Evolution and Immunity in Cancer and HIV

Evolution and Immunity in Cancer and HIV

a dissertation presented
by
Jeffrey M. Gerold
to
The Division of Medical Sciences

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biomedical Informatics

Harvard University
Cambridge, Massachusetts
May 2019

Thesis advisor: Professor Martin Nowak Jeffrey M. Gerold

Evolution and Immunity in Cancer and HIV

Abstract

Cancer and HIV are frequently-incurable diseases with high global burden. As evolving populations within a single individual, both exhibit dramatic expansions in size and diversity which interact with the immune system and complicate treatment. The advent of cheap and accessible DNA sequencing and quantification technologies has enabled detailed measurements of these diseases from samples often distributed sparsely in time. Integrating these data into a quantitative, dynamical view of disease progression in order to improve treatments remains an open challenge.

In this thesis, we explore examples of data integration with dynamical models in three areas: cancer evolution, cancer surveillance by the immune system, and HIV infection under an immune response. In the first chapter, we describe a tool for phylogenetic inference using DNA sequencing of spatially distinct samples from a cancer. In benchmarks, the tool overcomes noise introduced by sequencing to provide a picture of the evolutionary history of a tumor. In the second and third chapters, we describe the results of applying this tool to two datasets, first in primary pancreatic cancers with matched preneoplastic lesions and next in untreated metastases. We find many shared driver mutations among the primary tumor and preneoplastic lesions, suggesting preneoplastic cells can spread through the pancreas. In untreated metastases, we observe limited driver gene heterogeneity, consistent with a model of growth, mutation, and metastasis seeding from the primary tumor. In the fourth chapter, we describe a branching process model of neutral evolution in tumors and fit analytical predictions from it to cancer sequencing data. This neutral model explains

patterns in sequencing data for many tumors. In the fifth chapter, we propose and analyze a simple model of cancer immune surveillance. We find that tumors susceptible to immune clearance must have a rate of mutation higher than is usually observed clinically. In the final chapter, we propose a dynamical model of viral rebound and immune control and compare it to data from several studies in macaques infected with SIV and SHIV and treated with immunotherapy. These results are combined with data from HIV infection to make predictions for future trials in humans.

Contents

Listing of figures

For my teachers.

Don't know much about history
Don't know much biology
Don't know much about a science book
Don't know much about the french I took
But I do know that I love you
And I know that if you love me, too
What a wonderful world this would be

Sam Cooke, *What a Wonderful World*

Acknowledgments

This dissertation would not have been possible without the support of many people.

I am indebted to Martin Nowak for his support of me and the amazing environment that is the Program for Evolutionary Dynamics. His generosity, mentorship, and vision have been invaluable. Along the way, I encountered many other amazing mentors and collaborators, including Ivana Bozic, Hannes Reiter, and Alvin Makohon-Moore. Alison Hill provided exceptional insights and mentorship for all projects related to HIV. I benefited enormously from all their examples of scientific excellence and leadership. The larger community of the lab fostered an amazingly creative and dynamic environment in which it was a joy to be a graduate student. The ping-pong, happy hours, debates, and scientific dialogue with fellow PED graduate students and post-docs Chuck Noble, Sam Sinai, Carl Veller, Pavitra Muralidhar, Alex Heyde, Ben Adlam, Iulia Neagu, Anjalika Nande, Andrei Gheorghe, Julian Kates-Harbeck, Ski Krieger and everyone else at PED. The tireless work of administrators May Huang and Melinda Peterson and alumna Katherine Gallagher ensured the stresses of funding and travel rarely interfered with my experience in graduate school.

The members of my dissertation committee, including John Wakeley, Kenneth Kaye, and Shamil Sunyaev made many valuable suggestions for the direction of specific projects and how to navigate

the graduate student experience. I am especially grateful to Shamil for his tireless commitment to all the PhD students in the Bioinformatics and Integrative Genomics program. The program would not be possible without the support of the department chair Peter Park and the team of faculty assembled to support our community of students. Departmental administrators Cathy Haskell and Katherine Flannery made the process of navigating the program seamless and friendly. They also helped enormously to support all the departmental recruitment, seminars, and provide student guidance.

I was supported in my graduate studies by the generosity of the Landry Family through a Landry Cancer Biology Fellowship. I am grateful for their support and for the amazing community of students and faculty that Harvard has assembled under their banner. The seminars and informal discussions with students and faculty helped shape the direction of the research here.

I would not have encountered this graduate program at all had it not been for the timely intervention of Susanne Churchill, Isaac Kohane, and Martha Bulyk. Martha helped me, patiently and with characteristic rigor, to grasp the challenges and opportunities in computational biology when I was a lowly summer student. I am indebted to them for their influence on my trajectory in graduate school.

The support of my family and friends has been unflinching and invaluable. Graduate school is a long and challenging adventure which I could not have completed without their constant presence. The support of my parents, Jane and Marty, my brother, Michael, and all of my relatives ensured home never felt far away. Thank you to you all.

0

Introduction

This dissertation presents a collection of projects which explore the dynamics of two diseases: cancer and HIV. In both diseases, the ongoing replication of very many nearly identical units gives rise to disease pathology. The high amount of symmetry in each system suggests that a description which exploits it (i.e. a model) could be much less complex than a precise physical description of the system while remaining relatively accurate. Further, many natural questions about cancer and HIV lend themselves to modeling. For example, we cannot observe all the individual cells in a tumor, but we

might wish to know if are there any cells resistant to a particular treatment present in the tumor. A model of the evolution of a tumor will assign a probability to this event, and in some cases this approach will be more efficient than doing the very difficult work of improving measurement technology for cancer. We are also often interested in questions related to the past: When was someone first infected with HIV? How large was a tumor 10 years ago? Models provide a way for us to reason about these questions.

Next, we will briefly review the features of each disease essential to our modeling approaches.

0.1 Cancer

A cancer is a self-renewing population of asexually reproducing cells derived from healthy tissue of an organism. Cancer is a disease of the genome[58]. That is, a handful of genetic changes, perhaps 3[256], amidst a sea of genetic information ($3 * 10^9$ base pairs) are sufficient to transform a healthy human cell into a cancerous one. Genetic mutations can enter the DNA both before and during DNA replication (mitosis). In the modeling here, we group these rates together into the rate associated with replication for simplicity. Our picture of cancer is simplified additionally: the process of transformation in reality also interact with environmental factors in several ways. First, the rate of genetic change itself is influenced by the environment. Some environmental factors interfere with the faithful replication of DNA when cells divide, increasing the rate at which driver mutations arrive in cells. Second, the effects of genetic mutations are context-dependent, though in cancer there is a large set of mutations which seem to promote the development of cancer across a variety of contexts[11]. The genetic changes required to reliably induce a cancerous population of cells might be fewer in an individual with compromised immunity or local tissue damage. In some cases they also depend on the cell type in which the mutations arise.

A cancer starting from a single cell takes a very long time (sometimes decades) to grow to a clini-

cally meaningful size[279]. By this time, the cancer population has grown from one to 10^9 ($\sim 1cm^3$) cells. Because many cancer cells will die before having the chance to divide, a population of 10^9 cells will have undergone even more divisions, say 10^{11} [33]. The interaction of this extremely large number of divisions with very rare genetic processes such as the development of therapy-resistance mutations or additional driver gene mutations has been the focus of several modeling efforts (e.g. Refs.[37,215,77]). In chapter 5 we will explore how an immune system which recognizes foreign genetic material might interact with a population that has acquired so many genetic changes. The kinetics of cancer growth have been studied in great detail[276]. The early growth of a tumor is often approximately exponential, followed by slower growth as the tumor approaches a carrying capacity due to space constrains, nutrient limitations, or other forces. A branching process is a mathematical model of cancer cell behavior which gives rise to exponential growth and is conservative in the sense that it predicts cancer growth which does not slow down at a carrying capacity. In a branching process, a single cancer cell either divides into two or dies stochastically. In a growing tumor, the expected number of offspring of a cancer cell must be greater than 1. During division, daughter cells might also acquire mutations which distinguish their genomes from the parental genome. The branching process model provides a way to connect observable quantities in the tumor, like the rate of division, net growth rate, or the number of detectable mutations, to model parameters.

