Silesian University of Technology
Faculty of Automatic Control, Electronics and Computer Science

Models of cancer genome evolution used to evaluate the role of selection and occurrence of new mutations

Doctoral thesis

Paweł Kuś

Supervisor:
Prof. dr hab. inż. Marek Kimmel
Co-supervisor:
Dr inż. Roman Jaksik

Gliwice 2023

Acknowledgements

I would like to thank my supervisor, Prof. Marek Kimmel, for all his guidance in my scientific career and research. I would like to thank my co-supervisor, Dr. Roman Jaksik, who invited me into the world of science and whose advice was particularly insightful and constructive. I would also like to thank Prof. Joanna Polańska for her support during my Ph.D. studies.

Many thanks to all my colleagues, collaborators, and co-authors from Silesian University of Technology, National Oncology Institute of Maria Sklodowska-Curie - National Research Institute in Gliwice, Rice University in Houston, and Baylor Collage of Medicine in Houston. You invited me to be a part of your research and introduced me to new frontiers of science. Special thanks to the group of Dr. Bogdan Czerniak from MD Anderson Cancer Center in Houston, whose studies contributed to this thesis.

I would like to thank those who made it possible for me to pursue my studies. My parents, Stanisław and Jadwiga, who have always supported me and encouraged me to follow this path. My brother, Tomasz, for always being ready to listen and talk. And my girlfriend, Magdalena, for her love, presence and patience in the face of all my scientific doubts.

This work has been co-financed by the European Union through the European Social Fund (grant POWR.03.02.00-00-I029).

Calculations were carried out using the Ziemowit (`www.ziemowit.hpc.polsl.pl`), a computer cluster funded by the Silesian BIO-FARMA project No. POIG.02.01.00-00-166/08 in the Computational Biology and Bioinformatics Laboratory of the Biotechnology Centre in the Silesian University of Technology.

Models of cancer genome evolution used to evaluate the role of selection and occurrence of new mutations

Author: Paweł Kuś

Abstract

Cancer is one of the leading causes of death worldwide. Risk factors are often tied to lifestyle changes in developed countries, making cancer a significant research focus. The advent of Next Generation Sequencing (NGS) made molecular cancer data more available than ever and allowed scientists to unravel many molecular mechanisms that characterize cancer cells. Despite these advancements, effective anti-cancer therapy preventing the evolution towards drug resistance and relapse remains a challenge. Understanding the mechanisms that drive tumor evolution, mutagenesis, and selection may bring us closer to effective anti-cancer treatments.

Bulk DNA sequencing allows us to identify variants in tumor genomes and measure their allelic frequencies (VAF). It has been shown that processes of mutagenesis and selection shape the distribution of VAFs in the sample. Models were proposed that fit the VAF distribution with a mixture of power-law-shaped and binomial distributions. The power-law component models the neutral tail of variants, containing primarily neutral variants occurring in all cells, while the binomial components model the clones and selectively advantageous subclones. The parameters of these components reflect the evolutionary dynamics of the tumor.

We developed a new R package cevomod capable of fitting the mixture of the power-law and binomial components to the whole exome sequencing data, which previously could not be analyzed with other well-known algorithms due to the strict data quality requirements. cevomod allows one to choose between two types of models, a neutral-like one with the power-law exponent equal to 2 and an optimized model, in which the exponent is optimized to fit the data best. While the first model uses the assumptions of exponential tumor growth and constant mutation rate, the second one allows for validating these assumptions.

Using our new package and the collected data from 4 cancer types, we show that bulk DNA sequencing can be used to quantify the changes in the evolutionary dynamics of cancer upon progression, metastasis, and relapse. To prove that, we analyzed the DNA sequencing data from patients with Acute Myeloid Leukaemia, including samples from the time-points of diagnosis and relapse, patients with Breast Cancer and Laryngeal Cancer, including samples from the primary tumors and lymph node metastases, and two whole organ maps of Bladder Cancer, including the urothelial cancer samples along with the pre-malignant samples with different stage of disease progression.

We found significant differences in the evolutionary parameters between samples from the same tumor, such as the predominant increase of the mutation rate in lymph node metastases of laryngeal cancers, compared to the primary tumors or common upward and downward changes of mutation rate in the recurrent leukaemias.

Finally, we show that the assumptions underlying the most frequently used models used to estimate the parameters of tumor evolution may be violated in many cancers. We identified significant deviations of the neutral tail power-law exponent from the expected value of 2 that may indicate the non-exponential tumor growth, changing mutation rate, or presence of selectively advantageous micro-clones. We proposed a mathematical explanation for the observed phenomena, relating the deviations to the non-constant mutation rate.

We believe that our results can contribute to the understanding of processes responsible for the evolution of cancer.

Key words

cancer, evolution, mutagenesis, selection

Author's publications

1. Bondaruk, J., Jaksik, R., Wang, Z., Cogdell, D., Lee, S., Chen, Y., Dinh, K. N., Majewski, T., Zhang, L., Cao, S., Tian, F., Yao, H., **Kuś, P.**, Chen, H., Weinstein, J. N., Navai, N., Dinney, C., Gao, J., Theodorescu, D., Czerniak, B., et al. (2022). The origin of bladder cancer from mucosal field effects. IScience, 25(7), 104551. https://doi.org/10.1016/j.isci.2022.104551

2. Hormaechea Agulla, D., Matatall, K. A., Le, D. T., Kain, B., Long, X., **Kus, P.**, Jaksik, R., Challen, G. A., Kimmel, M., King, K. Y. (2020). IFNg signaling is a driver of Dnmt3a-mutant clonal hematopoiesis. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3564993

3. Hormaechea-Agulla, D., Matatall, K. A., Le, D. T., Kain, B., Long, X., **Kus, P.**, Jaksik, R., Challen, G. A., Kimmel, M., King, K. Y. (2021). Chronic infection drives Dnmt3a-loss-of-function clonal hematopoiesis via IFNγ signaling. Cell Stem Cell, 28(8), 1428-1442.e6. https://doi.org/10.1016/j.stem.2021.03.002

4. Janus, P., **Kuś, P.**, Vydra, N., Toma-Jonik, A., Stokowy, T., Mrowiec, K., Wojtaś, B., Gielniewski, B., Widłak, W. (2022). HSF1 can prevent inflammation following heat shock by inhibiting the excessive activation of the ATF3 and JUN&FOS genes. Cells (Basel, Switzerland), 11(16), 2510. https://doi.org/10.3390/cells11162510

5. Kurpas, M. K., Jaksik, R., **Kuś, P.**, Kimmel, M. (2022). Genomic analysis of SARS-CoV-2 Alpha, Beta and Delta variants of concern uncovers signatures of neutral and non-neutral evolution. Viruses, 14(11), 2375. https://doi.org/10.3390/v14112375

6. **Kuś, P.**, Jaksik, R., Kimmel, M. (2022). Analiza struktury klonalnej nowotworów - porównanie wyników różnych kombinacji metod. Wydawnictwo Politechniki Śląskiej. https://doi.org/10.34918/83571

7. Le, D., Florez, M. A., **Kus, P.**, Tran, B., Kain, B. N., Jain, A., Malovannaya, A., King, K. Y. (2022). BATF2 promotes HSC myeloid differentiation via amplification of the pro-inflammatory response during chronic infection. Blood, 140(Supplement 1), 5732–5733. https://doi.org/10.1182/blood-2022-164467

8. Le, D. T., Florez, M. A., **Kus, P.**, Tran, B. T., Kain, B., Zhu, Y., Christensen, K., Jain, A., Malovannaya, A., King, K. Y. (2023). BATF2 promotes HSC myeloid differentiation by amplifying IFN response mediators during chronic infection. IScience, 106059, 106059. https://doi.org/10.1016/j.isci.2023.106059

9. Vydra, N., Janus, P., **Kus, P.**, Stokowy, T., Mrowiec, K., Toma-Jonik, A., Krzywon, A., Cortez, A. J., Wojtas, B., Gielniewski, B., Jaksik, R., Kimmel, M., Widlak, W. (2021). Heat shock factor 1 (HSF1) cooperates with estrogen receptor α (ERα) in the regulation of estrogen action in breast cancer cells. ELife, 10. https://doi.org/10.7554/eLife.69843

Author's contributions to publications used in the dissertation

Parts of Section 5.2 describe the results published in the paper by Bondaruk et al. *The origin of bladder cancer from mucosal field effects* [10].

My contribution in [10]: Analyses of variants in α and β clusters and variants associated with the dormant and progressive phases of tumor evolution, including the analysis of mutational signatures of these groups of variants, validation of variants using the RNA sequencing data.

Conference abstracts

1. **Kuś, P.**, Jaksik, R., Kimmel, M.: *Measurement biases affecting RNA sequencing and methods of its normalization.* Bioinformatics in Torun 2019

2. **Kuś, P.**, Jaksik, R., Kimmel, M.: *Influence of measurement bias on the interpretation of RNA sequencing results.* qBio 2019 Conference, San Francisco, CA, US, DOI: 10.13140/RG.2.2.28633.19040

3. **Kuś, P.**, Jaksik, R., Kimmel, M., Widlak, W.: *Genomic action of estrogen receptor α in breast cancer patients may be enhanced by Heat Shock Factor 1.* The 5th Warsaw Conference on Perspectives of Molecular Oncology (2020) (online) DOI: 10.13140/RG.2.2.21083.44329

4. Janus, P., Vydra, N., **Kuś, P.**, Stokowy T., Mrowiec, K., Toma-Jonik, A., Krzywoń, A., Cortez, A. J., Wojtaś, B., Gielniewski, B., Jaksik, R., Kimmel, M., Widłak W.: *The role of HSF1 in the regulation of thetranscriptional response to estrogen at the level of chromatin organization.* Gliwice Scientific Meetings 2021

5. **Kuś, P.**, Jaksik, R., Kimmel, M.: *Analiza struktury klonalnej nowotworów - porównanie metod.* Computational Oncology and Personalized Medicine COPM2021 Conference, (online)

6. **Kuś, P.**, Kimmel, M.: *Sampling-oriented modelling of cancer evolution.* Gliwice Scientific Meetings 2022

7. Janus, P., **Kuś, P.**, Jaksik, R., Vydra, N., Kurpas, M., Kimmel, M., Widłak W.: *Differences in signaling induced by Transforming Growth Factor Beta in normal and malignant mammary epithelial cells.* Gliwice Scientific Meetings 2022

8. Toma-Jonik, A., Janus P., **Kuś P.**, Vydra N., Mrowiec K., Stokowy T., Wojtaś B., Gielniewski B., Widłak W.: *Heat shock-induced inflammatory response could be stronger in hsf1-deficient cells.* Gliwice Scientific Meetings 2022

9. Czerniak, B., Lee, S., Jung S., **Kus, P.**, Jaksik, R., Lee, J. G., Chen, H., Navai, N., Guo, C., Wei, P., Kimmel, M.: *Modeling of Bladder Cancer Evolution from Whole-Organ Mutational and Proteomic Profiles*, United States and Canadian Academy of Pathology USCAP, New Orleans 2023

Contents

List of abbreviations and symbols

AML	Acute Myleoid Leukaemia
BLCA	Bladder Urothelial Carcinoma
bp	base pair(s), a measure of DNA/RNA length, same as *nt*
BRCA	Breast Cancer
CNV	Copy Number Variation
DDR	DNA Damage Response/Repair
DNA	deoxyribonucleic acid
EGA	European Genome-Phenome Archive
HGIN	High-Grade Intraurothelial Neoplasia
HPC	High-Performance Computing
Indels	Insertions and Deletions
ITH	Intra-Tumor Heterogeneity
LGIN	Low-Grade Intraurothelial Neoplasia
LSCC	Larynx Squamous Cell Carcinoma
NGS	Next Generation Sequencing
NU	normal urothelium
nt	nucleotide(s), a measure of DNA/RNA length, same as *bp*
PCR	Polymerase Chain Reaction
RNA	ribonucleic acid
RNAseq	RNA sequencing
SD	standard deviation
SNV	Single Nucleotide Variation
SFS	Site Frequency Spectrum
SV	Structural Variant
T2T	Telomere-to-Telomere, first complete assembly of human genome without any gaps
TCGA	The Cancer Genome Atlas, large-scale project of cancer genome sequencing supervised by National Cancer Institure in United States
TMB	Tumor Mutational Burden
TNBC	Triple Negative Breast Cancer
VAF	Variant Allele Frequency
WXS	Whole Exome Sequencing

WGD	Whole Genome Doubling
WGS	Whole Genome Sequencing
UC	Urothelial Carcinoma

Chapter 1

Introduction

1.1 Motivation

Cancer is one of the leading causes of death worldwide, and as such, it receives particular attention from the research community. The advent of Next Generation Sequencing (NGS) made molecular cancer data more available than ever before. It empowered the scientists with the tools necessary for in-depth investigation of mechanisms driving the initiation and progression of cancer. In the last decade, most of this research was focused on unraveling the molecular mechanisms that allow the tumor to grow, which seemed the shortest path to finding the right cures. During that time, we learned that tumors continuously evolve, which allows them to develop resistance to the treatment, recur and metastasize. Understanding the mechanisms that drive tumor evolution, mutagenesis and selection may bring us closer to effective anti-cancer treatment.

As DNA sequencing has become popular and made the data more available, many interesting studies have been conducted. A number of them proposed great algorithms for reconstructing the tumor subclonal structure. They allow identifying the subpopulations of cells in the tumor with different sets of mutations and sometimes try to infer the phylogenetic tree of the tumor evolution. However, they usually do not estimate any parameters of tumor evolution, such as mutation rates or selection coefficients. Other studies proposed excellent mathematical models of tumor evolution that associate the tumor characteristics with the evolutionary parameters but did not develop software allowing other researchers to use it on their own data easily. Finally, some recent papers introduced the models and the appropriate software packages for fitting them. At least 3 of them were used to assess the role of selection across all the cancer cases in The Cancer Genome Atlas. Their assumptions and results were, however, questioned [117, 11, 83]. MOBSTER is probably the best-known of these algorithms. It estimates the parameters of tumor evolution based on assumptions of exponential tumor growth and constant mutation rate, which should result in a characteristic, power-law-shaped distribution of neutral mutations called the

neutral tail. These assumptions are not always fulfilled, though. Indeed, we observed and described in this thesis frequent deviations from the theoretical shape of the neutral tail and tried to investigate them. MOBSTER also requires deeply sequenced Whole Genome Sequencing input data to recognize the neutral tail properly; thus, its application is limited to top-quality data. We do not know any publicly-shared algorithm dedicated to the analysis of cheaper and more common Whole Exome Sequencing data.

The importance of mutagenesis and selection in cancer evolution, including recurrence and metastasis, is still an open question. In this thesis, we want to contribute to this area of cancer research.

1.2 Goals

The first goal of this thesis is to determine the role of mutagenesis and selection in tumor progression, metastasis, and recurrence. We used the sequencing data from 4 types of cancer, consisting of multiple samples obtained from each patient. Our data included the diagnostic and relapse samples from acute myeloid leukemia (AML), primary tumors, and lymph node metastases from breast cancers (BRCA) and laryngeal cancers (BRCA), and whole-organ mapping of bladder cancer (BLCA) specimens, including regions with early stages of disease progression.

Due to strict data quality requirements, the primary goal could not be achieved using the well-known MOBSTER algorithm. For this reason, the development of a new package became the second goal of the thesis. The new package should be applicable to Whole Exome Sequencing data and data with lower sequencing depth, both resulting in the under-representation of the low-frequency mutations in the expected neutral tails.

The third goal of the thesis was to check the validity of common model assumptions. Since the power-law-shaped neutral tails should follow the $1/f^2$ statistic under the assumptions of exponential growth rate, constant mutation rate, and lack of competing micro-clones, we compared the theoretical fits with the optimum ones. We also investigated the possible causes of the identified deviations.

1.3 Hypothesis

We state the following thesis:

Changes in the evolutionary dynamics of cancer upon metastasis and recurrence can be quantified from the bulk DNA sequencing data.

1.4 Plan of the thesis

Introduction. In this chapter, we introduce the motivation, hypothesis, and goals of the work.

Cancer evolution. The second chapter presents the fundamentals of molecular biology that underlie cancer evolution. We introduce the central dogma of molecular biology and its role in the maintenance of the cell state. Then, we describe the role of DNA, genes, and mutations and characterize the main categories of mutations. Finally, we describe the most important theories of tumor evolution, including Darwinian and non-Darwinian evolution, clonal evolution, punctuated evolution, field effect, and the theory of cancer stem cells.

Analysis of cancer genomes and cancer evolution. In this chapter, we first describe what underlies most of the modern cancer research: the development of DNA sequencing methods. We describe the 3 generations of sequencing methods, the fundamentals of sequencing data analysis, and the mathematical approaches in the analysis of cancer evolution. We also list the most important software algorithms and summarise the discussion raised in the scientific community by some publications.

Data and methods. This chapter describes the data and methods used in this thesis. We describe the two model fitting approaches that we implemented and present an R package *cevomod*, ***C**ancer **Evo**lutionary **Mod**els*, which we developed.

Results. We divided the Results chapter into two sections. The first one describes mostly the results of our work in the project *A systems approach to cancer progression and prognosis: New models and statistics for genomic data analysis* funded by the Polish National Science Center. In this project, we analyze the data from the primary tumors and lymph node metastases in BRCA and LSCC cohorts. We also parallelly included the study of the AML cohort in this section, although this data does not originate from our project and was downloaded from European Genome-Phenome Database. In the second section, we describe the results of our collaboration with the group of Dr. Bogdan Czerniak from MD Anderson Cancer Center in Houston. Together with his group, we investigated the origins of bladder cancer and showed how it develops from the mucosal field effect. This section is based on our paper Bondaruk et al. *The origin of bladder cancer from mucosal field effect* [10]; however, we added the more recent analyses made with our package *cevomod*, absent in the paper.

Summary. In the final chapter of the thesis, we summarise the study's achievements and limitations. We refer to the dissertation thesis we have proved and indicate the possible

paths for future research.

Chapter 2

Cancer evolution

Cancer is one of the leading causes of death in the world, even though more than seven decades have passed since Sidney Farber's early chemotherapy trials in Boston in the late 1940s [108]. According to the World Health Organization, cancer caused nearly 10 million deaths worldwide in 2020. In the same year, the five most common types of cancer (breast, lung, colorectal, prostate and skin) accounted for over 9 million of new cases, and about 400 000 cancer cases were diagnosed in children [16]. In Poland, over 170 000 new cancer cases and over 100 000 cancer-related deaths were reported in 2019 [42]. Many of the cancer risk factors are related to the recent changes in the human lifestyle: tobacco use, alcohol consumption, obesity, unhealthy diet and low physical activity, occupational exposure to carcinogens, or to infection with cancer-causing viruses, such as human papillomavirus (HPV) [16]. Efforts made in cancer research over the last few decades let scientists to unravel a number of molecular mechanisms underlying the biology of cancer [47, 48], and many advanced therapies have been introduced. Despite these advances, treatment of cancer patients which would prevent the development of therapy resistance and disease recurrence remains a challenge for the medicine. Our ability to treat or prevent cancer is limited, similarily as our understanding of the mechanisms underlying tumor initiation and progression. Since humans are more prone to cancer than many longer-living organisms with larger body-sizes than humans [9], cancer susceptability might be connected with complexity of human genome and its rapid evolution.

2.1 Biology of cancer

2.1.1 Hallmarks of cancer

Cancer is a general term that describes a number of diseases. Cancer cells can form tumors in different tissues and locations in the body, which can be solid or liquid, highly aggressive or benign, but all tumors tend to share a number of hallmarks. The hallmarks are directly related to the loss of the organism's control over cell growth and proliferation

and the tissue homeostasis. The first set of hallmarks was published in 2000 and consisted of: acquisition of the limitless replicative potential, self-sufficiency in growth signalling, resistance to anti-growth signaling, evading cell death, induction of angiogenesis, and ability to tissue invasion and metastasis [47]. However, a number of new mechanisms have been proposed as the new hallmarks in the following years, such as the deregulation of cellular energetics, evasion of immune destruction, genome instability, inflammation [48], or dysregulation of cell differentiation [25]. The list is still growing.

2.1.2 Central dogma of molecular biology

All the molecular mechanisms that control the cell cycle and maintain the tissue homeostasis depend on the interactions of proteins with proteins or other molecules. For this reason, protein expression is the key process in the regulation of the activity of molecular pathways. In this process, the genetic information stored in the nucleus, in the form of a double stranded sequence of DNA (deoxyribonucleic acid) is transcribed into messenger RNA (single stranded ribonucleic acid), transferred to the cytoplasm, and translated by ribosomes into an amino acid chain. The latter becomes a protein following additional post-translational modifications (PTM) and folding. The process is usually one-way, the fact known as the *central dogma of molecular biology* [26] (Fig. 2.1).

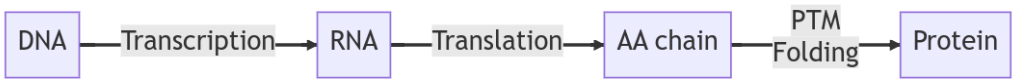


Figure 2.1: Central Dogma of Molecular Biology

The dogma implies among other that cell behaviour, which is mainly controlled by proteins and their interactions, can be influenced at any stage of the gene expression. Protein properties can be altered even at the very end of the process by the post-translational modifications, such as phosphorylation (addition of the phosphate group). Protein concentration levels can be modified earlier in the process, by controlling the transcription and translation processes. Transcription can be regulated, among other, by the action of transcription factors or transcription repressors, or by epigenetic DNA modifications such as methylation (binding of methyl groups to the DNA sequences), and translation can be silenced by the expression of certain micro-RNAs. Finally, the protein structure can be modified by mutations in the DNA, which can be inherited by all the descendant cells and are therefore fundamental for the evolution of the tumor genomes.

2.1.3 DNA, genome and genes

DNA is a double stranded helix consisting of two complementary sequences of nucleotides (G - guanine, C - cytosine, A - adenine, and T - thymine). In eukaryotic cells it is organised into a number of linear chromosomes. In humans, the genome (the total DNA sequence of each cell) consists of 23 pairs of chromosomes: 22 pairs of autosomal chromosomes numbered from 1 to 22, and 1 pair of sex chromosomes: X and Y in males and a pair of X chromosomes in females. The total length of the human diploid genome (consisting of pairs of chromosomes) exceeds 6×10^9 nucleotides (nt, or base-pairs, bp). However, not all of the entire DNA sequence encodes proteins - in human, the protein coding sequence constitutes only about 1 percent of the genome and is organized into approximately 25,000 of DNA fragments called genes. Genes vary in their length: from less than 1000, to more than 2 million nucleotides, and their coding sequences (exons) are usually interspersed with non-coding regions (introns). Non-coding sequences such as introns and intergenic regions can play an important role in the regulation of gene expression since they include regulatory sequences such as promoters, enhancers, and other.

Cancer driver genes

Some proteins, especially those involved in hallmarks of cancer, are important for the control of the cells' growth and proliferation. For this reason genes encoding these proteins are particularly often mutated in tumor cells, and are referred to as the *cancer driver genes*. Analyses of the mutation frequencies in genomes of different tumor types, such as these listed in *The Cancer Genome Atlas*, allowed to identify a number of such genes. Some of them are frequently mutated in the particular types of cancer, and some in all cancers. Cancer driver genes are divided into two groups: tumor suppressor genes and oncogenes.

Tumor suppressor genes encode proteins which action protects organisms from the developing cancer. For example, the TP53 gene plays a crucial role in the decision-making during the DNA damage response. If the DNA damage is too severe for the cell to divide, TP53 can block the cell-cycle until the damage is repaired. If repair fails, TP53 activates the apoptosis - programmed cell death. For this role in the DNA-status control, TP53 is called 'the guardian of the genome'. TP53 is also one of the most frequently mutated genes across all cancer types, approximately in 35% of all new cancer cases [85]. Other examples tumor suppressor genes are RB, which transduces the pro-growth signalling from the outside of cell [48], APC, or BRCA1/2, a pair of genes engaged in the homologous recombination - one of the basic pathways of the DNA damage repair. Often at least two mutations need to occur to deactivate both copies of the tumor suppressor gene (*Two-hit hypothesis* [64]). Tumor suppressor genes may be silenced by epigenetic mechanisms, with the most commonly reported being DNA methylation [61].

Second group of the cancer driver genes, the oncogenes, are genes whose abnormal expression shows a protumorigenic effect on the cell, for example via contribution to the pro-growth signaling. Oncogenes can be activated by point mutations or amplifications which may lead to their overexpression. Some examples of the oncogenes are: KRAS gene, frequently mutated in pancreatic, colorectal and lung cancers [135], which can be activated by point mutations and leads to cell proliferation and migration [35]; ERBB2 gene, whose malignant amplification and overexpression is frequent in breast cancers and is associated with proliferation, loss of cell polarity and invasion [89]; or MYC, gene most often mutated in ovarian or uterine cancers, whose activation is associated with the cancer growth and contributes to the evasion of immune system response [32].

The list of tumor suppressor genes and oncogenes in cancer types analyzed in the thesis is included in Appendix A.

Cancer essential genes

While some genes need to be mutated for cancer to growth, there are also genes which are necessary for both normal and tumor cells to survive. These genes called cell-essential genes are rarely mutated in tumor cells and their loss-of-function mutations lead to the fitness depletion and/or cell death. Genome-wide screening projects using gene editing systems like CRISPR allowed identifying many such genes, engaged in the regulation of cell cycle, protein homeostasis, DNA-damage response and other molecular mechanisms: CDK4, CDK6, MEK1, HDAC1, and a number of other [20, 124].

2.1.4 Mutations

Alterations in the DNA sequence of an organism are called *mutations*. Mutations can occur at any moment of the cell life and although most mutations do not result in a measurable change in the organism fitness, some of them can have beneficial or disadvantageous effect on the affected cells. Mutations that occur in germline cells (germline mutations) can affect all the cells of the offspring, and may increase the risk of cancer or other disorders. If a mutation occurs in non-germline cell (somatic mutations), it will affect only the cell and its descendants, but it can lead to the development of cancer if sufficient number of fitness-increasing mutations accumulate in one cell. If mutation affects the coding sequence of the gene, may lead to the loss, gain, or change of function of the protein. If it occurred in the non-coding sequence, it may alter gene expression when it affects regulatory sequences of the gene, such as promoters or enhancers. DNA sequence can be altered by many types of genetic alterations (Fig. 2.2):

- Single Nucleotide Variants (SNV) - one nucleotide is replaced by another one. SNVs may or may not lead to the alteration in the encoded protein because of the degeneracy of the human genetic code: the SNV is *silent* if the new codon encodes the

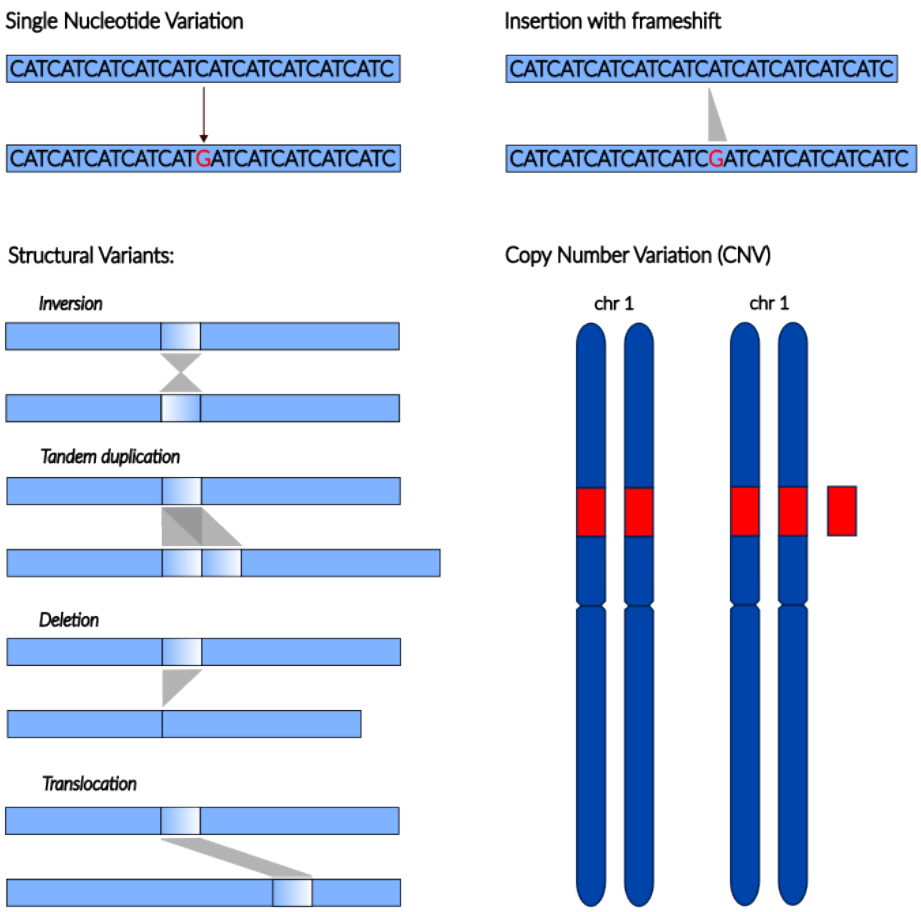


Figure 2.2: Types of mutations.

same amino acid, or *missense* if the encoded amino acid has changed. SNVs can also lead to the gain (or loss) of start or stop codons, and therefore result in severe or complete inactivation of the protein, or can occur in the gene regulatory elements (promoters, enhancers etc.) and affect the gene expression levels. Contiguous SNVs are often referred to as DNV (Double-Nucleotide Variants) or MNV (Multiple Nucleotide Variants)

- Insertions and Deletions (jointly called Indels) - cell gains or loses one or more nucleotides in its genome. Effect of an indel depends on the number of nucleotides inserted or lost, when this number is not a multiple of three and mutation occurs in the gene coding sequence, indel results in the *frameshift*: polymerase reads incorrect nucleotide triplets and the amino acid chain of downstream protein changes completely

- Copy Number Variants (CNV) - are large-scale events in which cell gains additional copies of large fragments of its DNA sequence, or loses them. CNVs alter the numbers

of gene copies in the cell, leading to the gene overexpression if copies are gained, or to gene silencing if copies are lost. CNVs can vary in size: from thousands of nucleotides to whole chromosome gains or loses and Whole Genome Doublings (WGD)

- Structural Variants (SV): duplications - when short DNA sequence is duplicated, translocations - when sequence is moved to another site in the genome, and inversions - sequence is placed in its own locus, but its orientation is changed

Different classes of variants are known to play different roles across the tumor types. For example, Copy Number changes drive early evolution of breast cancers [39], but occur late in the evolution of kidney or lung cancers and melanomas [40]. Reciprocal translocation of fragments of chromosomes 9 and 22, known as the Philadelphia chromosome, is a hallmark of chronic myeloid leukemia [59].

Mutations occur during the entire organism's lifetime and can be induced by various concurrent mutational processes. Recent advances in methods of tumor genome analysis allowed to gather collections of cancer associated somatic mutations larger than ever before, and identify a number of *mutational signatures*: fingerprints of mutational processes affecting the genome of cells. Works of Alexandrov et al. [3, 4] identified over 80 mutational signatures for SNVs, DNVs and Indels, and associated many of them with mutational processes. Those processes include endogenous mechanisms like spontaneous deamination of 5-methylcytosine, which results in numerous C to T substitutions and is correlated with patient's age at diagnosis; activity of APOBEC, family of cytidine deaminases, also resulting in C to T mutations; defective homologous recombination process (signature enriched in short Indels) and activity of other double strand break (DBS) repair mechanisms such as non-homologous end joining (NHEJ); but also the processes of exogenous origin, such as ultraviolet light exposure and tobacco smoking.

2.1.5 Tumor heterogeneity and evolution

Tumors differs not only among patients and cancer types, but also show significant intratumor heterogeneity, which is a result of ongoing mutational processes and tumor evolution. Tumor starts from a single mutated cell that gained proliferation advantage over the healthy tissue cells, but its descendant cells gain new mutations during their lifetime, giving rise to the new clones and subclones. As the time passes, some of the lineages become extinct, outcompeted by more successful clones or by chance, due to the genetic drift. If one cell lineage completely replaces others, it's founder cell becomes the most recent common ancestor and its mutations (unless mutation is lost due to another genetic event) are found present in all cells of the tumor, along with mutations newly obtained by the sub-lineages.

Both phenomena: tumor heterogeneity and constant evolution are key features of cancer that lead to the evolution of therapy resistance and disease relapses.

2.2 Modes of cancer evolution

2.2.1 Darwinian and non-Darwinian evolution

The process of species evolution proposed by Charles Darwin in 1859 is driven by the processes of mutagenesis and natural selection. Mutagenesis introduces new genetic variants to the population, increasing its genetic (and phenotypic) diversity. If the newly emerged cells or specimens are better suited to the environment, they perform better and have a better chance to proliferate, passing their genetic information to the offspring. This process, known as positive selection, increases the frequency of advantageous genetic variants in the entire population. Also, if the new cells bear mutations that are disadvantageous for their phenotype, they have a bigger chance to die and proliferate less frequently. This is called negative selection, and has a purifying effect on the population, clearing it of the unfavorable variants or keeping them rare. Thus, negative selection in its decreasing genetic diversity is complementary to the process of mutagenesis.

The importance of Darwinian selection in tumor evolution is an open question receiving increasing attention. Some works report an extensive genomic intra-tumor diversity that exceeds the diversity predicted under the Darwinian model with negative selection [79]. Other studies report the lack of positive selection as well. In asexually reproducing populations such as cancer cells, the positive selection leads to the emergence of new, distinct clones. As time passes, the fitter clones increase their frequency and eventually completely replace their ancestral populations, which is known as the selective sweep. A recent study of over 300 colorectal glands (structures of the glandular epithelium) reported the absence of selective sweeps in the 15 analyzed colorectal tumors [110].

Positive selection is not the only process that changes the frequencies of variants in populations. In neutrally evolving populations, especially if the population size remains constant, variant frequencies can increase or decrease by chance, due to random sampling of specimens to die or proliferate. This so-called genetic drift also affects the frequencies of non-neutral variants and might lead to the loss of the advantageous variants or the fixation of the disadvantageous ones. In some cases, it is not easy to dissolve the effects of genetic drift and positive selection.

As more and more sequencing data gets available, new methods for testing the hypothesis of evolution neutrality appear [130, 82, 131, 18], leading to discussions on their biological, mathematical and technical assumptions [117, 52, 11, 120]. An overview of modeling approaches will be described in Chapter 3.3.2.

2.2.2 Clonal cancer evolution

The model of clonal cancer evolution is a model of gradual evolution (Fig. 2.4a and 2.4b). In this model, cancer originates from a single malignant cell that starts to proliferate

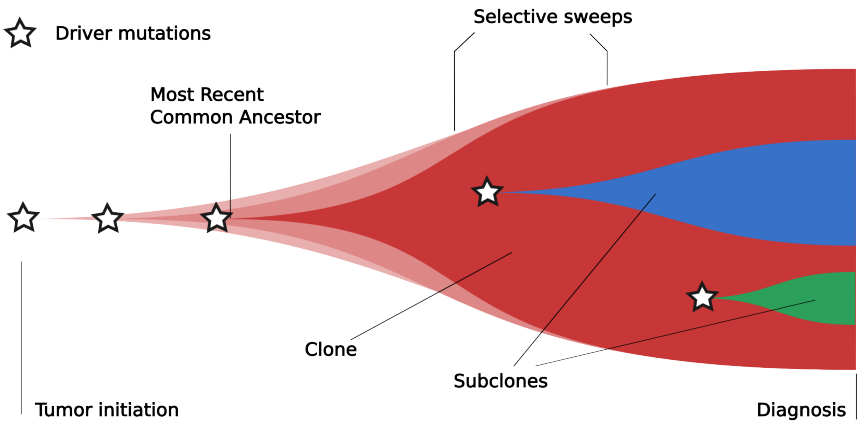


Figure 2.3: Model of clonal cancer evolution.

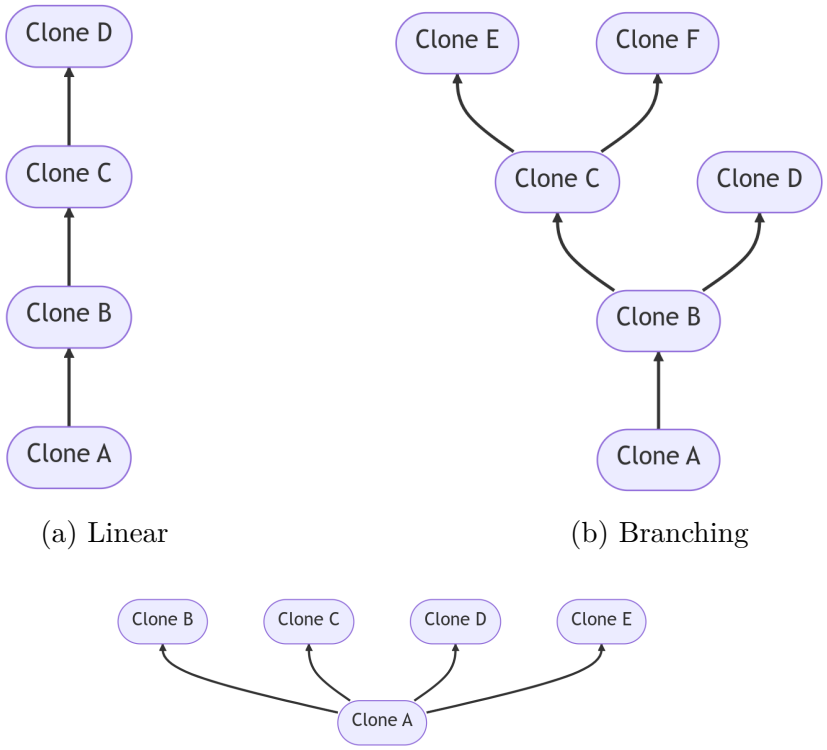
out of the organism's control and forms the original clone. New driver mutations occurring in the tumor initiate new subclones which grow faster than the original one, and eventually replace it (selective sweep). The evolution might be linear, if the new subclones originate from the most recent ones, or branching, if number of subclones originate from the same parental subclone (Fig. 2.4).

