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

Dissertation detailed abstract

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

mgr inż. Paweł Kuś

1 Introduction

Cancer is one of the leading causes of death worldwide. 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 new cases, and about 400 000 cancer cases were diagnosed in children [3]. Many cancer risk factors are linked to changes in people's lifestyles in developed countries. Cancer, therefore, receives considerable attention from the scientific community. Research over the past decades has uncovered a number of mechanisms that characterize cancer cells [8]. They have also led to the implementation of numerous advanced anti-cancer therapies, but despite these advances, treating cancer to prevent drug resistance and recurrence is still a challenge for medicine. The capabilities of medicine are limited in the treatment of cancer, similarly to our understanding of the role of mechanisms underlying its evolution: mutagenesis and selection.

Mutagenesis is the process by which the cells' genetic information is changed by mutation. It increases the genetic diversity of cells and can result in mutations in genes that are particularly important for the maintenance of homeostasis, related, for example, to the cell cycle or the repair of DNA damage. Mutations in these genes, referred to as cancer driver genes, can lead to cell selection: positive if the mutation increases the cell's fitness and the frequency of its progeny cells increases, or negative if it leads to the extinction of a population of cells affected by disadvantageous mutations.

The role of selection and mutagenesis mechanisms in cancer evolution is still under debate. Models have been proposed, such as the model of neutral evolution, in which newly occurring mutations do not provide the selective advantage, the fitness of all cells in a tumor is similar, and the tumor grows in a single expansion of early-emerged subclones, or the model of clonal evolution, in which new driver mutations initiate expansions of new clones that eventually completely replace earlier populations of cells with lower fitness.

The development of next-generation sequencing (NGS) methods has made it possible to study genomes, including cancer genomes, on an unprecedented scale. NGS has significantly reduced the cost of DNA sequencing, making it a common practice in cancer research. It also led to the creation of databases such as *The Cancer Genome Atlas*, which provide access to the molecular data of thousands of sequenced tumor samples. Bulk sequencing, in which genetic material is isolated from a sample containing millions of cells and then sequenced together, is particularly common.

DNA sequencing, after appropriate processing including quality control of the material, alignment of reads to a reference genome, detection of variants, and filtering of variants to reject false positives, provides information such as the variants' location and their variant allele frequencies (VAF), which are related to the frequency of cells with a particular variant in the sample.

The advent of widely available cancer DNA sequencing data has enabled the development of methods to analyze their evolution based on mutation frequencies. Popular algorithms such as SciClone [9] or PyClone [10] combine mutation data from multiple samples to determine the clonal structure of a tumor. The PhyloWGS algorithm [5] takes this a step further by examining the relationships between clones and constructing their phylogenetic tree. These algorithms, however, do not infer the evolutionary parameters of the tumor, such as the mutation rate or selection coefficients of subclones. The recently developed MOBSTER algorithm [4] identifies them, but it requires whole-genome sequencing data with at least 100x coverage to work effectively. For this reason, it does not work properly with data from the more commonly performed whole-exome sequencing. It also relies on certain assumptions, such as exponential population growth or a constant mutation rate, which may not be fulfilled in all tumors.

2 Thesis

We state the following thesis in this work: *Changes in the evolutionary dynamics of cancer upon metastasis and recurrence can be quantified from the bulk DNA sequencing data.* The main goals of the work involved:

1. analysis of evolutionary dynamics of cancer during progression, recurrence, and metastasis,

2. implementation of software capable of analyzing the evolutionary dynamics of cancer using whole exome sequencing and insufficient-coverage sequencing data,

3. validation of assumptions of popular models, such as the assumption of exponential population growth or the assumption of constant mutation rate.

3 Data

To address the thesis, we collected 4 datasets from bulk DNA sequencing experiments. The collected data represented 4 cancer types and included at least two tumor samples from each patient.

We investigated the evolutionary dynamics of recurrent cancers using the example of acute myeloid leukaemia (AML), which accounts for about 40% of all leukaemias in Poland. We used whole genome sequencing data published in the study by Shlush et al. [11]. The dataset included 11 patients, from whom two tumor samples were collected from each, representing the time of diagnosis and relapse, along with one control sample.