Because a cancer is often observed when it is very large, it is natural to imagine a branching process (or other tumor model), but moving *backwards* in time. This is called a coalescent process, because from this perspective, cells merge into their parental lineages ("coalesce") as time goes on. For a population with very large but fixed size, Kingman showed that a large class of models actually have essentially the same type of coalescent[133]. However, for an exponentially growing population, the situation is more complicated: the Kingman coalescent and the reverse-time branching process will be similar but never quite match[246]. Further, all coalescent processes can be thought of as trees. In a group of N individuals, two among them will have a most-recent common ancestor. After connect-

ing these two individuals to their most-recent common ancestor, among the now N-1 individuals (including the ancestor but not the original two individuals) there will again be a pair with a most-recent common ancestor. Individuals are joined in this way until only one remains: the most recent common ancestor of all the individuals in the population. The resulting graph is a tree which describes the history of reproduction in the population. In this thesis, when we say "phylogenetic tree" we are referring to this graph.

0.2 HIV

The emergence of HIV in the human population caused a public health crisis of global proportions. Until the late 1980s, an HIV diagnosis was essentially a death sentence[39]. But prolonged scrutiny of the virus from the biomedical research community, culminating in understanding of the mechanisms of HIV replication, enabled the development of small molecules which interfere specifically with the HIV life cycle. The change in an individual with HIV who received this antiretroviral therapy came to be known as the Lazarus effect.

In contrast to cancer, the process of reproduction in HIV is more complicated. HIV is a member of the Lentivirus genus, all of which are single-stranded RNA viruses which undergo reverse transcription into double-stranded DNA. In an HIV replication cycle, the HIV RNA genome, enveloped by a capsid protein, enters a cell and is converted into DNA by an enzyme called reverse transcriptase. This DNA integrates into the host genome and is then later transcribed back into RNA and translated into new viral particles. The HIV genome is only about 10^5 bases (compared to human genome $3 * 10^9$), but the rate of mutation during a cycle of replication is extremely high compared to humans ($3 * 10^-5$ per position per generation).

The very high rate of mutation in HIV contributes to the inability of a single antiretroviral therapy to reliably control an infection. Current treatment regimens typically combine three drugs—

often with disparate resistance mechanisms—in order to ensure that there is essentially no ongoing viral replication and, even if there were, combination resistance would remain quite unlikely[109]. However, several challenges remain in the fight against HIV. First, for infected individuals receiving treatment, therapy must be taken for life and is not without side effects. Second, for infected individuals not receiving treatment, challenges in global health pose barriers to the distribution and use of effective treatment regimens[of Health Statistics & Informatics]. Third, uninfected individuals remain at risk of disease transmission, due in part to the lack of an effective vaccine for HIV.

0.3 Chapters Overview

The chapters in this thesis are self-contained, each with its own introduction, description of methods and results, and discussion. Most of the work in this dissertation has been previously published. For these chapters, supplemental material has been omitted from the dissertation to avoid the length becoming unwieldy. Instead, the supplemental material is available online with the original publication. Because many of the projects involved large collaborations, in the forward before each chapter I have clarified some of the more practical details of the projects as well as my principal contributions.

In chapter 1, we describe a method, called Treeomics, for inferring the phylogenetic tree among several cancer samples taken from the same individuals. In this method we assume that each spatially distinct sample of a tumor represents a highly related collection of cells, so that we might imagine the phylogenetic tree we infer corresponds to the phylogenetic tree associated with the most recent common ancestor of each sample. The problem is complicated by the fact that the sequencing data considered is bulk tumor DNA sequencing, which has multiple sources of noise and destroys linkage information between mutations. The method combines information from each mutation in a model-based framework in order to infer the most consistent phylogeny.

In chapter 2, we describe an application of Treeomics to data collected from matched cancerous and precancerous lesions in the pancreas. The coincidence of these samples provides a perfect control for understanding what distinguishes the cancerous lesions, since in a particular individual the samples experienced a highly similar environment and they began on almost identical genetic backgrounds. We pay special attention to the distribution of driver mutations in these samples in order to assess the relatedness of these samples as well as whether or not precancerous lesions have experienced the genetic changes necessary for transformation.

In chapter 3, we describe another application of Treeomics to previously published data from untreated metastatic cancer samples. Many published datasets contain a few untreated individuals, and we collect these data in order to provide a reliable picture of the natural progression of metastasis. These data are particularly valuable because most treatments include some form of mutagenic therapy which obscures the underlying biological processes shaping the original metastasis formation. We again pay special attention to the distribution of driver mutation heterogeneity among metastases in a single individual in order to assess how much sampling would be required to develop a comprehensive picture of the driver gene mutations in a metastatic cancer.

In chapter 4, we turn to a more theoretical model of neutral mutations in an exponentially growing cancer. Starting from a model of cell division, death, and mutation, we derive the distribution of several observable quantities in terms of these parameters. We also derive an invariant of the model which can be used to assess deviations from neutrality and compare this result to cancer sequencing data.

In chapter 5, we describe unpublished work modeling the surveillance of a tumor by the immune system. This model builds on the features of the model in chapter 4 to account for an immune response which can distinguish self from non-self. Though the precise nature of immune surveillance remains poorly understood, our approach provides important bounds on its effectiveness for reducing cancer burden overall. We also derive the distribution of several observable quantities, including

the variant allele frequency spectrum, under different models of immunity.

In chapter 6, we analyze data from several trials in a model organism for HIV, the macaque. These trials investigate a combination of immunomodulatory treatments and therapeutic vaccination, that is, vaccination given to an infected individual as a form of treatment rather than prophylaxis. The data in these trials are some of the first examples of immunologic control of viremia, and therefore are very exciting to understand for the development of novel vaccination and therapy strategies.

1

Reconstructing metastatic seeding patterns of human cancers

1.1 Forward

This chapter describes an early method for inferring cancer phylogenies when multiple, spatially distinct samples are taken from a single tumor and subjected to DNA sequencing in bulk. The

work grew out of necessity from a collaboration between Martin Nowak and Christine Iacobuzio-Donahue. Hannes Reiter laid the algorithmic groundwork during his PhD in computer science at the Institute of Science and Technology Austria. Alvin Makohon-Moore, then a student of Christine Iacobuzio-Donahue, contributed valuable data and insights about the process of spatial dissemination. We realized that, under certain assumptions about the type of metastatic seeding which had occurred, we could infer not only the phylogeny but also the seeding graph, and developed a heuristic for assessing when these assumptions were violated. Later tools would relax these assumptions in a more principled way to handle cases in which more mixing is expected (e.g. Ref. [81]).

The problem posed to me by Hannes was to design a way of combining the data in a statistical model so that, after applying his algorithmic framework, the resulting phylogenetic tree had something like a maximum likelihood interpretation. As such, I designed the Bayesian inference section of this work and integrated it into the algorithmic approach. I am especially grateful to Hannes for his guidance in the project and the example he set for tool development.

This work was first published in Ref. [218]:

Reiter, J. G., Makohon-Moore, A. P., Gerold, J. M., Bozic, I., Chatterjee, K., Iacobuzio-Donahue, C. A., Vogelstein, B., and Nowak, M. A. (2017). Reconstructing metastatic seeding patterns of human cancers. *Nature Communications*, 8, 14114.

Supplemental materials can be found online at DOI 10.1038/ncomms14114

1.2 Abstract

Reconstructing the evolutionary history of metastases is critical for understanding their basic biological principles and has profound clinical implications. Genome-wide sequencing data has enabled modern phylogenomic methods to accurately dissect subclones and their phylogenies from noisy and impure bulk tumor samples at unprecedented depth. However, existing methods are not

designed to infer metastatic seeding patterns. Here we develop a tool, called Treeomics, to reconstruct the phylogeny of metastases and map subclones to their anatomic locations. Treeomics infers comprehensive seeding patterns for pancreatic, ovarian, and prostate cancers. Moreover, Treeomics correctly disambiguates true seeding patterns from sequencing artifacts; 7% of variants were misclassified by conventional statistical methods. These artifacts can skew phylogenies by creating illusory tumor heterogeneity among distinct samples. In silico benchmarking on simulated tumor phylogenies across a wide range of sample purities (15-95%) and sequencing depths (25-800x) demonstrates the accuracy of Treeomics compared to existing methods.

1.3 Introduction

Genetic evolution underlies our current understanding of cancer[196,263,173] and the development of resistance to therapies[75,37]. The principles governing this evolution are still an active area of research, particularly for metastasis[181,167,259], the final biological stage of cancer that is responsible for the vast majority of deaths from the disease. Although many insights into the nature of metastasis have emerged[251], we do not yet know how malignant tumors evolve the potential to metastasize, nor do we know the fraction of primary tumor cells that have the potential to give rise to metastases. Moreover, the temporal, spatial and evolutionary rules governing the seeding of metastases at spatially distinct sites distant from the primary tumor have mostly remained undetermined[181,171,112].