2.2.3 Big Bang model of cancer evolution

The big bang model of cancer evolution (also known as punctuated cancer evolution) (Fig. 2.4c and 2.5) claims that all the driver mutations occur early at the beginning of the tumor evolution. In this model, subclones born at the beginning of tumor progression coexist in the tumor mass and grow together in a single expansion, with no selective sweeps and no later emergence of new subclones. The model was proposed by Sottoriva et al. as the model of human colorectal tumor growth [110] and explains phenomena such as the presence of the same subclone on the opposite sides of the tumor. Subclones born early can easily mix in early malignancies with disrupted cell adhesion and thus be spread in various regions as the tumor expands.

In another study, Wang et al. found that most copy number changes in breast cancers also occur early in tumor development, in contrast to SNV mutations which arise later [126]. It makes the big bang model a correct description of some cases of CNV heterogeneity evolution and chromothripsis - events of complex chromosomal rearrangement affecting one or few chromosomes that occur in a single event.

The Big Bang model states, that mutations responsible for tumor invasion, metastasis, or evolution of the therapy resistance, may already be present in the malignancy from the beginning, although they might be too rare to be detected. In such a case, therapy which eliminates the most abundant clones, makes space for the future progression of the resistant ones.



(a) Linear

(b) Branching

(c) Big Bang

Figure 2.4: Different cancer phylogenies.

2.2.4 Cancer Stem Cells model

The Cancer Stem Cells (CSC) model postulates that not only a small fraction of tumor cells have a stem-like ability of self-renewal and limitless proliferation potential (Fig. 2.6). Those cells, called the cancer stem cells, give origin to the entire tumor mass [30] and are capable of reproducing the tumor after the therapy or transplantation into another organism. The CSC model mirrors the origin of many normal tissues, such as blood or intestinal epithelium, which originates from the haematopoietic stem cells and intestinal stem cells, respectively. CSCs have been identified in leukaemia, breast cancer, colorectal cancer, and glioblastoma [30, 70]. According to this model, successful treatment of cancer must target the niches of CSCs in order to prevent the relapse.

2.2.5 Evolution from a cancerization field

Not all cancers are found to be initiated by a single cell. Some cancers are recognized as multifocal, consisting of multiple independently initiated tumors; for such cases field cancerization (field effect) model was proposed. According to the field cancerization model cells in the histomorfologically normal tissue can carry severe mutations and experience clonal expansions of genetically mutated but healthy cells without the immediate manifestation of cancer. Such tissues are preconditioned for development of cancer and

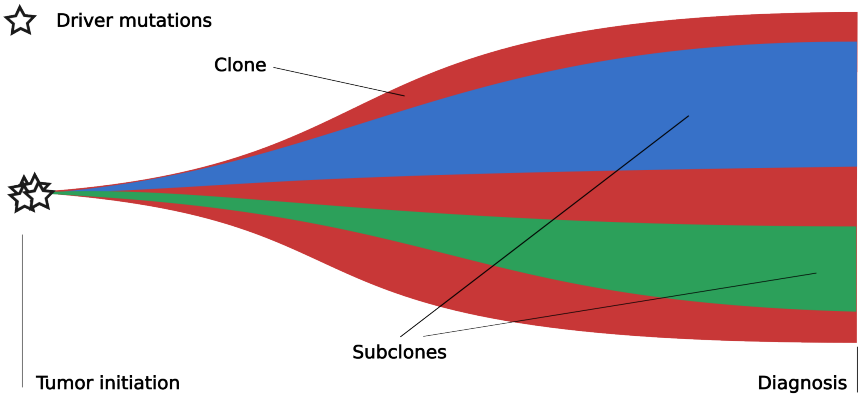


Figure 2.5: Big Bang model of cancer evolution.

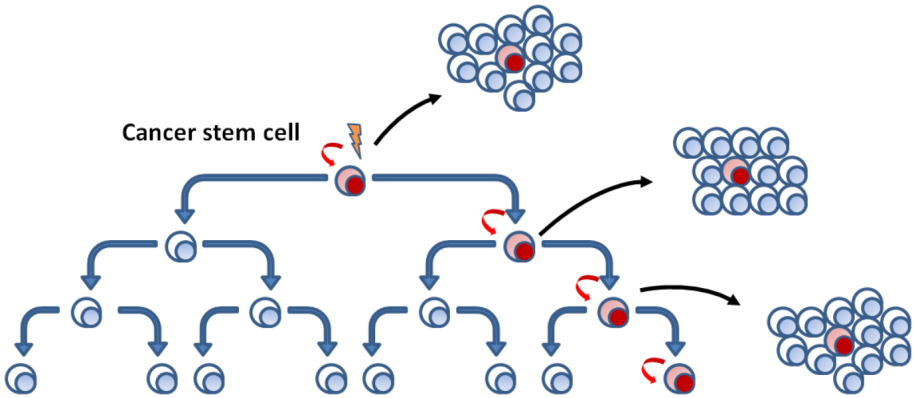


Figure 2.6: Cancer Stem Cells model. Source: *www.wikipedia.org* [17]

can promote it, leading to the emergence of many concurrent tumors. Field effect models apply particularly to the tissues with direct exposure to the mutagens and mutagenic environment: such as UV (skin) [91] or smoking (mouth skin and bladder mucosa) [129]. Evidence for the field effect model was identified, among others, in oral squamous cell carcinomas [109] and prostate cancers [14]. In [10] we also support the field effect model of cancer initiation in bladder cancers.

2.2.6 Clonal cooperation

The phenomena of clonal cooperation occurs when tumor clones are more tumorigenic when intermixed together than in the absence of the other one. An interesting case of clonal cooperation was described by Cleary et al. [23] in Wnt-driven mouse mammary tumors. Those tumors are composed of a mixture of the basal and luminal tumor cells; in some of them Cleary et al. observed that the basal cells carry driver mutation in the Hras gene and the luminal ones do not, thus both cell types represent different lineages. They found that both those lineages need to be transplanted to result in the tumor growth in the recipient mouse. Although the mouse model does not fully represent the complexity

of the human tumors which evolve much longer, it is an evidence that cancer clones might cooperate in general.

Chapter 3

Analysis of cancer genomes and cancer evolution

3.1 DNA sequencing

Rapid progress observed in the research on the cancer genome evolution during the last two decades was possible as the consequence of advancements made in the field of genome (and transcriptome) sequencing. DNA sequencing is the process of determining the order of nucleotides in the sequence of the DNA molecule. The ability to read the DNA sequence is crucial to the development of modern molecular sciences. It is the first step towards the identification of functional elements in the genome, such as the genes and their regulatory elements. Comparison of the normal and tumor genomes allows us to identify the genomic events that drive tumorigenesis. Comparison of the genomes of many related specimens allows us to infer their evolutionary tree. In the case of multiple cancer genomes - it allows tracking the evolution of the genome in time. The history of DNA sequencing dates back to the 1970s, and three major stages can be distinguished in its development [51].

3.1.1 The first generation DNA sequencing

Although the few-nucleotide long sequences of DNA could already be sequenced in the 1960s [53], the first big breakthrough came in 1977, when Sanger et al. developed a new sequencing method with the chain-terminating inhibitors [101]. In their technique, the analyzed DNA is placed in four samples along with the mixture of deoxyribonucleotides (dNTPs) of all bases: A, G, C, and T. Also, the radiolabelled dideoxynucleotides (ddNTPs) are intermixed in low concentrations, with one base type for each sample. dNTPs are the monomers of DNA, which in the reaction of DNA extension can form a chain of nucleotides complementary to the analyzed template. ddNTPs are modified dNTPs that lack the hydroxyl group necessary for further chain elongation. When the ddNTP is synthesized

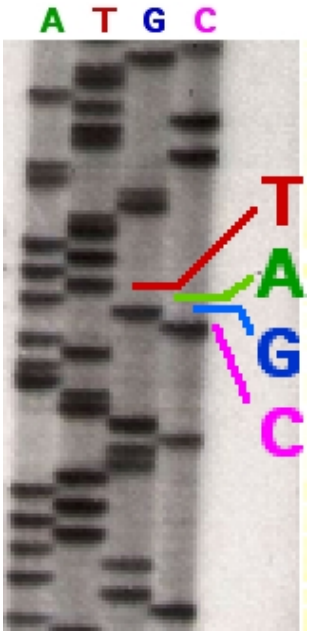


Figure 3.1: Sanger sequencing. Fragments of various length, terminated an known base type, are ordered by length in the process of electrophoresis. Source: *www.wikipedia.org* [102]

.

into the chain, the reaction stops, resulting in many sequences of various lengths in each sample, all terminated at the same type of base. Finally, the fragments from all 4 parallel reactions can be sorted by the length in the process of electrophoresis, and the order of nucleotides can be read (Fig. 3.1). The Sanger method allowed sequencing the DNA sequences as long as 1000 nt. It was also the primary method used in the Human Genome Project [24], which was run from 1990 to 2003, and resulted in the sequencing of about 92% of the human genome [55].

3.1.2 Next Generation Sequencing

The next breakthrough came with the development of Next Generation Sequencing (NGS), also called High-Throughput Sequencing (HTS)[51]. NGS methods are a massively parallel approach, enabling the simultaneous sequencing of millions of sequences, which drastically decreased the time and cost of sequencing. The Human Genome Project took 13 years to complete and cost $2.7 billion [119]; the introduction of NGS enabled sequencing of the human genome within a day, and decreased its price to approximately $400 (Fig. 3.2) [127].

In the last years, Illumina became the major provider of NGS [51]. In Illumina's method, the DNA is fragmented into smaller pieces, synthesized with the adapter sequences, amplified in the Polymerase Chain Reaction (PCR), and bound to the flow cells. Next, the sequences are amplified in the process called 'bridge PCR' to create clusters of identical sequences, spatially separated from other clusters. Finally, the fragments are

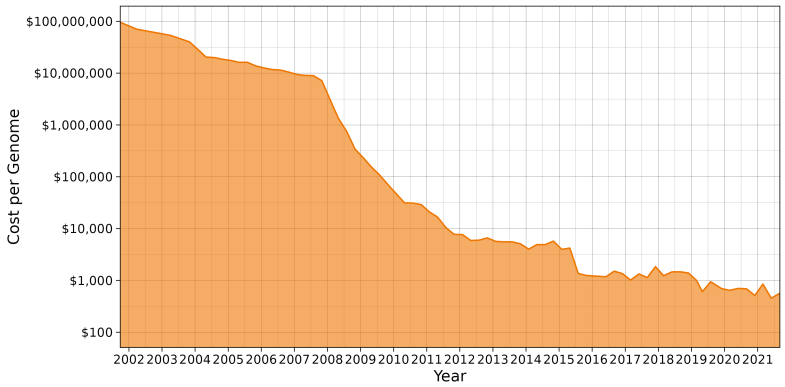


Figure 3.2: Sequencing cost of one human genome. Data: *www.genome.gov* [127]

.

synchronously sequenced by the synthesis of dNTPs with the fluorescent reversible terminators. In each cycle, the new dNTP is synthesized and scanned to recognize the base, before the terminator is removed to enable the synthesis of another dNTP. NGS usually allows the sequencing of up to several hundreds of nucleotides from one or both ends of the sequence. These sequences, known as reads, are then aligned to the reference genome, or assembled de novo to create a complete genome sequence.

Several types of NGS can be distinguished:

Whole Genome Sequencing (WGS) . It is the most general type of sequencing, in which the entire organism's genome undergoes sequencing. It allows the detection the mutations in all coding and non-coding sequences and provides a quite uniform read coverage along the sequence. The latter supports the CNV analysis since the amplified (gained) regions will show higher read coverage when aligned to the reference genome. Similarly, the coverage of lost sequences is lower than the average coverage of the genome.

Whole Exome Sequencing (WXS, or WES). This type of NGS is focused on the protein-coding sequence only, which constitutes 1-2% of the genome. This reduces the price of single-sample sequencing and allows to increase in the number of samples or the depth of sequencing - a number of reads covering each targeted region - to detect the rare mutations. However, it also requires an additional step of the target regions' capture. It makes the sequencing coverage less uniform due to sequence-specific biases and technological limitations [8], which makes the CNV analysis less accurate.

Target sequencing. Target sequencing is the most restricted type of NGS, in which only the selected regions, such as a set of genes of interest, are subjected to sequencing. Similarly to the WXS, target sequencing lowers the costs of sequencing a single sample., but requires the step of the target regions' capture.

RNA sequencing. NGS methods can be used to sequence the RNA as well. In RNA sequencing the main purpose of the experiment is the quantification of the gene expression levels. Although the variant-detection is still possible, it strongly depends on the expression of a given gene, and lowly expressed genes may be covered by an insufficient number of reads to perform the detection of variants.

3.1.3 Third generation DNA sequencing

The third-generation sequencing methods overcome the main limitations of NGS: the short read length, and the necessity of DNA amplification. One of the most popular third-generation methods is the Single Molecule, Real-Time (SMRT) sequencing by Pacific Biosciences [51, 98]. In SMRT, special hairpin adapters are synthesized to both ends of the double-stranded DNA, resulting in a library of single-stranded circular DNA templates. Then the primers and the polymerases are added, and the libraries are placed in special wells, called zero-mode waveguides (ZMW). There the sequencing by synthesis takes place, using the fluorescently labeled dNTPs. Since the diameter of ZMW is smaller than the wavelength of the laser light used to excite the fluorescent dNTPs, it is possible to excite only the one dNTP that is being synthesized. This allows to track the sequencing of particular molecules in real-time. Long reads provided by the third-generation sequencers make the proper assembly of highly repetitive regions of the genome possible, and the PacBio solution was used by the Telomere-to-Telomere (T2T) Consortium in the first complete, telomere-to-telomere assembly of the human genome [95].

3.1.4 Limitations of bulk sequencing

Bulk sequencing methods such as the Sanger methods and NGS provide information on the genome sequence which is averaged over the millions of cells from which the DNA was extracted and sequenced. When the reads are aligned to the reference genome (or assembled de novo), one can find the mutation sites by comparing the reads to the reference genome, or the sequence of the control sample. The ratio of the numbers of reads supporting the alternate and the reference alleles provides the statistics called Variant Allele Frequency (VAF), associated with the Mutated Cells Frequency (MCF) (see Section 3.2.2). However, bulk sequencing does not allow for determining which variants co-occur in the same sub-populations of cells. This information, necessary for the accurate reconstruction of the tumor clonal structure [113], is lost when the DNA from the population of cells is mixed and sequenced together. Clusters of mutations with similar VAFs may not be singular clones, but a mixture of clones with similar MCF, which cannot be separated based on the bulk sequencing results. Also, the proper recognition of cancer clones and subclones is further complicated by tumor purity issues (contamination of the tumor sample with the normal cells), and copy number changes [113], which can elevate or

decrease the observed VAFs .

3.1.5 Single Cell Sequencing

The limitations of bulk sequencing have led to the development of single-cell sequencing methods, alongside third-generation single-molecule sequencing methods. Most single-cell sequencing methods start with the isolation of the individual cells, followed by the barcoding of all cell DNA (or RNA) with the unique barcode sequence. The sequences can then be pooled (multiplexed) and sequenced together using NGS methods. The barcodes allow in silico separation of the sequences from different cells. The earlier the multiplexing is performed in the library preparation process, the less work needs to be done on each cell separately.

Single-cell RNA sequencing (scRNAseq) has rapidly become a popular method, enabling sequencing of RNA from millions of single cells after less than 10 years of scRNAseq techniques development [114]. In contrast, the development of single-cell DNA sequencing (scDNAseq) progressed more slowly. It requires an additional step of whole genome amplification (WGA), as there are only two copies of each DNA fragment present in each cell (in the diploid genomes), contrary to abundant RNA transcripts from actively expressed genes. All the biases and errors introduced by WGA affect the analysis of the sequencing results. Uneven sequence amplification complicates theCNV analysis. Sequencing errors (false positive mutations) that occurred during the WGA are hard to distinguish from the true variants [80], and many true variants can get lost when a sequence is not amplified [28].

Several different amplification methods have become particularly popular. The degenerate oligonucleotide-primed PCR (DOP-PCR) [118] is one of the oldest ones, and it has been successfully used in some CNV-focused studies [39, 19]. However, due to the large number of introduced false positives, it does not apply to the SNV-focused studies. Another method, the Multiple Displacement Amplification (MDA) [74], offers an accurate sequence amplification and can be used in the studies focused on SNVs, but the sequence amplification is not uniform enough for the CNV analysis.

Two more recent methods, the Multiple Annealing and Looping-Based Amplification Cycles (MALBAC) [136], and Linear Amplification via Transposon Insertion (LIANTI) [21] offer a more complete, accurate, and uniform genome amplification [125], and their popularization may greatly improve the quality of the scDNAseq data.

The final limitation of the scDNAseq is the low number of cells sequenced. The cell selection bias creates a risk, that the rare but important subclones are missed and undetected in the study [113].

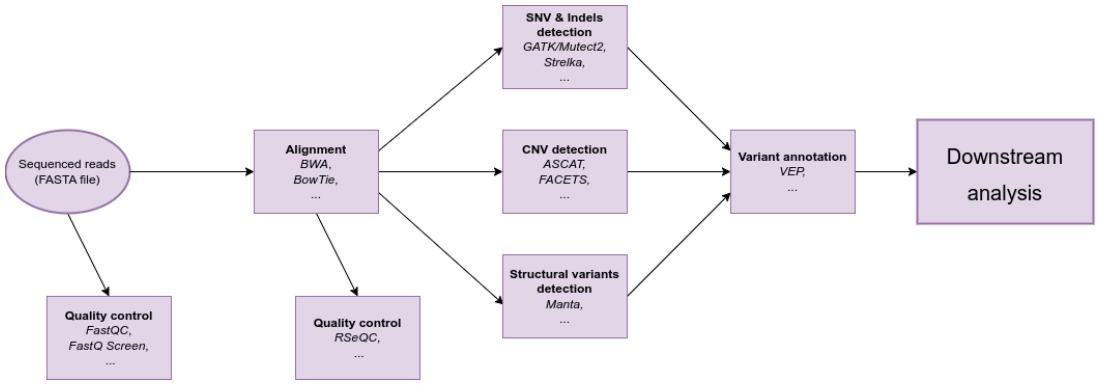


Figure 3.3: A generalized scheme of the *in silico* NGS data analysis.

3.2 NGS data analysis

The results of the DNA sequencing undergo a multistep *in silico* analysis 3.3:

Quality control. First, the results are checked for possible data quality issues. FastQC [5] and FastQ Screen [132] are among the most popular tools for the quality control of the NGS data. FastQC detects the problems such as low sequence quality scores, a high number of PCR duplicates, unexpected nucleotide composition, and others. FastQ Screen, in turn, was designed to detect the possible contamination of the sample with the foreign DNA (eg. bacteria).

Alignment. In the next step, the reads are aligned to the reference genome or assembled de novo. The alignment is a challenging process due to the huge number of reads to align, the short read length (50 to several hundred), the large genome size, and the presence of repetitive sequences, polymorphisms, and mutations. Several specialized aligners have been developed, such as BWA [78] and Bowtie [73]. The alignment may be followed by another step of quality control. Many tools from the RSeQC toolkit [123] are to be run on the aligned sequencing reads and can detect issues such as the short input sequencing size or the uneven coverage of the gene sequence.

Variant Calling (DNAseq). The aligned reads can be subjected to the process of variant calling. Algorithms such as GATK/Mutect2 [29], Strelka2 [62], or VarScan2 [65] can be used to compare pairs of tumor and normal samples, or tumor samples against the pool of normals, to detect the somatic SNVs and Indels present in the tumor. ASCAT [121] and FACETS [106] are popular choices for detecting CNVs in WGS and WXS data, respectively. Manta [22] is an example of software that offers structural variant detection. Variant calling algorithms try reliably distinguish the true algorithms from the sequencing errors and noise.

Variant annotation. Finally, algorithms such as the Variant Effect Predictor [84] can be used to predict the effects of variants. Prediction tools use the information on the structure and location of protein-coding and regulatory elements of the genome. They may also use tools such as PolyPhen [1] to predict the possible impact of amino acid substitution on the function of the protein and annotate the known variants with their frequencies from, for example. the 1000 Genomes Project [6].

The results of variant calling can be used in many different types of analysis, such as the analysis of mutational signatures, identification of driver mutations, reconstruction of the clonal structure, and others. The approaches focused on the main purpose of this thesis, the evaluation of the role of selection and mutagenesis will be described in Section 3.3.

3.2.1 Reference genomes

The accuracy of the reference genome is a factor that limits the accuracy of most if not all, the genome studies. Genome Reference Consortium is an international organization that assembles, improves, and releases reference genome assemblies. The main human genome assemblies are the GRCh37, released in 2009, and the GRch38, released in 2013, with its latest version patched in 2022 (GRCh38.p14). Both those releases are incomplete, with GRCh38 missing approximately 8% of the total genome length [95]. The Telomere-to-Telomere (T2T) genome, released by the Telomere-to-Telomere Consortium at the beginning of 2022 [95] is the first complete assembly of the human genome. It includes the genome regions unsequenced before, such as the centromeric regions or short arms of chromosomes 13, 14, 15, 21, and 22. The assembly of these regions has finally gotten possible with the development of the 3rd generation sequencing methods, providing long, continuous reads from the individual DNA molecules.

Although the T2T genome assembly has finally got available, it will take time until the bioinformatic databases get updated with the information for this new assembly. Also, many bioinformatic tools need to be updated to support the new genome. For this reason, the GRCh38 genome still is and will be used, in many ongoing studies.

Usage of an incomplete reference genome affects the accuracy of the variant detection [2]. Reads originating from the missing sequences, such as the unknown paralogs of the known genes, might be aligned to incorrect sites in the genome, increasing the number of false positive variants [95]. Also, the true variants present in the missing regions cannot be detected. This highlights the need for further updating of existing reference genomes.

3.2.2 Variant Allele Frequency (VAF)

The basic statistic that DNA sequencing provides for each mutation is the Variant Allele Frequency (VAF):

$$VAF = \frac{N_{alt}}{N_{alt} + N_{ref}} \tag{3.1}$$

where N_{alt} and $Nref$ are the numbers of reads supporting the alternate and reference alleles, respectively. The sum of N_{alt} and $Nref$ is the depth of variant sequencing:

$$DP = N_{alt} + N_{ref} \tag{3.2}$$

If we consider the bulk DNA sequencing in terms of drawing random sequences from the pool of DNA extracted from the wildtype and mutated cells, $_{alt}$ can be described by the binomial distribution:

$$N_{alt} \sim B(DP, p) \tag{3.3}$$

where p is the frequency of the mutated allele in the sample, equal to

$$p = \frac{MCF \cdot CN_{nut}}{MCF \cdot CN_{tot} \cdot (1 - MCF) \cdot CN_{norm}} \tag{3.4}$$

where MCF is the Mutated Cell Frequency, the fraction of cells in the sample that have the mutation, CN_{mut} is the number of copies of the mutant allele in the mutant cells, CN_{tot} is the total copy number in the mutated cells, and CN_{norm} is the ploidy of the normal cells. In the purely diploid population, where CN_{mut} is equal to 1, and CN_{tot} and CN_{norm} are equal to 2, p equals:

$$p = \frac{MCF}{2} \tag{3.5}$$

and

$$VAF \sim MCF/2 \tag{3.6}$$

Because of this, due to the complexity associated with the estimation of allele-specific copy numbers, many methods use VAF as an approximate measure of the MCF. In particular, VAF spectra, which represent the distributions of observed allelic frequencies for all variants in a sample, have been found useful in the modeling of neutrality and selection in cancer research.

3.3 Modelling of cancer growth and evolution

Mathematical modelling approaches are a powerful tool to test hypotheses and explain observations. Over the years many different types of models were applied in cancer research to understand the mechanisms underlying the initiation and progression of cancer, as well as to predict its future progression. In this section we will describe the selected models of tumor growth and evolution, including methods measuring the role of mutagenesis and selection.

3.3.1 Models of tumor growth

A number of models different models were proposed for the tumor growth. The simplest model of exponential tumor growth assumes constant growth rate, and infinite environment capacity. In this model number of cells in the population N in at time t can be described as:

$$N(t) = e^{\lambda t} \tag{3.7}$$

where λ is the growth rate. It was found to be the best fit for some breast cancers [115]. If the capacity of environment, or availability of nutrients is limited, population growth can be described with the logistic model:

$$N(t) = \frac{K}{1 + ae^{-b\lambda t}} \tag{3.8}$$

where M is the capacity of the environment, and b and c are the positive parameters, or the Gompertz model:

$$N(t) = ae^{-be^{-ct}} \tag{3.9}$$

where a, b, and c are the positive parameters of the model. In both those models, the growth rate decreases as the population size is approaching the upper limit. Both those models were found the best fit for other cases of breast cancers [111, 94]. If lack of nutrients availability and/or space limits the growth of cells mostly in the center of the tumor but cells on its surface grow freely (*surface-growth model*), population growth follows a power law [49]. Modes of tumor growth may also change depending on the stage of tumor development: slow at the beginning of cancer progression, accelerate after the vascularization (exponential law), and slow down again when the tumor becomes large (Gompertz/logistic model) [12]. Models of tumor population growth, particularly the exponential growth model, are key elements of some methods for modelling the neutrality and selection in the tumor evolution.

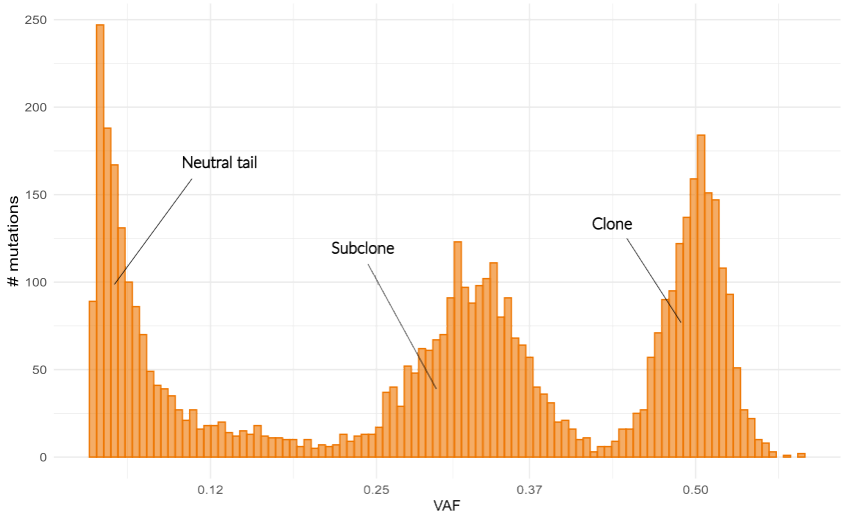


Figure 3.4: Example VAF spectrum

.

3.3.2 Modelling of selection and neutrality

The theory of neutral tumor evolution was proposed for the first time by Kimura, who found that the mutation rate of mammalian genomes would be too high to be tolerated by any species if all the mutations were non-neutral [63]. Kimura has proposed that the majority of mutations have a negligible effect on the organism, and many of them become fixed not by selection but by genetic drift. Kimura also has shown, that for selectively neutral sequences, the rate of substitutions per generation K is equal to the average mutation rate μ [63]:

$$K = 1/N * N\mu = \mu \tag{3.10}$$

where $1/N$ is the probability of fixation of a new variant in the population of size N, and $N\mu$ is the number of new variants occurring in the population. Since the advantageous mutations have a bigger chance of fixation, for sequences under the positive selection K is greater than μ. Similarly for the sequences under the negative selection, where most mutations would be disadvantageous and have a smaller chance of being fixed, K must be smaller than μ [36].

This property is being used by the methods that analyse the ratio of the nonsynonymous and synonymous mutations in sequences in order to detect the cancer driver genes, such as dNdScv [82].

3.3.3 VAF spectra and the neutral tails

Important methods detecting the selection and neutrality in tumor evolution are based on the VAF spectra. The spectra often present a multi-modal distribution of VAF in the

sample (Fig. 3.4).

High-frequency peaks usually consist of the clonal and subclonal variants, present in a fraction of cells two times higher than the mean VAF of variants in the peak (Eq. 3.6), if the genome is diploid. Peaks are called *clonal* if they represent the mutations of the Most Recent Common Ancestor (MRCA), which are present in all cancer cells in the sample. The mean VAF of those variants equals 0.5 in the pure cancer sample. However, if no selective sweep has occurred before, and the number of clonal mutations is low, this peak can be undetectable in the VAF spectrum. Subclonal peaks, in turn, consist of mutations possessed by the emerging subclones which have a selective advantage over the background of clonal cells. Those peaks can eventually merge with the clonal peaks (or create one) when the selective sweep is done. The random-sampling nature of bulk sequencing methods (Eq. 3.3) explains the binomial shape of the clonal and subclonal peaks.

The components of the spectrum with the lowest VAF are called the *neutral tails* and consist of the somatic mutations that occured at different times during the tumor growth. Most of those mutations are neutral or provide the very little selective advantage, and their MCF depends mostly on the number of tumor cells when the mutation occured. They may also contain the mutations of rare subclones whose low MCF results from the insufficient time that passed from the subclone initiation. Such subclones, indistinguishable from the neutral tail at present, can be responsible for the future progression, relapse, or therapy resistance [113].

3.3.4 Stochastic models of the neutral tail

A few models were proposed to describe the number of mutations and the shape of the neutral tail. Stochastic approaches utilize the Site Frequency Spectrum, an alternative to the VAF spectrum that counts the number of variants present in a given number of cells, contrary to the VAF spectrum utilizing the discrete intervals of VAF. In 2013, Durrett proposed a convenient approximation [37] of the earlier Griffiths and Tavaré [43] formula for the number of variants present in k-th bin of SFS. The formula uses the Infinite Sites Model (ISM), a statement that the number of sites where the mutation can occur is infinite, thus the same mutation is unlikely to occur twice. In Durrett's approximation, under the ISM and an assumption of exponential tumor growth, the number of variants $S(k)$ present in k cells is:

$$\mathbb{E}S_n(k) = \frac{\theta}{\lambda}\frac{n}{k(k-1)}, k = 2, ..., n-1 \tag{3.11}$$

and

$$\mathbb{E}S_n(1) \sim \frac{\theta n ln(\lambda N)}{\lambda} \tag{3.12}$$

where θ is the mutation rate, λ is the growth rate, n is the size of the sample, and N is the present population size.

In this thesis, we will call θ/λ the *reduced mutation rate* and denote it as μ.

3.3.5 Williams's test of neutrality.

Another approach was proposed by Williams et al. [130], who used the differential equation describing the number of new mutations that occur in exponentially growing tumor (Eq. 3.7) per unit time. Williams has shown that in neutrally evolving tumors the relationship between the number of mutations with VAF greater than f and the reciprocal of f is linear:

$$M(f) = \frac{\theta}{\beta}(\frac{1}{f} - \frac{1}{f_{max}}) \tag{3.13}$$

where θ is the mutation rate, β - fraction of successful cell divisions, and f_{max} - maximum f over which the model is fitted. The equation corresponds to the power-law distribution of neutral mutations in the mutation frequency spectrum, with the power coefficient equal to 2:

$$N(f) \sim \frac{\theta}{\beta}\frac{1}{f^2} \tag{3.14}$$

which is approximately equal to the Durrett's expression [37] (Section 3.3.4). Williams proposed an approach for the detection of selection in tumors that fits a linear model to the statistic described by Equation 3.13, and rejects the null hypothesis of the neutral evolution, if the R^2 metrics of the fit is lower than 0.98. This approach was implemented in an R package *neutralitytestr* and used to evaluate the mode of evolution of the tumors in The Cancer Genome Atlas (TCGA) Consortium [130]. Out of over 800 individual tumors, approximately 30% were found neutral, with the R^2 goodness-of-fit of the neutral model exceeding 0.98.

3.3.6 *neutralitytestr* model criticism

The results and the testing methodology of Williams [130] were met with significant criticism in the scientific community. Tarabichi et al. [117] have pointed 4 issues of the test, which were answered by Heide et al. [52]:

1. estimation of MCF requires the accurate estimation of local copy numbers and the purity of the sample, and the restriction of the test to the narrow range of VAF between 0.12 and 0.24, what the *neutralitytestr* does by default, is insufficient to

make the test robust to the copy number changes. Heide agreed that the proposed threshold is not universal and in some cases needs to be adjusted.

2. failure to reject the hypothesis of the neutral evolution does not prove that the null hypothesis is correct. While the statement is true, Heide claimed that the neutrality hypothesis is a reasonable null model in molecular evolution.

3. stochastic models of tumor growth are more realistic than the deterministic model used by Williams. However, Heide has shown that the stochastic simulations are not contradictory to their deterministic model, and the power-law shape of the neutral tail finds support the stochastic derivations [37].

4. *dNdScv* package [82], which detects the selection using the ratio of non-synonymous and synonymous mutations at the population-level, was able to find the significant positive selection in the tumors classified by *neutralitytestr* as neutral. Heide has shown, that *dNdScv* results can be due to the misclassification of particular samples and that both tools can complement each other in the evolutionary studies at the population and sample level.

Few other papers discussed also other aspects of the method proposed by Williams. McDonald et al. [83] have simulated a number of neutral and selective tumors and have shown, that the R^2 values of the linear $M(f) \sim 1/f$ fit were in many cases similar. For this reason, they argue that the linearity of this relationship cannot be used to distinguish the selective tumors from the neutral ones. In other paper, Bozic et al. [11] has shown that the time between the emergence of new selected subclone until it gets fixed in the population is short. In their simulations, they considered a two-type model of tumor that starts from a single transformed cells, and grows following the branching process with the birth rate b, death rate d, and growth rate $\lambda = b - d$ Cells in this model can gain a driver mutation at rate μ to become the type 1. Type 1 cells have the birth rate b_1, death rate d_1, and the growth rate $\lambda_1 = b_1 - d_1$, so that the $\lambda_1 > \lambda$. Bozic et al. have shown, that in most cases the frequency of the driver mutations is biased towards 0 or 1, and the probability of observing it at intermediate frequencies was in many cases below 30%, and never exceeded 60%.

3.3.7 Williams's model improvements

Williams et al. improved their modelling approach in the following years. New model was fitted to the entire VAF spectrum, instead of the $M(f) \sim 1/f$ statistic. It also consisted of the neutral power-law component complemented with the binomial components corresponding to the peaks of clonal and subclonal mutations, diluted by random sampling (see Eq. 3.3):

$$N(f) \sim \frac{A}{f^2} + \sum_{i=1}^{i=K} N_k \cdot binomial(n, f_k) \tag{3.15}$$

where
K - number of clones and subclones
N_k - number of mutations in (sub)clone k
f_k - true allelic frequency of mutations in (sub)clone k
A - constant proportional to the effective mutation rate defined in [130].

This improved model was first implemented by Williams in 2018 as Julia package SubClonalSelection.jl, using the computationally expensive and time consuming Bayesian approach [131] instead of previously criticized R^2 statistic. Two years later Williams and Caravagna developed much faster, machine learning based approach and implemented it in R package MOBSTER [18]. Both these packages fit the data with a number of neutral and selective models and evaluate them using the Bayesian Information Criterium (BIC), which addreses the second of the issues raised by Tarabichi [117] in response to *neutralitytestr* [130].

Both these packages also provide the equations to estimate the evolutionary parameters of fitted subclones: the emergence times and the selection coefficients. The selection coefficient s is defined as:

$$s = \frac{\lambda_s}{\lambda_c} - 1 \tag{3.16}$$

where λ_s is the subclone growth rate, and λ_c is the growth rate of the ancestral clone. Williams shows [131], that s can be calculated as:

$$s = \frac{\lambda_c t_1 + ln\left(\frac{f_{sub}}{1-f_{sub}}\right)}{\lambda_c(t_{end} - t_1)} \tag{3.17}$$

where f_{sub}, the subclone cell fraction, can be estimated from the parameters of the fitted binomial distribution, t_{end}, tumor age in population doublings, can be derived from the final population size N_{end}:

$$t_{end} = ln(1 - f_{sub} \times N_{end}) \tag{3.18}$$

and t_1, the subclone emergence time, can be calculated from the number of subclonal mutations N_s and tumor mutation rate:

$$t_1 = \frac{N_s}{2log(2) \times \mu} \tag{3.19}$$

3.3.8 Tung and Durrett's two-type model and the selection of micro-clones

The two-type model simulated by Bozic et al. [11] was further investigated by Tung and Durrett in their work published in 2021 [120]. They presented a mathematical proof for the observation of Bozic that the probability of observing the diver mutations after they emerge from the neutral tail and before they get fixed in the population is low. However, they also showed that the existence of the selected micro-clones in the tumor tail alters its shape. In the considered two-type model, cells in the host population (type 0) proliferate at rate λ_0, accumulate neutral mutations at rate μ, and mutate to type 1 at rate v. Type 1 cells proliferate at $\lambda_1 > \lambda_0$ and accumulate neutral mutations at the same rate μ. They proved that the power-law tail in this model is described by the equation:

$$SFS(f) = \frac{C}{f^{\alpha}} \tag{3.20}$$

where f is the variant frequency, and C is a positive constant, and α depends on the ratio of the types' growth rates:

$$\alpha = \frac{\lambda_0}{\lambda_1} + 1 \tag{3.21}$$

If λ_0 and λ_1 are equal, α equals 2, as in Williams [130] and Durrett [37]. However, if $\lambda_1 > \lambda_0$, the selection among micro-clones is manifested by $\alpha < 2$.

The presence of the selected micro-clones is not the only phenomenon that influences the power coefficient α, though. In Section 4.3, we describe another model in which the mutation rate is not constant, leading to α different from 2.

3.3.9 Reconstruction of the clonal tumor structure

In addition to the development of population genetics-based methods for estimating evolutionary dynamics parameters, a group of algorithms has emerged for reconstructing clonal tumor structure and phylogeny. Some of the best-known algorithms in this group are SciClone [88], PyClone [99], PhyloWGS [31], and DPclust [92]. The primary goal of these algorithms is to identify subpopulations of cells with shared genotypes in NGS data. To achieve this, they combine VAF information for short variants (SNVs, Indels) with copy number data from CNV callers, clustering mutations with similar cellular frequencies. Typically, these algorithms can simultaneously analyze multiple samples from the same tumor to increase the resolution. In addition, PhyloWGS also determines the phylogenetic structure of the tumor.

In our work [68], we compared the results of different combinations and settings of tools used in such analyses. We used 2 different CNV callers: FACETS [106] and TitanCNA

[46], and 2 algorithms for clonal structure reconstruction: PyClone [99] and PhyloWGS [31]. We also ran the analysis with and without the tumor purity estimates provided by CNV callers. We analyzed a subset of the data used in this thesis with all 8 combinations of these 3 elements (CNV caller, reconstruction algorithm, purity) and found very high variability in the obtained results. The differences were generated at each step of the analysis: FACETS and TitanCNA provided different estimates of purity and ploidy and varying sets of CNV calls; the numbers of clones identified by the pipelines were highly variable; there were substantial differences in mutation assignment to clones.

