# Computationally Reconstructing the Evolution of Cancer Risk Evolution

**Kefan Cao**

CMU-CS-25-105

April 2025

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

**Thesis Committee:**
Russell S. Schwartz, Chair
Oana Carja

*Submitted in partial fulfillment of the requirements for the degree of Master of Science.*

**Keywords:** Cancer, Machine Learning, Evolution, Risk Modeling

*For my advisor, whose patience and guidance made this journey. I am truly grateful.*

## Abstract

Understanding the evolution of cancer in its early stages is critical to identifying key drivers of cancer progression and developing better early diagnostics or prophylactic treatments. Early cancer is difficult to observe, though, since it is generally asymptomatic until extensive genetic damage has accumulated. In this study, we develop a computational approach to infer how once-healthy cells enter into and become committed to a pathway of aggressive cancer. We accomplish this through a strategy of using tumor phylogenetics to look backwards in time to earlier stages of tumor development combined with machine learning to infer how progression risk changes over those stages. We apply this paradigm to point mutation data from a set of cohorts from the Cancer Genome Atlas (TCGA) to formulate models of how progression risk evolves from the earliest stages of tumor growth, as well as how this evolution varies within and between cohorts. The results suggest general mechanisms by which risk develops as a cell population commits to aggressive cancer, but with significant variability between cohorts and individuals. These results imply limits to the potential for earlier diagnosis and intervention while also providing grounds for hope in extending these beyond current practice.

## Acknowledgments

I would like to thank my advisor, Russell Schwartz, for his guidance, support, and patience throughout the course of my research. I am also grateful to the members of Schwartz Lab for creating a supportive and encouraging research environment and making this journey both productive and enjoyable. Finally, I would like to thank my family for their unwavering support during my undergraduate and graduate studies.

This work is based in part on work previously submitted for publication as a preprint: Computationally reconstructing the evolution of cancer progression risk on bioRxiv [1].

[1] Available on https://www.biorxiv.org/content/10.1101/2024.12.23.629914v1

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Cancer remains a major source of mortality globally [7] despite many years of intensive research into prevention and treatment. Early screening was expected to improve outcomes by detecting cancers when they are treatable. While such efforts have saved lives, they underperformed early hopes in part due to overtreatment [3]: the statistical conclusion from reductions in mortality shows that as many early-detected cancers may never have posed a serious threat. In response, concerns about the harm to patients of overtreatment in turn has led to undertreatment [12] as evolving clinical practice towards more conservative treatment of cancers judged unlikely to be aggressive can lead to failure to aggressively treat some that need it. The challenge of navigating the complementary issues of overtreatment and undertreatment has led to numerous efforts to more accurately distinguish life-threatening from non-threatening cancers (c.f., [25]).

Much insight into how tumors evolve and progress has come from the discipline of tumor phylogenetics [19], i.e., the study of the evolutionary history of cancer cells. Phylogenetic analysis can reveal order of mutations in a cancer cell, the timing of these mutations, and the relationships between different subclones, e.g., whether the most deadly clones arise from a single lineage or through parallel events [10]. Work on cancer evolution has suggested that cancer is not normally a primary illness but rather usually a late-stage outcome of genetic instability, either intrinsic or due to environmental factors [2]. This instability could potentially indicate tumor risk long before the tissue becomes phenotypically abnormal [21]. Combined with the insight of overtreatment, these observations suggest that our primary goal in early cancer screening should focus not merely on detecting cancer *per se*, but also on identifying genetic lesions that will go on to be threatening to the patient.

Solving such prediction problems has led to a more prominent role for statistical inference and machine learning methods in personalized and precision cancer treatment. See, for example, [14] for a recent review. Machine learning tools have yielded ever better ability to identify those cancers that will threaten patient lives and predict which tumors will likely respond to what treatments [15]. Nonetheless, they are limited by our ability to gather data about tumors, our imperfect understanding of their biology, and the inherent stochasticity of the progression process.

Our prospects for predicting future cancer progression hinge on the question of when a cell lineage becomes committed to being a cancer, or a cancer with a bad progression outcome. If the risk of a cancer progressing is essentially constant until the moment it progresses, then the

prediction task is impossible. However, if progression risk gradually increases over a cancer's history then there is hope of predicting progression well before it occurs. There has been considerable prior work to answer variants of this question of how a tissue becomes committed to being an aggressive cancer, including the classic two-hit model [11] and the more recent "bad luck" model [23], which can be conceptualized as different ways of reasoning about how we expect measures of progression risk to vary over the history of a cell lineage. Does risk of progression rise suddenly, as from a single chance mutation shifting a tissue from low-risk to high-risk, or does it increase gradually, as from a series of mutations each slightly driving aggressiveness? Do these changes tend to occur early in tumor development, long before a cancer is typically detected, or late, once it is already advanced? The answers to these questions may be of great practical importance in understanding the limits of prospects for early detection of high-risk lesions or early interventions to keep incipient cancers off of high-risk pathways.

Our central goal in this study is to establish a model for reconstructing how progression risk develops over a tissue's history as it transitions from healthy to cancerous and potentially to lethality. We want to ask, if we could have observed the cells that eventually become cancers from their earliest development, how early could we have identified that they were on a trajectory to aggressive cancer? We emphasize that our goal here is not to provide the definitive answer to this question, but to show 1) that it is theoretically possible to ask it by computational methods applied to extant data sources and 2) that doing so will provide insight into basic cancer research and potentially actionable knowledge for improving early diagnosis and treatment. We recognize that we cannot definitively answer the question yet, largely because we do not have the ideal data for the proposed analysis, but proceed in the hope of inspiring future work to yield more definitive answers.