In order to better understand the evolutionary process of cancer, researchers have reconstructed the temporal evolution of patients' cancers from genome sequencing data[44,279,45,92,240]. Thus far, phylogenomic analysis has largely focused on the subclonal composition and branching patterns of primary tumors[61,68,285]. The evolutionary relationships among metastases are equally important but have less often been determined for several reasons[15,100,38,172]. First, comprehensive data sets of samples from spatially-distinct metastases in different organs are rarely available. Second, most ad-

vanced cancer samples are derived from patients who have been treated with toxic and mutagenic chemotherapies, imposing a variety of unknown constraints on genetic evolution, metastatic progression and its interpretation. Third, tumors are composed of varying proportions of neoplastic and non-neoplastic cells, and inferring meaningful evolutionary patterns from such impure samples is challenging[19,258]. Fourth, chromosome-level changes, including losses, are frequently observed in cancers, and previously acquired variants can be lost[172] (i.e., some variants are not "persistent"). Fifth, even when performed at high depth, next generation sequencing coverage is always non uniform, resulting in different amounts of uncertainty at different loci within the same DNA sample as well as among different samples at the same locus. Finally, evolutionarily informative genetic differences among the founding cells of distant metastases tend to be rare[139,161] and therefore the confidence in the inferred metastatic seeding pattern is often low.

The variety of methods that have recently been used to infer evolutionary relationships among tumors underscore these complicating factors and the need for a robust phylogenomic approach. The methods include those based on genetic distance[15,180], maximum parsimony[285,38,91], clonal ordering[173,92] and variant allele frequency[185,141,281]. Modern phylogenomic methods classify variants based on the observed variant allele frequencies (VAFs), account for varying ploidy and neoplastic cell content, and reconstruct comprehensive phylogenies[250,229,174,73,207,80,186,282,163]. In this study, however, as we will show below, in the case of reconstructing the evolution of metastases, these methods suffer from the low number of informative variants and may fail to identify the subclones that gave rise to the observed seeding patterns. Classical phylogenetics assumes that the individual traits are known with certainty[19]. Consequently, these methods struggle with noisy high-throughput DNA sequencing data and do not exploit the full potential of these data due to the error prone binary present/absent classification of variants. Furthermore, many of the methods used for inferring cancer evolutionary trees are based on those designed for more complex evolutionary processes involving sex and recombination[112]. The key conceptual difference between the new approach used here

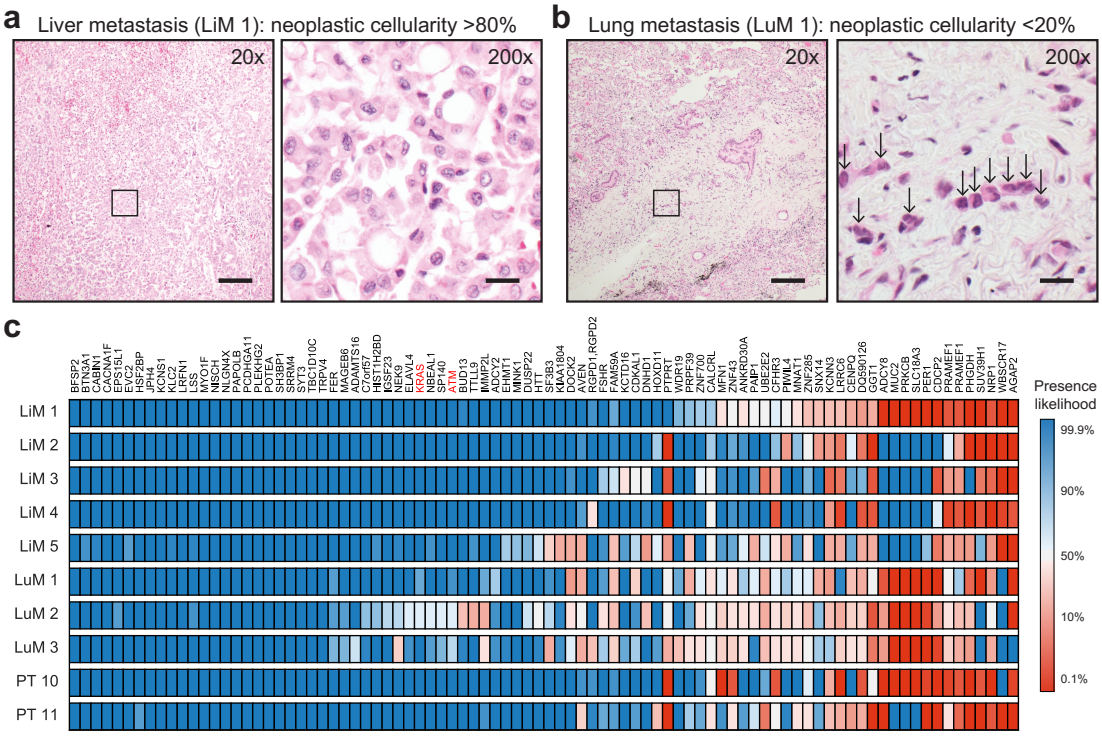


Figure 1.1: a, b | Histology at low (40x20x, scale bar 200 μm) and high (400x200x, scale bar 20 μm) power of liver metastasis LiM 1 and lung metastasis LuM 1, with estimates of neoplastic cellularity determined by pathological review. Arrows highlight the few cancer cells in LuM 1. c | Heatmap depicting the posterior probability (p) that a variant is considered as present in deep targeted sequencing data. Top five rows show samples from five distinct liver metastases (LiM 1-5); the following three rows show samples from three distinct lung metastases (LuM 1-3); the bottom two rows show different parts of the primary tumor (PT 10 11). Dark blue corresponds to a variant being present with probability >99.9% and dark red corresponds to being absent with probability >99.9%. In some samples the mutation status for the most likely clonal driver mutations in ATM and KRAS is unknown.

("Treeomics") and previous ones is that Treeomics reconstructs metastatic seeding patterns and infers the ancestral subclones that seeded metastases at various anatomic locations. Treeomics utilizes multiple samples from spatially distinct sites and assumes mostly monophyletic samples (i.e., monoclonal seeding; polyclonal seeding and reseeding of metastases only happens occasionally[259]).

1.4 Results

To illustrate our approach, we first focused on the data of a treatment-naïve pancreatic cancer patient Pam03[161] (Figure 1.1). WGS (whole-genome sequencing; coverage: median 51x, mean 56x) as well as deep targeted sequencing (coverage: median 296x, mean 644x) was performed on ten spatially-distinct samples: two from the primary tumor and eight from distinct liver and lung metastases (Methods and ref[161]). Estimated purities ranged from 21% to 48% per sample (Supplementary Figure 1), typical for low cellularity cancers (Figure 1.1). Founder variants (clonal in all samples) and unique variants (present in exactly one sample) are parsimony-uninformative in the sense that they do not provide any information about common ancestors of spatially-distinct samples (except the founding clone) and hence do not resolve metastatic seeding patterns. Nonetheless, unique variants can provide information about the subclonal composition and phylogeny within a sample. Parsimony informative variants (variants present in some but not in all samples) exhibited contradicting mutation patterns when we tried to reconstruct a phylogeny consistent with the evolutionary processes underlying tumor progression using conventional methods. Identifying the evolutionarily compatible variants is known as the "binary maximum compatibility problem" and has been widely studied for decades[66,28,84,179,231,103]. A strict binary present/absent classification can be very problematic due to the above described reasons. For example, likely clonal variants in the driver genes ATM and KRAS would be classified as absent in sample LuM 2 because both were sequenced only fourteen times and were mutated only once (Figure 1.1c; Supplementary Data 1). We developed a Bayesian inference model to determine the posterior probability of whether a variant was or was not found in each sequenced lesion rather than rely on a binary input ("present" or "absent"; Figure 1.1c; Methods). This generalization, formalized as a Mixed Integer Linear Program[182] (MILP), enabled us to simultaneously predict sequencing artifacts and infer phylogenies in a remarkably robust fashion.