Although these tools can be useful, the quality of their results depends on the quality of the input data. Usually, they do not distinguish between the true clones and the neutral tail, classifying the low-VAF variants as a single subclone with low cellular frequency. It is an oversimplification since the mutations occur in all cells, and bulk DNA sequencing does not allow distinguishing whether the rare variants coexist in the same subset of cells or if they occurred in different cell lineages. As Caravagna et al. show, identification of the correct tumor phylogenies requires the removal of neutral tail variants, which can be achieved using population-based algorithms, such as MOBSTER [18].

3.3.10 Summary

The emergence of the Next Generation Sequencing methods was a breakthrough in cancer research. The popularization of NGS methods, which followed the decrease of its price, provided data that enabled the studies of the dysregulated molecular mechanisms in cancer and tumor evolution. One fundamental statistic the DNA sequencing provides for each detected variant is its allelic frequency VAF, related to the frequency of mutated cells (MCF) in the sample and frequently treated as a proxy measure of it. As Durrett [37], Williams [130], and others have shown, the spectra of MCF (and its proxy measure, VAF) reflect the evolutionary dynamics of tumors and allow us to estimate the parameters of tumor evolution, such as the effective/reduced mutation rates, subclonal selection coefficients, and subclone emergence times.

In this thesis, we analyze the VAF spectra of the primary and secondary tumor samples and apply the model consisting of a mixture of power-law shaped and binomial components (Equation 3.15) to study the evolutionary dynamics in these tumors. Whereas the power-law component has a power coefficient α equal to 2 in this model, certain biological processes, such as the presence of competing micro-clones or varying mutation rates, can lead to different values of α (Sections 3.3.8 and 4.3). To investigate such cases, we implement a second type of model, in which α is not fixed, but we optimize it to find the value that most accurately describes the observed data.

In the following chapter, we introduce the datasets utilized in this work and outline the methods employed.

Chapter 4

Data and methods

4.1 Data

To address our thesis, we collected the Next Generation Sequencing data from 4 different cancer types, including the data from at least two tumor samples from each patient. We analyzed the evolutionary dynamics of recurrent cancers on the example of Acute Myeloid Leukaemia, using the whole genome sequencing data from the study of Shlush [107]. Evolutionary dynamics of metastatic cancers was investigated on example of breast and laryngeal cancers, using the whole exome sequencing data obtained from our experiments. Finally, the evolutionary dynamics of cancer during the tumor progression was studied on two specimens of bladder cancer, using the whole exome sequencing data of many samples with different stages of disease progression.

Breast Cancer and Larynx Cancer Cohorts

Breast Cancer (BRCA) and Larynx Cancer (LSCC) cohorts were based on the National Science Center-funded grant *A systems approach to cancer progression and prognosis: New models and statistics for genomic data analysis*, grant no. 2018/29/B/ST7/02550. BRCA cohort consists of 30 tumor samples collected from 15 female breast cancer patients, along with 15 control samples from the healthy tissue. To analyze the mechanisms responsible for the cancer metastasis, two tumor samples were collected from each patient: one sample from the primary tumor and one from the local lymph node metastasis.

LSCC cohort included sequencing data from 12 patients with laryngeal cancer. Again, three samples per patient were collected: one sample from the primary tumor, one metastatic sample, and one control sample.

All samples were subjected to Whole Exome Sequencing (WXS) to detect short variants (SNVs and Indels) in the protein-coding sequence, with targeted coverage 100x. Data on molecular cancer subtypes and patient's age and sex were included in Table 4.1. Histopathological tumor purity estimates were available only for a fraction of BRCA samples

	cohort	patient ID	molecular subtype	sex	age	samples
1	AML	A-1		F	35	Dx, Rx
2	AML	A-2		M	75	Dx, Rx
3	AML	A-3		F	71	Dx, Rx
4	AML	A-4		M	43	Dx, Rx
5	AML	A-6		M	60	Dx, Rx
6	AML	A-8		M	61	Dx, Rx
7	AML	A-9		M	60	Dx, Rx
8	AML	A-10		F	49	Dx, Rx
9	AML	A-11		F	27	Dx, Rx
10	AML	A-12		M	75	Dx, Rx
11	AML	A-15		M	62	Dx, Rx
12	BRCA	G02	HER2+	F		P1, L1
13	BRCA	G04	HER2+	F		P1, L1
14	BRCA	G30	HER2+	F		P1, L1
15	BRCA	G31	HER2+	F		P1, L1
16	BRCA	G32	TNBC	F		P1, L1
17	BRCA	G33	Luminal A	F		P1, L1
18	BRCA	G35	TNBC	F		P1, L1
19	BRCA	G36	Luminal A	F		P1, L1
20	BRCA	G40	Luminal A	F		P1, L1
21	BRCA	G41	Luminal A	F		P1, L1
22	BRCA	G43	Luminal A	F		P1, L1
23	BRCA	G45	Luminal A	F		P1, L1
24	BRCA	G46	Luminal A	F		P1, L1
25	BRCA	G47	Luminal A	F		P1, L1
26	BRCA	G48	Luminal A	F		P1, L1
27	LSCC	L01		M	48	P1, L1
28	LSCC	L03		M	62	P1, L1
29	LSCC	L04		M	23	P1, L1
30	LSCC	L05		M	59	P1, L1
31	LSCC	L07		F	68	P1, L1
32	LSCC	L10		M	57	P1, L1
33	LSCC	L14		M	43	P1, L1
34	LSCC	L15		M	68	P1, L1
35	LSCC	L16		M	53	P1, L1
36	LSCC	L19		M	63	P1, L1
37	LSCC	L20		M	55	P1, L1
38	LSCC	L22		M	62	P1, L1

Table 4.1: List of patients and samples in AML, BRCA, and LSCC cohorts. Two tumor samples were obtained and sequenced from each patient. *Dx* - diagnostic sample, *Rx* - relapse sample, *P1* - primary tumor sample, *L1* - lymph node metastasis sample, *F* - female, *M* - male.

Patient ID	sample	Purity [%]	molecular subtype
G30	L1	90.00	HER2+
G30	P1	90.00	HER2+
G31	L1	95.00	HER2+
G31	P1	80.00	HER2+
G32	L1	85.00	TNBC
G32	P1	80.00	TNBC
G35	L1	90.00	TNBC
G35	P1	90.00	TNBC
G40	L1	90.00	Luminal A
G40	P1	90.00	Luminal A
G45	L1	90.00	Luminal A
G45	P1	90.00	Luminal A
G46	L1	95.00	Luminal A
G46	P1	95.00	Luminal A
G48	L1	90.00	Luminal A
G48	P1	85.00	Luminal A

Table 4.2: Histopathological estimates of tumor purity in BRCA cohort. *P1* - primary tumor sample, *L1* - lymph node metastasis sample

(Table 4.2) and indicated high purity of tumor samples.

Acute Myleoid Leukaemia Cohort

Acute Myleoid Leukaemia dataset was obtained from the published study of Shlush [107]. Cohort includes 11 leukaemia patients, including 7 men and 4 women, from whom total number of 33 samples was collected, 3 samples per each patient: one sample at the diagnosis, one sample at relapse, and one control sample. All samples were subjected to whole genome sequencing, with the average sequencing depth of 100x. Sequencing results were downloaded as GRCh37-aligned BAM files from the EGA (ID: EGAD00001003234). Data on age and sex of patients in AML cohort was included in Table 4.1.

Bladder Cancer

Bladder Cancer (BLCA) data used in this work was obtained from the laboratory of Dr. Bogdan Czerniak at the MD Anderson Cancer Center in Houston, TX, and used in the study by Bondaruk et al. *The origin of bladder cancer from mucosal field effect* [10]. Two bladder cystectomy specimens no. 19 and 24 were selected to track the development of bladder cancer from the bladder mucosa, representing the basal and luminal tumors, respectively. Both specimens were opened along the anterior wall and divided into 1 x 2 cm areas of the mucosa, which were classified into four groups: the normal urothelium (NU), the low-grade intraurothelial neoplasia (LGIN), high-grade intraurothelial neoplasia (HGIN), and urothelial carcinoma (UC). Selected regions were subjected to multi-omic

experiments, including WXS, RNA sequencing, whole-genome methylation array hybridization, and whole genome polymorphism-based copy number analysis. In this work, we utilize the WXS data. Counts of samples subjected to WXS, divided by patient, molecular subtype, and sample classification, are presented in Table 4.3.

patient ID	molecular subtype	group	# samples
map19	basal	NU	9
map19	basal	LGIN	16
map19	basal	HGIN	2
map19	basal	UC	2
map24	luminal	NU	22
map24	luminal	LGIN	10
map24	luminal	HGIN	2
map24	luminal	UC	3

Table 4.3: Counts of sequenced samples in BLCA cohort by patient, molecular subtype, and sample classification.

4.2 Methods

4.2.1 Processing of NGS data

BRCA and LSCC cohorts. Quality control was conducted using FastQC and FastQ Screen. Raw reads were aligned to the GRCh38 reference genome using the BWA-MEM (v0.7.17) [78] in the alternative contigs aware mode.

AML, BRCA, and LSCC cohorts. All aligned reads were processed using MarkDuplicates algorithm from the Picard tool set and BaseRecalibrator which is a part of the Genome Analysis Toolkit (GATK v4.2.6.1) [29]. Somatic mutations were identified using MuTect2 (v4.2.6.1) [29] based on tumor-normal sample pairs. Variants were filtered using GATK's FilterMutectCalls based on MuTect2 results, as well as sample contamination estimates obtained using CalculateContamination tool and read orientation bias statistics obtained with LearnReadOrientationModel tool. The retained variants were annotated using Variant Effect Predictor (v107) [84]. Finally, we filtered out the lowest-coverage variants, whose coverage did not exceed 10 reads in any sample (18.5k variants, 2.16% of all variants).

BLCA cohort. All the BLCA cohort analyses in this thesis used the VCF files with sets of SNVs and Indels used in the paper of Bondaruk et al. [10]. The process of NGS raw data processing was similar to our pipeline described above: raw reads were aligned to the GRCh38 genome using BWA-MEM (v0.7.12), GATK and Mutect2 (v3.4.46) were used

to prepare the BAM files and for variant calling, and Oncotator (v1.8.0.0) was used to annotate the variants. Finally, we filtered out the variants covered by less than 25 reads.

4.2.2 Statistical analysis

Statistical analyses included in the thesis and described in the sections below were conducted in R v4.2.1 and Python v3.10.9.

4.2.3 Intra-Tumor Heterogeneity (ITH) measure

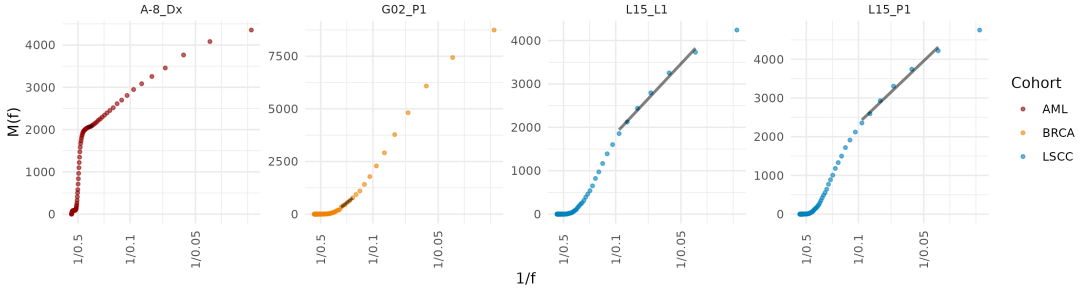
Intra-tumor heterogeneity (ITH) in AML, BRCA, and LSCC cohorts was assessed using the Jaccard Index (JI). JI equals to the number of elements in the intersection of two sets divided by the number elements in their union. For each patient, we calculated JI using the sets of variants detected in the pair of tumor samples obtained from the patient. The index is, therefore, negatively correlated with the ITH; the lower the index, the higher ITH.

4.2.4 MOBSTER model fitting

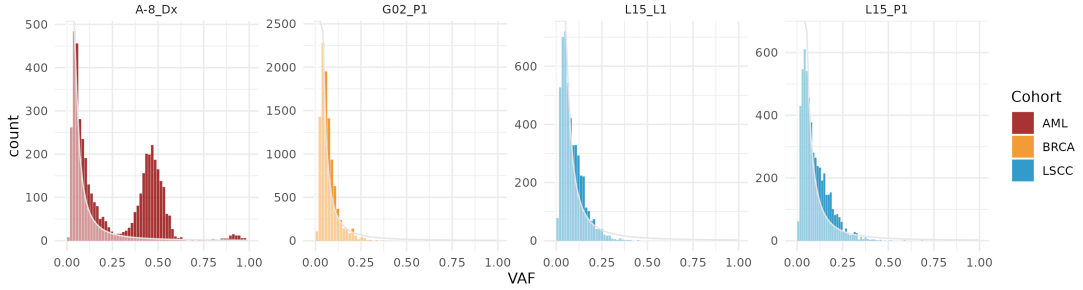
MOBSTER models were fitted using *MOBSTER* package v1.0.0 with the default parameters. The *auto_setup* option was left unset in order to conduct a more exhaustive analysis compared to the alternative, 'FAST' auto-setup option.

4.2.5 cevomod model fitting

We implemented our model fitting approaches in an R package cevomod (see more details in Section 4.4.1). Methods to fit two types of models were implemented: in the first model, we assume the neutral shape of the power-law component ($\alpha = 2$) and use Williams $M(f) \sim 1/f$ statistics to find the optimum mutation rate. In the second model, we use an optimization algorithm to fit both parameters: α and the mutation rate. Both methods are robust to the incompleteness of the neutral tail data due to variant filtration. Proving the presence of selection requires the rejection of the hypothesis of neutral evolution. For this reason, we fit the power-law components and binomial components sequentially. The power-law components are fitted first, maximizing their contribution to the model. Then, the binomial components are fitted to the residuals of the power-law components. Details on fitting the power-law and binomial components are presented in Sections 4.2.6, 4.2.7, and 4.2.8.



(a) *cevomod* fits $M(f) \sim 1/f$ relationship with a number of linear models in order to detect the minimal slope of the line.



(b) Identified minimum effective mutation rates can be used to calculate the power-law-distributed neutral tails in the VAF spectra.

Figure 4.1: Neutral model fitting in *cevomod*

4.2.6 Fitting neutral power-law component with *cevomod*

In the first type of models, the power coefficient α of the power-law component equals 2, and only the mutation rate μ needs to be estimated:

$$y(f) \sim \frac{\mu}{f^2}$$

Williams et al. have shown that this parameter equals the slope of the linear relationship between the $M(f)$ - number of mutations with VAF higher than f and the reciprocal of f [130]. In models with selection, this relationship is not linear, but (1) the slope of the curve at $1/f$ corresponds to the μ of a power-law component tangent to the VAF spectra at frequency f, and (2) the minimum slope of the $M(f) \sim 1/f$ corresponds to μ of the power-law component that does not detach from the spectra at any VAF. It is, therefore, the upper limit of μ, maximizing the contribution of the power-law component in the model. μ higher than in (2) results in intervals of VAF with fewer mutations than the model predicts.

In *cevomod*, we use principle (2) to fit the maximal potential power-law component to the spectra. The model is fitted in a few steps:

1. Shrink the VAF spectra by removal of 5% of variants from both ends of the spectra.

This excludes from the process the lowest VAFs, at which most mutations were lost due to insufficient variant support, and the highest VAFs, where the VAF spectrum is nearly flat due to the low count of variants.

2. Calculate the $M(f) \sim 1/f$ statistic for discrete f values rounded to two decimal places
3. Fit linear models to 0.05-wide sections of the statistic
4. Filter out the non-linear fits with $R^2 < 0.98$
5. Pick the model with the smallest slope, which is the final estimate of the μ
6. Power-law curves are calculated using the formula:

$$y(f) = \frac{A}{f^2}$$

where $A = \frac{\mu}{n}$ and n is the number of bins in the spectrum, equal to the median coverage of variants in the sample

Examples of the final linear $M(f) \sim 1/f$ fits with the corresponding power-law fits are shown the Figure 4.1.

4.2.7 Fitting the models with the best-fitting power coefficient

In our second type of models, the power coefficient α is also optimized, so the power-law component is described by:

$$y(f) = \frac{A}{f^\alpha}$$

We designed an optimization-based approach that finds both optimal parameters simultaneously, despite the dropout of low-frequency mutations. The optimization follows the three rules:

- the power-law can predict fewer mutations than exist in the VAF spectrum since it does not model the clones and subclones, but the count of low-VAF mutations (neutral tail) under the power-law curve is maximized
- the counts of mutations in the spectrum after the first peak (the neutral tail peak) constitute the upper bound for the power-law predictions. The power-law component cannot predict more mutations than we observe in the data, but
- the power-law curve should ignore the deficiency of variants before the first peak since they are largely lost during variant filtration

We prepare the data in two steps. First, we shrink the VAF spectra to cut off the bins separated from the main body of the spectrum by more than 2 empty bins (the bin is considered as 'empty' if it contains less than 1% of mutations of the highest peak). These

bins usually result from the data's noisiness or loss of heterozygosity. Then, we smooth the spectrum using the *stats::filter()* function with a vector of weights '*c(1/3, 1/3, 1/3)*'.

Next, the process of optimization is run multiple times using the *stats::optim()* function with the 'control' parameter set by default to '*list(maxit = 1000, ndeps = c(0.1, 0.01))*', a grid of initial α and A values: '*c(0.8, 1.2, 1.8, 2.5, 3.5)*' for α and '*c(1, 2, 4, 8, 16, 32)*' for A, and the minimized performance function equal to $-I$, where I is the difference between the mutation count reward (MCR) and the spectrum detach penalty (SDP).

$$I = MCR - SDP \tag{4.1}$$

The reward component of I is responsible for pushing the curve up and maximization of the number of mutations in the spectrum that lie under the power-law curve between the boundary index values i_{min} and i_{max}:

$$MCR = \sum_{i=i_{min}}^{i_{max}} \min\Big(S_i, y_i\Big) \tag{4.2}$$

where:
i - index of bin in the VAF spectrum
S_i - number of mutations in the i-th bin of spectrum
y_i - value of the power-law component for the i-th bin

and

$$i_{\min} = \max \begin{cases} \min i : y_i \leq S_i \\ \arg\max_i S_i \end{cases}$$

and

$$i_{\max} = \min \begin{cases} i : f_i < 0.4 \\ \text{number of bins in the spectrum} \end{cases}$$

The second component of I, the detach penalty, minimizes the number of mutations predicted by the power-law component, but not observed in the data. In other words, it does not allow the curve to detach significantly from the spectrum. It is defined as:

$$SDP = \sum_{i=i_{min}}^{N} \Big[|y_i - S_i| \cdot w_i\Big]$$

where:
i - index of bin in the VAF spectrum
N - number of bins in the VAF spectrum
S_i - number of mutations in the i-th bin of spectrum

y_i - value of the power-law component for the i-th bin
w_i - weight of the penalty for the i-th bin, equal to 0 if $y_i < S_i$, and to the length of the detached segment of y otherwise

and

$$i_{\min} = \max \begin{cases} \min\{i : y_i \leq S_i\} \\ \arg\max_i S_i \end{cases}$$

This definition of the SDP allows the curve to detach from the spectrum on specific bins, but as the length of the detached segment increases, the penalty grows dramatically. Finally, the solution with the minimal I is marked as the best one, and all the solutions are stored in the *cevodata* object.

4.2.8 Fitting the binomial components

In *cevomod*, binomial components for clonal and subclonal variants are fitted to the positive part of the power-law model residuals. We implemented two methods for fitting the binomial components. By default, we randomly subsample the SNVs and Indels in each spectrum bin to the number given by the power-law component residual. Then, we employ the *BMix* package [18] to fit the VAF distribution of these variants with a mixture of 1 to 3 binomial distributions (clone plus subclones), accounting for the variant's sequencing depth. The best model is selected based on the Bayesian Information Criterium (BIC).

In an alternative, approximate method, we generate artificial vector VAF values according to the power-law component residuals. We then use a more popular *mclust* package [103] to cluster them into 1 to 3 clones using Gaussian model-based clustering. Then, the binomial components are constructed using the number of variants in each cluster, their mean VAFs, and the median sequencing coverage of true variants with a given VAF. Finally, we remove the solutions with overlapping subclones and select the best one using BIC. Although this alternative method works approximately 3-4 times faster, the clustering step relies on the Gaussian distributions instead of the binomial ones and it does not use the true sequencing depths, resulting in a more approximate outcome. Therefore, we recommend using the default method for more accurate results.

4.2.9 Plotting

Most of the figures included in this thesis were prepared using R v4.2.1 and the following packages: *ggplot2* v3.4.0 [128], *patchwork* v1.1.2 [97], and *cowplot* v1.1.1. Heatmaps, including the oncoplot, were plotted using *ComplexHeatmap* v2.12.1 [44]. p-values from the statistical tests were annotated using *ggpubr* v0.5.0 [60]. A number of functions to

plot VAF spectra, model fits and residuals are included in our package *cevomod* described in details in section 4.4.1.

4.3 Mutation rate changes and the power-law exponent

It can be shown that the growing mutation rate results in the increased α coefficient of the power-law component. Consider an exponentially growing population with a growth rate λ, an initial mutation rate μ, and all mutations being neutral. Let $M(t)$ represent the number of mutations that have occurred up to time t. We assume that the rate at which mutations accumulate, $M'(t)$, depends on the population size with an additional coefficient κ:

$$M'(t) = \mu\lambda N(t)^{\kappa} \tag{4.3}$$

where $N(t)$ is the population size at time t, and $\kappa \in (0, \infty)$. The size of the population at time t is given by:

$$N(t) = e^{\lambda t} \tag{4.4}$$

Combining 4.3 and 4.4:

$$M'(t) = \mu\lambda e^{\lambda\kappa t} \tag{4.5}$$

Since all mutations are neutral, variant frequency f is constant and is equal to the reciprocal of the population size at the moment when the mutation occurred:

$$f = e^{-\lambda t_f} \tag{4.6}$$

which leads to:

$$t_f = -\ln(f)/\lambda \tag{4.7}$$

Number of variants with frequency greater than f is equal to the integral of $M'(t)$ from 0 to t_f:

$$M(t_f) = \int_0^{t_f} M'(t)dt = \mu\lambda \int_0^{t_f} e^{\lambda\kappa t}dt = \frac{\mu}{\kappa}\left(e^{\lambda\kappa t_f} - 1\right) \tag{4.8}$$

By substituting equation 4.7 into equation 4.8, we get:

$$M(f) = \frac{\mu}{\kappa}\left(f^{-\kappa} - 1\right) \tag{4.9}$$

Finally, we can derive the equation 4.9 and obtain the formula for frequency spectrum:

$$X(f) = \frac{\mu}{f^{\kappa+1}} \tag{4.10}$$

When $\kappa = 1$ (mutation accumulation rate does not depend on N(t)), equation 4.10 is consistent with Williams [130] and Durrett [37].

4.4 Software developed

4.4.1 cevomod

The modeling approach proposed in the thesis has been implemented in the R package cevomod, a shortcut for the Cancer Evolutionary Models. The package can be easily installed from its GitHub repository at `https://github.com/pawelqs/cevomod`.

cevomod works with objects of *cevodata* class, which can store the data on the cohort of samples, as well as cevomod analysis results. Keeping the data on many samples in a specific object rather than in a list of single-sample objects facilitates conducting larger studies of cohorts of samples. *cevomod* internally iterates over the samples if needed, and uses vectorized R functions where possible, which is much faster than classic loop-based approach. In addition, we implemented plotting methods which are cohort-oriented and allow the user to easily compare the results between samples and groups of samples.

All the data in the *cevodata* object are stored in tibbles, re-implementation of classic R data frames provided by package *tibble*. Internally, each *cevodata* object is a list (as all S3 class implementations in R) and its main components are:

- metadata - tibble that associates sample IDs to patient IDs and contains all the metadata on patients (such as sex, age, or molecular subtype of the tumor) and samples (such as purity estimations),
- SNVs - list of tibbles containing SNVs and Indels. *cevodata* can store variants called by multiple variant callers and easily switch between different sets of mutations,
- CNVs - list of tibbles containing CNVs. Similarly to SNVs, *cevodata* can store and switch between CNV calls from multiple CNV callers,
- models - list of tibbles describing the models fitted by *cevomod*, but also intermediate sample descriptors used by cevomod: VAF spectra, $M(f)$ $1/f$ statistics and cumulative tails counts,
- misc - list of tibbles used by *cevomod* to store for e.g. the model residuals

The user interface of *cevomod* was inspired by the *tidyverse* R packages ecosystem and is *pipe*-oriented. Most functions accept the cevodata as the first argument and return

modified *cevodata*. This convention allows building *pipelines*, for e.g., to compose the *cevodata* object by adding new data components step by step:

```
library(cevomod)

cd <- init_cevodata(name = "AML cohort", cancer = "AML") |>
    add_SNVs(snvs_tbl, name = "Mutect2") |>
    add_SNVs(snvs_tbl2, name = "Strelka") |>
    add_CNVs(cnvs_tbl, name = "FACETS") |>
    add_patient_data(clinical_data) |>
    add_sample_data(sample_purities)
```

Listing 4.1: cevodata object construction

When all data components are added, models can be fitted:

```
cd <- cd |>
    # calc optimum number of bins in the spectra
    prepare_SNVs() |>
    # fit models
    fit_williams_neutral_models() |>
    fit_subclones()
```

Listing 4.2: Fitting cevomod models

cevomod implements many methods supporting the analysis of cancer evolution, including the model fitting approaches described in Sections 4.2.6, 4.2.7, and 4.2.8. Results can be plotted using one of many implemented data visualization functions, such as the *plot_models()* function. Most plotting functions return a standard *ggplot2* object, which can be easily adjusted and modified:

```
plot_models(cd) +
    # color spectra by added metadata, e.g. cohort
    aes(fill = cohort) +
    # customize labels
    labs(title = "Neutral model fits for sample XYZ")
```

Listing 4.3: Plotting cevomod models

In addition, we implemented basic data-transforming methods in *cevomod*:

- *filter()* method which allows to filter the *cevodata* object and to narrow it to the subset of samples based on sample-metadata. *filter* method works in *dplyr*-like manner, treating *cevodata* object as it were an usual tibble,

- *merge()* method to merge multiple cevodata objects into one,
- *split_by()* method to split the object into the list of *cevodata* objects

Detailed documentation on package functions and all cevomod functionalities is available at `https://pawelqs.github.io/cevomod`.

4.5 Implemented workflows

Efficient execution of genomic analyses requires the building of scalable processing workflows and pipelines that facilitate the management of the processes and the data. Two popular frameworks for such purpose are Snakemake [90] and Nextflow [33]. Both tools allow to easily compose numerous programs (sequence aligners, quality control tools, mutation callers, etc.) into fully automated workflows, which can be quickly run on the new batches of the data or re-run on the entire set of data if needed. Workflow management systems (WMS) support running jobs on *High-Performance Computing* (HPC) clusters via workload managers, for example, SLURM [134]. Moreover, workflow management systems provide additional error-control mechanisms (for example, the processes that failed can be re-run with the allocation of additional resources) and support containerization with environments like Docker [86] or Apptainer [67] (formerly known as Singularity), which greatly improves science reproducibility.

For the purpose of scalable and reproducible research, all the computationally extensive tasks have been implemented as workflows and run on HPC Ziemowit [54]. All workflows were implemented in Snakemake, which is a Python-based framework and does not require knowledge of any other language. Snakemake can be easily installed as a Python package with package managers, such as Conda or its faster replacement Mamba [81]. During our work on this thesis we have created the following workflows:

- *Preprocessing workflow* - we use this workflow to preprocess the NGS data. It uses the FastQC tools and MultiQC for quality control, BWA-mem for reads alignment, and finally, samtools and GATK to prepare aligned-BAM files for variant calling.
- *Mutect2 workflow* - This workflow implements the GATK Best Practices workflow for Somatic short variant discovery (SNVs + Indels) [15] using Mutect2 [29]. It starts with the analysis-ready BAM files from the preprocessing workflow and results in VCF files containing the filtered list of short variants annotated with VEP [84].

These workflows, along with other workflows for the analysis of clonal tumor structure and RNAseq data analysis, used in our other studies based on the NGS data [68, 76, 122, 57], are available via the GitHub repository at `www.github.com/pawelqs/ngs_workflows`.

Chapter 5

Results

This chapter of the thesis consists of two main parts. Section 5.1 describes combined studies of the evolution of metastatic breast (BRCA) and laryngeal (LSCC) cancers, and recurrent acute myeloid leukaemia (AML). The BRCA and LSCC analysis is based on mostly unpublished data from our own study (4 samples of the cohort were used as an example in the paper by Kurpas and Kimmel [66]), whereas the AML analysis uses the published data from the study by Shlush et al. [107]. Section 5.2 describes the results of the investigation of bladder cancer evolution from the mucosal field effect. This ongoing study is being conducted in collaboration with Dr. Bogdan Czerniak's group at the MD Anderson Cancer Center in Houston, USA, and makes use of published [10] data sequenced by the group.

5.1 Evolution of metastatic breast and larynx cancers and recurring leukaemia

Breast cancer (BRCA) is the most common cancer type in the world, with 2.26 million new cases in 2020 almost exclusively in women (over 12% of all new cancer cases and 25% of all new cancer cases in women) [16, 133]. In the same year, it was the cause of nearly 700 000 deaths, making it the fifth most deadly cancer type. Most breast cancers cases are hormone-dependent and they usually express one or more hormone receptors: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). For this reason, the common molecular classification of breast cancers divides them into 4 groups/subtypes based on the status of the receptors: luminal A - presence of ER and PR receptors, luminal B - presence of ER and sometimes PR or HER2 receptors, HER2+ - overexpression of HER2 only, and triple-negative breast cancer (TNBC) - also called basal-like, characterised by the absence of all 3 receptors (Table 5.1). Hormone-independent TNBC, although considered more aggressive than other subtypes, is rare, comprising only 15% of all breast cancers [96].

Molecular Subtype	ER status	PR status	HER2 status
Luminal A	positive	positive	negative
Luminal B	positive	some cases	negative
HER2+	negative	negative	positive
Triple negative (TNBC)	negative	negative	negative

Table 5.1: Molecular subtypes of BRCA based on the status of hormone receptors. Positive status - receptor present, negative status - receptor absent. Based on [96]

Laryngeal squamous cell carcinoma (LSCC) is a type of head and neck squamous cell cancers (HNSC) which usually begin in the squamous epithelial cells of the mucosal surfaces of the head and neck. LSCC occurs approximately 5-times more often in males than females, and the main risk factors for LSCC are alcohol consumption and tobacco smoking. Although the LSCC incidence has declined significantly in Europe during the last 30 years, it has grown globally [93], resulting in over 180 000 new cases in 2020 (contributing 1% of all new cancer cases) [133]. Both BRCA and LSCC share similar origin from the epithelial cells. Also, both BRCA and HNSC in general, show high mRNA signatures of Epithelial–Mesenchymal Transition (EMT) [41], a mechanism that may support cancers metastasis. To study the evolution and metastasis of BRCA and LSCC, we performed the WXS of the data from 15 BRCA patients and 12 LSCC patients (see Section 4.1).

Acute leukaemia contributes 40% of all leukaemia cases in Poland, and 70% of these are diagnosed as acute myeloid leukaemia (AML) [104]. In the United States, AML is the second most common subtype of leukaemia after the chronic lymphocytic leukaemia. AML

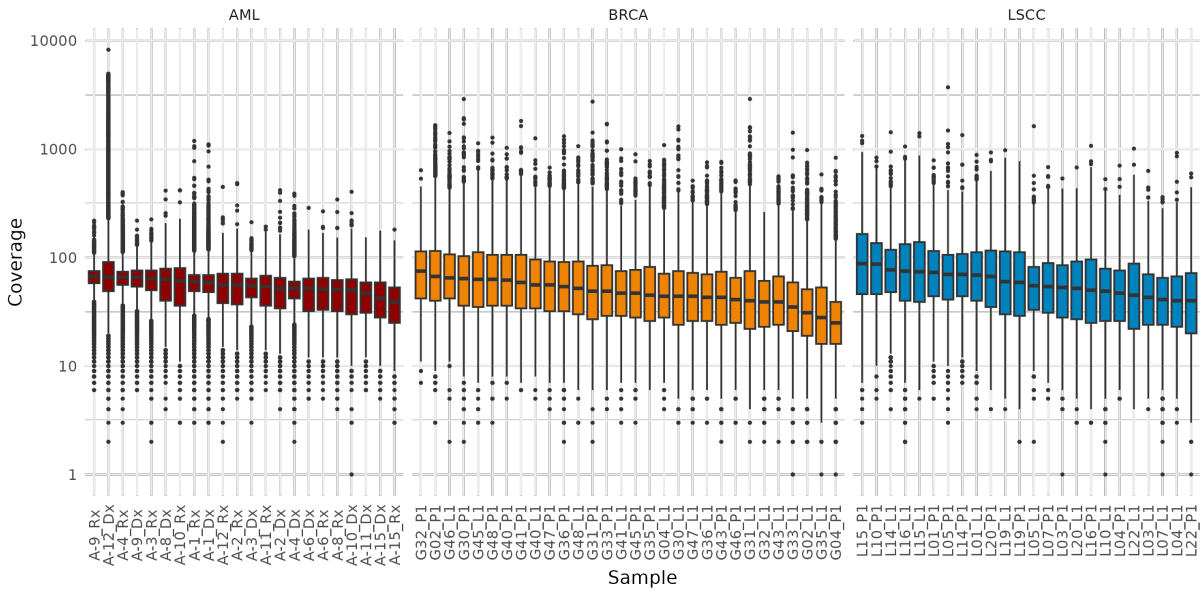


Figure 5.1: Sequencing coverage of SNVs and Indels across samples in all cohorts: AML (WGS), BRCA (WXS) and LSCC (WXS).

results from the clonal expansion of myeloid progenitor cells associated with the impaired hematopoietic stem cells differentiation and is often preceded by the clonal haematopoiesis [56]. AML is an example of cancer that fits the cancer stem cells model of evolution. To investigate the mechanisms of relapse evolution in AML we utilize the data published by Shlush et al. in 2017 [107]. The dataset consists of WGS results of 22 samples obtained from 11 patients at two time points: diagnosis and relapse (see Section 4.1).

5.1.1 Data overview

Variant Allele Frequency Spectra

After WES results for BRCA, LSCC, and AML cohorts were processed as described in Section 4.2.1, we calculated the Variant Allele Frequency Spectra for all the samples in cohorts. The expected sequencing coverage of 100x was not achieved in many samples, and median coverage of SNV and Indel variants varied from 25 (primary tumor sample of patient G04) to 74 (lymph node metastasis sample of patient L15) (Fig. 5.1). The resolution of VAF depends on the number of reads covering the variant; thus, variants with low coverage introduce additional noise during the binarization of VAF spectra (which is related to the aliasing phenomena). For this reason, we limited the number of bins in the calculated VAF spectra to the median coverage of variants in each sample. This step limited the noise in samples with low coverage and allowed us to fully utilize the resolution achieved in samples sequenced to a greater depth.

We observed two distinct shapes of VAF spectra in the data: bimodal spectra predominant among AML samples and unimodal spectra in BRCA/LSCC samples (fig. 5.2).

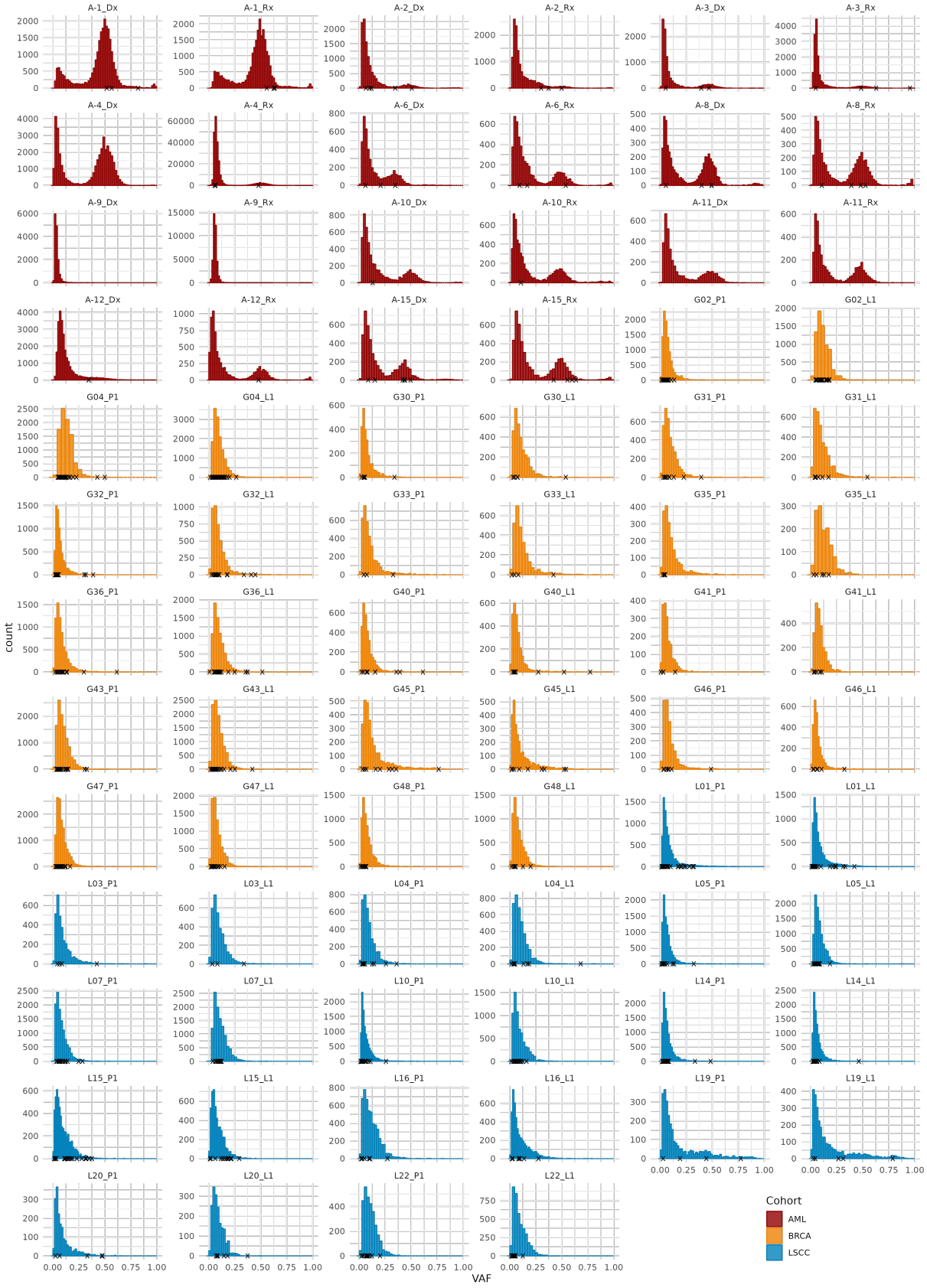
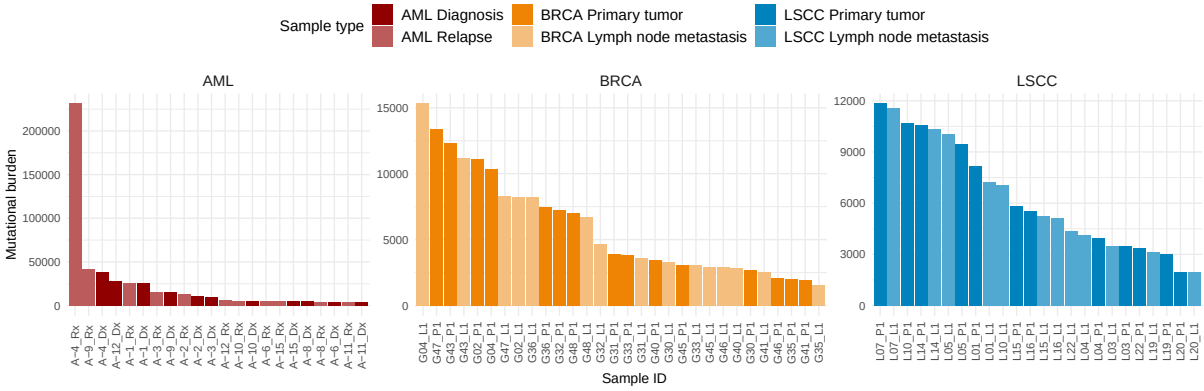


Figure 5.2: Variant Allele Frequency (VAF) spectra. While BRCA and LSCC spectra were unimodal and neutral-like, most of AML samples showed bimodal shape of spectra with clear clonal peaks. × - mutations in cancer driver genes (according to Bailey et al. [7]).