Our major contributions are:

- Developing a paradigm for tumor phylogenetics combined with machine learning to characterize how the landscape of cancer risk evolves over time.
- Implementing a realization of this paradigm using existing tools and data.
- Applying that realization to a pilot study of lung and colorectal cancers to suggest how cancer risk evolves and how it can vary between tumor types and individual patients.

The remainder of this paper describes how we accomplish these steps and examines the results and conclusions we can draw from them before returning in the Discussion to consider how this question might be asked better in the future.

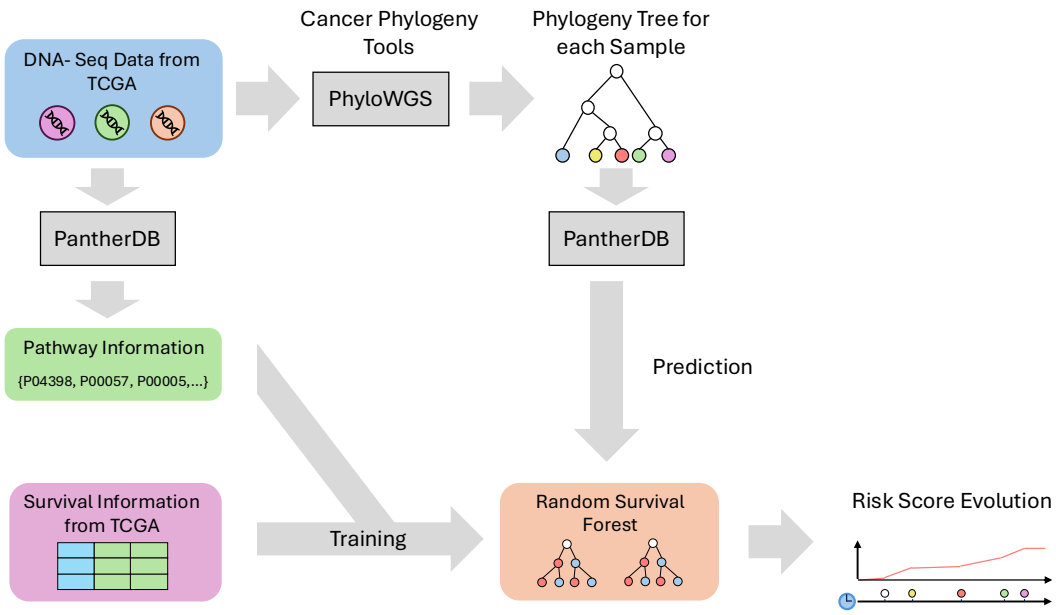


Figure 1.1: Summary figure describing the overall analysis pipeline.

# Chapter 2

# Methodology

## 2.1 Data Collection and Preprocessing

We apply our methodology to data from The Cancer Genome Atlas (TCGA) [4], which provides genomic data from large cohorts of thirty three cancer types with associated clinical and demographic metadata. We accessed TCGA data through the Genomic Data Commons (GDC) data portal [9], restricting analysis for the present study to DNA-seq single nucleotide variant (SNV) data. We also downloaded clinical data to extract survival information. We restricted our analysis to cohorts with at least 400 subjects and low class imbalance in survival outcomes due to our need for predictive regression models of censored survival. We therefore chose to focus the study on four cancers: lung adenocarcinoma (TCGA-LUAD), colorectal adenocarcinoma (TCGA-COAD), glioblastoma (TCGA-GBM), and head and neck squamous cell carcinoma (TCGA-HNSC).

## 2.2 Phylogenetic Analysis

A limitation of TCGA data is that it normally offers only one bulk DNA-seq sample per patient, which makes clonal phylogenetic inference challenging. To reconstruct the phylogenetic trees of the cohorts, we used PhyloWGS [5], a Bayesian method for tumor phylogeny inference from bulk SNV data, which our prior experience has shown to work comparatively well on single bulk samples. We applied PhyloWGS to the SNV data of each cohort to infer the evolution history of the cancer cell lineages. Where PhyloWGS inferred multiple possible trees, we selected the tree with the highest posterior probability. We then used these trees to estimate the time points of key mutations in the evolution of each cell lineage, as described below. To ensure a reasonable runtime and exclude samples with insufficient mutations, we limited our analysis to samples with mutation numbers ranging from 45 to 1200 SNVs. Additionally, we omitted a small number of samples for which PhyloWGS failed to return a result. Consequently, approximately 80% of the samples from each cohort were included in the study.

## 2.3 Pathway and Mutation Analysis

Due to large numbers of SNV mutations ($\geq 10,000$ genes for 400 to 1100 patients), we took a systems approach to reduce the problem dimension by aggregating mutations to key pathways relevant to tumor progression [17]. We used the PANTHER [16, 22] database to convert raw SNV data to affected pathways, assuming that an SNV in any gene in a pathway affects the pathway. We used random forest and Coxnet regression, via the scikit-survival [18] package, applied to the pathway data to identify key pathways that affect patient survival.

We used a random forest classifier to drive risk scores for each patient based on their pathway-mapped mutations and used random search to optimize the hyperparameters and a K-fold cross-validation with $k = 5$. Since the set of variants and affected pathways for the two cancers are different, we trained separate models for the two cancer types.