Two clonal variants are evolutionary compatible if there exists an evolutionary tree where each

variant is only acquired once and never lost. This condition is known as the perfect (the same variant is not independently acquired twice; infinite sites model[155]) and persistent (acquired variants are not lost; no back mutation) phylogeny assumption – the basic principle of modern tumor phylogeny reconstruction methods[229,174,73,207,80]. In our case the mutation pattern of a variant is given by the set of samples where the variant is present (Supplementary Figure 2). Therefore, two somatic variants α and β are evolutionarily incompatible if and only if samples with the following three patterns exist: (i) variant α is absent and β is present, (ii) α is present and β is absent, and (iii) both variants are present. Because somatic variants are by definition absent in the germline, α and β are evolutionarily incompatible and no perfect and persistent phylogeny can explain these data (Supplementary Figure 2). As expected, based on conventional binary present/absent classification of variants, a perfect and persistent tree consistent with the observed (noisy) data of Pam03 cannot be inferred. We show that such a phylogeny indeed exists but that it is hidden behind misleading artifacts, mostly resulting from insufficient coverage or low neoplastic cell content.

1.4.1 Identifying evolutionarily compatible mutation patterns

To account for inconclusive data, we utilize a Bayesian inference model to calculate the probability that a variant is present in a sample (Figure 1.1c; Methods). Using these probabilities for each individual variant, we calculated reliability scores combining the evidence for each possible mutation pattern across all variants and samples. We constructed an evolutionary conflict graph where the nodes were determined through analysis of all mutation patterns. Each node was assigned a weight provided by the calculated reliability scores (Supplementary Figure 3). If two nodes (mutation patterns) were evolutionarily incompatible, an edge between the corresponding nodes was added. We aimed to identify the set of nodes that maximized the sum of the weights (reliability scores) when no pair of nodes was evolutionarily incompatible. This maximal set represents the most reliable and evolutionarily compatible mutation patterns (Supplementary Methods). To evaluate the confidence

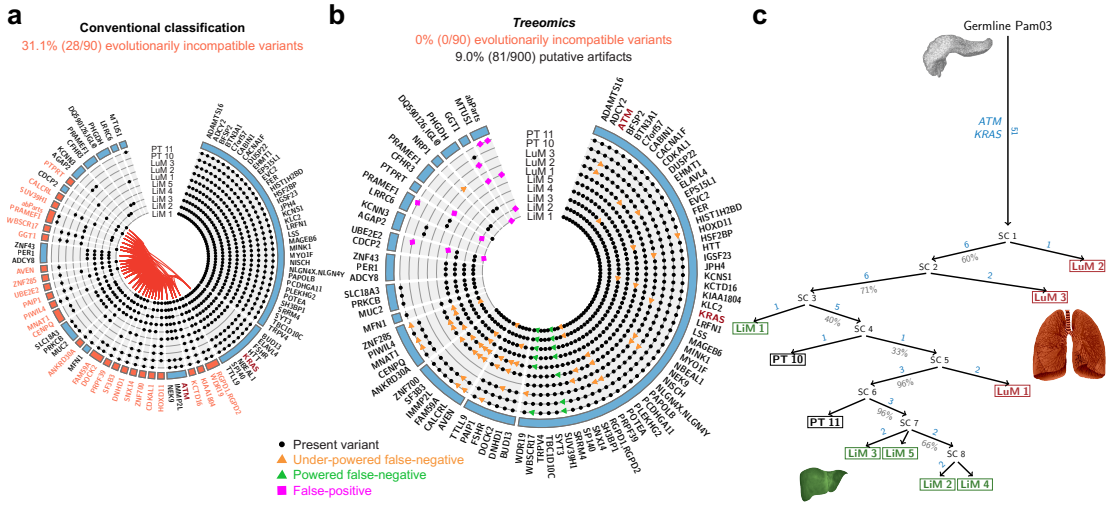


Figure 1.2: a, b | Variants shown in Figure 1.1c are organized as evolutionarily-defined groups ("nodes"). Blue colored nodes are evolutionarily compatible and red colored nodes are evolutionarily incompatible. Based on conventional present/absent classification, 31.1% of the variants were evolutionarily incompatible (a). The incompatibilities are demarcated by red lines ("edges") in the center of the circle that connect each pair of incompatible nodes. Based on a Bayesian inference model and an Integer Linear Program, Treeomics identified the most likely evolutionarily compatible mutation pattern for each variant (b; Methods). This method predicted that 9% (81/900) of the variants across all samples were misclassified and thereby caused the evolutionary incompatibilities shown in panel a. 75% of the predicted artifacts were validated in the WGS data, among those were driver mutations in ATM and KRAS. c | Reconstructed phylogeny from the identified evolutionarily compatible mutation patterns in panel b. Gray percentages indicate bootstrapping values from 1,000 samples. SC indicate predicted subclones. Lung metastases (LuM 1-3) are depicted in red; Liver metastases (LiM 1-5) are depicted in green; Primary tumor samples (PT 10-11) are depicted in black.

in the identified evolutionarily compatible mutation patterns, we performed bootstrapping on the given variants.

1.4.2 Predicting putative artifacts in sequencing data

The solution obtained with the MILP directly provided the most likely evolutionarily compatible mutation pattern for each variant. By comparing our inferred classifications to conventional binary classifications, Treeomics predicted putative sequencing artifacts in the data (Figure 1.2a,b). The conventional classifications differed in 9.0% of the variants in Pam03 (81 putative artifacts from 90

variants across 10 samples; Figure 1.2b). As expected, the majority (68) of the differences were caused by putative false negatives in the binary classification that were inferred to be present by Treeomics. Fifty-five of these putative false-negatives had relatively low coverage (mean: 21), explaining how they could easily be misclassified as absent given the low neoplastic cell content in the samples. Accordingly, many of these under-powered false-negatives occurred in samples with the lowest coverage (liver metastasis LiM 5, lung metastases LuM 2-3) or lowest neoplastic cell content (LuM 1; Supplementary Figure 1). In LuM 2, the driver gene mutation KRAS was incorrectly classified as absent by conventional means though it is most likely a clonal founding mutation and was present at a VAF of 19% in the original WGS sample (Supplementary Table 1). Similarly, the driver gene mutation ATM was incorrectly classified as absent in two samples (VAF 18% and 19% in the WGS data). Although manual review of these samples revealed mutant reads in KRAS, it is not scalable to manually review every putative variant detected by next generation sequencing. Some variants contained false negatives across many samples, indicating that these variants were generally difficult to call. Remarkably, 89% (49/55) of the predicted under powered false-negatives were either significantly present in the WGS data (38/49; mostly at higher coverage than in the targeted sequencing data), or the genomic region of the variant possessed a low alignability score[72] (28/49; Supplementary Table 1).

For two variants sequenced at high depth, Treeomics predicted 13 putative false negatives. The WGS data confirmed sequencing artifacts in these two variants but indicated that 4 likely false positives (all absent in the WGS data) induced Treeomics to predict 13 false negatives rather than 4 false positives (Supplementary Table 2). Of the 13 putative false positives (pink squares in Figure 1.2b), 92% (12/13) were classified as absent in the original WGS data and their mean VAF was 2.3% (Supplementary Table 3). In total, 75% (49 putative false-negatives + 12 putative false-positives; 61/81) of the predicted artifacts were successfully validated. Hence, we verified that at least 7% (61/900) of the variants were misclassified by conventional binary classification. If a phylogenomic method does not

account for sequencing artifacts, the mutation patterns of a large fraction of variants will often be inconsistent with any inferred evolutionary tree. In Pam03, the mutation patterns of 31.1% (28/90) of the variants would be evolutionarily incompatible (Figure 1.2a). These putative artifacts may also help to explain the observed high tumor heterogeneity in earlier studies and the recently reported intratumor similarity when sequencing depth is increased[285,139,161].

1.4.3 Inferring evolutionary trees

From the identified mutation patterns, Treeomics inferred an evolutionary tree rooted at the germline DNA sequence of the pancreatic cancer patient Pam03 (Figure 1.2c). We found strong support for an evolutionarily related group of geographically distinct lesions: samples LiM 2-5 (liver metastases) and PT 11 (primary tumor). This result suggests that a recent parental clone of PT 11 seeded these liver metastases. We also found the same evolutionary relationship by using the low-coverage WGS data (Supplementary Figure 4). In contrast to the targeted sequencing data, the WGS data indicated that lung metastasis LuM 1 was more closely related to LuM 2 and LuM 3. Though the low neoplastic cell content prevents a definite conclusion about the seeding subclone of LuM 1, the reconstructed phylogeny strongly suggests that the liver metastasis LiM 1 was seeded from a genetically different subclone than all other liver metastases. This diversity in seeding subclones and the origin of distinct metastases was also found in another treatment-naïve pancreatic cancer patient (Pam01) whose data similarly indicated that liver metastases were seeded from genetically distinct subclones (Supplementary Figure 5). The phylogeny of Pam01 suggested that distinct subclones of the primary tumor gave rise to not just different liver metastases but also different lymph node metastases. This observation suggests that spatially and genetically distinct subclones in the primary tumor have the capacity to seed metastases. Moreover, these subclones are not necessarily predisposed to seeding at a particular site. In contrast, the phylogeny of Pam02 revealed that all liver metastases except one (LiM 7 with low median coverage of 27) were very closely related to each other and to various

regions of the primary tumor–indicating recent divergence (Supplementary Figure 6). Pam02's pancreatic cancer might have expanded very rapidly with only 0.5 months from diagnosis to death compared to 7 and 10 months for Pam01 and Pam03. The observed genetic similarity across geographically distinct regions of the primary tumor and seven metastases could indicate high metastatic potential of large parts of the primary tumor leading to this very short survival.