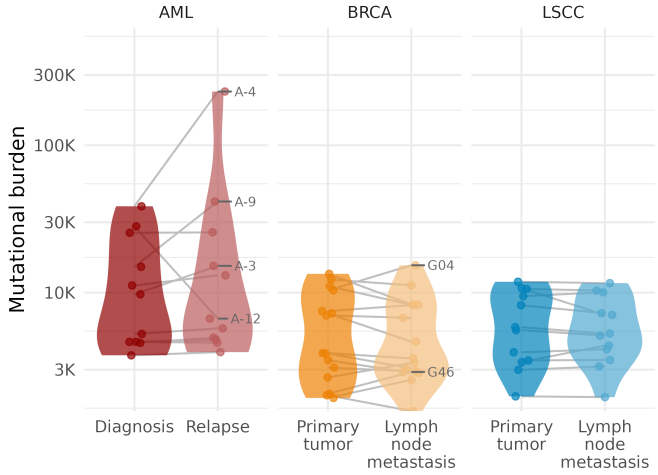
Binomial spectra of the majority of AML samples (all but: A-2_Rx, A-9_Dx, and A-9_Rx, A-12_Dx) contained clear high-frequency peaks of clonal mutations, shared by cancer cells in the sample. The allelic frequency of those peaks, oscillating around 0.5, suggests 100% purity of the samples (0% contamination of the tumor by normal cells). Low-frequency peaks of the bimodal spectra, as well as the peaks of the unimodal distributions, were positively skewed with the longer right tails. It is consistent with the concept of power-law shaped *neutral tails*, described in the literature [37, 130, 18]. The *neutral tails* contain mostly neutral mutations and/or mutations present in small subclones, indistinguishable from the neutral mutations due to low selection advantage or young age, both resulting in low cellular prevalence [113]. We identified numerous mutations in known cancer-driver genes in the low-VAF peaks of BRCA and LSCC tumors, which may indicate the presence of such subclones (fig. 5.2). Variants with the lowest VAFs are typically underrepresented in the spectra compared to the theoretical power-law predictions, which can result from the filtering applied by variant callers that remove variants with insufficient support in the sequencing results [34]. The filtering process can lead to the complete loss of the neutral tails and unimodal distribution of VAF spectra. However, the spectra containing only the clonal mutations should have a binomial shape. The power-law-like shape of spectra in the BRCA and LSCC cohorts indicates the mostly neutral origin of these mutations, a small number of clonal mutations, and the absence of previous selective sweeps.

Tumor Mutational Burden

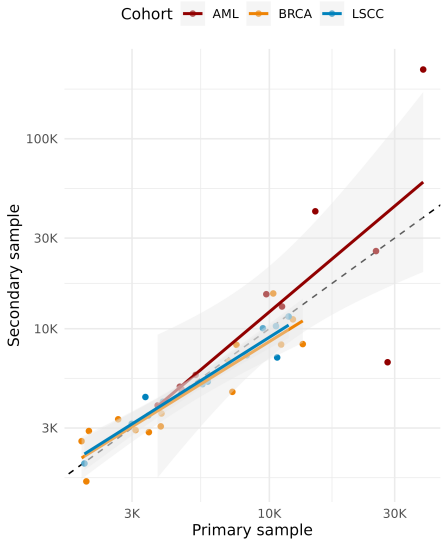
Next, we estimated the Tumor Mutational Burden (TMB, the total number of mutations detected in the sample) for all samples (Fig. 5.3). TMB was the highest in AML (WGS data, mean: 23,312 variants per sample, SD: 48,036), intermediate for LSCC (WXS data, mean: 6550 variants per sample, SD: 3217), and lowest for BRCA (mean: 6195 variants per sample, SD: 4107). The median TMB per megabase of exome was equal to 1.04 in AML, 4.98 in BRCA, and 8.37 in LSCC, within the ranges reported in the literature [58]. The differences between the mean TMB in primary samples (diagnosis or primary tumor) and secondary samples (relapse of lymph node metastasis) were not statistically significant (paired t-test, Fig. 5.3b). The TMB of the primary and secondary samples was significantly correlated (Fig. 5.3c): Pearson coefficient of correlation was equal to 0.74 in AML samples (p-value: 9.65×10^{-3}), 0.87 in BRCA (p-value: 3.13×10^{-5}), and 0.94 in LSCC (p-value: 4.74×10^{-6}). Although the changes in TMB were not common, there were particular patients in which the changes were substantial, such as patients A-4 and A-9 in the AML cohort, in which the TMB increased 3-fold and 6-fold, respectively, or A-12, in which the TMB decreased more than 4-fold (Fig 5.3b).



(a) Per sample TMB.



(b) Per sample TMB by cohort and sample type.



(c) TMB in primary versus secondary tumors.

Figure 5.3: Tumor Mutational Burden (TMB) in AML, BRCA, and LSCC cohorts, calculated as the total number of mutations detected in the sample. All the p-values in (b) were > 0.05 (paired t-test). TMB in primary and secondary tumor samples is strongly correlated (c). Pearson coefficients of correlation: AML: 0.74, BRCA: 0.87, LSCC: 0.94. All the p-values $\ll 0.001$.

Intra-tumor heterogeneity

We assessed the intra-tumor heterogeneity (ITH) using the Jaccard index. While the index is a similarity measure, it is inversely related to ITH; a lower index indicates higher ITH. For patients from BRCA and LSCC cohorts, Jaccard index values varied between 0.37 to 0.65, meaning that both tumor samples shared from 37 to 65 percent of all identified mutations. Jaccard index varied much more in the AML cohort: from 0.16 in patient A-4 to 0.98 in patient A-1, meaning that both A-1 samples contained a nearly identical set of variants (Fig. 5.4). The extreme similarity of the diagnostic and relapse samples in this patient has already been noted by Shlush et al. in the original study [107], in which they believe that the dominant clone must have survived the chemotherapy and regenerated

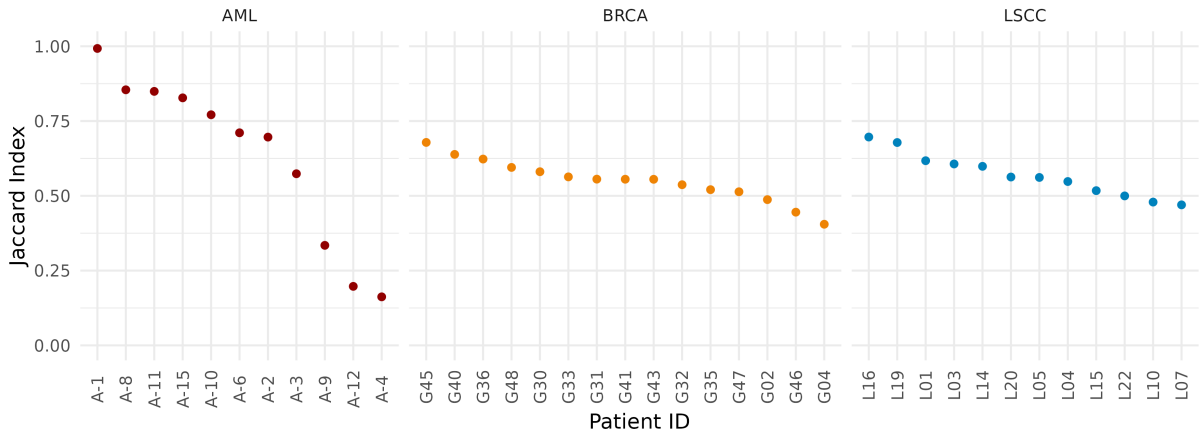


Figure 5.4: Genetic similarity between pairs of samples measured using Jaccard Index. ITH negatively correlates with the index: lower index values mean the higher tumor heterogeneity.

upon relapse. Mean values of the Jaccard index were 0.63 in AML (SD: 0.27), 0.52 in BRCA (SD: 0.07), and 0.53 in LSCC (SD: 0.07).

5.1.2 Mutations in Cancer Driver Genes

We used a list of cancer driver genes from Bailey et al. [7] to analyze the patterns of driver mutations in the data. For each cohort, we prepared a list of cancer driver genes consisting of all confirmed Pan-Cancer driver genes and both confirmed and supposed cancer-type specific driver genes. Then we filtered the lists of SNVs and Indels for variants with high or moderate impact (as predicted by *VEP* tool) in the selected driver genes. We also used the *genecards.org* database to annotate the functions of the most frequently mutated cancer driver genes.

The most commonly mutated driver genes in the AML cohort were associated with the regulation of haematopoiesis (FLT3, mutated in 5/11 patients, PTPN11, 4/11 patients), cell proliferation (NPM1, also mutated in 5/11 patients), and metabolism and energy production (IDH2, 3/11 patients) (Fig. 5.6). Interestingly, in 5 samples, we did not identify any high- or moderate-impact mutations in the analyzed known cancer driver genes. These samples included both samples from patients A-9 and A-11 and a diagnostic sample from patient A-4 (Fig. 5.5).

In the BRCA cohort, the most often mutated driver genes were involved in cellular polarization mechanisms (FAT1, mutated in 8/15 patients), methylation processes (KMT2C and KMT2D, mutated in 8/15 and 6/15 patients, respectively), regulation of WNT signaling pathway (APC, mutated in 7/15 patients), DNA damage response (BRCA1 and TP53, in 7/15 and 6/15 patients, respectively), chromatin remodeling (ATRX, 6/15 patients), Ras signal transduction pathway (NF1, 6/15 patients), PIK3K pathway (PIK3CA, 6/15 patients), and splicing mechanism (SF3B1, in 6/15 patients).

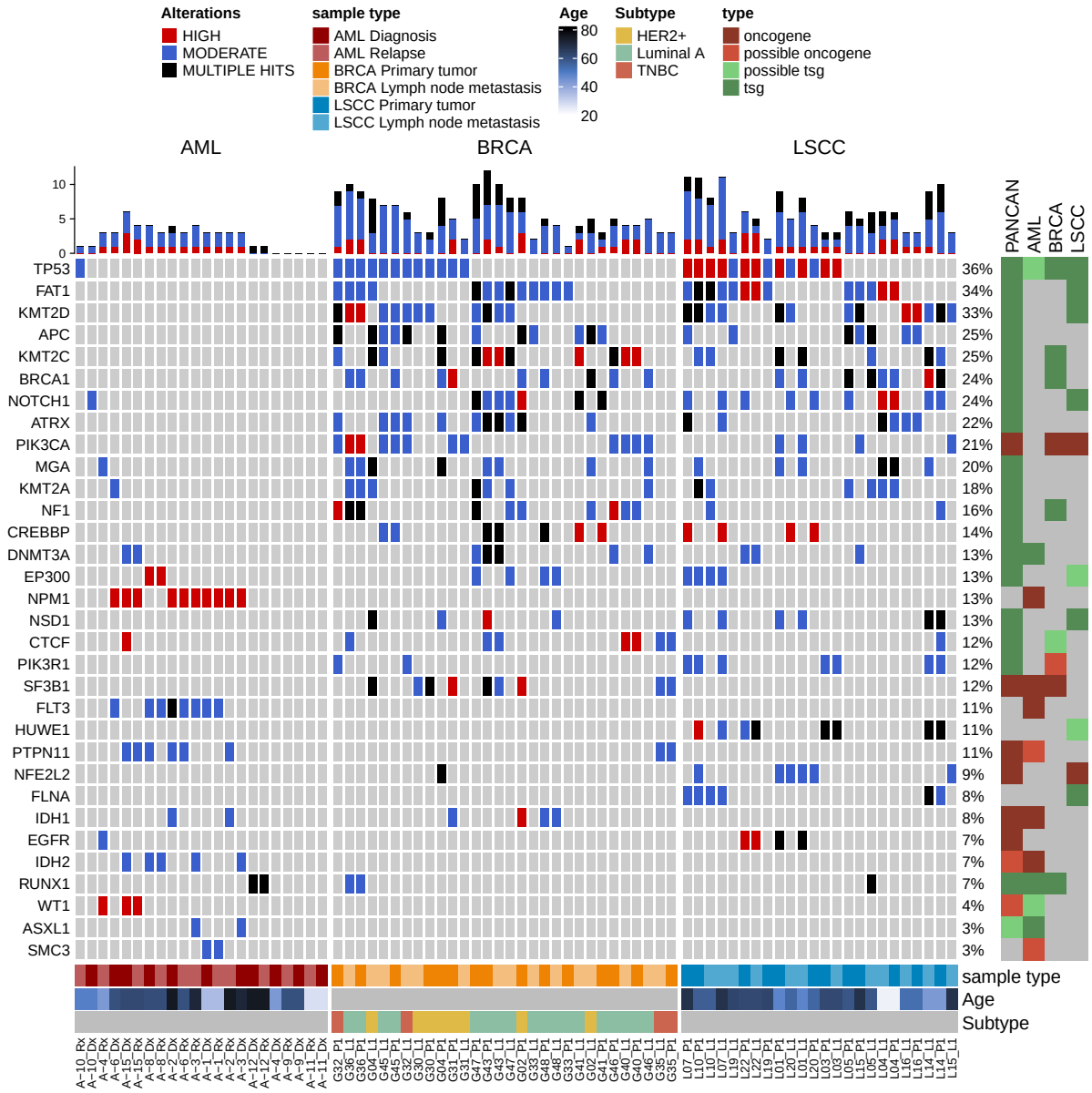


Figure 5.5: Landscape of mutations in Pan-Cancer and cancer-type specific cancer driver genes.

The most commonly mutated driver genes among LSCC patients were KMT2D (8/12 patients), FAT1 and TP53 (both mutated in 7/12 patients), NOTCH1 (mutated in 6/12 patients, involved in cell differentiation and proliferation), APC and HUWE1 (both mutated in 5/12 patients).

The numbers of mutations in driver genes varied from 0 to 10 in AML (median: 6), from 7 to 60 in BRCA (median: 17), and from 8 to 45 in LSCC (median: 27.5). In general, most of the driver mutations were present in both tumor samples (Fig. 5.7), and mutations present in the only single sample were more often detected in the primary samples (diagnostic/primary tumor) rather than in the secondary ones (relapse, metastasis).

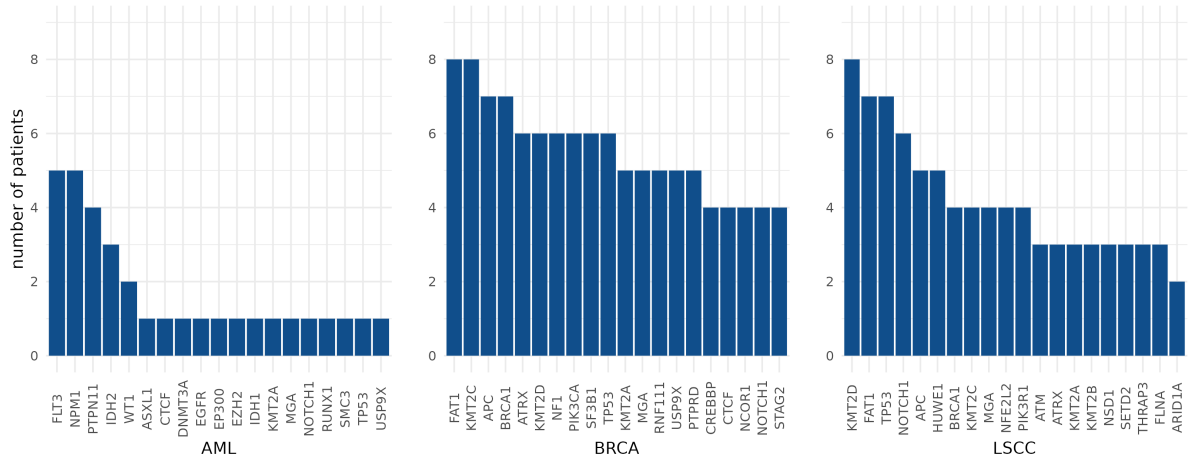


Figure 5.6: The most commonly mutated driver genes in AML, BRCA and LSCC cohorts.

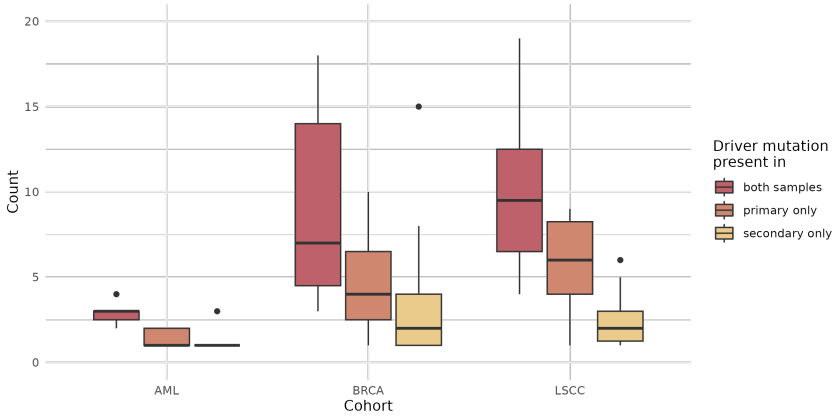


Figure 5.7: Numbers driver mutations detected in both tumor samples, only in primary samples (diagnostic/primary tumor), and only in secondary samples (relapse/lymph node metastasis).

5.1.3 Evolutionary parameters under exponential growth model

MOBSTER models

We used the MOBSTER package [18] to fit models to all 76 samples in our 3 cohorts. We found that MOBSTER was unable to correctly identify the neutral tails in our data (Fig. 5.8 and 5.9), which is necessary for the estimation of the evolutionary parameters. BRCA and LSCC samples were, in general, fitted with the uni-clonal models consisting of singular binomial peaks. AML models usually consisted of the main clone (denoted as C1 in the figure, red) and one or two subclones (denoted as C2, blue, and C3, green) (Fig. 5.8).

MOBSTER identified the neutral tails in 9 of 22 AML samples, 4 of 30 BRCA samples, and 1 of 24 LSCC samples. However, the neutral tails were only fitted correctly in 3 of these samples (both samples from patient A-1 and the diagnostic sample of A-4). Fractions of the neutral tail mutations ranged from 22% to 40% in these samples, but were typically

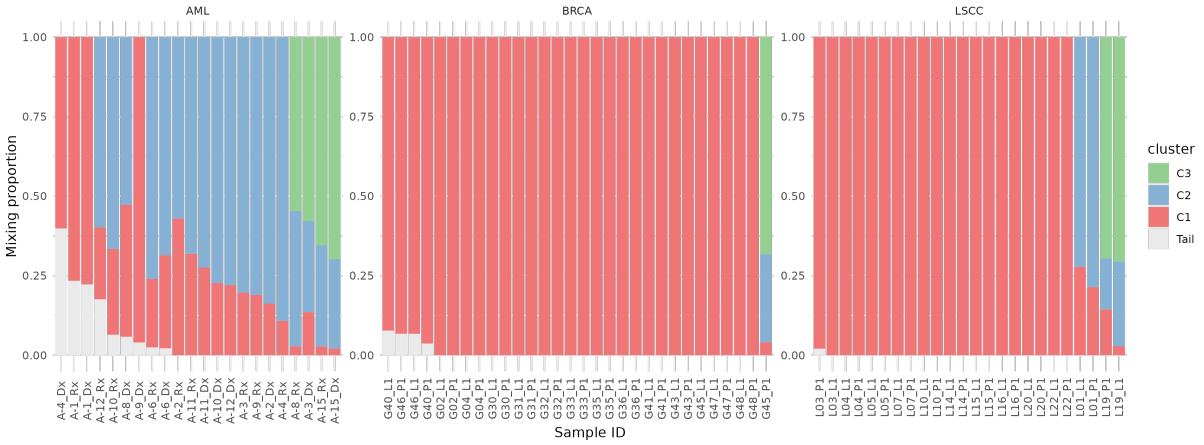


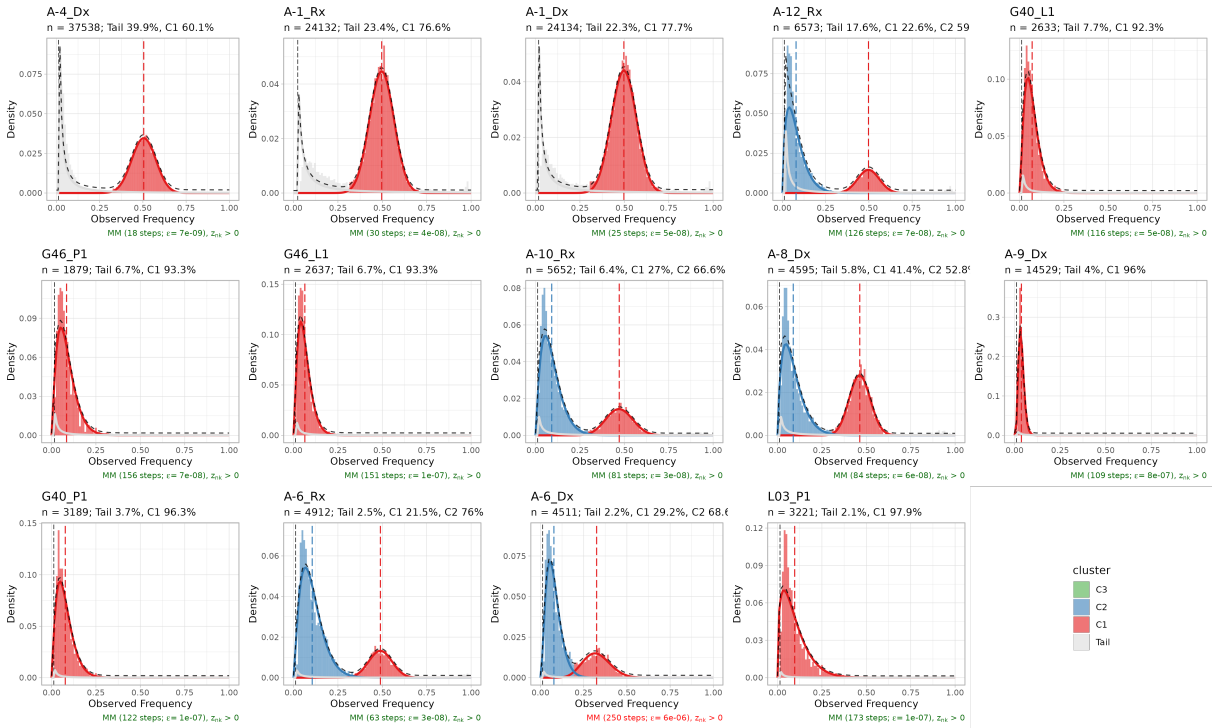
Figure 5.8: Summary of the MOBSTER fits. Mixing proportions correspond to fractions of variants classified as Neutral tail, clonal (C1), or subclonal (C2, C3). Not many models contained the Neutral tail component; most BRCA/LSCC models were uni-clonal, without a neutral tail nor subclones, and most AML samples were fitted with bi- or tri-clonal models.

below 10% in others. Median contributions of the neutral tail mutations in samples where the tails were fitted were equal to 14% in AML, 6% in BRCA, and 2% in LSCC samples. The shapes of these neutral tails did not match the true shapes of low-frequency peaks of spectra, and the numbers of the neutral tail mutations were clearly underestimated (Fig. 5.9a). Next, we found that all the three-clone models were erroneously inflating the number of clones (Fig. 5.9b). Low-frequency variants were assigned to the additional subclonal component instead of contributing to the neutral tails. Also, variants with VAFs close to 1.0, presumably due to loss-of-heterozygosity, resulted in an improper recognition of the clonal cluster in a few samples.

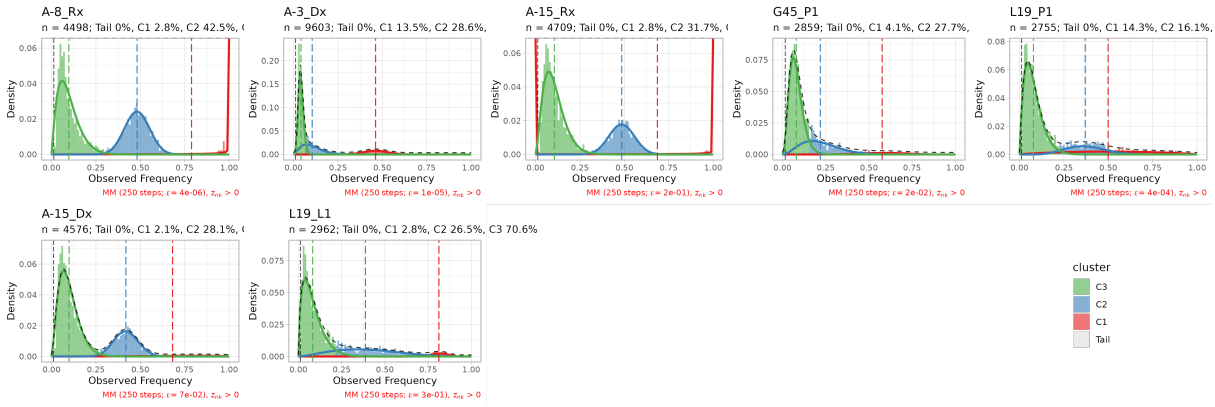
Because MOBSTER failed to properly identify neutral tails in both WGS and WXS data due to insufficient numbers of neutral tail variants, we developed a new model fitting approach robust to the lack of low-frequency variants and implemented it in an R package *cevomod* [69].

Fitting *cevomod* models

Using our approach described in Methods (Sections 4.2.6 and 4.2.8), we successfully fitted the power-law binomial models to the spectra of all samples in our cohorts despite the lack of many low-frequency variants (Fig. 5.10). Unlike in MOBSTER, most of the mutations in the low-frequency spectra peaks could be explained by the power-law neutral tail components. Only in 3 cases (relapse samples of patients A-3, A-4, and A-9) the power-law curves seemed to underestimate the true neutral tails, and most of the low-frequency mutations were above the neutral power-law curve. In most cases, the distributions of the surplus mutations were properly approximated using one or two binomial distributions.



(a) MOBSTER models with the highest contribution of the neutral tail mutations. Neutral tail contributions (white) were underestimated in all samples but one (A-4_Dx).



(b) MOBSTER models with the three clones. The third binomial component results from the nonrecognition of the neutral tail. In some cases, the presence of the variants with the loss-of-heterozygosity resulted in bad recognition of the clonal component.

Figure 5.9: Selected MOBSTER model fits. MOBSTER failed to correctly fit the neutral tail components, or overestimated the true number of clones in the several samples.

Mutation contributions of model components

We observed large variability in the contributions of the neutral tail and clonal or subclonal variants to the total mutational burden.

In the AML cohort, the contribution of the neutral tail variants varied from 6% in the relapse sample from patient A-4 up to 81% in the diagnostic sample of A-6 (Fig. 5.11a). In the 3 relapse samples (from patients A-3, A-4, and A-9), in which the neutral

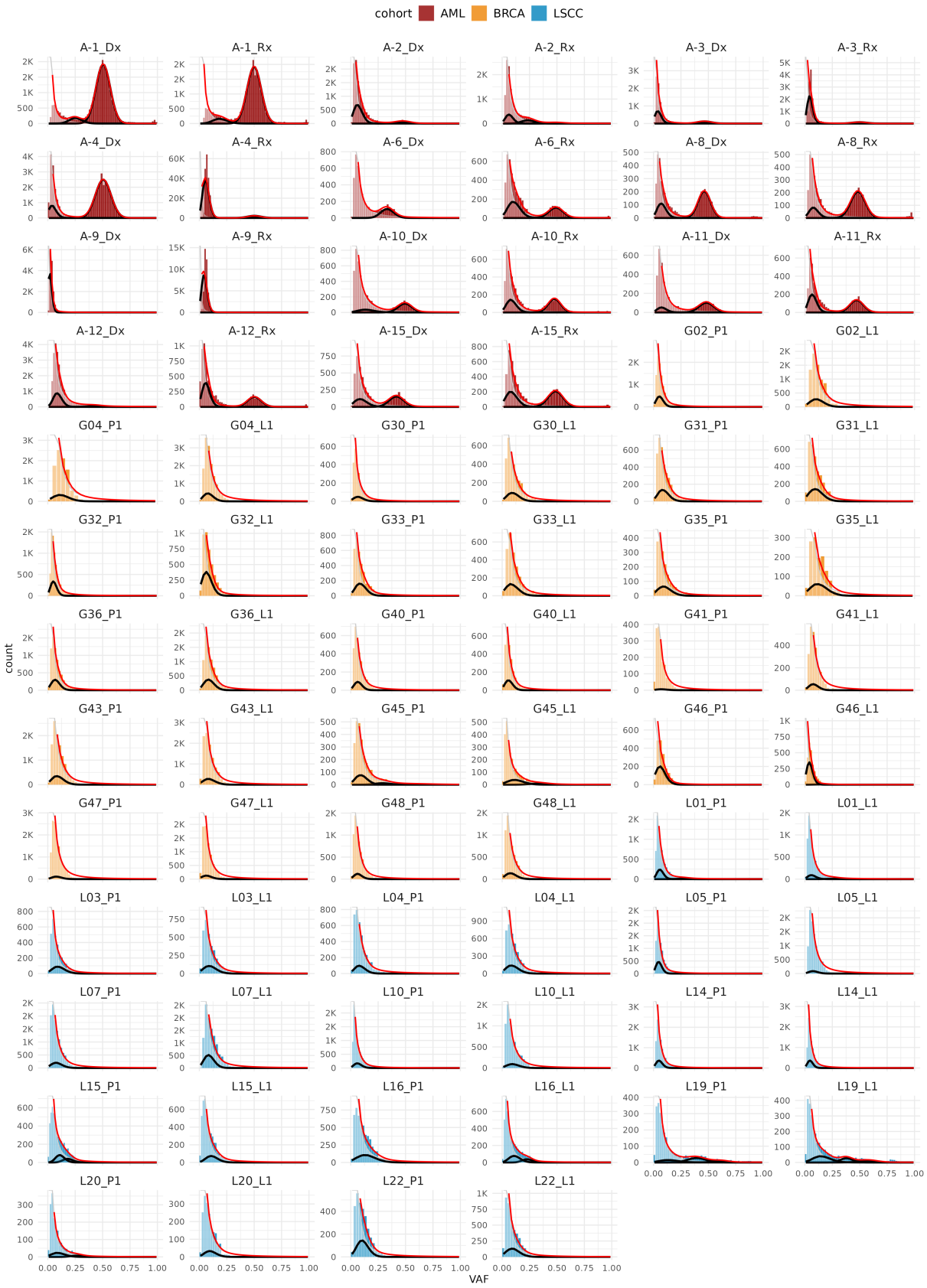
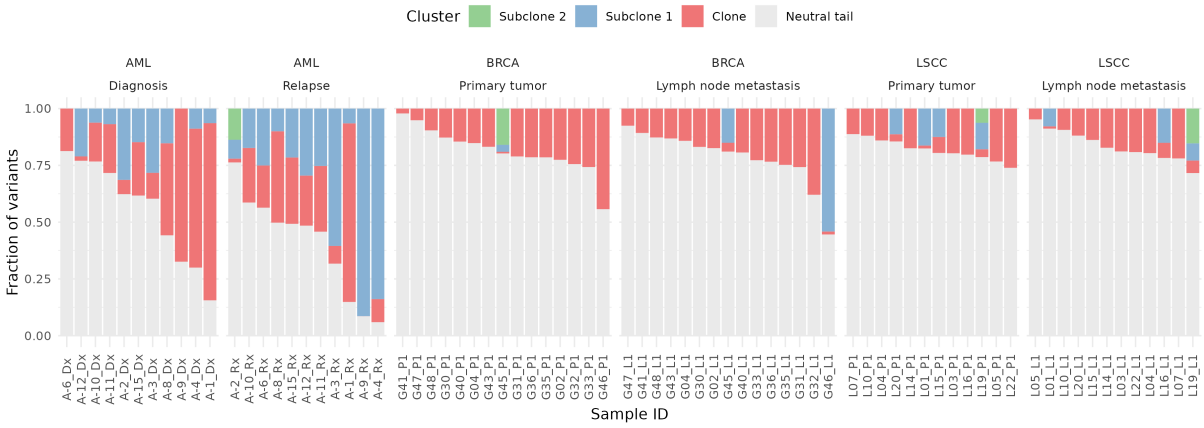
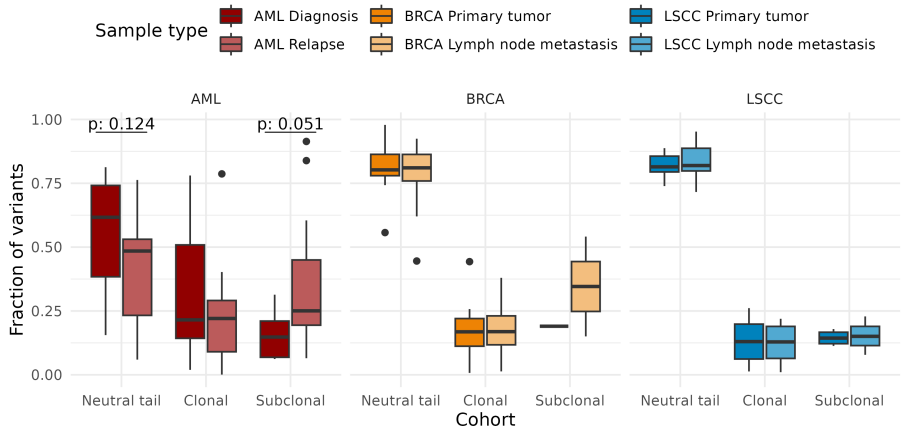


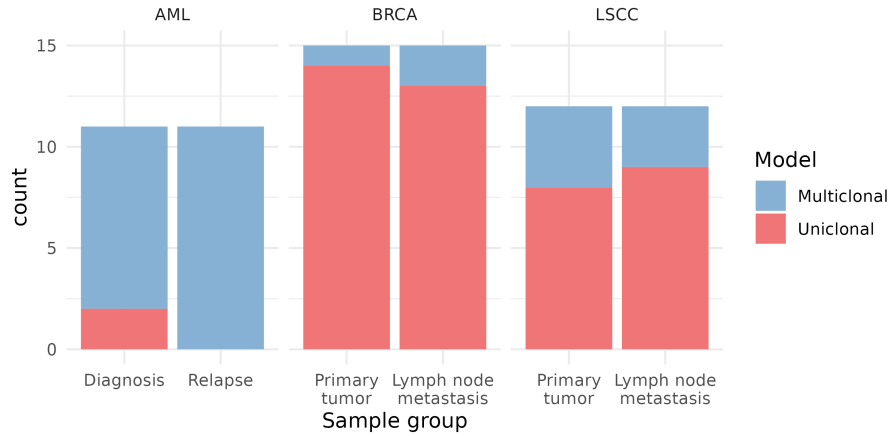
Figure 5.10: *cevomod* models with neutral power-law tails and subclones.



(a) Mutational contributions of model components in all cohorts.



(b) Comparison of mutational contributions of neutral tail, clonal and subclonal components in the primary and secondary tumor samples. AML relapse samples tend to have increased contribution of subclonal mutations, and decreased contribution of the neutral tail mutations, compared to the diagnosis samples. We did not observe any trends in other cohorts. *p*-values: t-test.



(c) Numbers of models with and without subclones in cohorts. Proportions were compared using *z-test*, all the p-values were greater than 0.05.

Figure 5.11: Mutational contributions and clonality across AML, BRCA, and LSCC cohorts

components were underestimated (Fig. 5.10), most of the neutral tail variants were fitted with additional binomial distributions. In these samples, contributions of subclonal variants are likely overestimated. Interestingly, in one of these samples (relapse sample of A-9), the subclonal component contained 36445 low-frequency variants, compared to 31 high-frequency variants of the clonal component. We pay more attention to these samples in Section 5.1.4, in which we fit the second type of model, with the α coefficient optimized for each sample.