The evolutionary dynamics of metastatic cancers was investigated using breast cancer (BRCA) and laryngeal cancer (BRCA) data. These 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, and the data was not published before. Breast cancer is the most commonly diagnosed cancer worldwide and occurs almost exclusively in women. Laryngeal cancer is a type of head and neck cancer (HNSC) and is five times more common in men than in women. Both breast and laryngeal cancers are tumors of epithelial origin, and both often show the activity of the epithelial-mesenchymal transition (EMT), a process associated with metastasis. The BRCA and LSCC cohorts included 15 female patients with breast cancer and 12 patients of both sexes with laryngeal cancer. One primary tumor sample, one local lymph node metastasis sample, and one control sample were collected from each patient. All samples were subjected to whole exome sequencing with a targeted coverage of 100x.

Changes in evolutionary dynamics during cancer progression have been studied in two bladder cancers using whole organ mapping with whole-exome sequencing of multiple samples at different stages of disease progression, creating maps of the entire bladder. These data were obtained from the laboratory of Dr. Bogdan Czerniak at MD Anderson Cancer Center in Houston, TX, USA, and were previously used in a study by Bondaruk et al. *The origin of bladder cancer from mucosal field effect* [1]. The data came from two patients, no. 19 and 24, representing two different molecular types of bladder cancer: luminal and basal. 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). In the paper by Bondaruk et al. [1], selected regions of both maps were subjected to multi-omics experiments, including whole exome sequencing, RNA sequencing, whole-genome methylation array hybridization, and whole genome polymorphism-based copy number analysis. In this study, we re-use the whole exome sequencing data.

4 Methods

NGS data processing. Quality control of raw sequencing results in BRCA and LSCC cohorts was conducted using FastQC and FastQ Screen. Raw reads were aligned to the GRCh38 reference genome using BWA-mem software. We used the GATK toolkit to prepare the BAM files for the variant calling, and Mutect2 to detect Single Nucleotide Variants (SNVs) and short Indels in the data. The detected variants were filtered using GATK's FilterMutectCalls and annotated using Variant Effect Predictor.

AML data were downloaded as GRCh37-aligned BAM files from the European Genome-Phenome Archive (ID: EGAD00001003234) and processed using the same processing pipeline as BRCA and LSCC data.

BLCA analyses were based on VCF files containing the sets of SNVs and Indels used in the paper by Bondaruk et al. [1]. However, the process of raw data processing was similar to our pipeline described above: raw reads were aligned to the GRCh38 genome using BWA-MEM, BAM file preparation and variant calling were performed using GATK and Mutect2, and the detected variants were annotated using the Oncotator.

Modeling. Presented analyses were performed in R v4.2.2.
Initially, the data were modeled using the MOBSTER algorithm. This algorithm fits a mixture of power-law-shaped and binomial distributions to the VAF spectrum:

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

where f is the allelic frequency of mutations, $N(f)$ specifies the number of mutations with frequency f, K number of clones and sub-clones in the tumor, N_k number of mutations associated with (sub)clone k, f_k allelic frequency of mutations in (sub)clone k, and A is a constant proportional to the mutation frequency per effective cell division, described in [15]. The power-law component of this model is called the *neutral tail* and describes the mostly neutral mutations occurring in all cells. The allelic frequency of these variants depends mainly on the timing of the mutation and is equal to the inverse of the number of cells in the tumor at the time of the mutation. As shown by Durrett [6], Williams [15], and others, the α exponent of the power-law curve describing the neutral tail is equal to 2, under the assumption of constant mutation rate and exponential population growth. The coefficient A then allows us to determine the average mutation rate per cell division [15]. The binomial components correspond to the variants associated with clones and sub-clones. The parameters of these distributions can be used to determine the subclonal emergence time and selection coefficients [16].

In the majority of samples examined in this study, the MOBSTER algorithm was unable to accurately fit the model due to an insufficient number of mutations representing the neutral tail of the distribution. To address this issue, we developed a novel R package, *cevomod*, as part of this research. This package includes the model fitting methods designed for whole-exome sequencing data and data with insufficient coverage that were previously challenging for the MOBSTER algorithm to handle.

We proposed two alternative methods for fitting the power-law component to the data. The first one assumes exponential population growth and a constant mutation rate. In this case, $\alpha = 2$, and A depends on the number of bins in the VAF spectrum and the mutation rate per effective division mu. μ can be estimated from the relationship:

$$M(f) \sim \mu/f \tag{2}$$

introduced in [15], where $M(f)$ is the number of mutations with a frequency greater than f. The second implemented method can be used to validate the model assumptions and detect the presence of selectively advantageous micro-clones in the tail, as described by Tung and Durrett [14]. In this method, both parameters of the power-law component A and α are fitted to the data, and the values of α different from 2 indicate that the assumptions' violation or the presence of selectively advantageous micro-clones. The power curve is fitted by an optimization process in which the maximized objective function I consists of two components: the reward for the count of mutations under the curve (MCR) and the spectrum detachment penalty (SDP).

$$I = MCR - SDP \tag{3}$$

In both methods, the binomial components of the model are fitted by clustering the variants not explained by the power component. In the first step, n_i variants are drawn in each bin of the VAF spectrum, where n_i is the positive part of the difference between the number of variants in the i-th bin and the number of variants predicted for that bin by the power-law component. Then, we use the BMix package [4] to cluster these variants using a mixture of 1 to 3 binomial distributions, taking into account the sequencing depth of the variants.