To further validate our approach, we reanalyzed data from high-grade serous ovarian cancers[15]. We were able to reproduce all phylogenetic trees of Bashashati et al.[15] except for Cases 1 and 5 (Supplementary Figure 7 and Figure 1D in[15]; Supplementary Figure 8). For case 5, the authors reported an early divergence of sample 5c while Treeomics suggested a later divergence (Supplementary Figure 7c). Comprehensive analysis of their data (reinterpreted in Supplementary Figure 7a,b) revealed that their tree either required that several variants (including two driver gene mutations and multiple indels) occurred independently twice or that two mutations in the driver genes ABL1 and MDM4 were lost. Both possibilities seem unlikely (Supplementary Figure 7 and Figure 1D in[15]); this discrepancy was also identified by Popic et al.[207]. Treeomics did not require these implausible scenarios to construct an otherwise similar tree. Distance based methods can be compromised by large differences in the number of acquired mutations among samples; sample 5c had twice as many mutations than all other samples. For case 1, Treeomics reported rather low bootstrap values and Popic et al. inferred yet another phylogeny such that no definitive conclusion could be obtained. This disagreement across methods highlights the importance of a confidence measure for the inferred branches as otherwise phylogenies are difficult to interpret in a conclusive fashion.

If multiple subclones with spatially distinct evolutionary histories (i.e, polyphyletic samples due to polyclonal seeding or reseeding of a metastasis) were present in the same sample at detectable frequencies, conventional phylogenetic approaches would be unable to separate their evolutionary trajectories. In these scenarios, evolutionarily incompatible mutation patterns with high reliability scores were utilized to detect these subclones and to infer separate evolutionary histories (Supple-

mentary Figure 9a; Methods). For the prostate cancer data of case 6 [61] (Supplementary Figure 9), Treeomics identified subclonal structures and separated their evolutionary trajectories without requiring high purity samples or deep sequencing data.

1.4.4 In silico benchmarking demonstrates high accuracy

We implemented a stochastic continuous-time multi-type branching process to imitate the genetics of distinct metastases seeded according to an evolving cancer [102,5] (Figure 1.3; Methods). We investigated a total of 90,000 independently simulated phylogenies comprised of 180 different combinations of sample purity, mean sequencing depth, point mutation rate, chromosome level changes and mono- and polyphyletic metastases. Based on the simulated ground truth data, we compared the performance of Treeomics with conventional phylogenetic methods (Maximum Parsimony and Neighbor Joining) and modern phylogenomic methods (LICHeE [207] and PhyloWGS [73]) across sample purities of 15% to 95% and sequencing depths of 25 to 800 (Figure 1.3c) representing the range of common sequencing data. A comparison of the mean branching error demonstrates that phylogenies reconstructed from low coverage WES data or from samples with very low neoplastic cell content exhibit high error rates independent of the used method. For mean coverages of 100 and above, the error rates drop dramatically and phylogenies can be accurately reconstructed (Figure 1.3c, Supplementary Figure 10).

Current subclone inference algorithms do not directly reconstruct phylogenies of distinct sites as Treeomics does but infer joint phylogenies of variants, which are sometimes simultaneously grouped into subclones [73,207,80,186,282]. To enable a comparison of these slightly different methodologies, we developed a mutation matrix error score (similar as in [207]) that checks (i) if variants of the same subclone were indeed assigned to the same subclone and (ii) if the ancestral relationship among variants was correctly determined (Methods). For example, in the simulated phylogeny illustrated in Figure 1.3a, the tested tools had to correctly assign the acquired variants to the founding subclone (PT

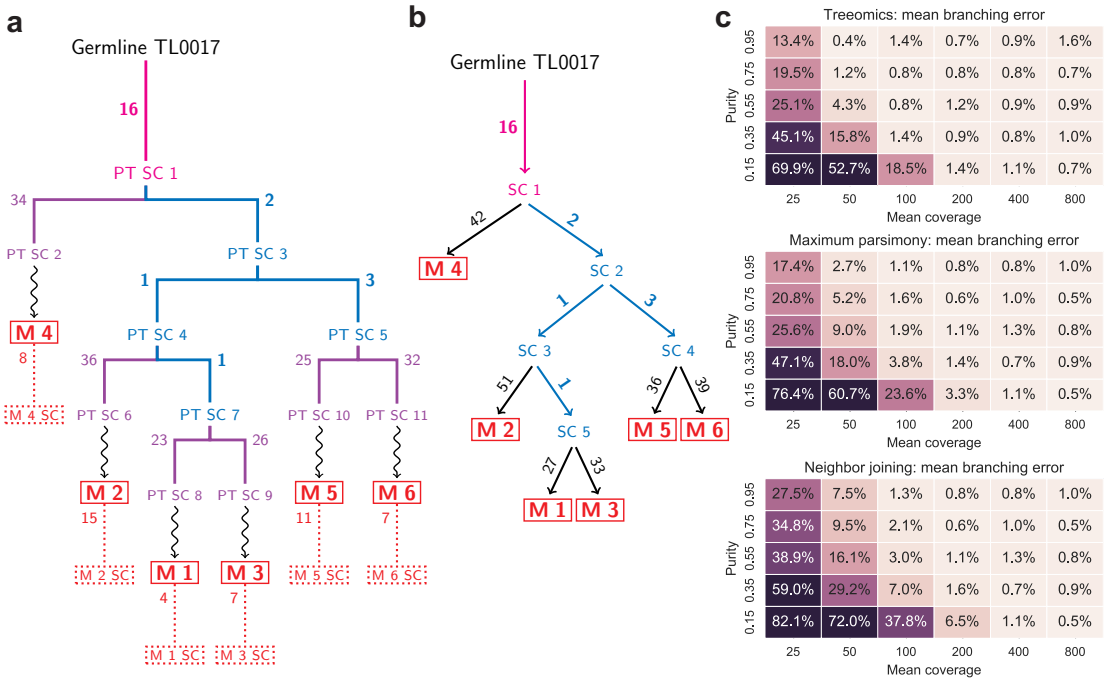


Figure 1.3: a | Simulated metastatic progression according to a stochastic branching process [102,5]. Metastases (M 1-6) are numbered in chronological order of their seeding. Purple and blue lines indicate evolution among lineages within the primary tumor. Pink numbers correspond to the founding variants present in all cancer cells and blue numbers correspond to the parsimony informative variants. Numbers in red denote subclonal variants acquired after the seeding of the metastasis. SC indicates subclone. Dotted boxes illustrate biopsies. b | Treeomics correctly reconstructed the simulated phylogeny in panel a by identifying the parsimony informative variants (blue). Private mutations (purple numbers in panel a) acquired in the primary tumor are indistinguishable from subsequently acquired mutations (red numbers in panel a). c | Benchmarking across 15,000 simulated phylogenies with six monoclonal monophyletic metastases depicting the mean branching error conditioned on at least one variant per branch. Phylogenies reconstructed from low coverage WES data or from samples with very low neoplastic cell content exhibited high error rates independent of the used method. Necessary binary present/absent classification for maximum parsimony and neighbor joining was based on Treeomics' Bayesian inference model (variant was present if p>50%).