In the BRCA and LSCC cohorts, with the clear high-frequency peak of clonal variants absent, neutral tails contributed to a higher proportion of variants than in AML samples. In most samples, they contributed 70% - 98% of all variants, with only 3 samples below this range: both samples from patient G46 and a lymph node metastasis sample of G32.

We compared the mutational contributions of the neutral tail, clonal, and subclonal components of the models in the primary (diagnostic or primary tumor) and secondary (relapse or metastatic) tumor samples using the *t*-test (Fig. 5.11b). Relapse samples in the AML cohort contained more subclonal mutations compared to the diagnostic samples and slightly fewer neutral tail/clonal mutations. In BRCA and LSCC cohorts, model component contributions were similar in both primary tumor and lymph node metastasis samples.

Subclonal composition of primary and secondary samples

We also compared the numbers of uni-clonal and multi-clonal models in the primary and secondary tumor samples (Fig. 5.11c). The *p*-values from *z-tests* of proportions were close to 1 for all 3 cohorts. Almost all models in the AML cohort were multi-clonal, except for 2 diagnostic samples (from patients A-6 and A-9). Uni-clonal models prevailed in BRCA and LSCC cohorts, with only 1 multi-clonal fit among BRCA primary tumor samples and 2 multi-clonal fits in the BRCA lymph nodes. In the LSCC cohort, there were 4 multi-clonal fits among the primary tumor samples and 3 multi-clonal fits among the lymph node metastasis samples.

Proportions of the detected and undetected mutations

At low frequencies, the number of variants estimated by the power-law fits (consistent with the estimated reduced mutation rates) differs significantly from the number of variants detected due to filtering applied by variant callers. We used our model fits to compare the theoretical and real numbers of variants with a VAF greater than 0.01 and found that the average numbers of mutations predicted by the models were approximately 4 times higher than the numbers of the mutations detected (Fig. 5.12). The undetected mutations contributed, on average, 62% of all predicted mutations (SD: 16%), but there were outliers in the AML cohort, with the fraction of undetected mutations as small as 10% in the

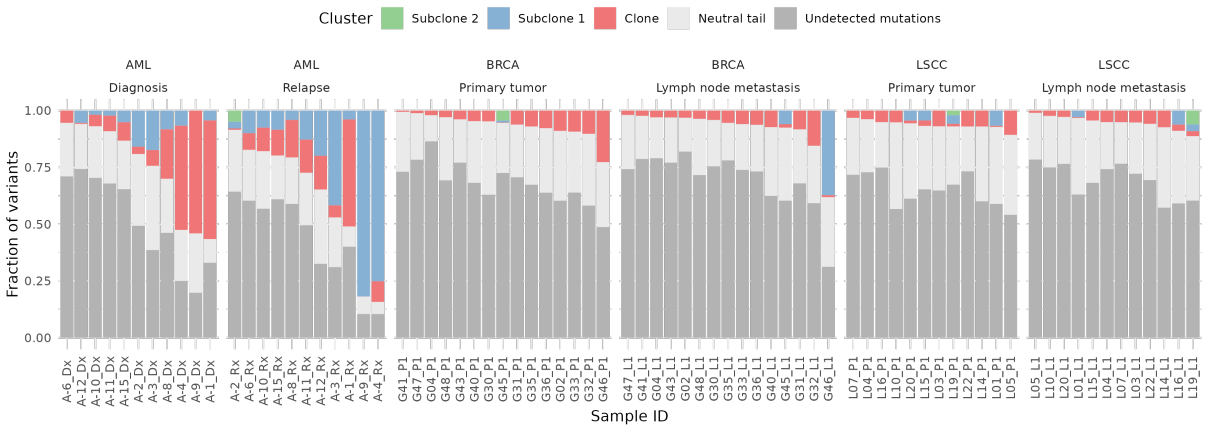


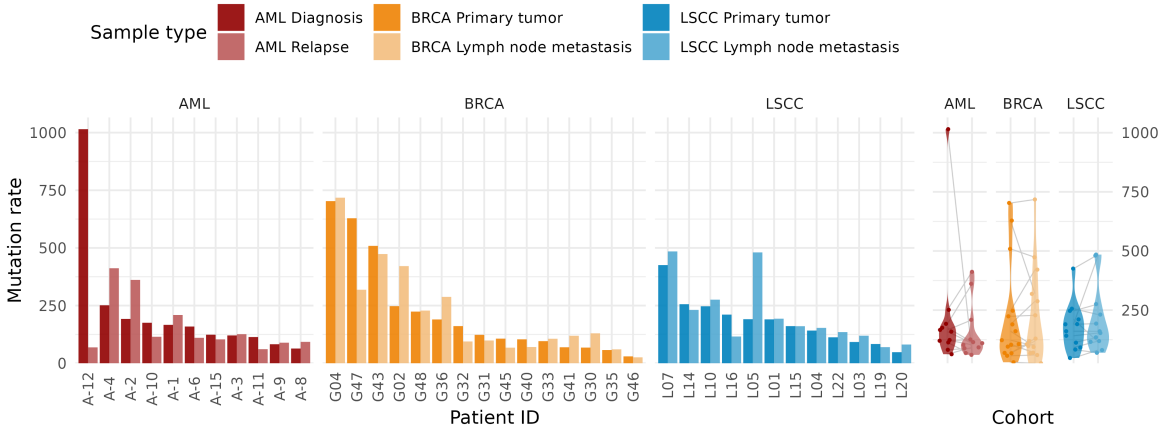
Figure 5.12: Estimated proportions of detected and undetected mutations with VAF higher than 0.01. According to our model, the numbers of undetected variants (or filtered out in variant calling) in are up to 4 times higher than the numbers of the detected ones.

relapse sample from patient A-4. Notably, the lowest proportions of undetected variants were predicted in samples with the most inaccurate fits: relapse samples of patients A-4 and A-9. The proportion of undetected mutations was usually higher in samples with more accurate fits.

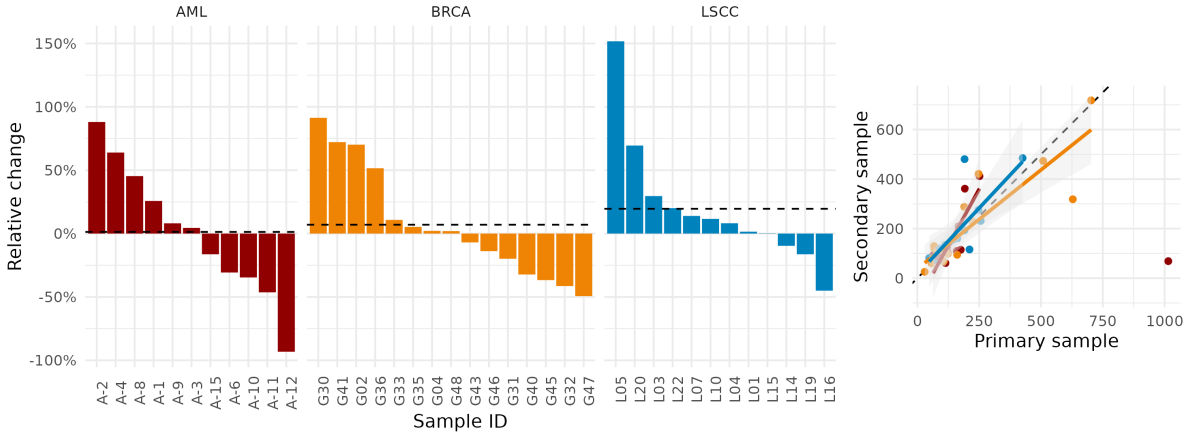
Estimated mutation rates

We detected over 10-fold variability in the estimated effective mutation rates [130] (MR, average number of mutations per successful cell division) in the samples (Fig. 5.13a). Minimum MR values in all the cohorts were smaller than 100 (lymph node sample of G46: 25, primary tumor sample of L20: 48, diagnostic sample of A-11: 61), median varied around 150 (AML: 122, BRCA: 121, LSCC: 160), and the maximum values were close to 500 or exceeded it (A-12, diagnostic sample: 1014, G04, lymph node sample: 717, L05, lymph node sample: 485). Average values of MR, scaled by the genome size, were about 5.98×10^{-8} in AML, 6.8×10^{-8} in BRCA, and 6.06×10^{-8} in LSCC. This is one order of magnitude higher than the somatic mutation rate in normal human cells reported by [87] (10^{-9}).

Mean values of MR were similar in the primary (diagnosis or primary tumor samples) and the secondary tumors (relapse or lymph node metastasis) in all cohorts (Fig. 5.13a, right panel). The p-values from the paired *t-test* were all greater than 0.05 (AML: 0.49, BRCA: 0.81, LSCC: 0.3). However, relative differences between the primary and secondary samples were significant in many patients (Fig. 5.13b). The highest increases of MR in secondary samples compared to the primary ones were noted in patients L05 (+151%), G30 (+91%), and A-2(+88%). On the other side of the spectrum, the most significant decreases in MR occurred in patients A-12 (-93%), G47 (-49%), A-11 (-46%), and L16 (-45%). There was no predominant direction of MR change in the AML and BRCA cohorts,



(a) Estimated mutation rates in all samples. Means of MR in primary and secondary samples were compared using the paired *t-test*. *ns*: $p > 0.05$.



(b) *Left*: Relative changes in the mutation rates in the secondary samples compared to the primary samples. Dashed line shows the mean relative change in the cohort. *Right*: Correlation of MR in primary and secondary samples. Patient A-12 (the left-most point) was omitted when fitting the linear trend lines.

Figure 5.13: Mutation rates under the model of exponential growth model. Mutation rates were calculated using Williams's formula [130] and cevomod approach (see Methods, section 4.2.6).

but upward changes prevailed among the LSCC samples, with an average 19.6% increase in the lymph node metastasis, compared to the primary tumor sample.

Evolutionary parameters of the subclones

We used the equations described in Section 3.3.7 to calculate the selection coefficients and the emergence times of the identified subclones.

In multi-clonal samples, we estimated these evolutionary parameters for all the identified subclones but not for the clonal peaks, which may have the highest cellular prevalence but do not signify an ongoing selection in the tumor. In monoclonal BRCA and LSCC samples, we assumed that the only identified binomial cluster is subclonal rather than

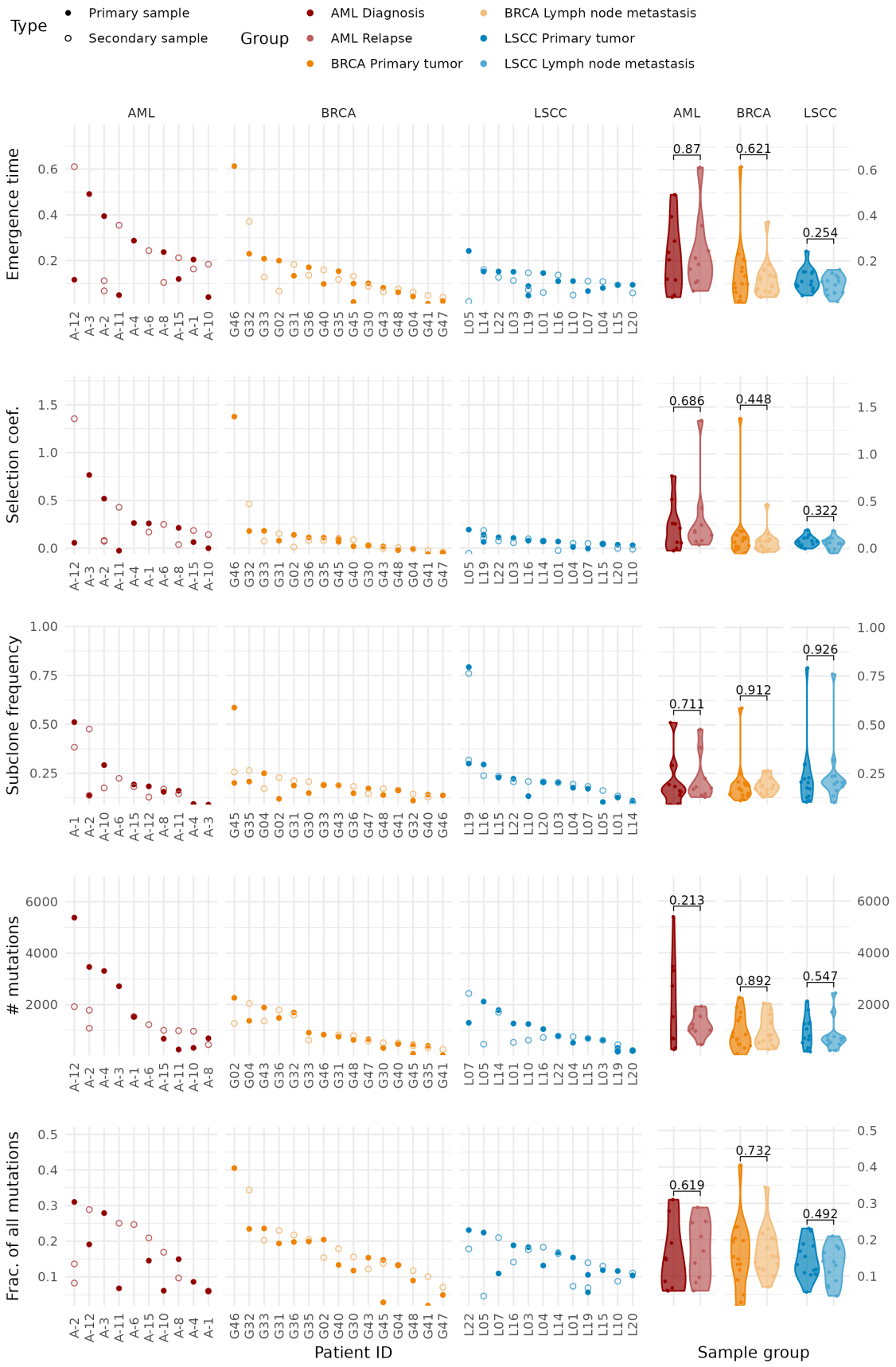


Figure 5.14: Evolutionary parameters of subclones in AML, BRCA, and LSCC cohorts.

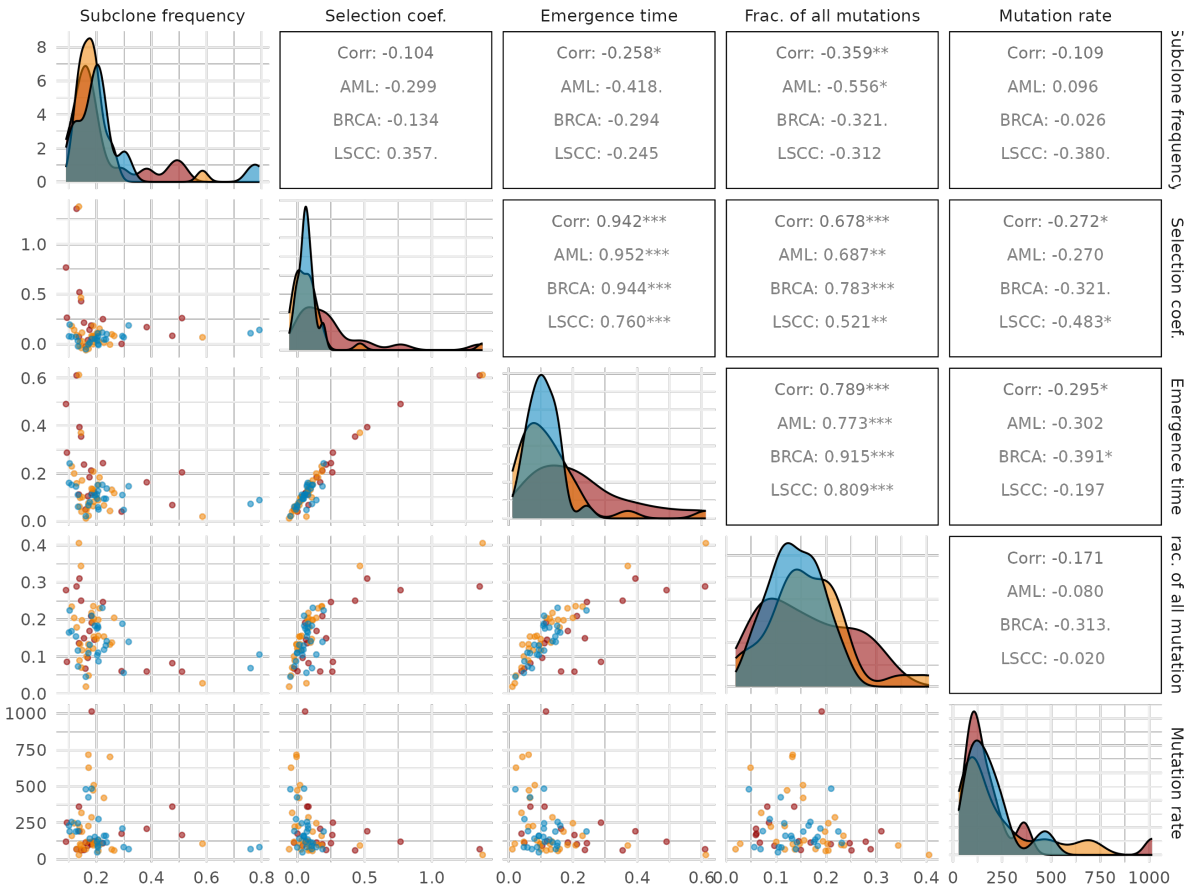


Figure 5.15: Correlations of the evolutionary parameters. Numbers above the diagonal represent the values of the Pearson correlation coefficient. *** - $p < 0.001$, ** - $p < 0.01$, * - $p < 0.05$, . - $p < 0.1$

clonal - the low mean allelic frequencies of mutations in that clones support this assumption. In the AML cohort, characterized by high purity, we assumed that the clonal components in monoclonal models were truly clonal. Accordingly, we did not calculate the evolutionary parameters for these clones (diagnostic samples from patients A-6 and A-9).

For the so-defined subclonal components, we estimated the subclone emergence times relative to the tumor age and the selection coefficients. We identified 4 outlying subclones with estimated selection coefficients much below. These included 3 relapse AML samples from patients A-3, A-4, and A-9, in which we previously recognized the power-law components to be inaccurate and the mutation rates underestimated (Fig. 5.10). In these samples, the selection coefficients were equal to -2.55 (A-3_Rx), -1.1 (A-4_Rx), and -1.11 (A-9_Rx). The last outlier was found in the lymph node metastasis sample from patient G46, with the selection coefficient equal to -5.17. Also, the estimated emergence times were out of the expected range of values in those samples, exceeding the calculated tumor age: in relapse samples of A-3, A-4, and A-9, the subclones emerged after 52, 331, and 295 tumor population doublings (tpd), respectively. In the lymph node sample of G46

the estimated emergence time was equal to 40 tpd. However, the estimated tumor age at the time of sequencing was approximately 33 tpd in all these cases. We excluded all these 4 outliers from the further analysis.

Emergence times. In all other samples, the scaled subclonal emergence times were up to 0.6 of the total tumor age. (Fig. 5.14). In all the cohorts, the minimal emergence time was similar: 0.01 in BRCA (primary tumor in G41), 0.02 in LSCC (lymph node metastasis in L05), and 0.04 in AML (relapse sample of A-10). The latest subclones emerged at time 0.61 in AML and BRCA (relapse sample of A-12, and primary tumor of G46, respectively) and 0.24 in LSCC (primary tumor of L05). The average time of emergence was similar in BRCA and LSCC (means: 0.13 and 0.11; SD: 0.12 and 0.05, respectively) and two times higher in AML (mean: 0.22, SD: 0.16).

In BRCA and LSCC, nearly all clones emerged not later than at time 0.2 tumor age, and the times were similar in both primary tumor and lymph node metastasis samples. In AML, in each patient, we found a subclone that emerged after that time-point in one sample and a subclone much younger in another one. Old subclones were equally frequent in both diagnostic and relapse groups of samples, and the times in both samples of the same patient were highly discordant.

We used the t-test to compare the average emergence times in the primary and secondary samples, but none of the p-values were smaller than 0.2 (Fig. 5.14).

Selection coefficients. The estimates of selection coefficients were strongly correlated with the estimated times of emergence (Pearson coefficient of correlation: 0.94, p-value $< 2.2 \times 10^{-16}$). The minimal values oscillated around 0 in all cohorts, and the highest values were: 1.35 in AML (relapse sample of A-12), 1.38 in BRCA (primary tumor sample of G46), and 0.2 in LSCC (primary tumor sample of L05). 2 subclones in the BRCA cohort and 5 in AML cohort showed the selection coefficient higher than 0.25.

For BRCA and LSCC, the selection coefficients were similar in both patient samples. For AML, we again observed the significant differences between the samples, but without an increasing trend in either the diagnostic or relapse samples. We tested for the significance of the selection coefficient differences between the primary and secondary samples, but all the p-values were above 0.3 (Fig. 5.14).

Subclone frequencies. In all cohorts, the majority of subclones contributed to up to 25% of the cells. However, we identified the individual outliers in all cohorts: in AML, both samples from patient A-1 had subclones that contributed to approximately 50% of the tumor cells; in BRCA, a primary tumor sample from patient G45 had a subclone contributing 59% of the tumor, and in LSCC, patient L19 had subclones contributing more than 75% of all tumor cells in both of his samples.

Subclonal frequencies were comparable between the primary and secondary samples (t-test, Fig. 5.14).

Frequencies of subclonal cells were negatively correlated with the fractions of subclonal mutations (Pearson coefficient of corelation (r): -0.359; p: 0.0017) and with the emergence time (r: -0.258; p: 0.02663) (Fig. 5.15).

Mutations in subclones. Subclones usually had 5% to 25% of all mutations in the tumor, with two much higher values observed in the BRCA cohort: subclones in the primary tumor sample of G46 and lymph node metastasis of G32 had 40% and 35% of all mutations, respectively. Fractions of mutations in subclones were similar in the patient's primary and secondary samples in BRCA and LSCC cohorts but showed high variability in patients in the AML cohort. We also did not observe any significant difference between the mean fractions of subclonal mutation between the primary and secondary samples across the patients (t-test, Fig. 5.14).

The fraction of mutations in subclones was correlated with the subclone frequency (negatively, r: -0.36, p: 0.0016), selection coefficient (r: 0.68, p $< 3.1 \cdot 410^{-11}$) and emergence time (r: 0.79, p $< 2.2 \cdot 10^{-11}$). However, no correlation was observed between the fraction of mutations in subclones and the mutation rate(Fig. 5.15).

5.1.4 Optimization of the power-law exponent.

In some samples, including the relapse samples of patients A-3, A-4, and A-9, the neutral tail slopes were, particularly step, and the power-law components with exponents α equal to 2 did not fit the data correctly. Tung and Durrett have demonstrated that the competition between the micro-clones, indistinguishable from the power-law neutral tail, may alter the shape of the neutral tail, decreasing its power-law exponent [120]. To investigate the samples with inaccurate power-law fits and seek the examples of selectively advantageous micro-clones described by Tung and Durrett, we fitted new models to the data, in which the power-law exponents α were optimized to fit the data best.

We fitted the models without any boundary restrictions for the α values and observed common deviations of α from the expected value of 2. Approximately half of the AML and LSCC samples showed a downward deviation of α, possibly pointing to the ongoing selection between the two types of cells (Fig. 5.17). However, α values greater than 2 were also common in these cohorts and prevailed in the BRCA cohort, possibly indicating the violation of the model assumptions, such as non-exponential tumor growth or changing mutation rate. In Section 4.3, we present a mathematical explanation of how the non-constant mutation rate can increase or decrease α.

3 samples had the optimum α values surprisingly high: relapse samples of patients A-4 and A-9, and lymph node metastasis sample from patient G41, where α varied between 3

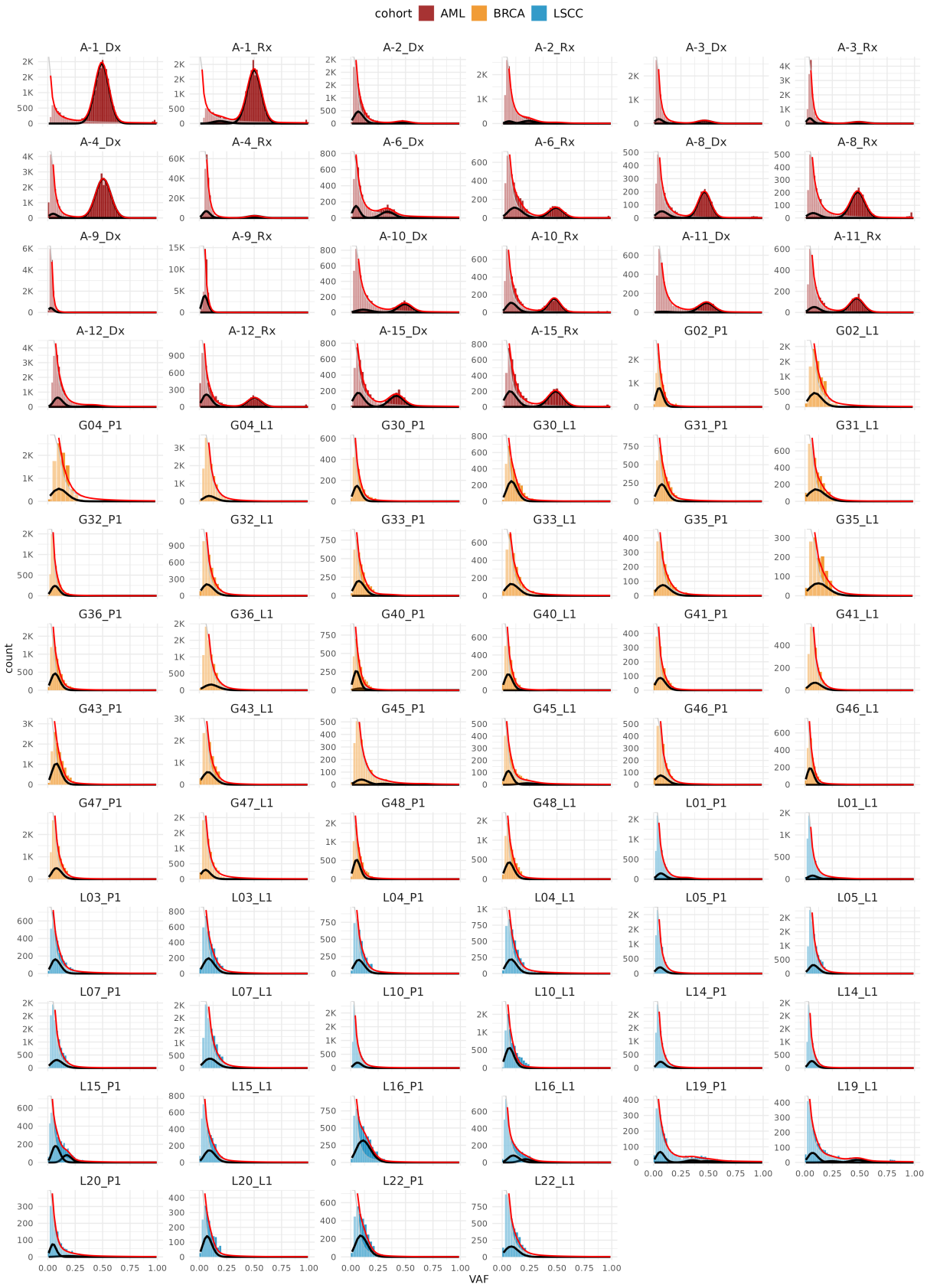
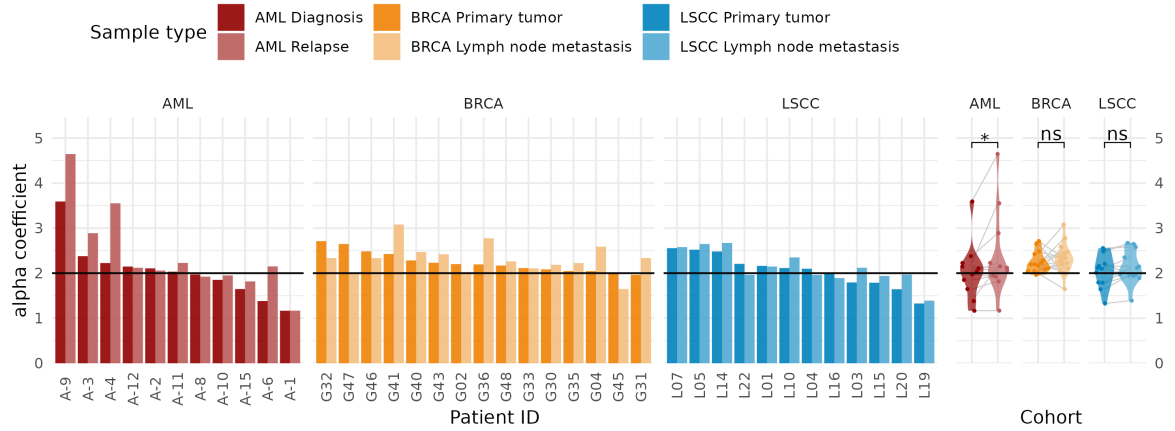
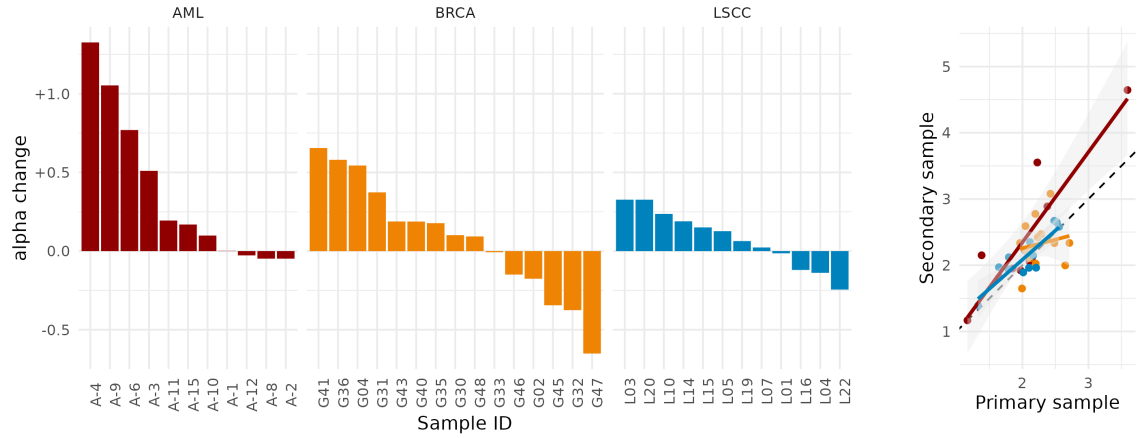


Figure 5.16: Model fits with optimized power-law exponents α.



(a) α coefficients of the optimized model fits. *Left:* α values by sample, *Right:* α values by cohort and sample type. While many AML and LSCC samples shows signs of selection ($\alpha < 2$), it is rare in BRCA. α coefficients greater than two were more common that α less than 2.



(b) Comparison of the α values between the primary and secondary tumor samples. *Left:* Differences between the α in the secondary and primary tumor samples. *Right:* α in secondary sample versus α in the primary tumor sample. Pearson correlation coefficients: 0.89 (AML), 0.18 (BRCA), and 0.88 (LSCC).

Figure 5.17: α coefficients of the optimized model fits.

and 5. The new, optimized models fit the data much more accurately than the previous models from Section 5.1.3 (Fig. 5.18).

Interestingly, the α coefficients were often higher in the secondary samples (relapse of metastases) than in the primary ones (primary tumors or diagnostic samples) (Fig. 5.17b, left). α values in the primary and secondary tumor samples were correlated in AML and LSCC cohorts, with the Pearson coefficient of correlation equal to 0.89 in AML, 0.18 in BRCA, and 0.88 in LSCC (Fig. 5.17b, right).

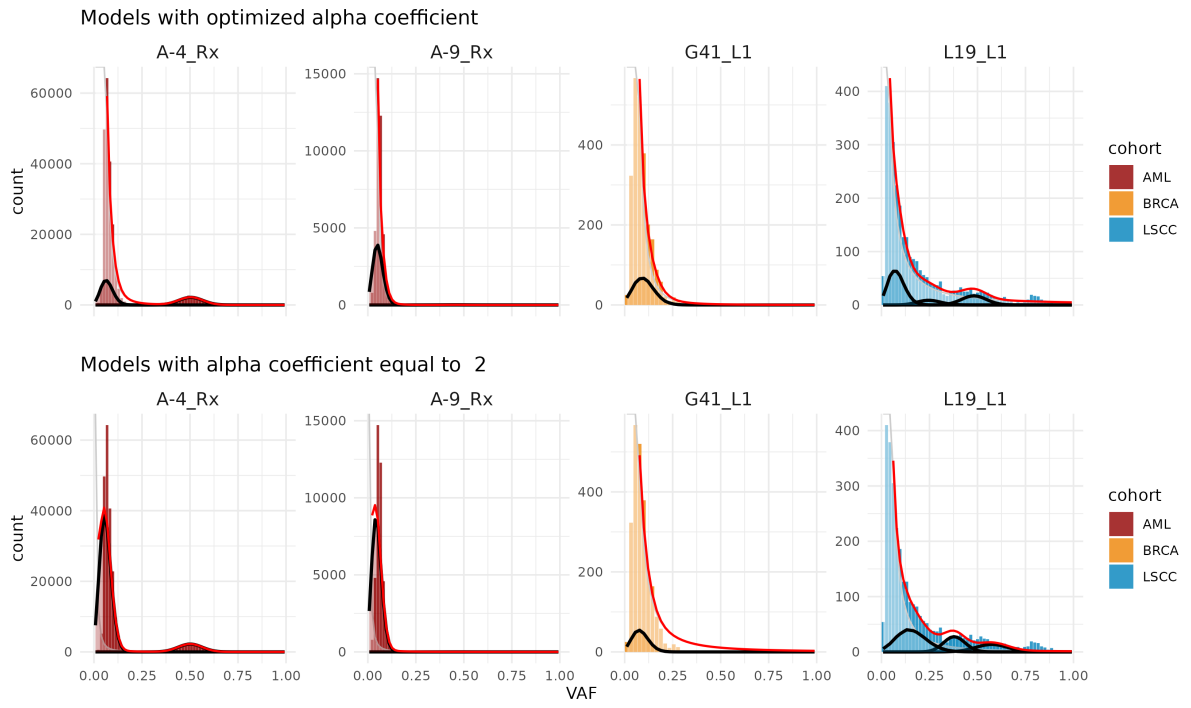


Figure 5.18: Comparison of the models with high and low optimum α values with the models based on the power-law exponent equal to 2 from Section 5.1.3.

5.1.5 Comparison of runtimes

We measured the times needed to fit the models to all samples in AML, BRCA, and LSCC cohorts. cevomod fitting methods were an order of magnitude faster than MOBSTER with *auto_setup="FAST"* pre-configuration and 2 orders of magnitude faster than MOBSTER with the default setup. The mean runtime for cevomod models was 3 seconds, compared to 37 seconds in fast MOBSTER setup and nearly 5 minutes in the default mode. cevomod was even faster when the alternative, approximate method of binomial model fitting was used, with a mean runtime of 1.3 seconds.

5.1.6 Discussion

To investigate the evolutionary dynamics of cancer during metastasis and recurrence, we studied the results of bulk DNA sequencing data from 38 patients representing 3 cancer types: 11 patients with acute myeloid leukaemia (AML), 15 patients with breast cancer (BRCA), and 12 patients with laryngeal cancer (LSCC). We analyzed the total number of 76 tumor samples, 2 per patient. In the AML cohort, we used the whole genome sequencing data published by Shlush [107]. This dataset included the diagnostic and relapse time points data, which we used to analyze the evolutionary dynamics in recurrent tumors. In BRCA and LSCC cohorts, the data included the whole exome sequencing results of 2 samples from the same time point: the primary tumor sample and the lymph node

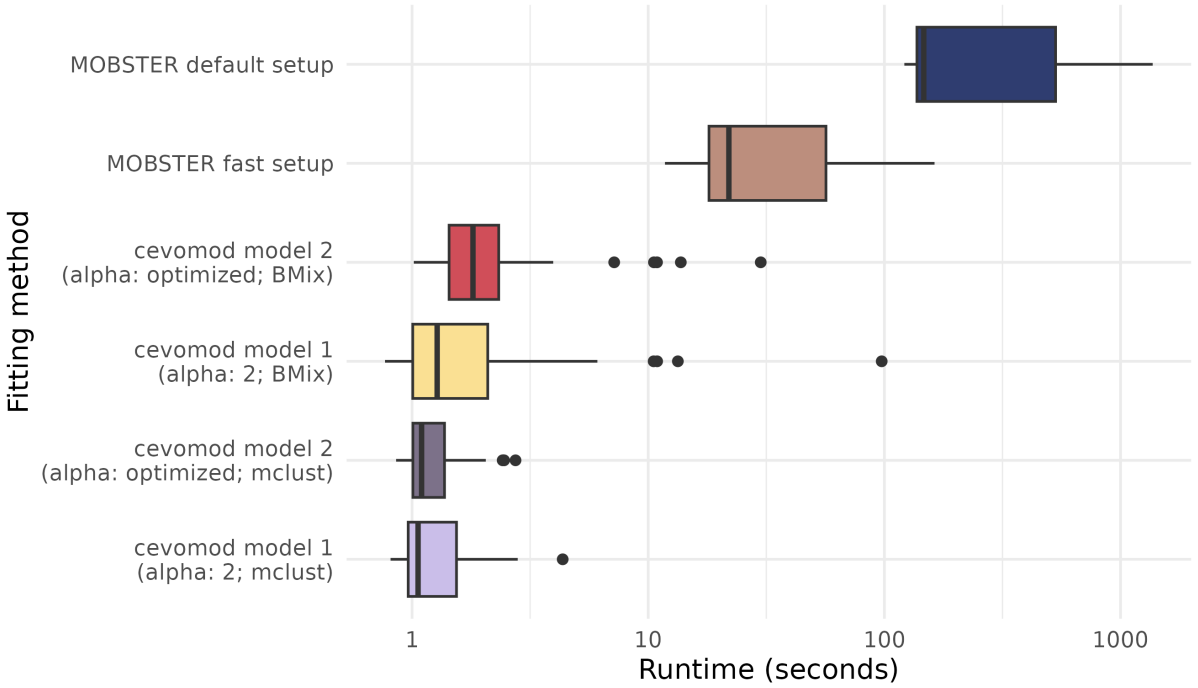


Figure 5.19: Runtimes of cevomod and MOBSTER. cevomod performs an order of magnitude faster than MOBSTER with the *fast* pre-configuration. Using an approximate method of fitting subclones, the mean runtime of cevomod was 1.3 seconds per sample.

metastasis.