# Chapter 3

# Results

Figure 3.1 shows how predicted risk varies over time for the four cohorts, LUAD (Fig 3.1a-b), COAD (Fig 3.1c-d), GBM (Fig 3.1e-f), and HNSC (Fig 3.1g-h), separating subjects by survival outcome. Both cohorts show a qualitatively similar portrait of slowly increasing risk over time. For all four cohorts, mean risk scores are somewhat elevated in individuals with bad versus good survival outcomes throughout the tumor history, suggesting that there are at least sometimes intrinsic differences predictive of outcome from the earliest stages of cancer development. However, the variability patient-to-patient is substantially larger than the difference between good- and bad-outcome subgroups. In the case of LUAD, there is minimal separation by outcome early on, but they diverge in the latter half of their evolutionary trajectory, indicating that a significant portion of the determination of outcome occurs late in the tumor's development. In the case of COAD, the separation between surviving and deceased subjects is more consistent across the tumor's timeline. The case of GBM is similar to that of COAD, but they diverge at the last time point. The separation of the scores in HNSC is small from early on, and it fluctuates as the cancer progresses, and the risk scores also grow faster compared with the three other cohorts in this study.

For the four types of cancers we studied, risk scores are significantly different between living and deceased patients, as assessed by a t-test (p-value $< 0.05$). In these cases, however, the ability to distinguish patients by outcome improves sharply over the latter half of the trajectory as risk for both outcome groups increases. The separation only becomes statistically significant late in the progression process, although this is in part a function of the cohort size among other study variables and not just an intrinsic property of the system.

To better understand what drives the evolution of the risk scores, we plotted for each tumor type the mean fraction of mutations for the five most frequently mutated driver genes of the highest-risk clones. These appear as Fig 3.2 (LUAD) and Fig 3.3 (COAD). Each figure provides two subplots to separate patients surviving versus deceased over the course of the TCGA study followup. To map mutations to an approximate time axis, we assumed a molecular clock with mutations accumulating at a uniform rate across the tree and the length of a tree branch proportional to the number of mutations it contains. The figures show that while the overall risk evolution appears similar for the two cohorts, they exhibit quite different patterns of accumulation of key mutations in the evolution of cancer cells. LUAD involves a more gradual increase in inferred mutations over the course of a tumor's development. It is also less dominated by any