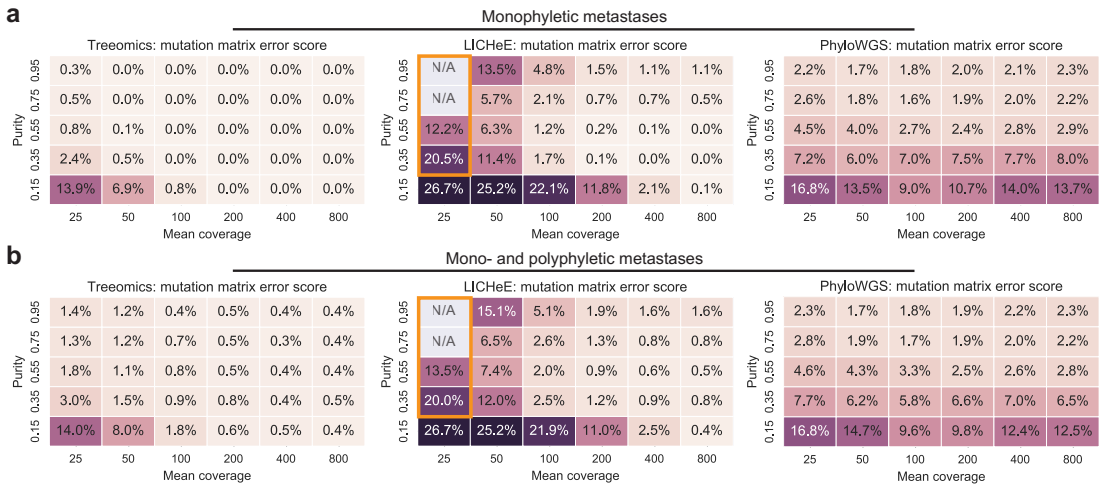


Figure 1.4: a | Benchmarking across 15,000 simulated phylogenies with six monophyletic metastases (no reseeding). Treeomics greatly outperformed LICHeE in all considered scenarios. In the orange framed scenarios, LICHeE was unable to infer a valid solution for the majority of cases. PhyloWGS exhibited mean error scores more than 10-fold higher than those of Treeomics in most considered scenarios. b | Benchmarking across 15,000 simulated phylogenies with three monophyletic and three polyphyletic metastases imitating patients with reseeded metastases21,23,53. Treeomics exhibited the lowest mean error score across all scenarios. The performance of PhyloWGS did not significantly change compared to monophyletic metastases (possibly due to the advantageous input). The error scores of Treeomics and LICHeE slightly increased.

SC 1) and the parsimony informative subclones (PT SC 3-5, 7). Since the runtime of PhyloWGS increases significantly with the number of variants, we removed all private variants in the input for PhyloWGS (purple and red variants in Figure 1.3a). Treeomics and LICHeE were provided with all detected variants and therefore had to distinguish between parsimony informative variants and private variants as well as sequencing artifacts. All tools accurately identified ancestral subclones and their variants for mean coverages above 200 and a neoplastic cell content greater than 35% (Figure 1.4a). Treeomics outperformed LICHeE and PhyloWGS in all considered scenarios (Figure 1.4a). In the majority of scenarios, the error score of PhyloWGS was more than 10-fold higher than the error score of Treeomics. For mean coverages below 50, the error score of LICHeE increased notably while PhyloWGS was mostly struggling with low neoplastic cell content (<35%).

In the case of reseeded metastases[100,172,157] leading to multiple evolutionary trajectories and therefore polyphyletic lesions, the error score of Treeomics and LICHeE slightly increased while the performance of PhyloWGS did not change significantly (possibly due to the advantageous input; Figure 1.4b). Treeomics exhibited the lowest error score across methods in all scenarios. Interestingly both Treeomics and LICHeE performed best in the case of high sequencing depth but low or medium purity – suggesting that there is further room for improvement (Figure 1.4b). We hypothesize that the higher purity leads to more detected private variants and hence to more potential sequencing artifacts. In the case of an elevated point mutation rate (e.g. due to mismatch repair deficiency) or highly chromosomally unstable cancers[55], Treeomics continued to have the lowest mutation matrix error score in 119 of 120 considered scenarios (Supplementary Figs. 11, 12). The runtime of PhyloWGS was around 5-8 hours per simulated phylogeny (in total 300,000 core computing hours; elevated mutation rate could not be evaluated due to the high runtime), while LICHeE needed on average a few minutes (4,000 hours) and Treeomics less than a minute per case (in total 800 core computing hours).

1.5 Discussion

The new approach described here efficiently reconstructs the evolutionary history, detects potential artifacts in noisy sequencing data, and finds the ancestral subclones giving rise to the distinct metastases. The evolutionary theory of asexually evolving populations combined with Bayesian inference and Integer Linear Programming enabled us to infer detailed phylogenomic trees with significantly fewer errors than existing methods (Figure 1.3, Figure 1.4, Supplementary Figs. 10-12). In contrast to other tools, Treeomics accounts for putative artifacts in sequencing data and can thereby infer the branches where somatic variants were acquired as well as where some may have been lost during evolution, presumably through losses of heterozygosity resulting from chromosomal instability[172,145]. The branching in the inferred trees shed new light on the origin and the seeding patterns of particular metastatic lesions[181,112]. For example, in contrast to colon cancer, where liver metastases are assumed to seed lung metastases[261], our results suggest that this may not be the case in pancreatic cancer. The reconstructed phylogenies also indicate that distinct subclones in the primary tumor were equally capable to seed metastases in the same and in different organs (Supplementary Figure 5). However, we did not find any evidence for polyphyletic metastases which confirms findings in a mouse model of pancreatic cancer where the large majority of lung and liver metastases were monophyletic[157]. The evolutionary rules of natural metastatic cancers leading to the highly non-random pattern of metastases in Pam03 are just beginning to emerge.

Despite these detailed reconstructed phylogenies, there are several limitations that should not be neglected. A low mutation matrix error score does not directly imply correctly reconstructed seeding patterns and vice versa (compare Figure 1.3c and Figure 1.4a). A method can exhibit low mutation matrix error scores while exhibiting high branching errors and vice versa. Moreover, without additional data, even correctly inferred cancer phylogenies do neither directly provide information about the temporal ordering in which metastases were seeded nor about the anatomic location of the seed-

ing subclones. For example, metastasis M4 diverged first in the simulated phylogeny but was seeded rather late (Figure 1.3a). Furthermore, a single seeding event cannot be distinguished from multiple seeding events from the topology of the reconstructed tree alone (see[112]). Only sufficient sampling of all sites can provide evidence about the location of the seeding subclone and the likely timing of the seeding event. For example, the genetic similarity of the primary tumor sample PT 11 and the liver metastases LiM 2-5 suggests multiple seeding events from a recent ancestor of PT 11. Future phylogenomic approaches could incorporate estimated growth rates and mutation rates to better quantify the probability of metastasis to metastasis spread.

We have designed Treeomics from first principles to directly handle ambiguity in high-throughput sequencing data, including samples with low neoplastic cell content or coverage. The mutation patterns and their evolutionary conflict graph form a robust data structure and consequently the painful task of semi-automatic filtering becomes unnecessary. As a result of the Bayesian confidence estimates for the individual variants, this method can infer more robust results than traditional phylogenetic methods, which employ a binary representation of sequencing data (Figure 1.2a). Furthermore, as shown above, distance-based methods can produce results inconsistent with the evolutionary theory of cancer as they often ignore knowledge of biological phenomena specific to neoplasia (Supplementary Figure 7). We note that PhyloWGS, LICHeE and other subclone inference methods have not been designed to reconstruct phylogenies based on these few genetic variants that determine the evolutionary history of metastases. The key difference between these approaches is that Treeomics assumes that mixing of subclones from two spatially distinct sites and hence polyphyletic samples are rare[172,139,157]. Treeomics therefore works extremely well among metastases but is not applicable for liquid cancers. On the contrary, tools like PhyloWGS work extremely well in liquid cancers. Last, we compared our results to AncesTree[80], which roughly identified the evolutionarily related samples in Pam03 but excluded 70% (63/90) of the variants (among them the driver gene mutations in KRAS and ATM) in the inferred phylogeny due to evolutionary incompatibilities

(Supplementary Figure 13).

At present, Treeomics only employs nucleotide substitutions and short insertions and deletions – a subset of the available information. The benchmarking results demonstrate that a single mutation varying in two samples is typically sufficient for Treeomics to infer the correct evolutionary history (Figure 1.3a,b); a crucial property given the high genetic similarity of metastases[139,161]. Other types of data, such as copy number alterations, structural variations and DNA methylation, could be incorporated into Treeomics to further improve the accuracy of the inferred results.

1.6 Methods

1.6.1 DNA sequencing design and validation

Sequencing data were generated in two stages (see[161]). First, genomic DNA from 26 tumor samples of three subjects (20 metastases and 6 primary tumor sections) was evaluated by 60x whole genome sequencing (WGS) using an Illumina Hi-Seq 2000 (see Figure 1.1, Supplementary Figs. 5, 6 for anatomic locations of the individual samples). Importantly, genomic DNA from the normal tissue of each patient was used to facilitate identification of somatic variants. We obtained an average coverage of 69x with 97.5% of bases covered at >10x, revealing a total of 127,597 putative coding and noncoding somatic mutations, (average of 4,908 per sample). To limit the artifacts generated by WGS and alignment, we filtered the putative variants using several quality parameters, including read directionality, mutant allele frequency detected in the normal, known human SNPs, and the number of independent tags at each site. This analysis, combined with manual inspection of the raw data, yielded a total of 2,105 potential mutations for subsequent validation.