We observed the bimodal Variant Allele Frequency (VAF) spectra predominant among the AML samples and unimodal spectra in BRCA and LSCC cohorts. The binomial shape of the high-VAF peaks in the AML cohort is consistent with the random-sampling-like nature of DNA sequencing (see Section 3.2.2). VAF values oscillating around 0.5 indicate that these peaks contain the clonal mutations present in all cancer cells. The shape of the BRCA and LSCC spectra and the low-VAF peaks in AML samples was more power-low-like, compatible with the neutral tail models described by Durrett [37], Williams [130, 131], and others. The absence of clear clonal peaks may indicate the low number of true clonal mutations, the lack of previous selective sweeps, and the predominance of neutral evolution in these samples. The presence of mutations in known cancer driver genes at high VAF in many of these samples is in line with the high histopathology estimates of tumors' purities.

Values of Tumor Mutational Burden (TMB) per megabase were higher in the analyzed samples than reported in some other studies. Median TMB per Mb was equal to 1, 5, and 8, in AML, BRCA, and LSCC, respectively, compared to 0.3, 1, and 3 in these tumor types reported by Kandoth [58]. The differences between our results and the results of Kadoth might result from different variant calling and filtering methods applied; differences between the tumor types were preserved, though. TMB was similar in the primary and secondary tumor samples in most patients. There were only 3 AML patients in which

TBM significantly differed in the relapse sample; in 2 cases increased, and in 1 case decreased. In these patients, only about 25% variants were shared by both samples, contrary to 40% to 60% in most other patients.

Mutation rates

We were not able to model our data using the well-known R package MOBSTER due to the insufficient number of neutral tail variants. Caravagna et al. [18] recommend using MOBSTER with WGS data with a sequencing depth of at least 100x. However, MOBSTER failed not only to fit the neutral tail components in BRCA and LSCC WXS but also in the WGS AML cohort. For this reason, we developed our own package, cevomod, for fitting the mixture of power-law and binomial components to the incomplete data, in which many neutral tail mutations were lost. cevomod successfully fitted the power-law and binomial components and estimated the evolutionary parameters in all our samples.

In BRCA and LSCC, the neutral tails contributed to about 75% of all mutations. In the AML cohort, this fraction varied between 25% and 75%, whereas the remaining 75% to 25% variants contributed to clonal peaks. However, we showed that the true counts of neutral tail variants were underestimated due to filtering applied by variant calling algorithms. The numbers of neutral tail variants predicted by the power-law component of the model were up to 4 times higher than the number of detected variants in these components, even in the WGS data.

The estimated mutation rates per effective cell division (MR) varied significantly within the cohorts, from approximately 50 mutations per effective cell division in all 3 cohorts, up to 500 in LSCC, 700 in BRCA, and 1000 in AML. Mean values of MR per base were similar in all cohorts: 5.98×10^{-8} in AML, 6.8×10^{-8} in BRCA, and 6.06×10^{-8} in LSCC. It is an order of magnitude higher than the somatic mutation rate in normal human cells according to [87] (10^{-9} per bp), but lower than the MR in normal human lymphocytes (5-24 $\times 10^{-8}$ per bp), and in the malignant lymphocytes (5.3 - 66 $\times 10^{-7}$ per bp) reported by Seshadri [105]. We did not observe the 100-fold difference in MR between AML and BRCA reported by Williams (AML: 10^{-9}, BRCA: 10^{-7}) [131]; however, the AML cohort consisted of the WGS data from 100% tumor-pure samples, and BRCA data originated from WXS of 80-90% tumor-pure samples, which might affect the analysis. For this reason, caution is needed when comparing the AML results with the other cohorts. [110]

Although the overall MR values were similar in the groups of primary and secondary tumor samples, we identified significant differences between these samples in many patients. In most patients, the MR in the secondary tumor sample changed from -50% (2-fold decrease) to +100% (2-fold increase) compared to the primary tumor sample. Only in two patients were these limits exceeded: in one AML patient, we observed a 90% drop of MR in the relapse sample, and in one LSCC patient, a 150% increase of MR in the

lymph node metastasis. In AML and BRCA, there was no predominant direction of MR change, but in LSCC, upward changes prevailed.

Evolutionary parameters of subclones

In nearly all samples, the low-frequency peaks were fitted with the mixture of power-law and binomial components, indicating the deviation from the truly neutral shape of the tail. The mean fraction of the subclonal variants was increased in relapse AML samples compared to the diagnostic samples. Using the equations from Williams et al. [131], we calculated the emergence times and selection coefficients for all the subclones.

In BRCA and LSCC, subclones in most samples emerged early, during the first 20% of tumor volume doublings. The selection coefficients associated with these subclones were low, exceeding 0.2 only in two samples. BRCA and LSCC samples also lacked the clear clonal peaks. Both these observations indicate a close-to-neutral evolution of these cancers, with short variants (SNVs and Indels) providing a limited selective advantage. In such tumors, selective sweeps might be rare or not occurring at all. Tumors with the subclonal selection coefficients closest to 0 may undergo a punctuated cancer evolution, in which all crucial genomic events occur early and are followed by a neutral-like evolution of the growing population. This type of evolution was recently identified in colorectal cancers [110]. Single-cell DNA sequencing studies, such as [39], show that the punctuated evolution followed by the stable tumor expansion may correctly describe the evolution of CNV in breast cancers. Tumors with subclonal selective coefficients closer to 0.2 and lacking the clonal peaks might have been sequenced before the first selective sweep occurred. The simulation study of Bozic et al. [11] shows that drivers' frequencies in small tumors (10^7 cells) are strongly biased towards 0, indicating that selective sweeps have not occurred yet. In BRCA and LSCC cohorts, we indeed observed numerous mutations in known cancer driver genes at frequencies between 0 and 0.1. It supports the thesis that these tumors might not have been sufficiently large and have not anticipated the selective sweep yet.

In the AML cohort, the emergence times of subclones varied significantly, and the latest subclones emerged after approximately 60% of all tumor volume doublings. The selection coefficients were higher in these samples, and the presence of clear clonal components indicated past selective sweeps. However, in some samples, the slopes of the neutral tails were too steep to be approximated by the power-law exponent equal to 2. In these samples, the power-law components were not fitted accurately, underestimating the mutation rate and overestimating the fraction of subclonal variants. Consequently, the subclone emergence times and the selection coefficients might have been overestimated. In some samples, the subclone emergence times were 10-fold greater than the estimated tumor age. For this reason, we developed another model in which the power-law exponent is optimized to fit the data most accurately.

Theoretical and actual power-law exponents

Tung and Durrett demonstrated that the presence of the selectively advantageous micro-clones results in the power-law exponent *alpha* less than 2 [120]. We found that in approximately half of the AML and LSCC samples, the optimum α was less than 2. However, we have also observed optimum α values greater than 2, which were common in AML and LSCC cohorts, and predominant in the BRCA cohort. These unexpected values of α indicate that the model assumptions, such as the exponential population growth or the constant mutation rate, might often be violated in the actual data. In Section 4.3, we showed how the changing mutation rate during the cancer progression alters the power-law exponent. Interestingly, we found that most secondary tumors (relapse samples and lymph node metastases) had α increased compared to the primary tumors (diagnostic and primary tumor samples). It might indicate that the increase of the mutation rate might be even faster, or the selectively advantageous micro-clones might be less common in secondary tumors. Determining which of these phenomena indeed alters the power-law exponent is hard using bulk DNA sequencing, which cannot distinguish rare subclones [113]. We further explore these phenomena in Section 5.2.4.

cevomod R package

We implemented our model fitting approaches in an R package cevomod. The package can be installed from its GitHub repository at `https://github.com/pawelqs/cevomod`, and the detailed documentation of the most recent version is available at `https://pawelqs.github.io/cevomod/`. cevomod can quickly fit the mixtures of the power-law and binomial components to the whole exome sequencing data or low-coverage data, in which the neutral tails are severely incomplete. It allows one to select between two types of models, a 'neutral' model with the power-law exponent equal to 2 and the 'optimized' model, in which the power-law exponent is fitted to the data as well.

5.2 Evolution of Bladder cancer from mucosal field effects

Bladder Cancer (BLCA) is epithelial cancer that develops in the epithelial tissue lining the interior of the urinary bladder, known as urothelium. Urothelium forms a barrier between the urine and the underlying tissues, exposing it to various metabolic products and environmental factors, many of which are oncogenic. Most BLCA cases are induced by chemical carcinogens, such as those found in tobacco smoke [27]. The carcinogens, along with chronic infections, can induce changes in the bladder mucosa (a *field effect*), leading to the development of cancer.

Recent development in the research on BLCA let to distinguish two main molecular subtypes of BLCA, which differ in their expression signature, clinical behaviors and, as the research shows, originate from different progenitor cells [27, 45]. The luminal subtype expresses a signature of the luminal cells which line the lumen of the bladder (such as uroplakins, CK20, or CK18), and it was shown to originate from the more differentiated luminal progenitor cells and the luminal field effect. Similarly, the basal subtype expresses the genes which characterize the basal cells that form the basal membrane of the urothelium (such as KRT, KRT14, or CD44). This subtype originates from the less differentiated basal progenitor cells and is induced by the basal field effect [27].

Based on the experimental results from the laboratory of Dr. Bogdan Czerniak at the MD Anderson Cancer Center in Houston, TX, we investigated the origins of BLCA. The results were published in the paper by Bondaruk et al. *The origin of bladder cancer from mucosal field effect* [10].

Paragraph *Transcriptomic and DNA methylation analysis* of this Section, and paragraphs *Mutational landscape*, *Mutations' ages and selection coefficients*, and *Mutational signatures of dormant and progressive phase mutations* of Section 5.2.1 summarize the results of Bondaruk et al. [10]. Paragraph *Mutational signatures of dormant and progressive phase mutations* describes my contribution to the paper. Other parts of this Section are new analyses not used in the paper.

Transcriptomic and DNA methylation analysis. It was found that the dysregulation of mRNA expression and changes in DNA methylation were widespread in the mucosal effects, whereas the mutational changes were numerous but usually restricted to the particular fields of the maps. In particular, the process of luminal differentiation was altered but maintained during the progression of luminal cancer in map 24. In basal map 19, the differentiation process was clearly suppressed, but the genes associated with the epithelial-mesenchymal transition, such as the target genes of TGFB1 or TP53, were activated. It is consistent with the previous findings that EMT is an important factor in the development of basal cancer [45].

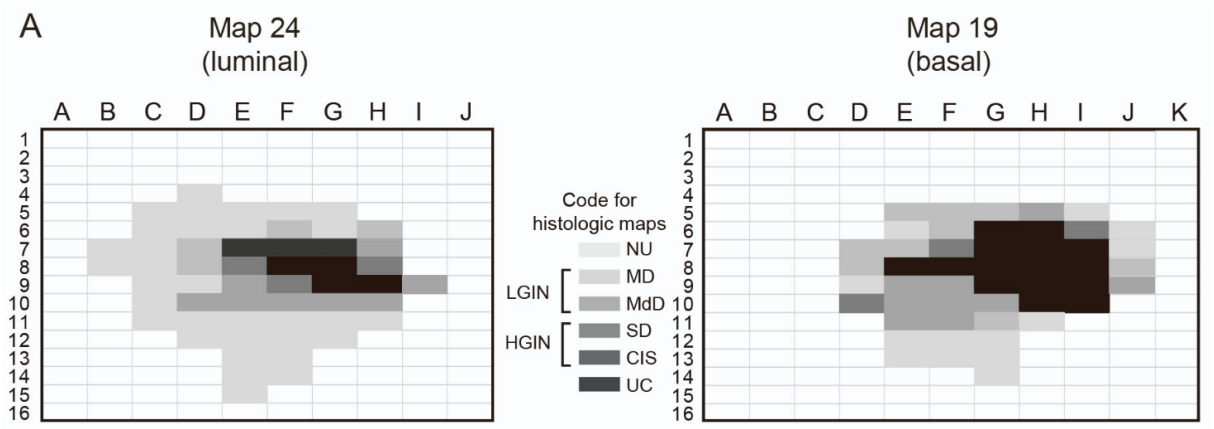


Figure 5.20: Map of analyzed BLCA specimens. Source: Bondaruk et al. *The origin of bladder cancer from mucosal field effect* [10], Figure S2, panel A, on Creative Commons Attribution – NonCommercial – NoDerivs (CC BY-NC-ND 4.0) license.

5.2.1 Mutational landscape of the maps.

Driver mutations in the maps. Using WXS data, numerous mutations were detected in all samples of the mucosa. Nearly all samples, including the NU and LGIN, had mutations in known cancer driver genes (Fig. 5.21). In most NU/LGIN samples, these mutations had VAFs up to 0.2, except for samples H7 and G6 in map24, which had driver mutations at VAFs comparable to HGIN and UC samples. Most of the driver mutations were restricted to particular samples of the maps, and only a few were spread across many fields of the map: 18 in map 19 and 7 in map 24. Interestingly, we observed 4 different spread patterns in the basal map 19: mutation in RHOA gene was detected in two non-adjacent fields in the upper part of the map; mutations in ELF3, KMT2D, and RHOB were present in adjacent fields at the bottom of the map; mutations in TLR4 and NIPBL were widely spread in non-neighboring fields of the map, and 12 mutations in different genes were present in HGIN field I6 and one of the UC fields: H8, but not in the other UC field I10. (Fig. 5.22). None of the analyzed driver mutations from the sample I10 (UC) was spread to other fields of the map. In map 24, 6/7 mutations (in RB1, FBXW7, BRAF, CDKN1A, APC, and BAP1) showed a similar spread pattern on the right side of the map. Only one driver mutation (in CACNA1A) was widely spread across other fields of the map 24.

Mutational landscape. In the mutational landscape of the two maps, 3 groups of variants we recognized, which were called A, α, and β (Figures 6 and S12 in paper, Bondaruk et al. [10]). The A group was the most abundant one and contained 1303 out of 1379 nonsynonymous variants in the map24, and 2176 out of 2678 nonsynonymous variants in map19. Variants in this group were usually restricted to a single sample of the map, and their allelic frequencies were low, usually below 10%. The α group of variants contained variants that were spread across the mucosa in all groups of samples (NU, LGIN, HGIN,

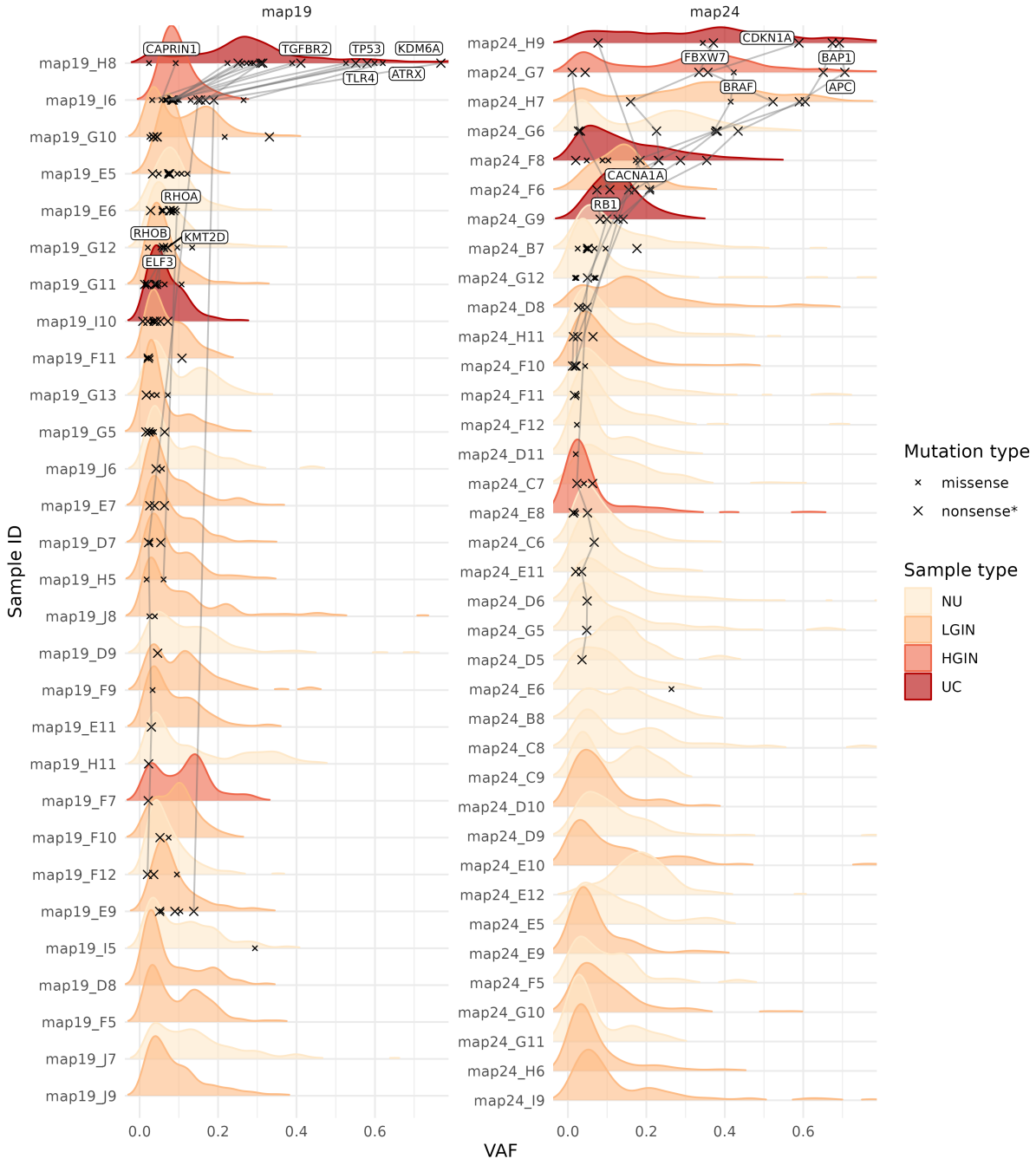


Figure 5.21: Mutations in known cancer driver genes in maps 19 and 24. The ridge plot shows the densities of mutations with given VAFs. Colors denote the sample classification, and mutations in the PanCancer and BLCA driver genes (according to Bailey et al. [7]) are shown with x marks. The size of the mark indicates the impact of mutation: x - missense mutations, X - nonsense mutations, indels, and mutations in splice-sites and start/stop codons. Samples were ordered using VAF matrix clusterization. Many NU/LGIN samples contained one or more mutations in the driver genes. Mutations spread across multiple samples are joined by lines. Selected spread mutations are labeled in the sample where the mutation VAF was the highest.

UC); however, their allelic frequencies were still low, similar to the frequencies of mutations in group A. Although the cells with these mutations underwent a clonal expansion,

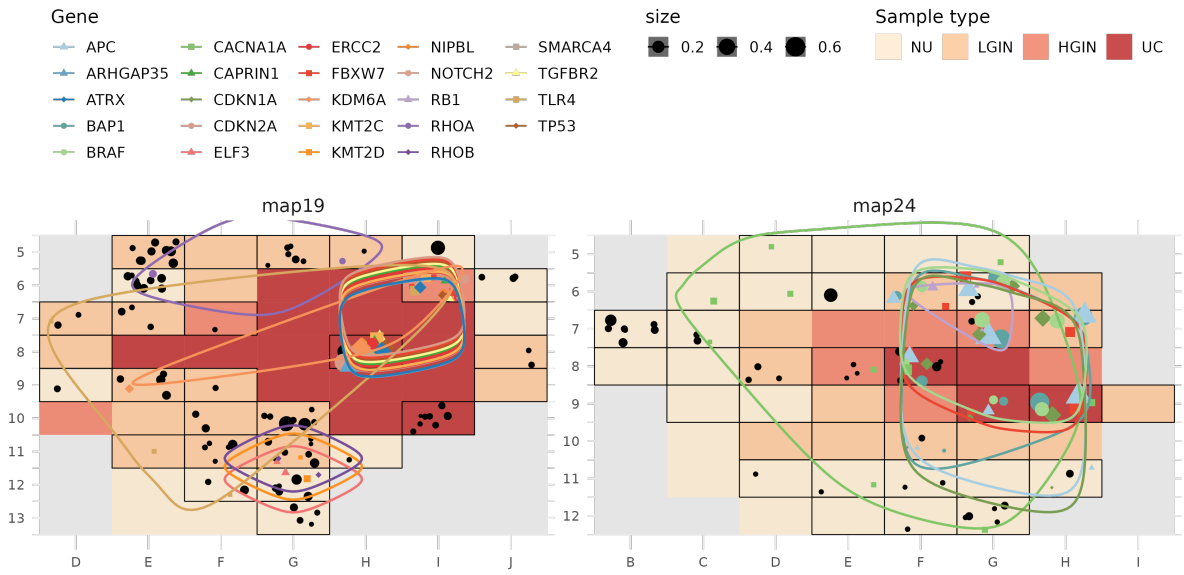
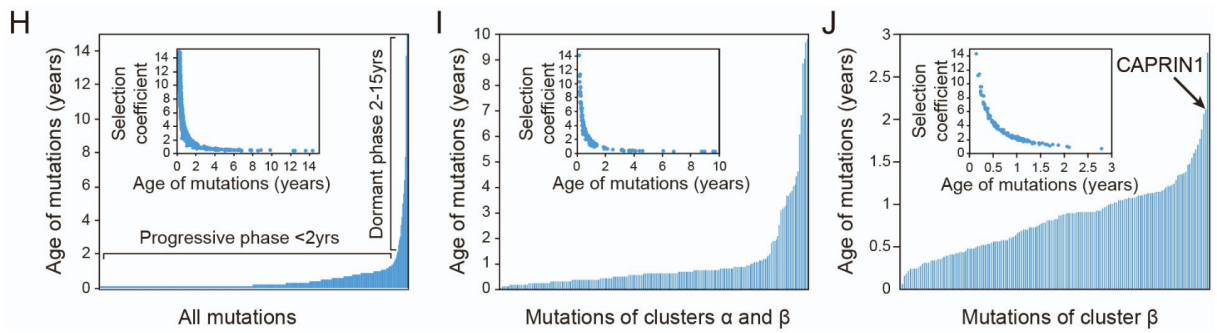


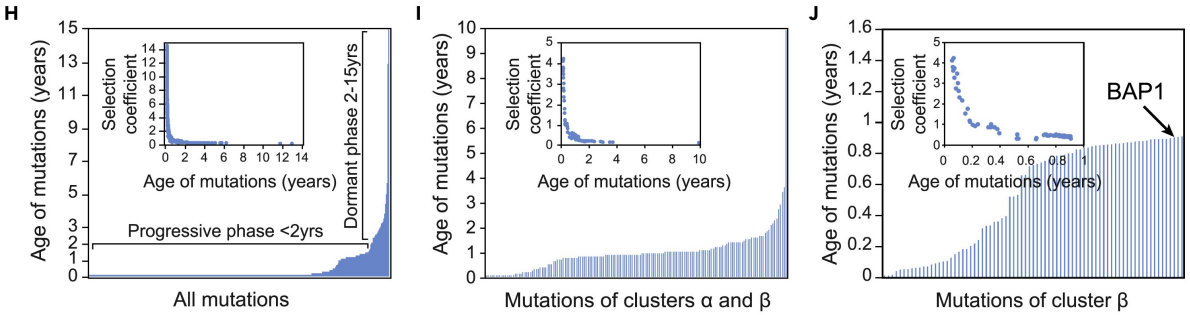
Figure 5.22: Spread of mutations in known cancer driver genes (according to Bailey et al. [7]) across the BLCA maps. Mutations restricted to the particular samples are shown in black, the spread mutations are marked in colors and shapes, and encircled. In *map19*, 4 groups of mutations with different spatial patterns are visible, two of them limited to the NU/LGIN samples. Clonal expansion in the normal-appearing urothelium supports the thesis of the *field effect* origins of BLCA. Fields not surrounded by a black frame were not sequenced.

they also coexisted within the apparently normal urothelium. This group contained 80 mutations in map24 and 43 mutations in map19. The final group β contained variants that were highly abundant in a subset of samples, including most of the HGIN and UC samples. These variants were responsible for the final progression of cancer, and at least some of them must have provided a selective advantage to the cells. This group counted 77 nonsilent variants in map24, and 155 variants in map19.

Mutations' ages and selection coefficients. In reference [10], an approach was applied that models the spread of the mutation across the fields of mucosa as a function of mutation age and its selective advantage. A section of the original paper describing the modeling process and details is attached in Appendix B. According to the model, the age of mutations varied from 0 to more than 15 years, with a dramatic increase in the number of mutations that occurred two years before the cystectomy. For this reason, two phases in the tumor progression were distinguished: the dormant phase, which started approximately 15 years before the cystectomy, and a progressive phase, which began around 2 years before the cystectomy. Mutations in the progressive phase were characterized by much higher selection coefficients than these in the dormant phase. It also included all the β mutations, characterized by high VAFs and presence in the HGIN and UC samples.



(a) Map 19



(b) Map 24

Figure 5.23: Age and selection coefficients of: all mutations, α and β mutations, and solely β mutations. Source: Bondaruk et al. *The origin of bladder cancer from mucosal field effect* [10], Figure 7H-J and S21H-J, on Creative Commons Attribution – NonCommercial – NoDerivs (CC BY-NC-ND 4.0) license.

Mutational signatures of dormant and progressive phase mutations. The mutations in the progressive phase were enriched in C>T substitutions and associated with increased diversity of mutational processes compared to the dormant phase (Fig. 5.24). In map19 (basal), many mutations were associated with the activity of signatures 1 (aging), 2 (APOBEC activity), 3 (dispaired DNA repair by homologous recombination), 5 (eti-

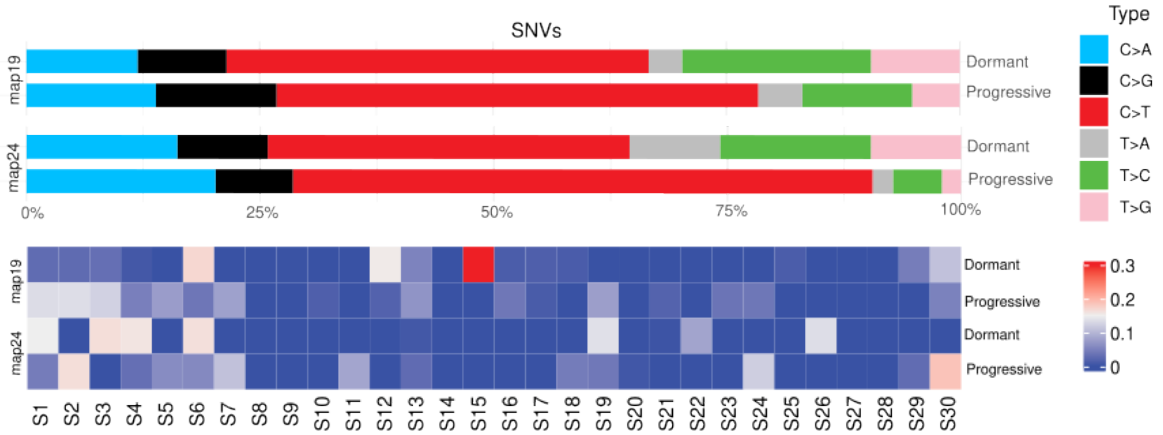
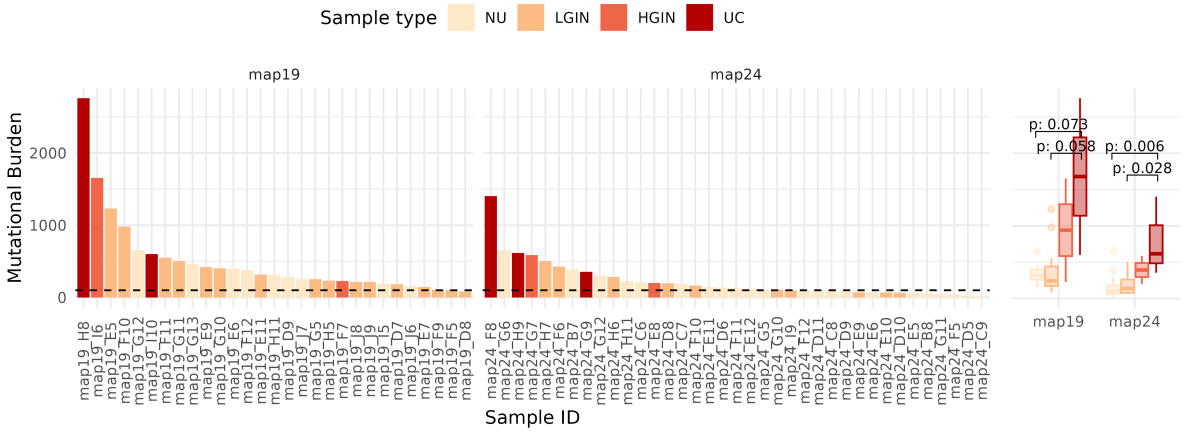
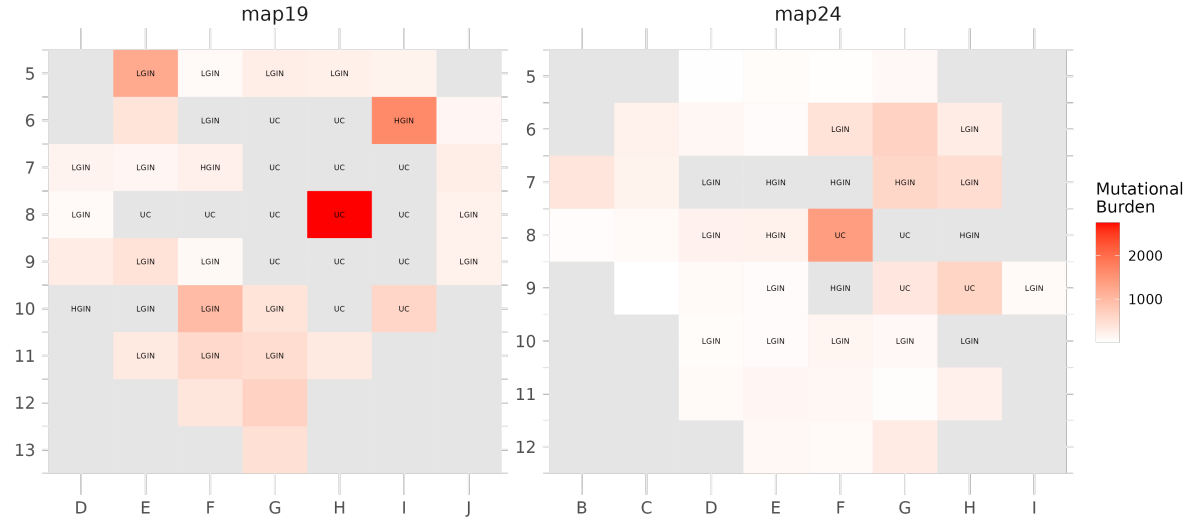


Figure 5.24: Substitution types (top) and mutational signatures (bottom) of dormant and progressive phase mutations. Progressive phase mutations are enriched in C>T substitutions.



(a) TMB values across all samples. *p*-values: two sample *Wilcoxon* test.

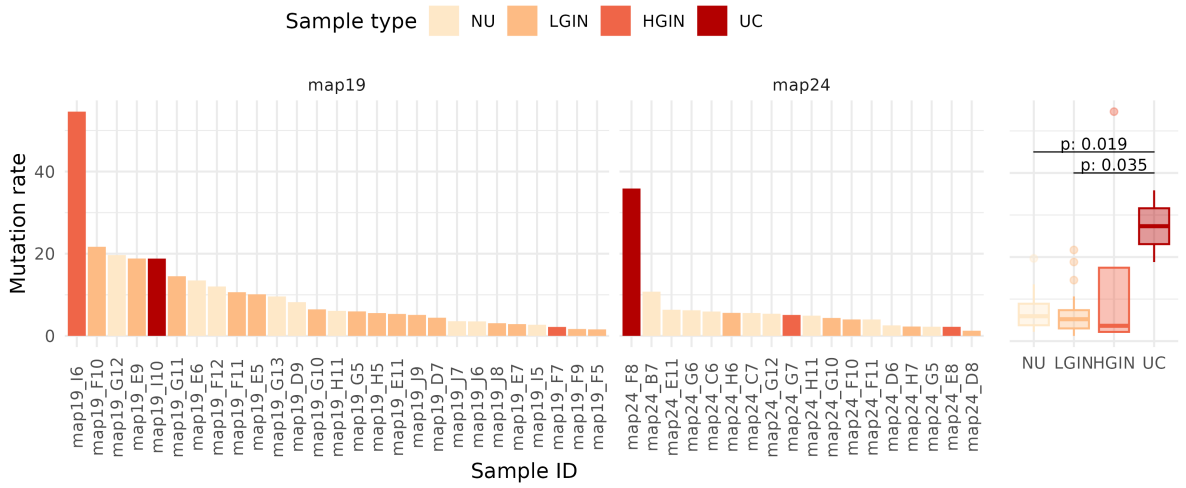


(b) TMB spread across the maps. *gray* - no WXS data; LGIN, HGIN and UC fields are labelled on the map, unlabelled fields - NU.

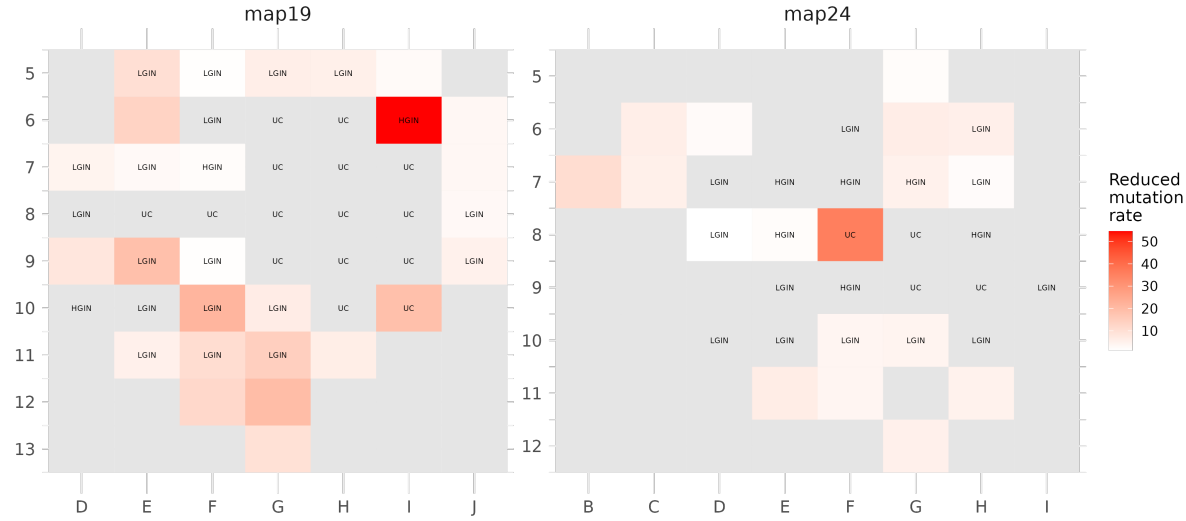
Figure 5.25: Tumor Mutational Burden across BLCA specimens.

ology unknown), 7 (UV exposure), and 19 (etiology unknown). Activities of signatures 6, 15 (defective DNA mismatch repair), and 12 (etiology unknown) were significantly lower than in the dormant phase. In map24 (luminal), the activities of signatures 2 (APOBEC activity), 7 (UV exposure), 11 (alkylating agents), 24 (exposure to aflatoxin), and 30 (etiology unknown) were increased in the progressive phase. Signatures 1 (aging), 3 (dispaired DNA repair by homologous recombination), 4 (smoking), 6, 26 (both associated with defective DNA mismatch repair), 19 (etiology unknown), and 22 (exposure to aristolochic acid) were less active than in the dormant phase.

Tumor Mutational Burden Total mutational burden (TMB) was rising progressively, with median values increasing from 344 and 149 in NU to 941 and 390 in HGIN and 1679 and 791 in UC in maps 19 and 24, respectively (Fig. 5.25a). TMB was consistently higher



(a) Mutation Rates values across all samples. *p*-values: *Wilcoxon* test.



(b) Mutation Rates across the maps. *gray* - no WXS sequencing data or insufficient number of mutations; unlabelled fields are normal urothelium (NU).

Figure 5.26: Mutation Rates per cell division across BLCA specimens under the exponential growth model.

in the basal map 19 compared to the luminal map 24 at each stage of progression. There was substantial variability of TMB within each group of samples. Both maps contained NU, LGIN, and HGIN samples with similar TMB values, and at least one NU sample with greater TMB than at least one UC sample. In map 19, only 1 out of 29 samples had fewer than 100 detected mutations, in contrast to 15 out of 37 samples in map 24, which corresponds to a greater bladder area affected by the disease. In map 19, high TMB was observed even in the most distant NU samples at the bottom of the map (Fig. 5.25b). The most mutated samples, map19_H8 and map24_F8, contained 2758 and 1655, respectively, which gives approximately 46 and 27 mutations per exome Mb. This value is within the range of values reported in the literature [58].