5 Results

We have described the results of our work in two parts. The first part presents the analyses of the AML, BRCA, and LSCC datasets, in which two tumor samples were collected from each patient. The second part presents the results of the analysis of the BLCA dataset, containing a total number of 66 samples from two patients.

5.1 Evolution of metastatic breast and larynx cancers and recurring leukaemia

Tumors from the AML and BRCA/LSCC cohorts showed different shapes of VAF spectra. In most AML samples, two groups of variants were recognized. The VAF distribution of high-frequency variants had a symmetrical shape with a mean VAF value oscillating around 0.5, which in diploid cells (with two copies of each gene) corresponds to clonal variants present in all cells. The high abundance of mutations in this group indicates the past *selective sweeps*, in which more recent and selectively advantageous subclones with new mutations have completely replaced earlier clones. The VAF distributions of low-frequency variants were asymmetric, with the right-hand tails longer, consistent with the *neutral tail* model. In the BRCA and LSCC cohorts, VAF distributions consisted of single, asymmetric distributions at low frequencies with no clear clonal mutation peaks.

Using our *cevomod* package, we fitted mixtures of power-law-shaped and binomial distributions to the data. Using a model with exponent $\alpha = 2$, we determined the mutation rates for each sample and emergence times along with the selection coefficients for all subclones. In all cohorts, we identified common changes in mutation rate between samples from primary tumors (diagnostic samples in AML and primary tumor samples in BRCA and LSCC) and secondary tumors (the relapse samples in AML and lymph node metastases in BRCA and LSCC). The increases in mutation rate in the secondary tumor prevailed in the LSCC cohort, with an average increase of 14%. In the AML and BRCA cohorts, mutation rate changes in both directions were equally frequent, most often no more than 2-fold. In only one AML patient, we observed a nearly 10-fold decrease in mutation rate in the relapse sample.

In BRCA and LSCC, most of the detected subclones appeared early, during the first 20% doublings of tumor volume. Their selection coefficients were most often (with the exception of two samples) between 0% and 20%. The early emergence times of subclones, their low to moderate selection coefficients, and the absence of clear peaks of clonal mutations indicate a close-to-neutral evolution of the analyzed BRCA and LSCC tumors from tumor initiation to the time of sequencing.

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 proposed for some colorectal cancers [12] and for the evolution of copy number changes in breast cancers [7]. Samples with clones with higher selection coefficients, close to +20%, may have been sequenced before the first *selective sweep* took place. The low allelic frequencies of mutations in known driver genes in these samples are consistent with the simulations of young tumors in the study by Bozic et al. [2].

In the AML cohort, emergence times and clonal selection coefficients were much more variable. The emergence times of subclones in most samples were up to 60% of the tumor volume doublings, with the outliers in 3 samples which exceeded the estimated age of the tumor. In these samples, the model fits for $\alpha = 2$ were extremely inaccurate. Using a second model with an optimized α coefficient, we identified significant deviations of α from the expected value, indicating that the model assumptions were not fulfilled. We compared the optimum α values in primary and secondary tumors and identified an overall increase in the α in recurrent or metastatic samples.

Tung and Durrett [14] have shown that $\alpha < 2$ may result from numerous micro-clones under positive selection. However, this solution only explains values of α less than 2. In this thesis, we proposed a mathematical explanation for this phenomenon by showing that if the rate of accumulation of M' new mutations in a tumor depends on the tumor size:

$$M' = \mu \lambda N(t)^{\kappa} \tag{4}$$

where μ is the initial mutation rate, λ the growth rate, $N(t)$ the tumour size at time t, and $kappa$ the positive constant $in(0, \infty)$, then the number of mutations with frequency f can be expressed by:

$$X(f) = \frac{\mu}{f^{\kappa+1}} \tag{5}$$

If the mutation rate does not depend on the tumor size ($\kappa = 1$), then $\alpha = \kappa + 1 = 2$, and the equation 5 is consistent with the equations proposed by Durrett [6] and Williams [15]. Equation 5 shows that α different than 2 can result from changes in the tumor mutation rate.