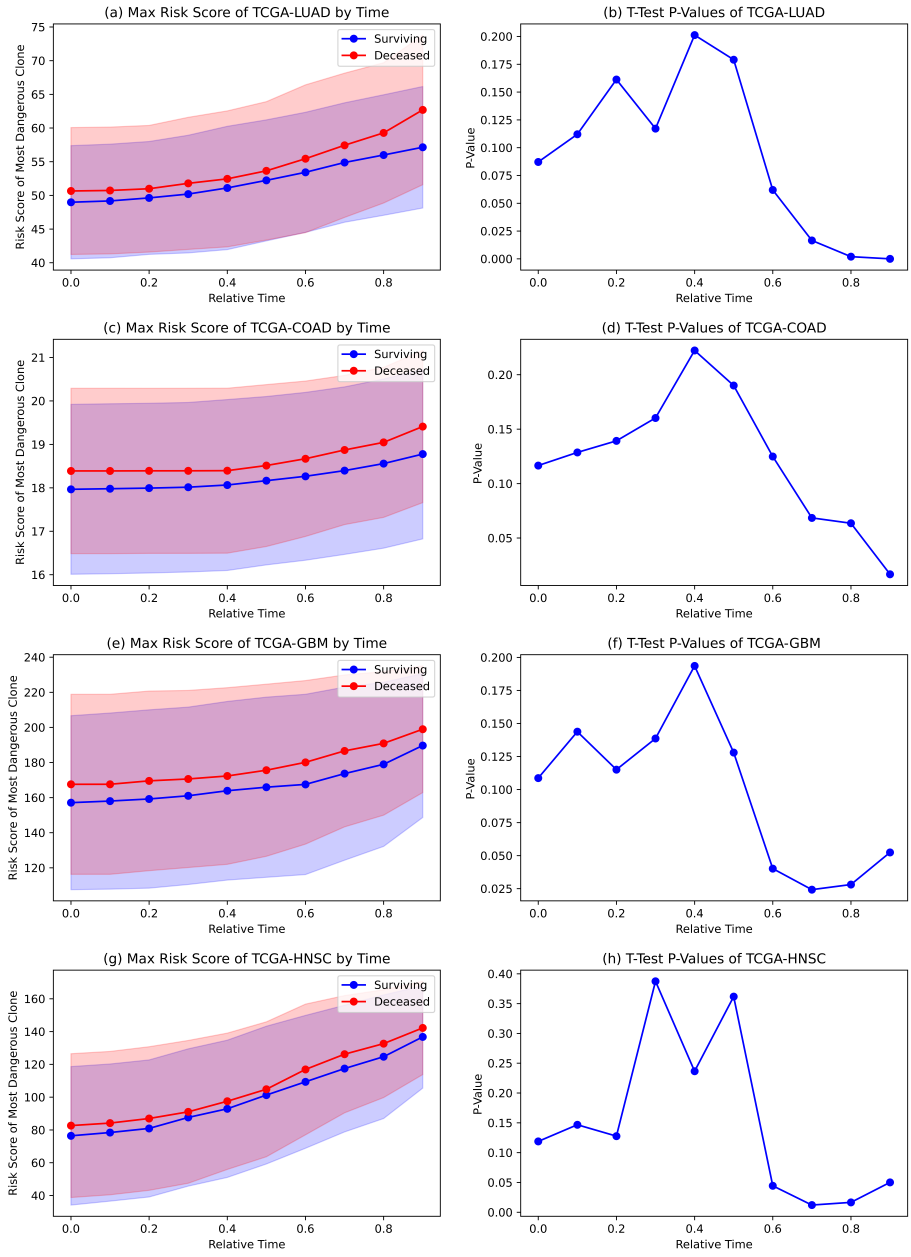


Figure 3.1: Risk scores and p-values versus inferred time for LUAD and COAD cohorts. (a) Risk score of the most high-risk clone vs. inferred time for LUAD subjects. (b) p-value for distinguishing by survival outcome in LUAD subjects. (c) Risk score of the most high-risk clone vs. inferred time for COAD subjects. (d) p-value for distinguishing by survival outcome in COAD subjects.

one mutation, although TP53 emerges early as the most common mutation in each. There is little evident qualitative difference between the good and bad outcome plots aside from higher levels

of TP53 mutation and lower levels of CSMD3 in bad-outcome cancers. COAD, by contrast, shows a sharper increase and higher rate of mutation, especially for the two most mutated genes (TP53 and APC). While mutations are predicted to accumulate in these genes from early times, there is a notable later increase in their mutation frequencies.

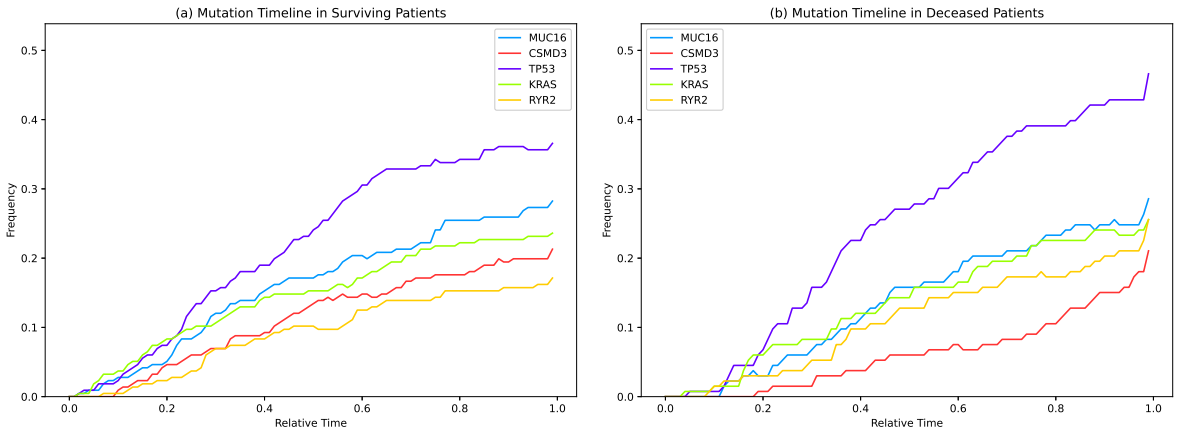


Figure 3.2: Fractions of patients with key mutations in the evolution of cancer cells for LUAD. For each subfigure, the X-axis shows the relative time of the mutation and the Y-axis is the fraction of patients with the mutation. (a) Accumulation of mutations over time in surviving patients. (b) Accumulation of mutations over time in deceased patients.

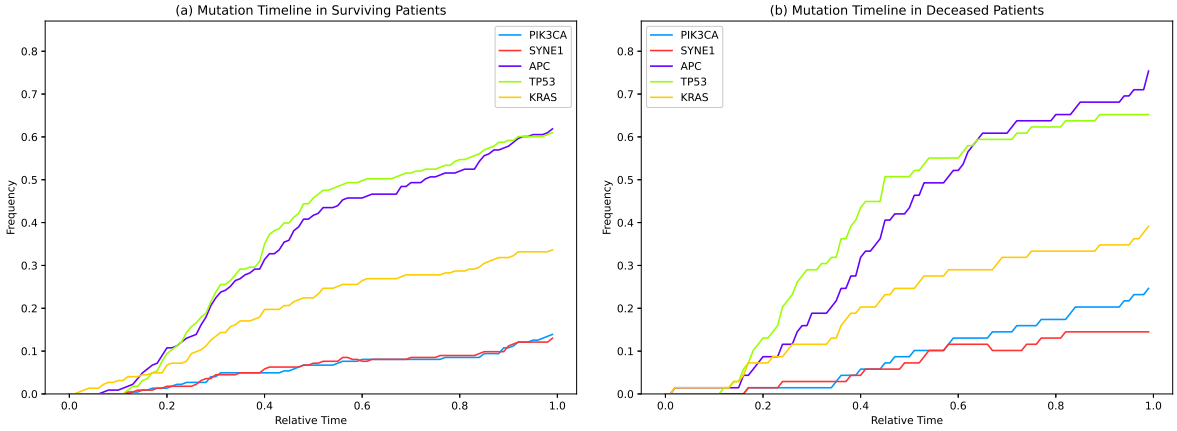


Figure 3.3: Fractions of patients with key mutations in the evolution of cancer cells for COAD. For each subfigure, the X-axis shows the relative time of the mutation and the Y-axis is the fraction of patients with the mutation. (a) Accumulation of mutations over time in surviving patients. (b) Accumulation of mutations over time in deceased patients.