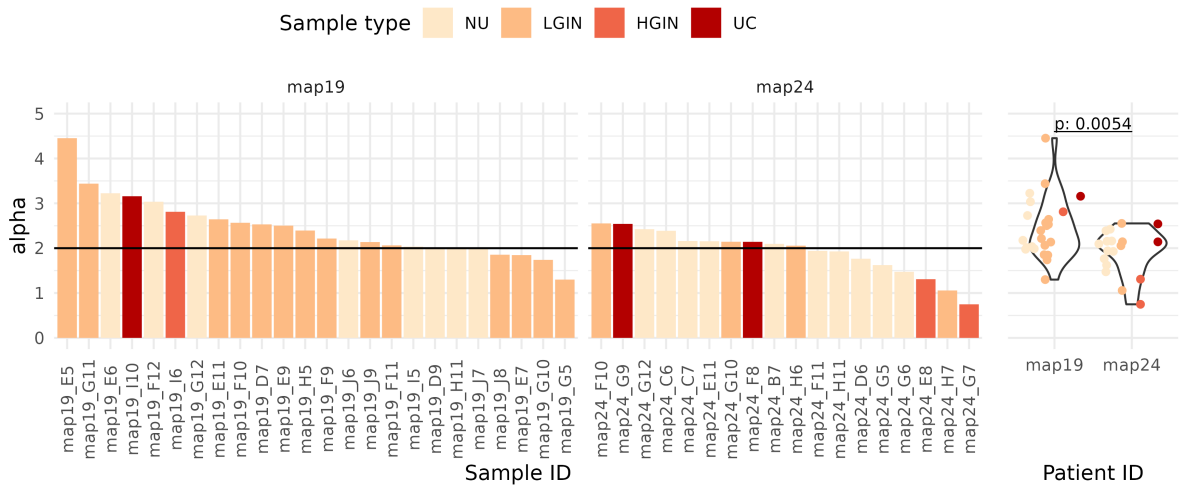
5.2.2 Mutation rates under the exponential growth model

Based on the data described in previous sections, we estimated the mutation rates per cell division (MR) across the BLCA maps under the neutral-power-law models. We excluded from the analysis samples with the lowest number of mutations using an arbitrary threshold of 100 mutations, and fitted the neutral power-law models using *cevomod* package and the approach described in Section 4.2.6. Models in 3 samples of map 19 (regions E5, E6, and I6) were significantly detached from the spectra, resulting in overestimated MR values. We refitted these models without trimming the highest-frequency variants, which resulted in improved fits. We also discarded the models of 5 other samples: from field H8 in map 19 and fields E12, F6, G9, and H9 in map 24, in which we could not visually detect the true neutral tails and evaluate the correctness of the fits.

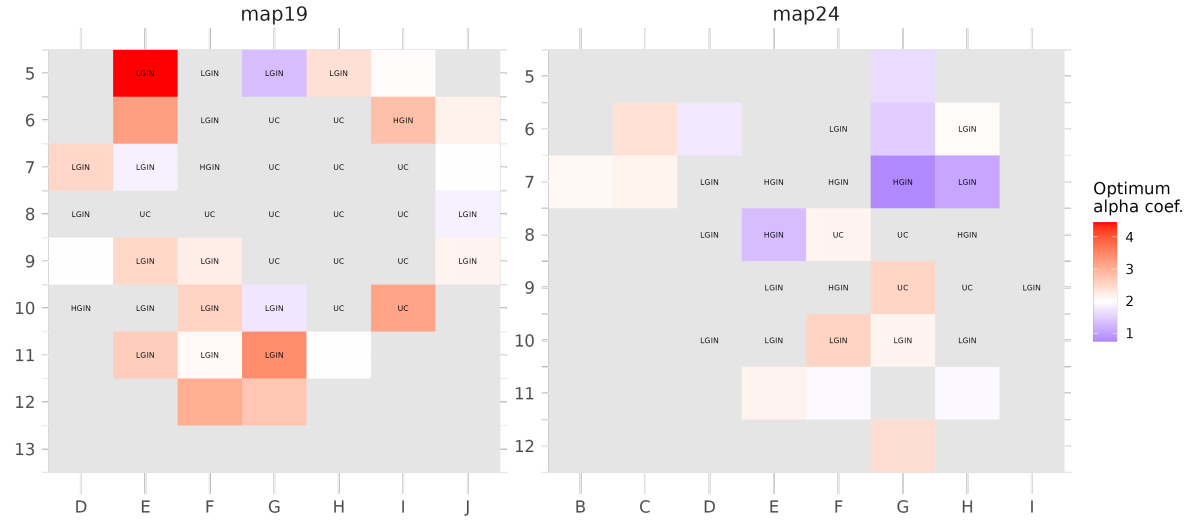
Estimated MR values for most samples were in the range of 1 to 20 mutations per cell division in map 19 and 1 to 10 in map 24 (Fig. 5.26a). Only in two fields: I6 (HGIN) in map 19 and F8 (UC) in map 24, the MR values were higher: 55 and 36, respectively. In general, MR values were similarly low in NU and LGIN, high in UC, and highly variable in HGIN, with the lowest values below the median NU in LGIN, and the highest value in sample I6 of map 19, mentioned before. MR estimate for this sample was nearly 3 times higher than for the UC field I10. In association with the higher TMB (Fig. 5.25a) and more remarkable similarity of the driver mutations to the other urothelial cancer field (H8, Fig. 5.22), it shows that the I6 field is more transformed than I10, despite its classification by a pathologist as HGIN. It is also worth noting that at least three NU and LGIN samples in map 19 exhibited higher MR values than UC sample I10. These samples contributed to a larger area of NU/LGIN fields, located further from the main UC body, with elevated MR (Fig. 5.26b) and several driver mutations present, which were not found in the primary UC samples (Fig. 5.22).

5.2.3 Optimization of the power-law exponent.

Next, we checked if the best-fitting power coefficients α of the power-law component are equal to 2, as predicted under the assumptions of exponential tumor growth, constant mutation rate, and in the absence of selectively advantageous micro-clones. For this purpose, we used our *cevomod* package to fit the data with our second type of model, optimizing both coefficients of the power-law component, A and α. Upon visual evaluation, we excluded 8 fits from the analysis: F5, F7, G13, and H8 in map 19, and D8, E12, F6, and H9. In these fields, the neutral tails contained too few bins to be unambiguously fitted with the model containing two parameters. Furthermore, we refined 2 inaccurate fits in each map. In fields G6 and H7 in map 24, having significant binomial peaks at intermediate frequencies, the power-law curves were erroneously fitted to the right slopes of binomial peaks. We refitted the power-law components of these models after truncat-



(a) α values across all samples. p-value: *t-test*.

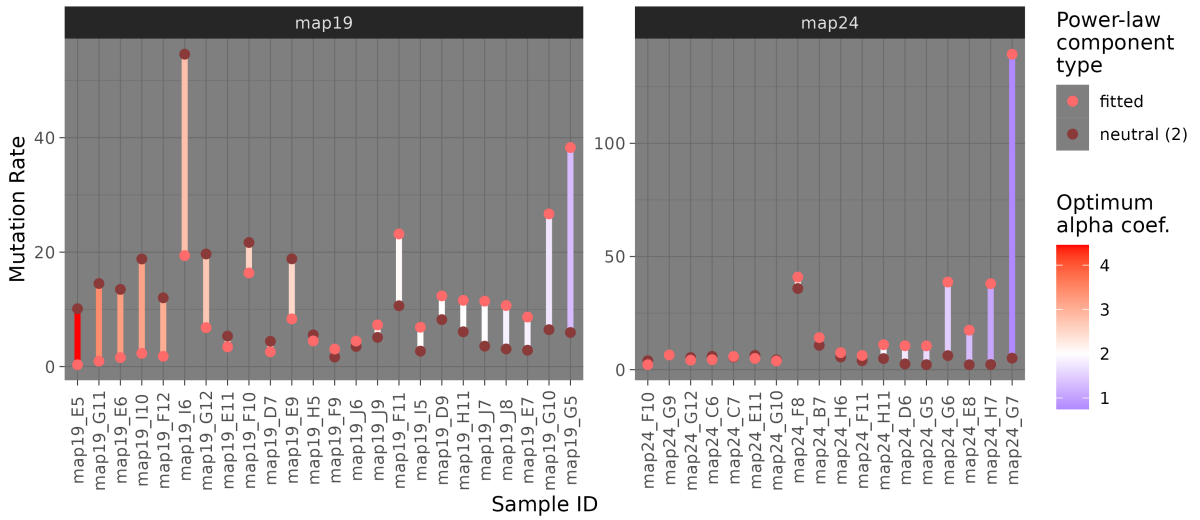


(b) α values across the maps. *gray* - no WXS sequencing data or insufficient number of mutations; unlabelled fields are normal urothelium (NU).

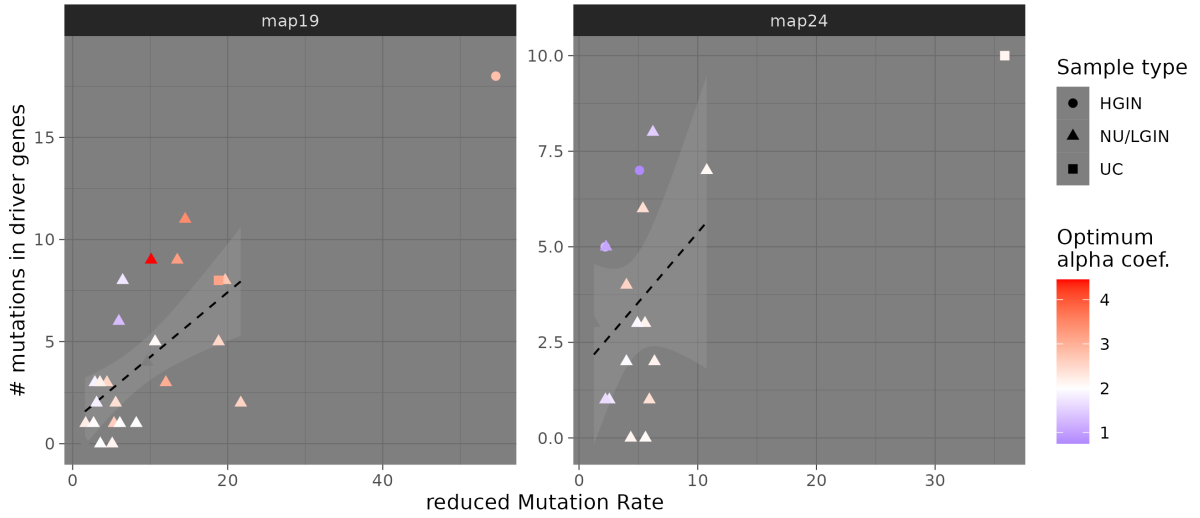
Figure 5.27: Optimum α coefficients in second-type models.

ing the spectra above the frequency of 0.15. In samples E5 and F10 in map 19, the fitted models were detached from the spectra. We refitted these two models with the threshold under which the bin of the VAF spectrum is considered empty (*y_threshold_pct* argument in *cevomod*) lowered to 0 .

Similarly to the previously analyzed cohorts, the optimum α values frequently deviated from the value of 2 corresponding to neutrality (Fig. 5.27a). Distinct trends were observed in both maps: in map 19, the α coefficients more frequently exceeded 2, with an average value of 2.45 (SD: 0.67). In map 24, the numbers of samples with α values greater and lower than 2 were similar, but the downward deviations were more significant. An average α value in map 24 was 1.91 (SD: 0.5). Interestingly, the most significant deviations in both maps occurred in the intermediate stages of LGIN and HGIN. In map 19, the highest α



(a) α values and the estimated MR in both types of models. Dark red dots mark the MR estimates under the neutral-like $\alpha = 2$; light red dots show the estimates of the initial MR under the optimum α coefficient, shown by the color of the connector.



(b) MR under the neutral-like model ($\alpha = 2$) versus the number of mutations in known driver genes. Shapes show the sample classification and color the optimum α coefficient.

Figure 5.28: Optimum α coefficients versus the mutation rates and driver mutation counts.

value greater than 4 appeared in LGIN sample E5. In map24, the lowest α was below 1 and appeared in HGIN sample G7. Most of the negative α values in map 24 were concentrated right to the main UC body (Fig. 5.27b).

5.2.4 Optimum power-law exponent versus the mutation rates and selection.

As it was shown by Tung [120] and in Section 4.3, α values can be altered by both the presence of selected micro-clones (downward) and the change of the mutation rate (any direction) and the distinguishing between these two phenomena is not trivial. The

positive selection among micro-clones and increasing mutation rate have opposite effects on α and might leave it unchanged in many samples if these forces are correlated. We used our experimental data to check what low and high α values are associated with. First, we found that in models with the optimized α parameters greater than 2, the estimated MR under the optimum α was significantly lower than in models with the neutral-like α equal to 2 (Fig. 5.28a). In other words, the greater alpha, the higher MR under the neutral model, compared to MR under the optimized alpha. Second, we found that samples with alpha values below 2 were characterized by low MR (according to the neutral-like model) and an increased number of mutations in driver genes (Fig. 5.28b). Confirming whether they do represent Tung-Durrett two-type model with selected micro-clones might not be possible using the bulk sequencing data.

5.2.5 Discussion

Study achievements. In collaboration with the group of Dr. Bogdan Czerniak from MD Anderson Cancer Center in Houston, US, we studied the evolutionary dynamics in two whole organ maps representing two distinct molecular subtypes of BLCA. In the paper by Bondaruk et al. [10], it was shown that molecular changes are widely present in the normal-appearing bladder mucosa. These changes include altered DNA methylation, transcriptional reprogramming, and early clonal expansions of urothelial clones bearing mutations in known cancer driver genes. The presence of the mutations in cancer driver genes in the normal urothelial tissue was also reported in other studies [50]. We believe these changes are associated with the so-called mucosal field effect, eventually initiating cancer growth.

The analysis of the mutational landscape of the two specimens revealed the presence of 3 groups of mutations: A, α, and β, which present different spread patterns, VAFs, ages, and selection coefficients. These types of mutations were associated with 2 phases of the tumor progression: low-frequency A mutations and low-frequency, but spread across many samples, α mutations drove the dormant phase, which lasted more than 10 years. Approximately 2 years before the cystectomy, the progressive phase began, driven by β mutations with a high selective advantage. We have shown that the mutations in the progressive phase resulted from different mutational processes, including APOBEC activity in both maps and dispaired homologous recombination in basal map 19.

We used our first-type model with the neutral-like power coefficient to estimate the mutation rates per cell division (MR) under the model of exponential tumor growth. MR was increasing with the progression of the disease; however, it was already elevated in many NU and LGIN samples, especially in the basal map 19. HGIN samples revealed the greatest variability of MR; in many HGIN samples, MR was comparable to NU/LGIN, but in sample I6 of map 19, MR was significantly higher than in one of the UC samples,

I10, in this map. It shows that the morphological appearance of the tissue may not always directly reflect the molecular state of the cells. Advanced mutational processes may be highly active even in non-cancerous tissues.

Using our second-type model, we checked how well the spectra followed the expected power-law shapes of the neutral tail distributions. We found that in many samples, the optimum power coefficient α deviated from the expected value of 2. The deviation trends were different in both maps. In map 19, the optimum α coefficient was greater than 2 in the majority of samples, contrary to map 24, in which the proportions of samples with α greater and lower than 2 were similar; however, the deviations downward were more significant. Areas with high and low α formed spatial patterns in both maps, with two high-α regions in map 19 and one low-α region in map24, which shows that the processes that altered α were similar in many adjacent regions.

Deviations of α can result from both the incidence of selection among many small clones and the changes in the mutation rate. Our analysis showed that MR estimates under the power-law component with α greater than 2 are smaller than the estimates in models with α fixed and equal to 2. Thus, high alpha can signal high actual MR in the sample. A downward deviation of α can result from the presence of selectively advantageous micro-clones [120] however, the MR increase may compensate for the impact of competing micro-clones (see the Section 4.3). Indeed, a downward deviation was less common than an upward deviation, and we identified the increase of MR with the progression of the disease. Low α values were identified in some samples with low MR but a high count of mutations in cancer driver genes. The ultimate answer if these are the samples with true competition among micro-clones may be out of reach of the bulk sequencing data analysis since it does not allow for distinguishing numerous small clones [113].

Fitting the evolutionary models to the data with low mutational burden where neutral tail, clonal and subclonal components are hard to identify is challenging. Using our cevomod package, we were able to fit the power-law-shaped components to the most spectra in both maps, despite the very low mutational burden of many NU and LGIN samples.

Limitations. The presented study was limited to the two extensively studied specimens. Thus we cannot state whether the differences between the basal and luminal maps described in this section are characteristic of these molecular subtypes. Also, this study outlines the molecular characteristics of bladder cancer progression via the dysplasia carcinoma in situ sequence but does not explore its development through the more frequently observed low-grade papillary pathway. More specimens need to be analyzed to better characterize the progression of basal and luminal bladder cancers.

Chapter 6

Summary

The subject of our study was the evaluation of the role of mutagenesis and selection in the evolution of cancer genomes. We stated a thesis that *Changes in the evolutionary dynamics of cancer upon metastasis and recurrence can be inferred from the bulk DNA sequencing data.* To prove this thesis, we collected 4 datasets, each including bulk DNA sequencing data from multiple tumors and multiple samples per tumor. We used two published datasets: the acute myeloid leukaemia data from the diagnosis and relapse time points published by Shlush et al. [107] and bladder cancer data from multiple sites of bladder urothelium with different stages of the disease published by Bondaruk et al. [10]. Two other datasets contained our unpublished data for breast and laryngeal cancers and included the data from the primary tumors and lymph node metastases.

6.1 Study achievements

***cevomod* R package.** It was shown that the Variant Allele Frequency Spectra of tumor samples can be approximated with a mixture of power-law and binomial components, which model the distributions of neutrally occurring variants and variants with frequencies higher than expected due to the selection [37, 130]. These models can be used to estimate the evolutionary parameters of cancer evolution. However, the existing tools are not suitable for the analysis of whole exome sequencing or low-coverage data due to the insufficient number of low-frequency, *neutral tail* variants. We developed a new R package, *cevomod*, which can model this data despite the incompleteness of the neutral tail VAF distributions. cevomod can be installed from its GitHub repository at `https://github.com/pawelqs/cevomod`, and its full documentation is available at `https://pawelqs.github.io/cevomod/`. cevomod also allows one to choose between two types of models, a neutral-like one with the power-law exponent equal to 2 and an optimized model, in which the exponent is optimized to fit the data best. Our package is fast; the time required to fit the model in one sample rarely exceeds a few seconds.

Evolutionary dynamics in tumor progression, metastasis, and relapse. With our modeling approach, we were able to fit the population genetics models [37, 18] to our 4 sets of data and to estimate the evolutionary parameters in the primary and secondary tumor samples. The identified differences in the mutation rates were up to 2-fold in AML and BRCA, without the predominant direction up or down. In LSCC, the upward changes were more common, although they were usually not as significant as in the other cohorts. We detected minor subclones in nearly all samples. Most of them were characterized by early emergence times, small cellular frequencies, and small selective advantages, especially in BRCA and LSCC cohorts. Small selection coefficients of these clones and lack of clear clonal peaks, and strong bias of driver mutations VAFs towards 0 indicated that the evolution of most of the BRCA and LSCC samples from the tumor initiation to sequencing could be nearly neutral, and the tumors were sequenced before the first selective sweep was achieved.

In bladder cancers, we fitted the power-law models to samples with different stages of disease progression. We found an increase in the tumor mutational burden and the mutation rate along the progression from normal urothelium samples through intraurothelial neoplasia to urothelial cancer. However, the mutation rates were already elevated in many normal urothelium samples, which were not adjacent to the urothelial cancer samples. These elevated mutation rates coincided with the expansions of cells with mutations in known cancer driver genes.

Most samples in the AML cohort showed clear peaks of clonal mutations, indicating the past selective sweeps. In this cohort, we recognized subclones with higher selection coefficients and later emergence times compared to the BRCA and LSCC cohorts. However, we also identified samples in which the neutral tail could not be accurately fitted with the power-law component with an exponent equal to 2. To investigate these cases, we developed a second model-fitting approach, in which the power-law exponent α is also fitted to the data.

Optimum exponent coefficients. We optimized the power-law components in all samples and found common deviations from the theoretical exponent value of 2. In many samples, we observed the optimum exponent values less than 2, which might indicate the presence of selectively advantageous micro-clones, described by Tung and Durrett [120]. However, we discovered that the upward deviations of α were more common than the downward ones. To explain this phenomenon, we demonstrated that mutation rate changing during the tumor progression can result in neutral tails with exponent values different than 2 (Section 4.3). In samples with high α values, the initial mutation rates estimated by the second-type models were smaller than the mutation of the first-type models, assuming the constant mutation rates. This observation supports our hypothesis that increased power-law exponents indicate an increase in the mutation rate. In some

tumor types, the low α values were associated with low overall mutation rates and a high number of mutations in known driver genes. Determining whether these samples represent the selectively advantageous micro-clones of Tung and Durrett [120] might not be possible with the bulk DNA sequencing data.

In AML, BRCA, and LSCC cohorts, most of the secondary tumors (relapse samples and lymph node metastases) had α increased compared to the primary tumors (diagnostic and primary tumor samples). An increase in mutation rate can be supporting in these samples, accelerating cells' adaptation [112] in the new environment.

6.2 Conclusion

Using the bulk sequencing data from over 140 samples, we have shown that evolutionary parameters of cancer evolution can be estimated from the bulk sequencing data and that there are quantifiable differences in the evolutionary dynamics of the primary and secondary tumor samples. We have thus proved our thesis that we stated, that *changes in the evolutionary dynamics of cancer upon metastasis and recurrence can be quantified from the bulk DNA sequencing data*, which was the first goal of this thesis. We have implemented an R package that can be used by other researchers with the WXS data, completing our second goal. Finally, we have shown that the assumptions underlying most frequently used models used to estimate the parameters of tumor evolution may be violated in many cancers and proposed a mathematical explanation for the observed phenomena, fulfilling the third goal of the thesis.

6.3 Limitations.

We are aware of some limitations of *cevomod* and this study.

First, *cevomod*, contrary to the better-known MOBSTER, does not prove or reject the hypothesis of neutrality or selection in the tumor. It was shown by Bozic [11], McDonald [83], that selection not always leaves clear trace in the spectrum. Also, selection is not a zero-or-one phenomenon, and weak selection might be difficult to distinguish from the neutrality. For this reason we prefer to report close-to-zero selection coefficients instead of rejecting the hypothesis of the non-neutral evolution.

Second, in cevomod, the binomial components are fitted to the residuals of the power-law component. Simultaneous fitting of both components could result a more accurate fit in some cases. However, we think that it is reasonable to try to explain maximum variability using the neutral model first, before claiming the presence of selection.

Third, the CNV-based correction of the variant frequencies was not a part of this study. Copy-number alterations and contamination of sample with the normal cells are known confunders in the bulk DNA sequencing data [100, 116]. The correction can be performed

using an approach described by Dentro, Wedge and Van Loo [28] and requires estimates of tumor purity and allele-specific copy numbers. However, the reliable estimation of these parameters from the WXS data is challenging, as we have shown in [68]. The purity estimates by algorithm FACETS [106] were highly discordant with the estimates provided by histopathologists. We decided not to use these estimates, and excluded variants with extremely high or low sequencing coverage. Variants most affected by the CNVs are more likely to fall into those outliers than variants from the diploid regions of the genome. However, we plan to include the variant frequency correction in cevomod, as we recently get the shallow whole exome sequencing results for LSCC cohort, and we expect to get these results for our future studies as well.

Finally, equations underlying our first model are based on the assumptions of exponential tumor growth and constant mutation rate [77]. The exponential growth model was found true for some breast cancer, [115], but does not necessarily apply to all tumors. In fact, some other studies identified the logistic or Gompertz models as more accurate for many tumors. The second assumption of the constant mutation rate is also not always fulfilled. Our results indicate that a mutation rate increase might be common in cancer.

6.4 Future work

Our study of the evolution of breast and laryngeal cancers finds continuation in our grant funded by Polish National Science Center, *Evolutionary genomics: modeling and predicting breast and lung cancer progression*, no. 2021/41/B/NZ2/04134. In this new project, we include a new cohort of lung cancer patients in the study and complement the WXS with shallow whole genome sequencing (sWGS) for reliable detection of copy number variants. sWGS experiments were recently performed also on the LSCC samples analyzed in this work.

Our collaboration on studying bladder cancer origins is continued. The actual works include more specimens: map no. 26 is the subject of the current group study, and samples from 6 more maps have been recently sequenced. All these specimens were subjected to multi-omic experiments, including RNA sequencing, proteomic analysis, and deep WXS, exceeding 300x, of multiple regions of the bladder mucosa. Also, single-cell RNA sequencing experiments were performed to study the role of CAB39L and LPAR6, the forerunner genes driving the evolution of BLCA along the basal and luminal paths.

There are a few ways of possible cevomod improvements. First, implementing the CNV-based correction of variant frequencies would allow us to fully utilize the sWGS data available for our lung and laryngeal datasets. This feature is missing in the current version of the package. Second, the subclonal structure of tumors could be resolved more precisely using the information from multi-sample of the same patient. Future releases could use the clustering algorithms such as PyClone-VI instead of the currently used

BMix. Third, the implementation of a web application would make cevomod available for researchers not familiar with R. Such an application could be implemented using a Shiny framework.

List of Figures

List of Tables

Appendices

A Cancer driver genes

Cancer	Type	Cancer Driver Genes
BLCA	oncogene	ERBB3, ERCC2, GNA13, RHOB, RXRA, SF1
BLCA	possible oncogene	DIAPH2, KLF5
BLCA	possible tsg	ASXL2, ELF3, FOXA1, FOXQ1, TXNIP, ZFP36L1
BLCA	tsg	SPTAN1
BRCA	oncogene	CHD4, FOXA1
BRCA	possible tsg	CBFB, CDKN1B, GATA3, TBX3
BRCA		PTPRD
HNSC	oncogene	MYH9
HNSC	possible oncogene	KEAP1
HNSC	possible tsg	CYLD, HUWE1, ZNF750
HNSC	tsg	FLNA
LAML	oncogene	FLT3, IDH2, KIT, NPM1, PTPDC1, SMC1A
LAML	possible oncogene	CEBPA, SMC3
LAML	possible tsg	PHF6, WT1
LAML	tsg	ASXL1
LAML		EZH2, RAD21, TET2
PANCAN	oncogene	AKT1, BRAF, CDK4, CTNNB1, CUL1, EGFR, ERBB2, FGFR2, FGFR3, GNA11, GNAQ, GTF2I, HRAS, IDH1, KRAS, MAP2K1, MAPK1, MYC, NFE2L2, NRAS, PCBP1, PIK3CA, PIK3R2, PPP2R1A, PTPN11, RAC1, RHOA, SF3B1, SOS1, SPOP, U2AF1
PANCAN	tsg	AJUBA, APC, ARHGAP35, ARID1A, ARID2, ATF7IP, ATM, ATRX, AXIN1, BAP1, BCOR, BRCA1, BRD7, CASP8, CDH1, CDK12, CDKN1A, CDKN2A, CHD8, CREBBP, CTCF, CTNND1, CUL3, DDX3X, DNMT3A, EP300, EPHA2, FAT1, FBXW7, FUBP1, GPS2, HLA-A, HLA-B, IRF2, JAK1, KANSL1, KDM6A, KMT2A, KMT2B, KMT2C, KMT2D, MAP2K4, MAP3K1, MGA, NCOR1, NF1, NF2, NIPBL, NOTCH1, NSD1, PBRM1, PIK3R1, PSIP1, PTEN, RASA1, RB1, RBM10, RNF111, RNF43, RPL5, RUNX1, SCAF4, SETD2, SMAD4, SMARCA1, STAG2, STK11, TCF12, TGFBR2, THRAP3, TP53, TRAF3, TSC1, USP9X, VHL

Table 1: Lists of driver genes for cancer types analysed in the thesis. *tsg* - tumor suppressor gene, *PANCAN* - Pan-Cancer *BLCA* - Bladder Urothelial Carcinoma, *BRCA* - Breast Cancer, *LAML* - Acute Myeloid Leukemia, *HNSC* - Head and neck squamous cell carcinoma, supertype of *LSCC* - Larynx Squamous Cell Carcinoma. Source: Bailey et al. [7]

B Modeling of bladder cancer evolution from field effects

This section describes the model used to estimate the mutation ages and selection coefficients in BLCA cohort, and originates from the paper Bondaruk et al. *The origin of bladder cancer from mucosal field effect* [10].

To reconstruct the time of evolution from mucosal field effects to bladder cancer, the time-continuous Markov branching process with immigration and parsimonious principles was used [72]. In brief, a mutation j appears at time t_0^j in a progenitor cell of the urinary bladder urothelial lining and gives rise to a mutant clone. Mutant cells divide at rate λ (1/year), and after division, one cell enters self-renewal and the other differentiates with probability $1 - s_j$ or both cells enter self-renewal with probability s_j. As a consequence, the mutant clone grows exponentially as $exp(\lambda s_j t)$, where t is the age of the j-th mutant's clone counted from t_0^j. The secondary clones expand, involving different areas of bladder mucosa at times t_i^j, $i \geq 0$ modeled by a stochastic Poisson process with intensity ν (1/yr) [75]. If the expected cell counts in the successive j-th mutant clones are denoted by $X_i^j(t)$, $i = 0, 1, 2, \ldots$, and the number of haploid genomes in normal uroprogenitor cells are denoted by $2N$, the corresponding VAFs $V_i^j(t)$ are defined as the ratios $V_i^j(t) = X_i^j(t)/(2N)$ and are computed as follows [72]:

$$E\left[V_i^j(t)\right] = \exp(\lambda s_j t)\left(\frac{\nu}{\nu + \lambda s_j}\right)^i \int_0^{(\nu+\lambda s_j)t} \frac{u^{i-1}}{(i-1)!}\exp(-u)du/(2N), i = 0, 1, 2, \ldots$$

For any mutation j of age t_j, the sequence of expectations $E\left[V_i^j(t_j)\right], i = 0, 1, 2, \ldots$, was computed to estimate the coefficients $a_j = \lambda s_j t_j$ and $b_j = \mu t_j$. The coefficient $c = 2N$ is a constant parameter representing an estimate of the number of uroprogenitor cells in the sampled area. The computations were performed for $10^2 - 10^5$ uroprogenitor cells in the sampled mucosal area, which did not significantly change the time modeling results, but the best fit was obtained with 5×10^3 uroprogenitor cells, for which the data are presented. With a cell division rate λ and migration rate μ, the parameter b_j is the proxy for mutation age t_j, whereas the ratio a_j/b_j is the proxy for selection coefficient s_j. A fitting algorithm with the optimization programs fminsearch and fminbnd in the MATLAB programming language was used to estimate the sequence of mutations in tumor development [71, 13, 38]. The resulting time estimates were presented as bar diagrams representing the ages of mutations and point charts of the corresponding selection coefficients.

C Supplementary figures

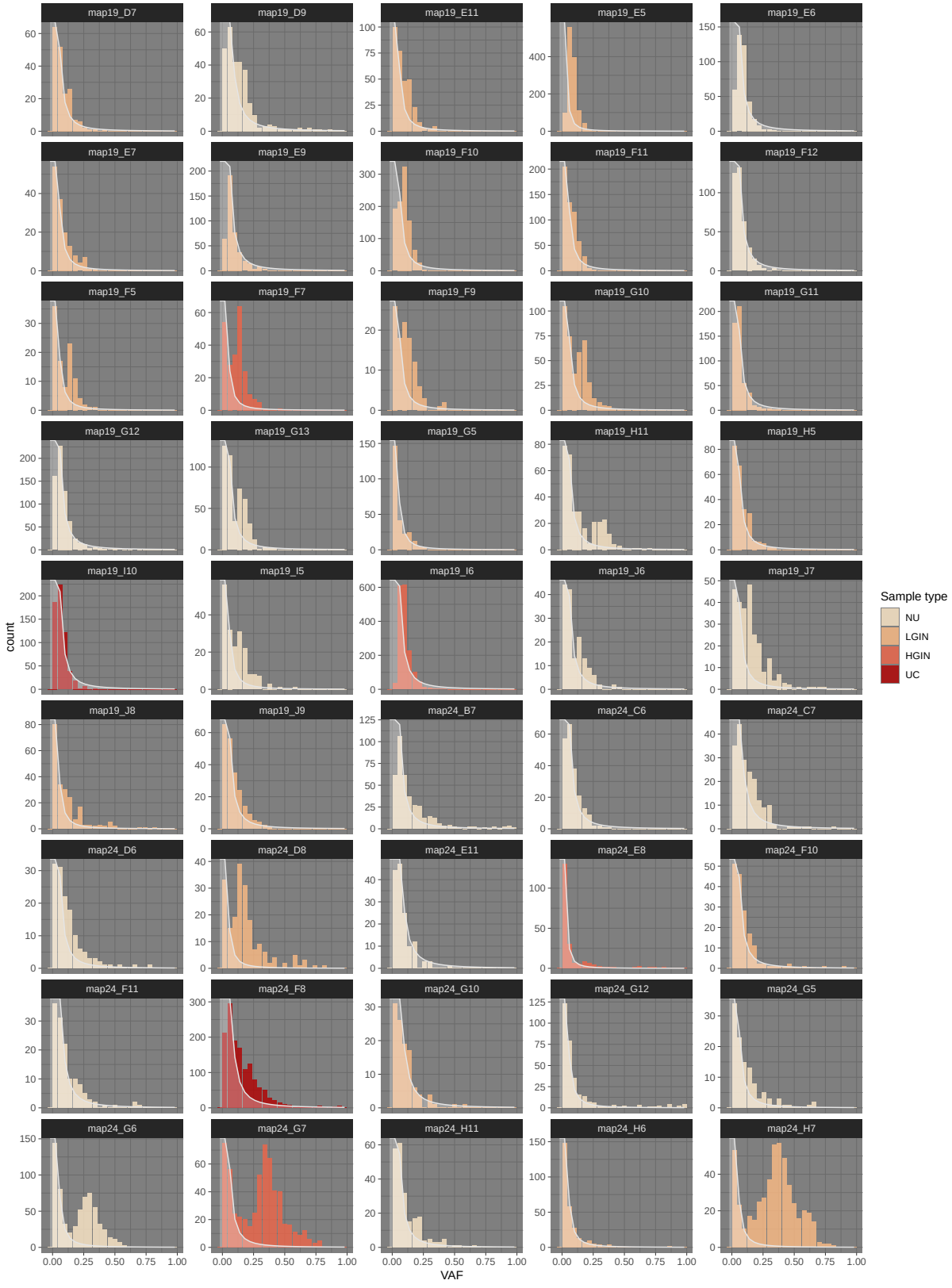


Figure 1: Neutral power-law component ($\alpha = 2$) fits in the BLCA samples.

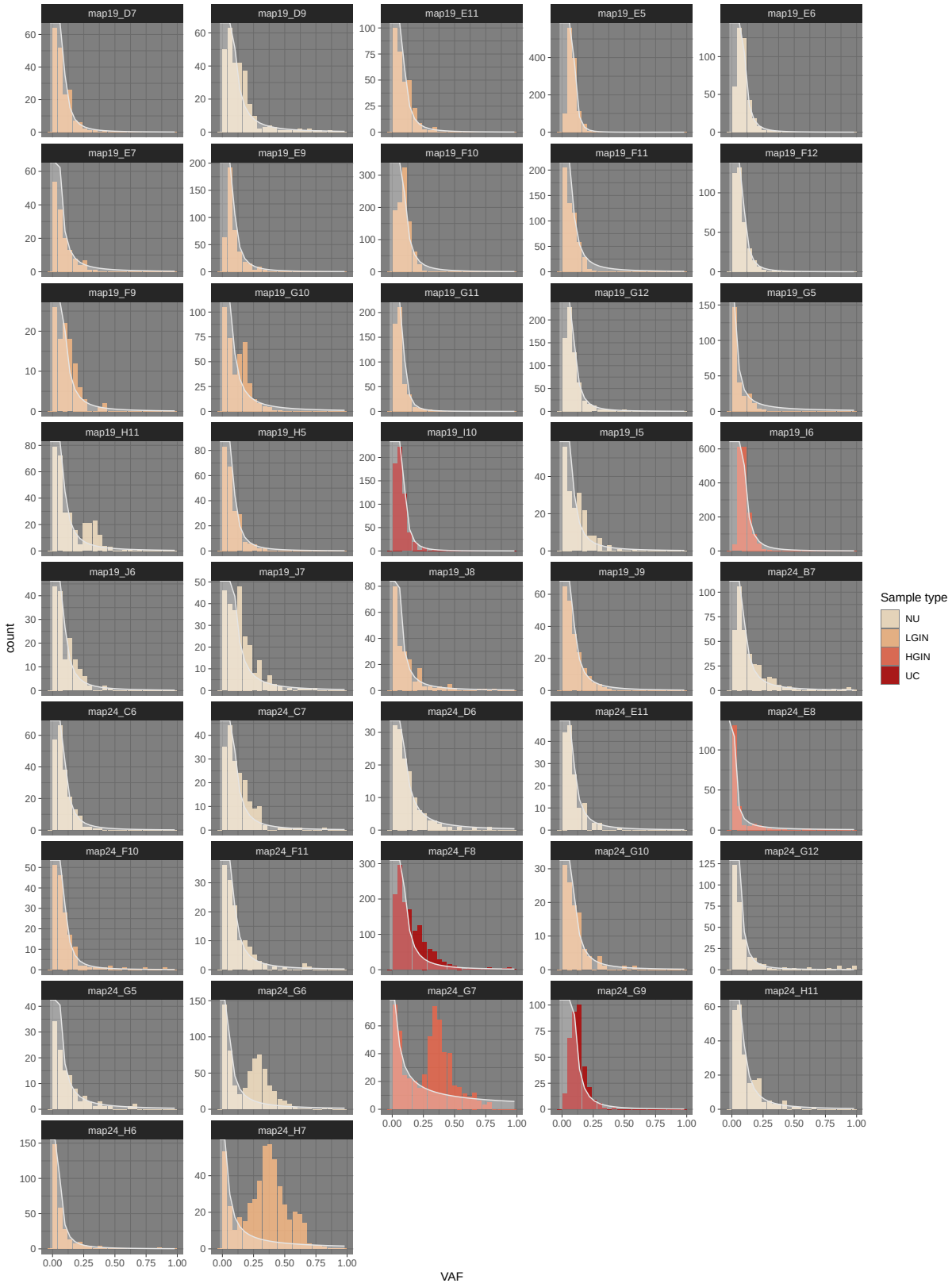


Figure 2: Optimized power-law component (α fitted) fits in the BLCA samples.

Bibliography

[1] Ivan Adzhubei, Daniel M. Jordan and Shamil R. Sunyaev. 'Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2'. In: *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]* 0 7 (Jan. 2013), Unit7.20. ISSN: 1934-8266. DOI: 10.1002/0471142905.hg0720s76. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4480630/ (visited on 26/03/2023).

[2] Sergey Aganezov et al. 'A complete reference genome improves analysis of human genetic variation'. In: *Science* 376.6588 (Apr. 2022), eabl3533. DOI: 10.1126/science.abl3533. URL: https://www.science.org/doi/10.1126/science.abl3533 (visited on 26/03/2023).