5.2 Evolution of Bladder cancer from mucosal field effects

In most samples of both maps, we identified numerous mutations in known cancer *driver* genes, including many mutations in fields classified as normal urothelium or low-grade neoplasia. While the majority of these mutations were exclusive to specific fields of the map, there were 18 driver mutations spread across multiple fields in map 19 and 7 such mutations in map 24. Interestingly, there were 4 different patterns of mutation spread in map 19, with only two of them being linked to the cancer-affected area of the bladder. The 2 spread patterns were mainly related to fields classified as NU or LGIN and indicated expansions of mutant cells not associated with the main tumor.

We used our *cevomod* package to fit the power-law models to the neutral tails in most fields of both maps. The estimated coefficients revealed not only the high mutation rates in urothelial cancer fields but also elevated mutation rates in numerous fields classified as normal urothelium and low- or high-grade neoplasia. In map 19, we identified a non-cancerous area with an elevated mutation rate coinciding with an area of local clonal expansion of mutations in the ELF3, KMT2D, and RHOB genes undetected in cancer samples.

Then, using the second proposed model, we measured the deviations of the α from the predicted value of 2. Values greater than two predominated in map 19 and often involved samples classified as normal urothelium or low-grade neoplasia. In map 24, the upward and downward deviations were similarily frequent, but the downward deviations were more significant. The α values were similar in neighboring map fields, suggesting a propagation of processes that influenced these values into neighboring areas.

Deviations of α can result from various processes, such as the occurrence of micro-clones with a selection advantage, or the changes in the mutation rate. We have shown that mutation rate estimates for $\alpha > 2$ are lower than in models with $\alpha = 2$, high α values may therefore signal higher mutation rates. The reduction in α values resulting from the presence of micro-clones may therefore be compensated by an increase in the mutation rate. Indeed, α values less than 2 were observed less frequently than greater than 2, and the mutation rate was higher in samples with more advanced disease stages. Low α values were observed in some samples with a low mutation rate but a high number of mutations in cancer driver genes. The ultimate answer, if these samples indeed represent the case of the selection among the micro-clones described by Tung and Durrett [14], may be beyond the reach of *bulk* sequencing data analysis, which does not allow for distinguishing numerous small clones [13].

6 Summary

In this thesis, we used the bulk DNA sequencing data from over 140 tumor 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 samples representing primary tumors, metastases, tumor recurrences, or different stages of tumor progression. 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.

The MOBSTER algorithm could not correctly fit the models to the data used in the study due to an insufficient number of mutations in the neutral tail. For this reason, we proposed new methods to fit the model to data from whole-exome sequencing or sequencing with insufficient coverage. The methods have been implemented in a new R package, *cevomod*, fulfilling the second aim of the dissertation. The package is publicly available in the GitHub repository at `https://github.com/pawelqs/cevomod`. It allows the user 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. The first model allows the determination of evolutionary dynamics parameters such as mutation rate, clone emergence time, or clonal selection coefficients under the assumptions of exponential tumor growth and constant mutation rate. The second model allows validating these assumptions: deviations of the optimal α value are indicative of the failure to meet any of them.

Using the proposed methods, we have shown that the assumptions underlying popular models used to estimate the parameters of tumor evolution, such as the exponential population growth or constant mutation rate, may be violated in many cancers. We observed and described frequent deviations of the exponent describing the neutral tail of the distribution from the expected value of 2. We also proposed a mathematical explanation for the observed phenomena, linking it to changes in the mutation rate during tumor growth. Thus, we achieved the third and final goal of the dissertation.

Acknowledgements. This work was co-financed by the European Union through the European Social Fund (grant POWR.03.02.00-00-I029) and National Science Centre (Poland) grant 2018/29/B/ST7/02550. 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.

References

[1] 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 01/18/2023).

[2] 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 01/18/2023).

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

[4] 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 12/15/2022).

[5] 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 04/12/2023).

[6] 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 12/15/2022).

[7] 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 01/12/2023).

[8] 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 12/15/2022).

[9] 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 04/12/2023).

[10] 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 04/12/2023).

[11] 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 12/15/2022).

[12] 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 01/18/2023).

[13] 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.

[14] 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 01/18/2023).

[15] 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 12/15/2022).

[16] 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 12/15/2022).