Second, we utilized a targeted sequencing approach to independently screen every mutation that we observed to be of high quality in at least one WGS tumor sample. Briefly, probes for capture were designed to flank each potential mutant base (2,105) and libraries were prepared for the origi-

nal 26 WGS samples of the three subjects. Using an Illumina chip-based approach, we successfully aligned, processed, and validated 381 mutations (range 106-164 per patient) at an average sequencing depth of 731x (Supplementary Data 1-3). In addition to the increased coverage and sensitivity of targeted sequencing, both sequencing approaches generated independent datasets in which we could directly compare putative variants in silico among many tumors within a patient. Additional details regarding patient selection, processing of tissue samples and DNA extraction and quantification can be found in[161].

1.6.2 Bayesian inference model

To compute reliability scores for each mutation pattern, we extract posterior probabilities for the presence and absence of a variant in a sample from a Bayesian binomial likelihood model of error-prone sequencing. If f is the true fraction of variant reads in the sample, π is our prior belief about f, and e is the sequencing error rate, the posterior distribution P of f given N total reads and K variant reads is

$$P(f|N,K) = \frac{1}{Z}\binom{N}{K}\left(f(1-e)+(1-f)e\right)^{K}\left(fe+(1-f)(1-e)\right)^{N-K}\pi(f) \tag{1.1}$$

where Z is a normalizing constant (see Supplementary InformationMethods). A priori, the variant allele frequency in a sample is exactly zero ($f = 0$) with some positive probability c_0. The prior π is then of the following form

$$\pi(f) = c_0\delta(f) + (1-c_0)g(f) \tag{1.2}$$

where $\delta(f)$ denotes the Dirac delta function and $g(f)$ denotes a prior given the variant is present. We use a sample-specific prior function to account for the by multiple fold varying neoplastic cell content across samples (Supplementary InformationMethods; Supplementary Figure 2). The posterior probability that a variant is absent in a sample with low neoplastic cell content will be lower than in

a sample with high neoplastic cell content despite the same K and N (Supplementary Information-Methods). The posterior probability that a variant is absent, denoted by q, and the probability that a variant is present, denoted by p, are

$$q = P(f \leq f_{absent}\gamma_s|N, K), p = 1 - q \tag{1.3}$$

where γ_s is the estimated neoplastic cell content in sample s and f_{absent} is the maximal frequency threshold for an absent SNV (Supplementary InformationMethods). A variety of more sophisticated variant detection algorithms can be used here as long as the output can be converted to posterior probabilities of presence and absence. We obtained robust results across all investigated scenarios with the frequency threshold of $f_{absent} = 0.05$, however other thresholds can be used. We calculate the probability of each mutation pattern for a particular variant by multiplying the corresponding posterior probabilities for each sample. Each mutation pattern has some positive probability, but those supported by the data are given much more weight. A mutation pattern ν is denoted as a binary vector of length $|S|$ (total number of samples) where ν_s is 1 if the variant is present in sample s and 0 if absent. The likelihood $L_\mu(\nu)$ that a variant μ exhibits pattern ν is

$$L_\mu(\nu) = \prod_{s \in S} p_{\mu,s}^{\nu_s} q_{\mu,s}^{1-\nu_s} \tag{1.4}$$

If the presence or absence of a variant in some samples is uncertain, the likelihood of any individual mutation pattern will generally be lower. The reliability score ω_ν of each mutation pattern ν (corresponding to a node in the evolutionary conflict graph; Supplementary Figure 3) is given by

$$\omega_\nu = \frac{-log(\prod_\mu \left(1 - L_\mu(\nu)\right)}{m} \tag{1.5}$$

Assuming mutations are independent across each other and across samples, the argument of the log-

arithm denotes the likelihood that no mutation has pattern ν and hence leverages the full sequencing information from all variants. With these scores (weights) normalized by the number of considered variants m, the minimum weight vertex cover of the evolutionary conflict graph corresponds to identifying the most reliable and evolutionarily compatible mutation patterns (see Supplementary Information Methods for further details).

1.6.3 Identifying evolutionarily compatible mutation patterns

Given the calculated reliability scores, we efficiently find the most reliable and evolutionarily compatible mutation pattern for all variants via solving a Mixed Integer Linear Program[182] (MILP). In the Supplementary Information we prove that finding these mutation patterns is equivalent to solving the Minimum Vertex Cover problem; one of Karp's original 21 NP-complete problems[66,124]. In the Minimum Vertex Cover problem one wants to find the minimum set of nodes in an undirected graph such that each edge in the graph is adjacent to one of the nodes in the minimum set. Therefore, by definition all edges are covered by the nodes in the minimum set. Similarly, we try to find the weighted set of nodes (here mutation patterns) with the minimal sum of reliability scores such that no evolutionary incompatibilities in the conflict graph remain. After this minimal set of nodes and their adjacent edges have been removed from the graph, we can easily infer an evolutionary tree since evolutionary conflicts no longer exist among the remaining nodes (i.e., all edges were covered and removed with the minimal set). The remaining set of mutation patterns is by definition the maximal set of evolutionarily compatible patterns (Supplementary Methods).

In the evolutionary conflict graph $G = (V, E)$, each node $i \in V$ represents a different mutation pattern. For n samples, the number of nodes $|V|$ is given by 2^n. For each pair of evolutionarily incompatible mutation patterns i and j, there exists an edge $(i, j) \in E$. The weight (c_i) of each node i is given by the reliability scores ω_i described in the Bayesian inference model section (Supplementary Figure 3).

The MILP to find the minimal-weighted set of evolutionarily incompatible mutation patterns is defined by the following objective function and constraints:

$$\begin{aligned} &\text{objective function} && \text{minimize } \textstyle\sum_{i \in V} c_i x_i \\ &\text{constraints} && \text{subject to } x_i + x_j \geq 1 \quad \text{for all } (i,j) \in E \\ & && x_i \in \{0,1\},\, c_i > 0 \qquad \text{for all } i \in V \end{aligned}$$

This formulation guarantees that the MILP solver finds the minimal value of the objective function such that all constraints are met and hence the nodes in the selected set cover all edges. The evolutionarily compatible and most reliable mutation patterns $\{i | x_i = 0\}$ are given by the complement set of the optimal solution $\{i | x_i = 1\}$ to the MILP.

Day and Sankoff showed that inferring the most likely evolutionary trajectories is a computationally challenging problem (NP-complete[66]). Sophisticated approximation algorithms have been developed in the context of language and cancer evolution[28,179,231]. However, medium-sized instances of NP-complete problems are no longer intractable due to the enormous engineering and research effort that has been devoted to ILP solvers. The MILP[182] formulation enables an efficient and robust analysis of large datasets. We prove that an approximation algorithm that would guarantee that its solution is at most 36.06% worse than the optimal solution cannot exist unless the complexity class P=NP (Supplementary Methods, Theorem 1). Salari et al.[231] explored a related approach but approximated two NP complete problems, possibly leading to suboptimal results. Treeomics produces a mathematically guaranteed to be optimal result without convergence or termination issues. Note that a mathematical optimal solution is not necessarily equivalent to the biological truth, especially in the case of low neoplastic cell content or coverage (Figure 1.3, Figure 1.4). MILPs may also be useful in other areas of phylogenetic inference where methods with strong biological assumptions (e.g. constant mutation rates or specific substitution profiles) are not applicable or are computationally too expensive to obtain guaranteed optimal solutions.

1.6.4 Inferring evolutionary trees

After the evolutionarily compatible mutation patterns $\{i|x_i = 0\}$ have been identified and variants are assigned to their most likely evolutionarily compatible pattern based on the maximum likelihood weights given by the Bayesian inference model, the derivation of an evolutionary tree is a trivial computational task. In quadratic time ($\mathcal{O}(mn)$) of the input size we construct a unique phylogeny where n is the number of samples and m is the total number of distinct variants[101]. The branches where the individual variants are acquired follow from the inferred tree.