[3] Ludmil B. Alexandrov et al. 'Signatures of mutational processes in human cancer'. en. In: *Nature* 500.7463 (Aug. 2013), pp. 415–421. ISSN: 1476-4687. DOI: 10.1038/nature12477. URL: https://www.nature.com/articles/nature12477 (visited on 15/01/2023).

[4] Ludmil B. Alexandrov et al. 'The repertoire of mutational signatures in human cancer'. en. In: *Nature* 578.7793 (Feb. 2020), pp. 94–101. ISSN: 1476-4687. DOI: 10.1038/s41586-020-1943-3. URL: https://www.nature.com/articles/s41586-020-1943-3 (visited on 15/01/2023).

[5] Simon Andrews. *FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]*. 2010.

[6] Adam Auton et al. 'A global reference for human genetic variation'. en. In: *Nature* 526.7571 (Oct. 2015), pp. 68–74. ISSN: 1476-4687. DOI: 10.1038/nature15393. URL: https://www.nature.com/articles/nature15393 (visited on 26/03/2023).

[7] Matthew H. Bailey et al. 'Comprehensive Characterization of Cancer Driver Genes and Mutations'. eng. In: *Cell* 173.2 (Apr. 2018), 371–385.e18. ISSN: 1097-4172. DOI: 10.1016/j.cell.2018.02.060.

[8] Yury A. Barbitoff et al. 'Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage'. en. In: *Scientific Reports* 10.1 (Feb. 2020), p. 2057. ISSN: 2045-2322. DOI: 10.1038/

s41598-020-59026-y. URL: https://www.nature.com/articles/s41598-020-59026-y (visited on 15/04/2023).

[9] Amy M. Boddy. 'The need for evolutionary theory in cancer research'. en. In: *European Journal of Epidemiology* (Nov. 2022). ISSN: 1573-7284. DOI: 10.1007/s10654-022-00936-8. URL: https://doi.org/10.1007/s10654-022-00936-8 (visited on 22/12/2022).

[10] Jolanta Bondaruk et al. 'The origin of bladder cancer from mucosal field effects'. In: *iScience* 25.7 (June 2022), p. 104551. ISSN: 2589-0042. DOI: 10.1016/j.isci.2022.104551. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9209726/ (visited on 18/01/2023).

[11] Ivana Bozic, Chay Paterson and Bartlomiej Waclaw. 'On measuring selection in cancer from subclonal mutation frequencies'. en. In: *PLOS Computational Biology* 15.9 (2019), e1007368. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007368. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007368 (visited on 18/01/2023).

[12] Ivana Bozic and Catherine J. Wu. 'Delineating the evolutionary dynamics of cancer from theory to reality'. en. In: *Nature Cancer* 1.6 (June 2020), pp. 580–588. ISSN: 2662-1347. DOI: 10.1038/s43018-020-0079-6. URL: https://www.nature.com/articles/s43018-020-0079-6 (visited on 03/01/2023).

[13] Richard P. Brent. *Algorithms for Minimization Without Derivatives*. en. Google-Books-ID: AITCAgAAQBAJ. Courier Corporation, June 2013. ISBN: 9780486143682.

[14] Claudia Buhigas et al. 'The architecture of clonal expansions in morphologically normal tissue from cancerous and non-cancerous prostates'. In: *Molecular Cancer* 21.1 (Sept. 2022), p. 183. ISSN: 1476-4598. DOI: 10.1186/s12943-022-01644-3. URL: https://doi.org/10.1186/s12943-022-01644-3 (visited on 18/01/2023).

[15] Derek Caetano-Anolles. *Somatic short variant discovery (SNVs + Indels)*. en-US. Mar. 2023. URL: https://gatk.broadinstitute.org/hc/en-us/articles/360035894731-Somatic-short-variant-discovery-SNVs-Indels- (visited on 10/04/2023).

[16] *Cancer*. en. Feb. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/cancer (visited on 22/12/2022).

[17] *Cancer stem cell*. en. Page Version ID: 1146463528. Mar. 2023. URL: https://en.wikipedia.org/w/index.php?title=Cancer_stem_cell&oldid=1146463528 (visited on 28/03/2023).

[18] Giulio Caravagna et al. 'Subclonal reconstruction of tumors by using machine learning and population genetics'. en. In: *Nature Genetics* 52.9 (Sept. 2020), pp. 898–907. ISSN: 1546-1718. DOI: 10.1038/s41588-020-0675-5. URL: https://www.nature.com/articles/s41588-020-0675-5 (visited on 15/12/2022).

[19] Anna K. Casasent et al. 'Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing'. en. In: *Cell* 172.1 (Jan. 2018), 205–217.e12. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.12.007. URL: https://www.sciencedirect.com/science/article/pii/S0092867417314496 (visited on 25/03/2023).

[20] Liang Chang et al. 'Targeting pan-essential genes in cancer: Challenges and opportunities'. en. In: *Cancer Cell* 39.4 (Apr. 2021), pp. 466–479. ISSN: 1535-6108. DOI: 10.1016/j.ccell.2020.12.008. URL: https://www.sciencedirect.com/science/article/pii/S1535610820306565 (visited on 20/12/2022).

[21] Chongyi Chen et al. 'Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI)'. eng. In: *Science (New York, N.Y.)* 356.6334 (Apr. 2017), pp. 189–194. ISSN: 1095-9203. DOI: 10.1126/science.aak9787.

[22] Xiaoyu Chen et al. 'Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications'. eng. In: *Bioinformatics (Oxford, England)* 32.8 (Apr. 2016), pp. 1220–1222. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv710.

[23] Allison S. Cleary et al. 'Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers'. en. In: *Nature* 508.7494 (Apr. 2014), pp. 113–117. ISSN: 1476-4687. DOI: 10.1038/nature13187. URL: https://www.nature.com/articles/nature13187 (visited on 18/03/2023).

[24] Francis S. Collins and Leslie Fink. 'The Human Genome Project'. In: *Alcohol Health and Research World* 19.3 (1995), pp. 190–195. ISSN: 0090-838X. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6875757/ (visited on 24/03/2023).

[25] Zachary Compton et al. *A Missing Hallmark of Cancer: Dysregulation of Differentiation*. arXiv:2210.13343 [q-bio]. Oct. 2022. DOI: 10.48550/arXiv.2210.13343. URL: http://arxiv.org/abs/2210.13343 (visited on 15/12/2022).

[26] Francis Crick. 'Central Dogma of Molecular Biology'. en. In: *Nature* 227.5258 (Aug. 1970), pp. 561–563. ISSN: 1476-4687. DOI: 10.1038/227561a0. URL: https://www.nature.com/articles/227561a0 (visited on 15/12/2022).

[27] Bogdan Czerniak, Colin Dinney and David McConkey. 'Origins of Bladder Cancer'. In: *Annual Review of Pathology: Mechanisms of Disease* 11.1 (2016), pp. 149–174. DOI: 10.1146/annurev-pathol-012513-104703. URL: https://doi.org/10.1146/annurev-pathol-012513-104703 (visited on 31/03/2023).

[28] Stefan C. Dentro, David C. Wedge and Peter Van Loo. 'Principles of Reconstructing the Subclonal Architecture of Cancers'. eng. In: *Cold Spring Harbor Perspectives in Medicine* 7.8 (Aug. 2017), a026625. ISSN: 2157-1422. DOI: 10.1101/cshperspect.a026625.

[29] Mark A. DePristo et al. 'A framework for variation discovery and genotyping using next-generation DNA sequencing data'. en. In: *Nature Genetics* 43.5 (May 2011), pp. 491–498. ISSN: 1546-1718. DOI: 10.1038/ng.806. URL: https://www.nature.com/articles/ng.806 (visited on 05/01/2023).

[30] Amar Desai, Yan Yan and Stanton L. Gerson. 'Concise Reviews: Cancer Stem Cell Targeted Therapies: Toward Clinical Success'. en. In: *Stem Cells Translational Medicine* 8.1 (Jan. 2019), p. 75. DOI: 10.1002/sctm.18-0123. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6312440/ (visited on 18/03/2023).

[31] Amit G. Deshwar et al. 'PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors'. In: *Genome Biology* 16.1 (Feb. 2015), p. 35. ISSN: 1465-6906. DOI: 10.1186/s13059-015-0602-8. URL: https://doi.org/10.1186/s13059-015-0602-8 (visited on 12/04/2023).

[32] Renumathy Dhanasekaran et al. 'The MYC oncogene — the grand orchestrator of cancer growth and immune evasion'. en. In: *Nature Reviews Clinical Oncology* 19.1 (Jan. 2022), pp. 23–36. ISSN: 1759-4782. DOI: 10.1038/s41571-021-00549-2. URL: https://www.nature.com/articles/s41571-021-00549-2 (visited on 17/12/2022).

[33] Paolo Di Tommaso et al. 'Nextflow enables reproducible computational workflows'. en. In: *Nature Biotechnology* 35.4 (Apr. 2017), pp. 316–319. ISSN: 1546-1696. DOI: 10.1038/nbt.3820. URL: https://www.nature.com/articles/nbt.3820 (visited on 05/01/2023).

[34] K. N. Dinh et al. *Statistical inference for the evolutionary history of cancer genomes*. en. Aug. 2019. DOI: 10.1101/722033. URL: https://www.biorxiv.org/content/10.1101/722033v1 (visited on 20/01/2023).

[35] Matthias Drosten et al. 'Genetic analysis of Ras signalling pathways in cell proliferation, migration and survival'. In: *The EMBO Journal* 29.6 (Mar. 2010), pp. 1091–1104. ISSN: 0261-4189. DOI: 10.1038/emboj.2010.7. URL: https://www.embopress.org/doi/full/10.1038/emboj.2010.7 (visited on 17/12/2022).

[36] Laurent Duret. 'Neutral Theory: The Null Hypothesis of Molecular Evolution | Learn Science at Scitable'. en. In: *Nature Education* (2008). URL: http://www.nature.com/scitable/topicpage/neutral-theory-the-null-hypothesis-of-molecular-839 (visited on 18/03/2023).

[37] Rick Durrett. 'POPULATION GENETICS OF NEUTRAL MUTATIONS IN EXPONENTIALLY GROWING CANCER CELL POPULATIONS'. In: *The annals of applied probability : an official journal of the Institute of Mathematical Statistics* 23.1 (2013), pp. 230–250. ISSN: 1050-5164. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3588108/ (visited on 15/12/2022).

[38] George Forsythe. 'Computer Methods For Mathematical Computations'. In: *Books by Alumni* (Jan. 1977). URL: https://works.swarthmore.edu/alum-books/1955.

[39] Ruli Gao et al. 'Punctuated copy number evolution and clonal stasis in triple-negative breast cancer'. en. In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1119–1130. ISSN: 1546-1718. DOI: 10.1038/ng.3641. URL: https://www.nature.com/articles/ng.3641 (visited on 12/01/2023).

[40] Moritz Gerstung et al. 'The evolutionary history of 2,658 cancers'. In: *Nature* 578.7793 (2020), pp. 122–128. ISSN: 0028-0836. DOI: 10.1038/s41586-019-1907-7. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7054212/ (visited on 12/01/2023).

[41] Don L. Gibbons and Chad J. Creighton. 'Pan-cancer survey of epithelial–mesenchymal transition markers across The Cancer Genome Atlas'. In: *Developmental dynamics : an official publication of the American Association of Anatomists* 247.3 (Mar. 2018), pp. 555–564. ISSN: 1058-8388. DOI: 10.1002/dvdy.24485. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5503821/ (visited on 16/02/2023).

[42] Jarosław Gośliński. *Nowotwory złośliwe w Polsce. Krajowy Rejestr Nowotworów*. pl-PL. Jan. 2022. URL: https://www.zwrotnikraka.pl/nowotwory-zlosliwe-w-polsce-krajowy-rejestr-nowotworow/ (visited on 22/12/2022).

[43] R.C. Griffiths and Simon Tavaré. 'The age of a mutation in a general coalescent tree'. In: *Communications in Statistics. Stochastic Models* 14.1-2 (Jan. 1998), pp. 273–295. ISSN: 0882-0287. DOI: 10.1080/15326349808807471. URL: https://doi.org/10.1080/15326349808807471 (visited on 27/03/2023).

[44] Zuguang Gu, Roland Eils and Matthias Schlesner. 'Complex heatmaps reveal patterns and correlations in multidimensional genomic data'. eng. In: *Bioinformatics (Oxford, England)* 32.18 (Sept. 2016), pp. 2847–2849. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw313.

[45] Charles C. Guo et al. 'Assessment of Luminal and Basal Phenotypes in Bladder Cancer'. en. In: *Scientific Reports* 10.1 (June 2020), p. 9743. ISSN: 2045-2322. DOI: 10.1038/s41598-020-66747-7. URL: https://www.nature.com/articles/s41598-020-66747-7 (visited on 31/03/2023).

[46] Gavin Ha et al. 'TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data'. In: *Genome Research* 24.11 (Nov. 2014), pp. 1881–1893. ISSN: 1088-9051. DOI: 10.1101/gr.180281.114. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4216928/ (visited on 12/04/2023).

[47] D. Hanahan and R. A. Weinberg. 'The hallmarks of cancer'. eng. In: *Cell* 100.1 (Jan. 2000), pp. 57–70. ISSN: 0092-8674. DOI: 10.1016/s0092-8674(00)81683-9.

[48] Douglas Hanahan and Robert A. Weinberg. 'Hallmarks of Cancer: The Next Generation'. English. In: *Cell* 144.5 (Mar. 2011), pp. 646–674. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2011.02.013. URL: https://www.cell.com/cell/abstract/S0092-8674(11)00127-9 (visited on 15/12/2022).

[49] D. Hart, E. Shochat and Z. Agur. 'The growth law of primary breast cancer as inferred from mammography screening trials data'. eng. In: *British Journal of Cancer* 78.3 (Aug. 1998), pp. 382–387. ISSN: 0007-0920. DOI: 10.1038/bjc.1998.503.

[50] Yujiro Hayashi et al. 'Targeted-sequence of normal urothelium and tumor of patients with non-muscle invasive bladder cancer'. en. In: *Scientific Reports* 12.1 (Oct. 2022), p. 16642. ISSN: 2045-2322. DOI: 10.1038/s41598-022-21158-8. URL: https://www.nature.com/articles/s41598-022-21158-8 (visited on 10/04/2023).

[51] James M. Heather and Benjamin Chain. 'The sequence of sequencers: The history of sequencing DNA'. en. In: *Genomics* 107.1 (Jan. 2016), pp. 1–8. ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2015.11.003. URL: https://www.sciencedirect.com/science/article/pii/S0888754315300410 (visited on 24/03/2023).

[52] Timon Heide et al. 'Reply to 'Neutral tumor evolution?'' en. In: *Nature Genetics* 50.12 (Dec. 2018), pp. 1633–1637. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0256-z. URL: https://www.nature.com/articles/s41588-018-0256-z (visited on 15/12/2022).

[53] Robert W. Holley, James T. Madison and Ada Zamir. 'A new method for sequence determination of large oligonucleotides'. en. In: *Biochemical and Biophysical Research Communications* 17.4 (Nov. 1964), pp. 389–394. ISSN: 0006-291X. DOI: 10.1016/0006-291X(64)90017-8. URL: https://www.sciencedirect.com/science/article/pii/0006291X64900178 (visited on 24/03/2023).

[54] *HPC Ziemowit | Strona użytkowników klastra obliczeniowego Ziemowit*. pl-PL. URL: https://www.ziemowit.hpc.polsl.pl/pl/ (visited on 07/01/2023).

[55] *Human Genome Project Fact Sheet*. en. Sept. 2022. URL: https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project (visited on 24/03/2023).

[56] Siddhartha Jaiswal and Peter Libby. 'Clonal haematopoiesis: connecting ageing and inflammation in cardiovascular disease'. In: *Nature reviews. Cardiology* 17.3 (Mar. 2020), pp. 137–144. ISSN: 1759-5002. DOI: 10.1038/s41569-019-0247-5. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9448847/ (visited on 06/04/2023).

[57] Patryk Janus et al. 'HSF1 Can Prevent Inflammation following Heat Shock by Inhibiting the Excessive Activation of the ATF3 and JUN&FOS Genes'. en. In: *Cells* 11.16 (Jan. 2022), p. 2510. ISSN: 2073-4409. DOI: 10.3390/cells11162510. URL: https://www.mdpi.com/2073-4409/11/16/2510 (visited on 10/04/2023).

[58] Cyriac Kandoth et al. 'Mutational landscape and significance across 12 major cancer types'. In: *Nature* 502.7471 (2013), pp. 333–339. ISSN: 0028-0836. DOI: 10.1038/nature12634. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3927368/ (visited on 03/04/2023).

[59] Zhi-Jie Kang et al. 'The Philadelphia chromosome in leukemogenesis'. In: *Chinese Journal of Cancer* 35 (May 2016), p. 48. ISSN: 1000-467X. DOI: 10.1186/s40880-016-0108-0. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4896164/ (visited on 13/01/2023).

[60] Alboukadel Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*. URL: https://rpkgs.datanovia.com/ggpubr/ (visited on 23/02/2023).

[61] Anna Kazanets et al. 'Epigenetic silencing of tumor suppressor genes: Paradigms, puzzles, and potential'. en. In: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1865.2 (Apr. 2016), pp. 275–288. ISSN: 0304-419X. DOI: 10.1016/j.bbcan.2016.04.001. URL: https://www.sciencedirect.com/science/article/pii/S0304419X16300294 (visited on 17/12/2022).

[62] Sangtae Kim et al. 'Strelka2: fast and accurate calling of germline and somatic variants'. eng. In: *Nature Methods* 15.8 (Aug. 2018), pp. 591–594. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0051-x.

[63] Motoo Kimura. 'Evolutionary Rate at the Molecular Level'. In: *Nature* (1968).

[64] Alfred G. Knudson. 'Mutation and Cancer: Statistical Study of Retinoblastoma'. In: *Proceedings of the National Academy of Sciences of the United States of America* 68.4 (Apr. 1971), pp. 820–823. ISSN: 0027-8424. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC389051/ (visited on 25/02/2023).

[65] Daniel C. Koboldt et al. 'VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing'. eng. In: *Genome Research* 22.3 (Mar. 2012), pp. 568–576. ISSN: 1549-5469. DOI: 10.1101/gr.129684.111.

[66] Monika K. Kurpas and Marek Kimmel. 'Modes of Selection in Tumors as Reflected by Two Mathematical Models and Site Frequency Spectra'. In: *Frontiers in Ecology and Evolution* 10 (2022). ISSN: 2296-701X. URL: https://www.frontiersin.org/articles/10.3389/fevo.2022.889438 (visited on 14/02/2023).

[67] Gregory M. Kurtzer et al. *hpcng/singularity: Singularity 3.7.3*. Apr. 2021. URL: https://zenodo.org/record/4667718 (visited on 05/01/2023).

[68] Pawel Kus, Roman Jaksik and Marek Kimmel. 'Tumor subclonal reconstruction pipelines - comparison of results ; Analiza struktury klonalnej nowotworów - porównanie wyników różnych kombinacji metod'. In: *Recent advances in computational oncology and personalized medicine. Vol. 1, Here and now! ; Postępy w onkologii obliczeniowej i spersonalizowanej medycynie. T. 1, Tu i teraz!* (2021), pp. 137–148. DOI: 10.34918/83571.

[69] Paweł Kuś and Marek Kimmel. *Sampling-oriented modelling of cancer evolution*. Nov. 2022.

[70] Alican Kuşoğlu and Çığır Biray Avcı. 'Cancer stem cells: A brief review of the current status'. en. In: *Gene* 681 (Jan. 2019), pp. 80–85. ISSN: 0378-1119. DOI: 10.1016/j.gene.2018.09.052. URL: https://www.sciencedirect.com/science/article/pii/S0378111918310163 (visited on 18/03/2023).

[71] Jeffrey C. Lagarias et al. 'Convergence Properties of the Nelder–Mead Simplex Method in Low Dimensions'. In: *SIAM Journal on Optimization* 9.1 (Jan. 1998), pp. 112–147. ISSN: 1052-6234. DOI: 10.1137/S1052623496303470. URL: https://epubs.siam.org/doi/10.1137/S1052623496303470 (visited on 06/04/2023).

[72] Kenneth Lange, Michael Boehnke and Richard Carson. 'Moment computations for subcritical branching processes'. en. In: *Journal of Applied Probability* 18.1 (Mar. 1981), pp. 52–64. ISSN: 0021-9002, 1475-6072. DOI: 10.2307/3213166. URL: https://www.cambridge.org/core/journals/journal-of-applied-probability/article/abs/moment-computations-for-subcritical-branching-processes/6669B41C9C6FD145BB77C754023A5893 (visited on 06/04/2023).

[73] Ben Langmead et al. 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome'. In: *Genome Biology* 10.3 (Mar. 2009), R25. ISSN: 1474-760X. DOI: 10.1186/gb-2009-10-3-r25. URL: https://doi.org/10.1186/gb-2009-10-3-r25 (visited on 25/03/2023).

[74] Roger S. Lasken. 'Genomic DNA amplification by the multiple displacement amplification (MDA) method'. eng. In: *Biochemical Society Transactions* 37.Pt 2 (Apr. 2009), pp. 450–453. ISSN: 1470-8752. DOI: 10.1042/BST0370450.

[75] Günter Last and Mathew Penrose. *Lectures on the Poisson Process*. en. Google-Books-ID: JRs3DwAAQBAJ. Cambridge University Press, Oct. 2017. ISBN: 9781107088016.

[76] Duy Le et al. 'BATF2 Promotes HSC Myeloid Differentiation Via Amplification of the Pro-Inflammatory Response during Chronic Infection'. In: *Blood* 140.Supplement 1 (Nov. 2022), pp. 5732–5733. ISSN: 0006-4971. DOI: 10.1182/blood-2022-164467. URL: https://doi.org/10.1182/blood-2022-164467 (visited on 10/04/2023).

[77] Haiyang Li et al. *Mutation divergence over space in tumour expansion*. en. Dec. 2022. DOI: 10.1101/2022.12.21.521509. URL: https://www.biorxiv.org/content/10.1101/2022.12.21.521509v1 (visited on 20/01/2023).

[78] Heng Li. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. arXiv:1303.3997 [q-bio]. May 2013. DOI: 10.48550/arXiv.1303.3997. URL: http://arxiv.org/abs/1303.3997 (visited on 05/01/2023).

[79] Shaoping Ling et al. 'Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution'. In: *Proceedings of the National Academy of Sciences* 112.47 (Nov. 2015), E6496–E6505. DOI: 10.1073/pnas.1519556112. URL: https://www.pnas.org/doi/full/10.1073/pnas.1519556112 (visited on 15/12/2022).

[80] Sten Linnarsson and Sarah A. Teichmann. 'Single-cell genomics: coming of age'. In: *Genome Biology* 17.1 (May 2016), p. 97. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0960-x. URL: https://doi.org/10.1186/s13059-016-0960-x (visited on 25/03/2023).

[81] *Mamba documentation*. URL: https://mamba.readthedocs.io/en/latest/ (visited on 07/01/2023).

[82] Iñigo Martincorena et al. 'Universal Patterns of Selection in Cancer and Somatic Tissues'. In: *Cell* 171.5 (Nov. 2017), 1029–1041.e21. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.09.042. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5720395/ (visited on 15/12/2022).

[83] Thomas O. McDonald, Shaon Chakrabarti and Franziska Michor. 'Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution'. en. In: *Nature Genetics* 50.12 (Dec. 2018), pp. 1620–1623. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0217-6. URL: https://www.nature.com/articles/s41588-018-0217-6 (visited on 24/02/2023).

[84] William McLaren et al. 'The Ensembl Variant Effect Predictor'. eng. In: *Genome Biology* 17.1 (June 2016), p. 122. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0974-4.

[85] Gaurav Mendiratta et al. 'Cancer gene mutation frequencies for the U.S. population'. en. In: *Nature Communications* 12.1 (Oct. 2021), p. 5961. ISSN: 2041-1723. DOI: 10.1038/s41467-021-26213-y. URL: https://www.nature.com/articles/s41467-021-26213-y (visited on 16/12/2022).

[86] Dirk Merkel. 'Docker: lightweight Linux containers for consistent development and deployment'. In: *Linux Journal* 2014.239 (Mar. 2014), 2:2. ISSN: 1075-3583.

[87] Brandon Milholland et al. 'Differences between germline and somatic mutation rates in humans and mice'. en. In: *Nature Communications* 8.1 (May 2017), p. 15183. ISSN: 2041-1723. DOI: 10.1038/ncomms15183. URL: https://www.nature.com/articles/ncomms15183 (visited on 27/01/2023).

[88] Christopher A. Miller et al. 'SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution'. en. In: *PLOS Computational Biology* 10.8 (2014), e1003665. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003665. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003665 (visited on 12/04/2023).

[89] Mark M. Moasser. 'The oncogene HER2; Its signaling and transforming functions and its role in human cancer pathogenesis'. In: *Oncogene* 26.45 (Oct. 2007), pp. 6469–6487. ISSN: 0950-9232. DOI: 10.1038/sj.onc.1210477. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3021475/ (visited on 17/12/2022).

[90] Felix Mölder et al. *Sustainable data analysis with Snakemake*. en. Tech. rep. 10:33. Type: article. F1000Research, Jan. 2021. URL: https://f1000research.com/articles/10-33 (visited on 05/01/2023).

[91] Manabu Muto et al. 'Field Effect of Alcohol, Cigarette Smoking, and Their Cessation on the Development of Multiple Dysplastic Lesions and Squamous Cell Carcinoma: A Long-term Multicenter Cohort Study'. en. In: *Gastro Hep Advances* 1.2 (Jan. 2022), pp. 265–276. ISSN: 2772-5723. DOI: 10.1016/j.gastha.2021.10.005. URL: https://www.sciencedirect.com/science/article/pii/S2772572321000212 (visited on 18/03/2023).

[92] Serena Nik-Zainal et al. 'The Life History of 21 Breast Cancers'. en. In: *Cell* 149.5 (May 2012), p. 994. DOI: 10.1016/j.cell.2012.04.023. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3428864/ (visited on 12/04/2023).

[93] Riccardo Nocini et al. 'Updates on larynx cancer epidemiology'. In: *Chinese Journal of Cancer Research* 32.1 (Feb. 2020), pp. 18–25. ISSN: 1000-9604. DOI: 10.21147/j.issn.1000-9604.2020.01.03. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7072014/ (visited on 14/02/2023).

[94] L. Norton. 'A Gompertzian model of human breast cancer growth'. eng. In: *Cancer Research* 48.24 Pt 1 (Dec. 1988), pp. 7067–7071. ISSN: 0008-5472.

[95] Sergey Nurk et al. 'The complete sequence of a human genome'. In: *Science* 376.6588 (Apr. 2022), pp. 44–53. DOI: 10.1126/science.abj6987. URL: https://www.science.org/doi/10.1126/science.abj6987 (visited on 19/12/2022).

[96] Erasmo Orrantia-Borunda et al. 'Subtypes of Breast Cancer'. eng. In: *Breast Cancer*. Ed. by Harvey N. Mayrovitz. Brisbane (AU): Exon Publications, 2022. ISBN: 9780645332032. URL: http://www.ncbi.nlm.nih.gov/books/NBK583808/ (visited on 14/02/2023).

[97] Thomas Lin Pedersen. *patchwork: The Composer of Plots*. 2022. URL: https://github.com/thomasp85/patchwork.

[98] Anthony Rhoads and Kin Fai Au. 'PacBio Sequencing and Its Applications'. eng. In: *Genomics, Proteomics & Bioinformatics* 13.5 (Oct. 2015), pp. 278–289. ISSN: 2210-3244. DOI: 10.1016/j.gpb.2015.08.002.

[99] Andrew Roth et al. 'PyClone: statistical inference of clonal population structure in cancer'. en. In: *Nature Methods* 11.4 (Apr. 2014), pp. 396–398. ISSN: 1548-7105. DOI: 10.1038/nmeth.2883. URL: https://www.nature.com/articles/nmeth.2883 (visited on 12/04/2023).

[100] Adriana Salcedo et al. 'A community effort to create standards for evaluating tumor subclonal reconstruction'. en. In: *Nature Biotechnology* 38.1 (Jan. 2020), pp. 97–107. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0364-z. URL: https://www.nature.com/articles/s41587-019-0364-z (visited on 15/12/2022).

[101] F. Sanger, S. Nicklen and A. R. Coulson. 'DNA sequencing with chain-terminating inhibitors'. In: *Proceedings of the National Academy of Sciences* 74.12 (Dec. 1977), pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463. URL: https://www.pnas.org/doi/abs/10.1073/pnas.74.12.5463 (visited on 24/03/2023).

[102] *Sanger sequencing*. en. Page Version ID: 1126432348. Dec. 2022. URL: https://en.wikipedia.org/w/index.php?title=Sanger_sequencing&oldid=1126432348 (visited on 24/03/2023).

[103] Luca Scrucca et al. 'mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models'. en. In: *The R Journal* 8.1 (2016), pp. 289–317. ISSN: 2073-4859. URL: https://journal.r-project.org/archive/2016/RJ-2016-021/index.html (visited on 09/04/2023).

[104] Ilona Seferyńska and Krzysztof Warzocha. 'A registry report from the Institute of Hematology and Transfusion Medicine on adult morbidity for acute leukemias between 2004–2010 in Poland made on behalf of the Polish Adult Leukemia Group (PALG)'. In: *Hematologia* (2014). URL: https://journals.viamedica.pl/hematology_in_clinical_practice/article/download/39426/27402 (visited on 16/02/2023).

[105] R. Seshadri et al. 'Mutation rate of normal and malignant human lymphocytes'. eng. In: *Cancer Research* 47.2 (Jan. 1987), pp. 407–409. ISSN: 0008-5472.

[106] Ronglai Shen and Venkatraman E. Seshan. 'FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing'. In: *Nucleic Acids Research* 44.16 (Sept. 2016), e131. ISSN: 0305-1048. DOI: 10.1093/nar/gkw520. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5027494/ (visited on 22/03/2023).

[107] Liran I. Shlush et al. 'Tracing the origins of relapse in acute myeloid leukaemia to stem cells'. en. In: *Nature* 547.7661 (July 2017), pp. 104–108. ISSN: 1476-4687. DOI: 10.1038/nature22993. URL: https://www.nature.com/articles/nature22993 (visited on 15/12/2022).

[108] *Sidney Farber, MD - Dana-Farber Cancer Institute | Boston, MA*. URL: https://www.dana-farber.org/about-us/history-and-milestones/sidney-farber,-md/ (visited on 15/12/2022).

[109] D. P. Slaughter, H. W. Southwick and W. Smejkal. 'Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin'. eng. In: *Cancer* 6.5 (Sept. 1953), pp. 963–968. ISSN: 0008-543X. DOI: 10.1002/1097-0142(195309)6:5<963::aid-cncr2820060515>3.0.co;2-q.

[110] Andrea Sottoriva et al. 'A Big Bang model of human colorectal tumor growth'. en. In: *Nature Genetics* 47.3 (Mar. 2015), pp. 209–216. ISSN: 1546-1718. DOI: 10.1038/ng.3214. URL: https://www.nature.com/articles/ng.3214 (visited on 18/01/2023).

[111] J. A. Spratt et al. 'Decelerating growth and human breast cancer'. eng. In: *Cancer* 71.6 (Mar. 1993), pp. 2013–2019. ISSN: 0008-543X. DOI: 10.1002/1097-0142(19930315)71:6<2013::aid-cncr2820710615>3.0.co;2-v.

[112] Kathleen Sprouffske et al. 'High mutation rates limit evolutionary adaptation in Escherichia coli'. In: *PLoS Genetics* 14.4 (Apr. 2018), e1007324. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1007324. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5942850/ (visited on 14/04/2023).

[113] Xianbin Su et al. 'Accurate tumor clonal structures require single-cell analysis'. eng. In: *Annals of the New York Academy of Sciences* 1517.1 (Nov. 2022), pp. 213–224. ISSN: 1749-6632. DOI: 10.1111/nyas.14897.

[114] Valentine Svensson, Roser Vento-Tormo and Sarah A. Teichmann. 'Exponential scaling of single-cell RNA-seq in the past decade'. en. In: *Nature Protocols* 13.4 (Apr. 2018), pp. 599–604. ISSN: 1750-2799. DOI: 10.1038/nprot.2017.149. URL: https://www.nature.com/articles/nprot.2017.149 (visited on 25/03/2023).

[115] Anne Talkington and Rick Durrett. 'Estimating tumor growth rates in vivo'. In: *Bulletin of mathematical biology* 77.10 (Oct. 2015), pp. 1934–1954. ISSN: 0092-8240. DOI: 10.1007/s11538-015-0110-8. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4764475/ (visited on 09/01/2023).

[116] Maxime Tarabichi et al. 'A practical guide to cancer subclonal reconstruction from DNA sequencing'. en. In: *Nature Methods* 18.2 (Feb. 2021), pp. 144–155. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01013-2. URL: https://www.nature.com/articles/s41592-020-01013-2 (visited on 13/04/2023).

[117] Maxime Tarabichi et al. 'Neutral tumor evolution?' en. In: *Nature Genetics* 50.12 (Dec. 2018), pp. 1630–1633. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0258-x. URL: https://www.nature.com/articles/s41588-018-0258-x (visited on 15/12/2022).

[118] H. Telenius et al. 'Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer'. eng. In: *Genomics* 13.3 (July 1992), pp. 718–725. ISSN: 0888-7543. DOI: 10.1016/0888-7543(92)90147-k.

[119] *The Cost of Sequencing a Human Genome*. en. Sept. 2022. URL: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost (visited on 24/03/2023).

[120] Hwai-Ray Tung and Rick Durrett. 'Signatures of neutral evolution in exponentially growing tumors: A theoretical perspective'. en. In: *PLOS Computational Biology* 17.2 (2021), e1008701. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008701. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008701 (visited on 18/01/2023).

[121] Peter Van Loo et al. 'Allele-specific copy number analysis of tumors'. In: *Proceedings of the National Academy of Sciences* 107.39 (Sept. 2010), pp. 16910–16915. DOI: 10.1073/pnas.1009843107. URL: https://www.pnas.org/doi/10.1073/pnas.1009843107 (visited on 26/03/2023).

[122] Natalia Vydra et al. 'Heat shock factor 1 (HSF1) cooperates with estrogen receptor (ER) in the regulation of estrogen action in breast cancer cells'. In: *eLife* 10 (Nov. 2021). Ed. by Maureen E Murphy et al., e69843. ISSN: 2050-084X. DOI: 10.7554/eLife.69843. URL: https://doi.org/10.7554/eLife.69843 (visited on 10/04/2023).

[123] Liguo Wang, Shengqin Wang and Wei Li. 'RSeQC: quality control of RNA-seq experiments'. In: *Bioinformatics* 28.16 (Aug. 2012), pp. 2184–2185. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts356. URL: https://doi.org/10.1093/bioinformatics/bts356 (visited on 25/03/2023).

[124] Tim Wang et al. 'Identification and characterization of essential genes in the human genome'. In: *Science* 350.6264 (Nov. 2015), pp. 1096–1101. DOI: 10.1126/science.aac7041. URL: https://www.science.org/doi/10.1126/science.aac7041 (visited on 20/12/2022).

[125] Xiaoyu Wang et al. 'Recent advances and application of whole genome amplification in molecular diagnosis and medicine'. eng. In: *MedComm* 3.1 (Mar. 2022), e116. ISSN: 2688-2663. DOI: 10.1002/mco2.116.

[126] Yong Wang et al. 'Clonal evolution in breast cancer revealed by single nucleus genome sequencing'. en. In: *Nature* 512.7513 (Aug. 2014), pp. 155–160. ISSN: 1476-4687. DOI: 10.1038/nature13600. URL: https://www.nature.com/articles/nature13600 (visited on 17/03/2023).

[127] KA Wetterstrand. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. en. Sept. 2022. URL: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (visited on 24/03/2023).

[128] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. 2016. Use R! Cham: Springer International Publishing : Imprint: Springer, 2016. ISBN: 9783319242774.

[129] Tyler J. Willenbrink et al. 'Field cancerization: Definition, epidemiology, risk factors, and outcomes'. English. In: *Journal of the American Academy of Dermatology* 83.3 (Sept. 2020), pp. 709–717. ISSN: 0190-9622, 1097-6787. DOI: 10.1016/j.jaad.2020.03.126. URL: https://www.jaad.org/article/S0190-9622(20)30791-X/fulltext (visited on 18/03/2023).

[130] Marc J. Williams et al. 'Identification of neutral tumor evolution across cancer types'. en. In: *Nature Genetics* 48.3 (Mar. 2016), pp. 238–244. ISSN: 1546-1718. DOI: 10.1038/ng.3489. URL: https://www.nature.com/articles/ng.3489 (visited on 15/12/2022).

[131] Marc J. Williams et al. 'Quantification of subclonal selection in cancer from bulk sequencing data'. en. In: *Nature Genetics* 50.6 (June 2018), pp. 895–903. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0128-6. URL: https://www.nature.com/articles/s41588-018-0128-6 (visited on 15/12/2022).

[132] Steven W. Wingett and Simon Andrews. 'FastQ Screen: A tool for multi-genome mapping and quality control'. In: *F1000Research* 7 (Sept. 2018), p. 1338. ISSN: 2046-1402. DOI: 10.12688/f1000research.15931.2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6124377/ (visited on 25/03/2023).

[133] *Worldwide cancer data | World Cancer Research Fund International*. en-US. URL: https://www.wcrf.org/cancer-trends/worldwide-cancer-data/ (visited on 14/02/2023).

[134] Andy B. Yoo, Morris A. Jette and Mark Grondona. 'SLURM: Simple Linux Utility for Resource Management'. en. In: *Job Scheduling Strategies for Parallel Processing*. Ed. by Dror Feitelson, Larry Rudolph and Uwe Schwiegelshohn. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2003, pp. 44–60. ISBN: 9783540397274. DOI: 10.1007/10968987_3.

[135] Chunxiao Zhu et al. 'Targeting KRAS mutant cancers: from druggable therapy to drug resistance'. In: *Molecular Cancer* 21.1 (Aug. 2022), p. 159. ISSN: 1476-4598. DOI: 10.1186/s12943-022-01629-2. URL: https://doi.org/10.1186/s12943-022-01629-2 (visited on 17/12/2022).

[136] Chenghang Zong et al. 'Genome-wide detection of single-nucleotide and copy-number variations of a single human cell'. eng. In: *Science (New York, N.Y.)* 338.6114 (Dec. 2012), pp. 1622–1626. ISSN: 1095-9203. DOI: 10.1126/science.1229164.