Fig 3.2(a-b), Fig 3.3(a-b), Fig 3.5(a-b), and Fig 3.4(a-b) examine the overall accumulation of mutations for the top genes and pathways for each cancer, broken down by good and bad outcome cancers. Table 3.1 provides descriptors for the pathways identified by this analysis. The

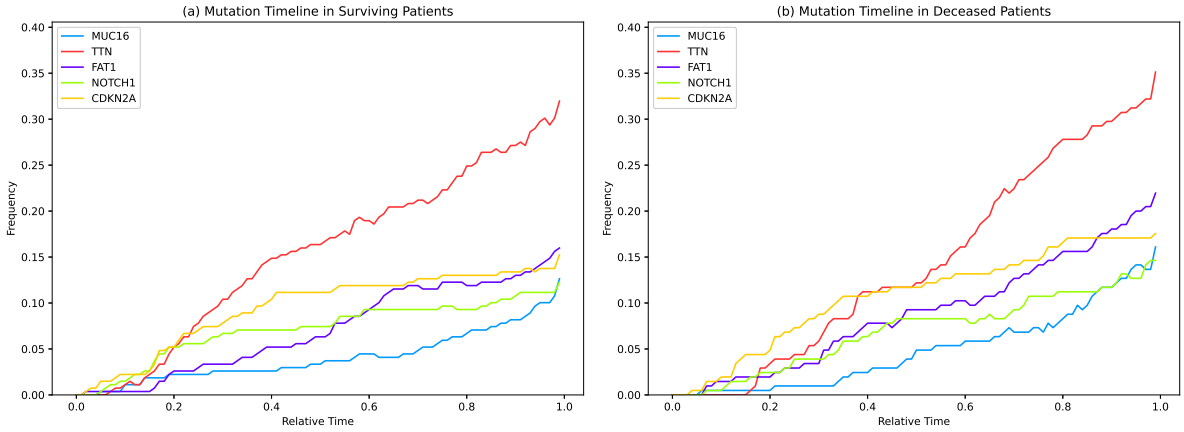


Figure 3.4: Fractions of patients with key mutations in the evolution of cancer cells for HNSC. For each subfigure, the X-axis shows the relative time of the mutation and the Y-axis is the fraction of patients with the mutation. (a) Accumulation of mutations over time in surviving patients. (b) Accumulation of mutations over time in deceased patients.

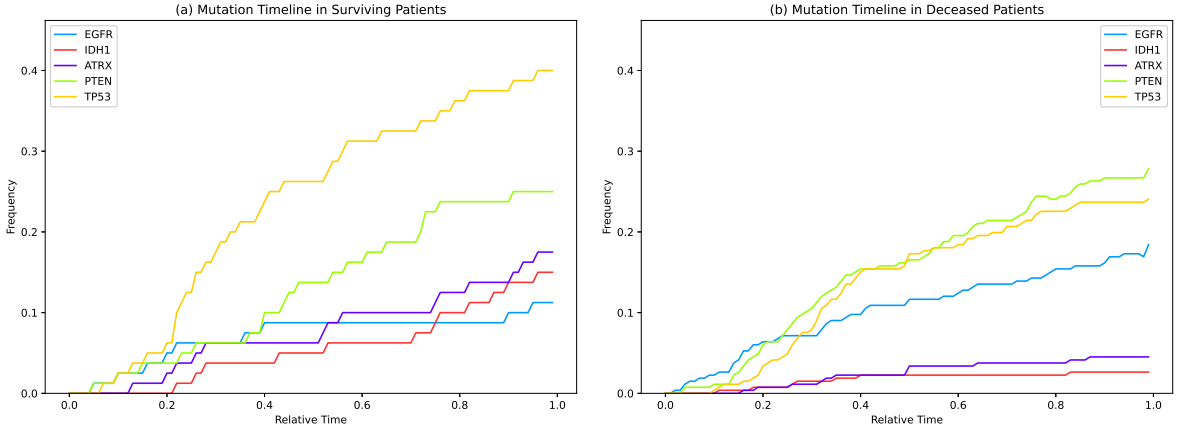


Figure 3.5: Fractions of patients with key mutations in the evolution of cancer cells for GBM. For each subfigure, the X-axis shows the relative time of the mutation and the Y-axis is the fraction of patients with the mutation. (a) Accumulation of mutations over time in surviving patients. (b) Accumulation of mutations over time in deceased patients.

top-ranked genes and pathways are similar as are frequencies between cohorts. The high-ranking pathways themselves are largely consistent with prior expectations, for example in prominent representation of Wnt signaling, TP53 signaling, and inflammatory pathways. The most striking difference for LUAD is the relatively higher rate of TP53 mutations and somewhat high rates of mutations across top-scoring pathways in subjects with poor outcomes. COAD shows similar results, with notably higher frequencies of mutations in TP53 and APC in deceased versus surviving groups and a qualitatively similar difference in pathways affected despite somewhat different pathways showing up as relevant. We also noticed that the frequency of mutations in

TP53 is higher in the surviving groups of GBM, starting at approximately 0.2 in relative time.

