):1–35, 2022.

[241] Kyu-Tae Kim, Hye Won Lee, Hae-Ock Lee, Sang Cheol Kim, Yun Jee Seo, Woosung Chung, Hye Hyeon Eum, Do-Hyun Nam, Junhyong Kim, Kyeung Min Joo, et al. Single-cell mrna sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biology*, 16(1):1–15, 2015.

[242] Sören Müller, Ara Cho, Siyuan J Liu, Daniel A Lim, and Aaron Diaz. Conics integrates scrna-seq with dna sequencing to map gene expression to tumor sub-clones. *Bioinformatics*, 34(18):3217, 2018.

[243] Kieran R Campbell, Adi Steif, Emma Laks, Hans Zahn, Daniel Lai, Andrew McPherson, Hossein Farahani, Farhia Kabeer, Ciara O'Flanagan, Justina Biele, et al. clonealign: statistical integration of independent single-cell rna and dna sequencing data from human cancers. *Genome biology*, 20(1):1–12, 2019.

[244] Chi-Yun Wu, Anuja Sathe, Jiazhen Rong, Paul R Hess, Billy Lau, Susan M Grimes, Hanlee P Ji, and Nancy R Zhang. Cancer subclone detection based on dna copy number in single cell and spatial omic sequencing data. *bioRxiv*, 2022.

[245] Shadi Darvish Shafighi, Agnieszka Geras, Barbara Jurzysta, Alireza Sahaf Naeini, Igor Filipiuk, Łukasz Rączkowski, et al. Tumoroscope: a probabilistic model for mapping tumor clones in cancerous tissues. *soon to be published*, 2022.

[246] John F Geweke et al. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Technical report, Federal Reserve Bank of Minneapolis, 1991.

[247] William Jay Conover. *Practical Nonparametric Statistics*, volume 350. john wiley & sons, 1999.

[248] Gabriela S Kinker, Alissa C Greenwald, Rotem Tal, Zhanna Orlova, Michael S Cuoco, James M McFarland, Allison Warren, Christopher Rodman, Jennifer A Roth, Samantha A Bender, et al. Pan-cancer single-cell rna-seq identifies recurring programs of cellular heterogeneity. *Nature genetics*, 52(11):1208–1218, 2020.

[249] Artem Lomakin, Jessica Svedlund, Carina Strell, Milana Gataric, Artem Shmatko, Gleb Rukhovich, Jun Sung Park, Young Seok Ju, Stefan Dentro, Vitalii Kleshchevnikov, et al. Spatial genomics maps the structure, nature and evolution of cancer clones. *Nature*, pages 1–9, 2022.