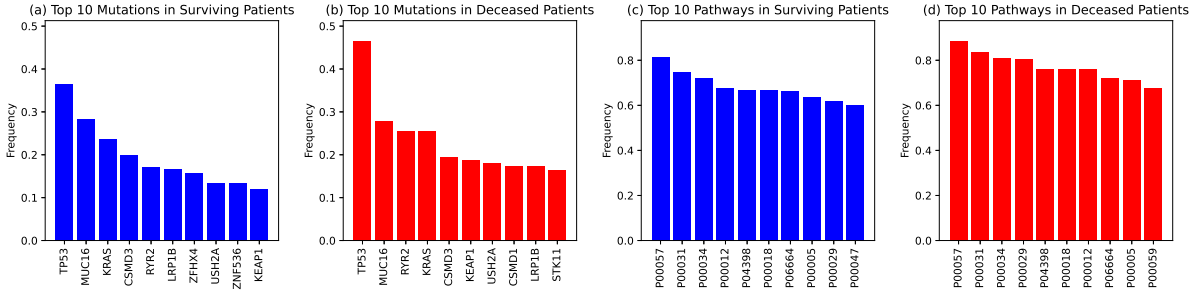


Figure 3.6: Drivers of risk evolution for LUAD.

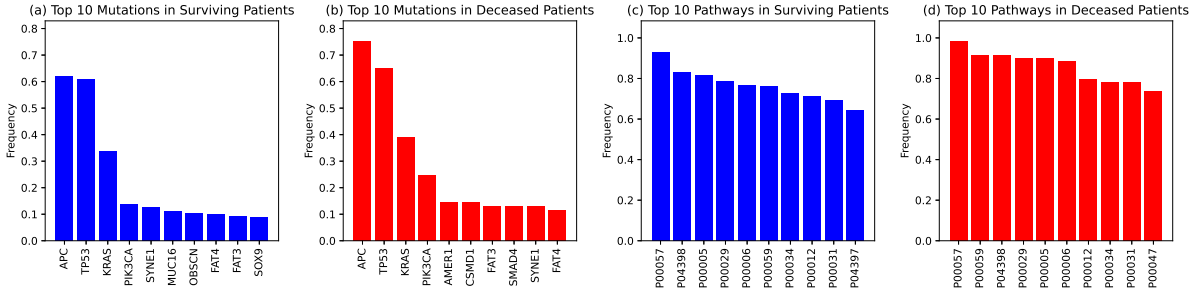


Figure 3.7: Drivers of risk evolution for COAD.

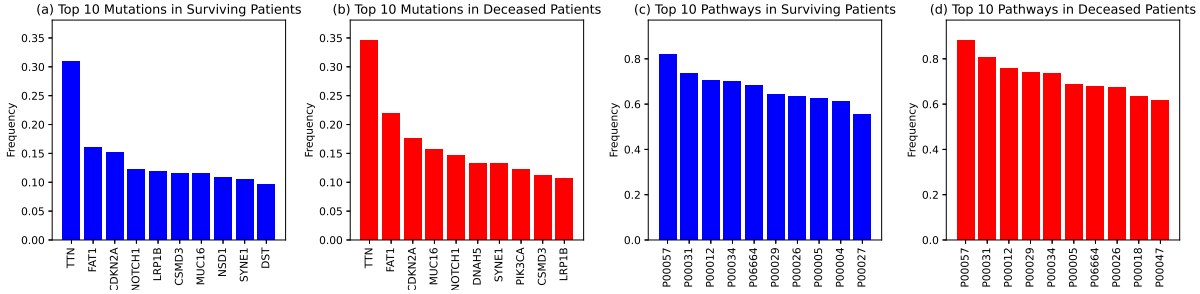


Figure 3.8: Drivers of risk evolution for HNSC.

Individual genes and pathways arising in these analyses are mostly consistent with expectations from prior literature. For example, EGFR is known to be important in early lung cancers [26] and although it does not show up as one of the five most frequently mutated genes, examining it individually shows that it is indeed predicted by our analysis to accumulate predominantly early in poor-outcome LUAD cancers (Figure 3.10). Some other key mutations, such as ERBB2, occur mostly through copy number alterations (CNAs) and would not be seen in this study. Other key mutations in defining cancer risk, such as TP53 [26] have been suggested to

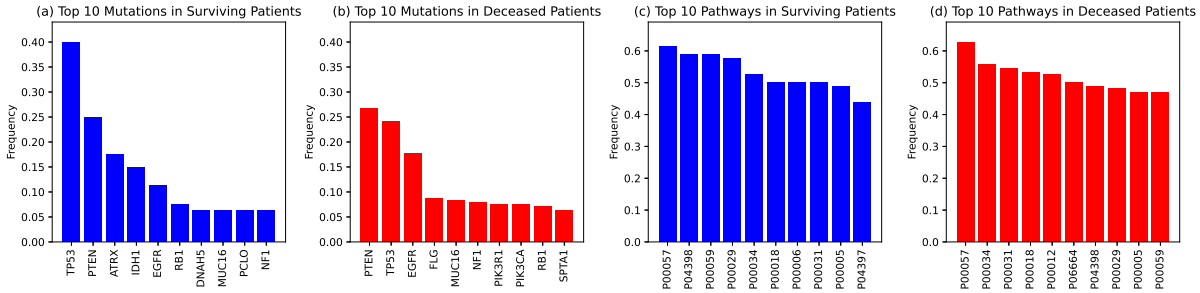


Figure 3.9: Drivers of risk evolution for GBM.

| Pathway ID | Pathway Name |
|---|---|
| P00057 | Wnt signaling pathway |
| P04398 | p53 pathway feedback loops 2 |
| P00031 | Inflammation mediated by chemokine and cytokine signaling pathway |
| P00034 | Integrin signalling pathway |
| P00012 | Cadherin signaling pathway |
| P00018 | EGF receptor signaling pathway |
| P06664 | Gonadotropin releasing hormone receptor pathway |
| P00005 | Angiogenesis |
| P00029 | Huntington disease |
| P00047 | PDGF signaling pathway |
| P00059 | p53 pathway |
| P00006 | Apoptosis signaling pathway |
| P04397 | p53 pathway by glucose deprivation |

Table 3.1: Most frequently mutated PANTHER pathways in the LUAD and COAD cohorts

occur more often in the later stages of cancer evolution, although our results suggest high patient-to-patient variability. In contrast, key risk-driving mutations in COAD [24] appear less likely to be seen in the root node of the phylogeny but rather appear to accumulate more gradually. We also noticed a few unexpected results, for example, the frequency of TP53 mutation is higher in the surviving group of GBM compared with the deceased group (Figure 3.9).

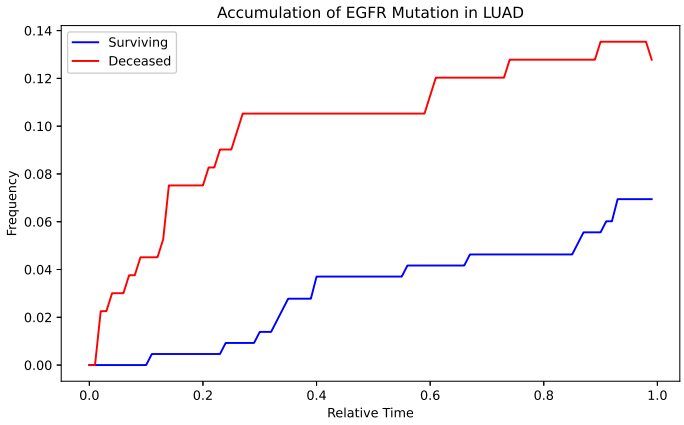


Figure 3.10: Fraction of mutated EGFR versus estimated time for LUAD.

# Chapter 4

# Conclusion and Discussion

We sought to characterize how risk of progression to aggressive cancer develops over time. We established a computational inference model for answering this question with extant cancer genomic and outcome data. We applied an implementation of that model to two cancer cohorts, revealing qualitatively similar portraits of risk accumulation although with some variability by tumor type and large variability by subject within each type. The results suggest that there are average differences predictive of outcome, captured in machine learning-derived risk scores, present from the earliest stages of tumor development although these differences are dominated by high variability within each cohort. Statistically significant separation of good and bad outcome subjects by risk score becomes possible only relatively late, in part reflecting the high variability by subject. However, this separation might be pushed back to earlier times with larger cohorts, more comprehensive variant data, or more accurate phylogeny inferences, among other factors. The results suggest that there is potential, for at least subsets of patients, for earlier diagnostics or interventions before cancer is commonly detected — and potentially even before it is cancer — but further advances in methodology will be needed to judge the absolute limits.

As noted in the Introduction, our goal is not to definitively answer the question we pose about the evolution of cancer risk but rather to demonstrate that it is feasible and worthwhile to answer it, in the hope of inspiring future work. The biggest obstacle to answering this question better is data. The present work used TCGA data because the nature of the question requires relatively large cohorts with consistent data types, and TCGA was a landmark in making such data available. TCGA sequencing was not designed with tumor phylogenetics in mind, though, and did not have access to many technologies available to us today. We further expect that we are missing a substantial portion of risk-driving mutations by focusing here only on SNV mutations, ignoring copy number alteration (CNA) and various kinds of structural variation (SV) mutations important in cancers. Methods exist to map SNV, CNA, and SV mutations to a tumor lineage tree tree [6, 8], but they perform poorly on single-sample data such as is available with TCGA. Sufficiently large cohorts with multiple samples per patient, single-cell genomic data, longitudinal data, and protocols for profiling somatic variation in non-cancerous tissues [13] might improve the effectiveness of future studies along these lines. How to choose among many possible options in designing such a study is itself an emerging question for which tools are beginning to become available [20]. The computational tools used in this study are also older, in part because they were selected to be suitable for the data currently available; better methods

for phylogenetic risk prediction might also be chosen or developed to accommodate richer data sources.

Finally, we consider what one could do with answers from a study such as is presented here. Better models of cancer risk progression could have particular value to early cancer screening and public health interventions. As we now appreciate that somatic variation is ubiquitous [1] and usually asymptomatic, it is important to develop better ways to identify when seemingly harmless variation poses a risk and better respond to problems of both overtreatment and undertreatment. Studies such as that prototyped here may also inspire better ways to respond to cancers prophylactically before they become threatening. Better characterizing the features defining high-risk lesions may also help during treatment in tailoring personalized therapies to high-risk clones or predicting when and how a tumor under treatment might recur or become more aggressive.

# Bibliography


[1] A Acha-Sagredo, P Ganguli, and F D Ciccarelli. Somatic variation in normal tissues: friend or foe of cancer early detection? *Ann Oncol*, 33(12):1239–1249, September 2022. 4

[2] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *nature*, 500 (7463):415–421, 2013. 1

[3] Jaimin R Bhatt and Laurence Klotz. Overtreatment in cancer – is it a problem? *Expert Opinion on Pharmacotherapy*, 17(1):1–5, 2016. doi: 10.1517/14656566.2016.1115481. URL https://doi.org/10.1517/14656566.2016.1115481. PMID: 26789721. 1

[4] Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas Pan-Cancer analysis project. *Nat Genet*, 45(10):1113–1120, October 2013. 2.1

[5] Amit G. Deshwar, Shankar Vembu, Christina K. Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. Phylowgs: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, Feb 2015. ISSN 1465-6906. doi: 10.1186/s13059-015-0602-8. URL https://doi.org/10.1186/s13059-015-0602-8. 2.2

[6] Jesse Eaton, Jingyi Wang, and Russell Schwartz. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics*, 34(13):i357–i365, 2018. 4

[7] Clara Frick, Harriet Rumgay, Jérôme Vignat, Ophira Ginsburg, Ellen Nolte, Freddie Bray, and Isabelle Soerjomataram. Quantitative estimates of preventable and treatable deaths from 36 cancers worldwide: a population-based study. *The Lancet Global Health*, 11(11): e1700–e1712, Nov 2023. ISSN 2214-109X. doi: 10.1016/S2214-109X(23)00406-0. URL https://doi.org/10.1016/S2214-109X(23)00406-0. 1

[8] Xuecong Fu, Haoyun Lei, Yifeng Tao, and Russell Schwartz. Reconstructing tumor clonal lineage trees incorporating single-nucleotide variants, copy number alterations and structural variations. *Bioinformatics*, 38(Supplement_1):i125–i133, 2022. 4

[9] Robert L. Grossman, Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt. Toward a shared vision for cancer

genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016. doi: 10.1056/NEJMp1607591. 2.1

[10] Woo Suk Hong, Max Shpak, and Jeffrey P. Townsend. Inferring the origin of metastases from cancer phylogenies. *Cancer Research*, 2015. doi: 10.1158/0008-5472.CAN-15-1889. 1

[11] Alfred G Knudson Jr. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823, 1971. 1

[12] Janusz Kocik. Undertreatment or overtreatment – how far from each other in geriatric oncology? 2(1):18–25, 2020. doi: 10.36553/WM.15. 1

[13] Iñigo Martincorena and Peter J. Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015. doi: 10.1126/science.aab4082. URL https://www.science.org/doi/abs/10.1126/science.aab4082. 4

[14] Michele Masucci, Claes Karlsson, Lennart Blomqvist, and Ingemar Ernberg. Bridging the divide: A review on the implementation of personalized cancer medicine. *Journal of Personalized Medicine*, 14(6), 2024. ISSN 2075-4426. doi: 10.3390/jpm14060561. URL https://www.mdpi.com/2075-4426/14/6/561. 1

[15] Hiba Mechahougui, James Gutmans, Gina Colarusso, Roumaïssa Gouasmi, and Alex Friedlaender. Advances in personalized oncology. *Cancers (Basel)*, 16(16):2862, August 2024. 1

[16] Huaiyu Mi and Paul Thomas. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol*, 563:123–140, 2009. 2.3

[17] Yongjin Park, Stanley Shackney, and Russell Schwartz. Network-based inference of cancer progression from microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(2):200–212, 2008. 2.3

[18] Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. URL http://jmlr.org/papers/v21/20-729.html. 2.3

[19] Russell Schwartz and Alejandro A. Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4):213–229, Apr 2017. ISSN 1471-0064. doi: 10.1038/nrg.2016.170. URL https://doi.org/10.1038/nrg.2016.170. 1

[20] Arjun Srivatsa and Russell Schwartz. Optimizing design of genomics studies for clonal evolution analysis. *Bioinformatics Advances*, 4(1):vbae193, 12 2024. ISSN 2635-0041. doi: 10.1093/bioadv/vbae193. URL https://doi.org/10.1093/bioadv/vbae193. 4

[21] Yifeng Tao, Ashok Rajaraman, Xiaoyue Cui, Ziyi Cui, Haoran Chen, Yuanqi Zhao, Jesse Eaton, Hannah Kim, Jian Ma, and Russell Schwartz. Assessing the contribution of tumor mutational phenotypes to cancer progression risk. *PLoS computational biology*, 17(3): e1008777, 2021. 1

[22] Paul D. Thomas, Dustin Ebert, Anushya Muruganujan, Tremayne Mushayahama, Laurent-Philippe Albou, and Huaiyu Mi. PANTHER: Making genome-scale phylogenetics accessi-

ble to all. *Protein Science*, 31(1):8–22, 2022. doi: https://doi.org/10.1002/pro.4218. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4218. 2.3

[23] Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81, 2015. 1

[24] Ziwei Wu, You Li, Yibin Zhang, Hui Hu, Tangwei Wu, Shuiyi Liu, Weiqun Chen, Shenggao Xie, and Zhongxin Lu. Colorectal cancer screening methods and molecular markers for early detection. *Technol Cancer Res Treat*, 19:1533033820980426, January 2020. 3

[25] Bo Zhang, Huiping Shi, and Hongtao Wang. Machine learning and AI in cancer prognosis, prediction, and treatment selection: A critical approach. *J. Multidiscip. Healthc.*, 16:1779–1791, June 2023. 1

[26] Chao Zhang, Jianjun Zhang, Fang-Ping Xu, Yin-Guang Wang, Zhi Xie, Jian Su, Song Dong, Qiang Nie, Yang Shao, Qing Zhou, Jin-Ji Yang, Xue-Ning Yang, Xu-Chao Zhang, Zhi Li, Yi-Long Wu, and Wen-Zhao Zhong. Genomic landscape and immune microenvironment features of preinvasive and early invasive lung adenocarcinoma. *J Thorac Oncol*, 14(11):1912–1923, August 2019. 3