1.6.5 Detecting subclones of distinct origin

Evolutionary incompatible mutation patterns with high reliability scores may indicate mixed subclones with distinct evolutionary trajectories (Supplementary Figure 9). Recall that evolutionary incompatibility requires that the conflicting variants need to be present together in at least one sample. However, even if both variants are mutated in a statistically significant fraction in the same sample, these variants may not be present in the same cells and the evolutionary laws of an asexually evolving population may not be violated. If an evolutionarily incompatible mutation pattern exhibits a reliability score higher than expected from noise, Treeomics utilizes this evidence to infer subclones with distinct evolutionary trajectories and unidirectional spreading. A detailed pseudo code is provided in the Supplementary Methods. Subsets (descendants) and supersets (ancestors) of the conflicting mutation pattern are simultaneously identified and a comprehensive evolutionary tree is inferred. We performed extensive benchmarking of the subclone detection algorithm for various scenarios described in the following section (Figure 1.4, Supplementary Figure 9). Furthermore, we tested the method on sequencing samples from the same prostate. After two subclones were separated in mixed samples from a prostate tumor17, 12643 (out of 12645) variants supported the inferred evolutionary tree (Supplementary Figure 9). The remaining two variants were predicted

to be false positives by Treeomics.

1.6.6 *In silico* benchmarking

To assess the performance of Treeomics, we simulated metastatic progression according to a stochastic multi-type continuous-time branching process[102,10,275,32,215,216] where metastases are seeded independently at random. Cells divide with birth rate $b = 0.16$, die with death rate $d = 0.1555$, and can leave the current site to successfully colonize a new site with probability $q = 10 - 9$[102,88]. When a cell divides, a point mutation is acquired with probability $u = 0.045$ (assuming a point mutation rate of $5 * 10^{-10}$ per basepair and 45 megabases covered by Illumina exome sequencing[121]) and a Copy Number Variant (CNVs) is acquired with a rate of 0.1% per division. The evolutionary process is initiated by a single advanced cancer that already accumulated driver gene mutations. Subsequently accumulated mutations, Single Nucleotide Variants (SNVs) and CNVs, are assumed to be neutral[272,34]. Variants are acquired randomly across all chromosome pairs such that no two copy number events overlap along the same lineage. SNVs and CNVs may overlap, in which case the timing of the events is used to determine the allele fraction of SNVs at the affected locus. CNV length is sampled from the observed length distribution in[23]. After m spatially distinct metastases reached the detection size $M = 10^8$, the simulation is stopped. Note that new metastases can also be seeded from previously seeded metastases.

To model the biopsy and sequencing process, a single sample consisting of one million cells of each of the m metastases consistent to the considered purity (15%, 35%, 55%, 75%, 95%) is subject to in silico sequencing. Metastases with a mixture of ancestries (polyphyletic samples) are simulated by random sampling from two distinct sites proportional to the tumor sizes at these sites (size of the second site possibly still below the detection limit). Sequencing depth is negative binomially distributed with a given mean (25, 50, 100, 200, 400, 800). A sequencing error rate of $e = 0.5\%$ is assumed. The simulation output is the number of variant and reference "reads" in each metas-

tasis sample for each mutated locus present with a VAF of at least 5% and supported by at least 4 variant reads (2 in the case of a coverage of 25) in any of the sampled metastases. An example for a simulated phylogeny is depicted in Figure 1.3a. Simulated phylogenies are available on github: https://github.com/johannesreiter/treeomics.

We compared Treeomics to standard phylogenetic reconstruction (Maximum Parsimony[239], Neighbor Joining[239]) and modern tumor phylogeny reconstruction methods (LICHeE[207], PhyloWGS[73]). Two different error metrics demonstrate the performance of Treeomics against existing methods: branching error and mutation matrix error score. The branching error quantifies the accuracy of the reconstructed coalescent relationships among distinct sites. From the true coalescent tree among metastatic sites, the collection of coalescent events among the sites is computed and compared to those predicted by the method. The branching error is defined as the fraction of true coalescent events missed by the reconstruction method. Since maximum parsimony and neighbor joining trees do not infer the evolutionary relationships among individual variants, the branching error metric was used to compare these methods (Figure 1.3).

The mutation matrix error score quantifies the accuracy of the reconstructed sequence of mutations acquired during an evolutionary process. For a tumor with k parsimony-informative mutations across m metastases, a k by k matrix A is constructed where $A_{i,j} = 1$ if mutation i is parental to mutation j and 0 otherwise. If two mutations are acquired on the branch in the true phylogeny, the correct evolutionary ordering among this pair of mutations is not required and $A_{i,j} = 0.5$. In PhyloWGS, where many phylogenies are sampled, this reconstructed phylogeny mutation matrix $\hat{A}$ is averaged over all samples. If a tool did not provide any information about a pair of mutations i, j, $\hat{A}_{i,j}$ is set to $A_{i,j} - 0.5$. For the reconstructed matrix A ̂, the normalized error score is computed as $\sum_{i,j} \left(A_{i,j} - \hat{A}_{i,j} \right)^2 / (k^2 - k)$. Because LICHeE and PhyloWGS do not directly infer the coalescent relationship among sites, the mutation matrix error score was used in the benchmarking (Figure 1.4, Supplementary Figs. 11-12). Recall that only founder and parsimony informative mutations were

provided as input to PhyloWGS while LICHeE and Treeomics also had to deal with noisy private mutations. PhyloWGS was run with 2,500 MCMC iterations and 5,000 inner Metropolis-Hastings iterations for a maximum of 15 hours for each individual case. Increasing the number of samples and iterations did not significantly decrease the mutation matrix error score. LICHeE was run with the default parameter values except that we set maxVAFAbsent and minVAFPresent to 0.05 as well as minClusterSize and minProfileSupport to 1. These parameter changes significantly improved the performance of LICHeE in our data set.

1.6.7 Binary present/absent classification

We perform conventional binary present/absent classification of each variant to allow a comparison to the inferred classification used in our new approach. We scored each variant by calculating a p-value in all samples (one-tailed binomial test): $Pr(X > K|H_0, K, N) = 1 - \sum_{i=0}^{K-1} \binom{N}{i} p_{fpr}^{i}(1 - p_{fpr})^{N-i}$ where N denotes the coverage, K denotes the number of variant reads observed at this position, and X denotes the random number of false-positives. As null hypothesis H_0, we assume that the variant is absent. Similar to Gundem et al.[100], we assumed a false-positive rate (p_{fpr}) of 0.5% for the Illumina chip-based targeted deep sequencing. We used the step-up method[21] to control for an average false discovery rate (FDR) of 5% in the combined set of p-values from all samples of a patient. Variants with a rejected null hypothesis were classified as present. The remaining variants were classified as absent.

1.6.8 Code availability

The source code and a manual for Treeomics, as well as multiple examples illustrating its usage, are provided at https://github.com/johannesreiter/treeomics as well as in Supplementary Software. Treeomics v1.5.2 was used for the entire analysis. The tool is implemented in Python 3.4. The inputs

to the tool are the called variants and the corresponding sequencing data, either in tab-separated-values format or as matched tumor-normal VCF files. As output, Treeomics produces a comprehensive HTML report (see github repository) including statistical analysis of the data, a mutation table plot and a list of putative artifacts (false-positives, well-powered and under-powered false negatives). Additionally, Treeomics produces evolutionary trees in LaTeX/TikZ format for high-resolution plots in PDF format. If circos is installed, Treeomics automatically creates the evolutionary conflict graph and adds it to the HTML report. Treeomics also supports various filtering (e.g., minimal sample median coverage, false-positive rate, false-discovery rate) for an extensive analysis of the sequencing data. Detailed instructions for the filtering and analysis are provided in the readme file in the online repository. For solving the MILP, Treeomics makes use of the common CPLEX solver (v12.6) from IBM.

1.6.9 Data availability

Targeted sequencing data of subjects Pam01, Pam02, and Pam03 have been deposited in the github repository in the directory /src/input/Makohon2016 and are also provided in Supplementary Data 1-3. All other relevant data are available within the article and its Supplementary Files or available from the corresponding authors.

1.7 Funding and support

We thank Martin Chmelik, Alison Hill and Adeeti Ullal for valuable discussions. This work was supported by the European Research Council (ERC) start grant 279307: Graph Games (J.G.R., C.K.), Austrian Science Fund (FWF) grant no P23499-N23 (J.G.R., C.K.), FWF NFN grant no S11407-N23 RiSE/SHiNE (J.G.R., C.K.), a Landry Cancer Biology Fellowship (J.M.G.), National Institutes of Health grants CA179991 (C.I.-D., I.B.), F31CA180682 (A.M.-M.), CA43460 (B.V.), the

Lustgarten Foundation for Pancreatic Cancer Research, the The Sol Goldman Center for Pancreatic Cancer Research, the The Virginia and D.K. Ludwig Fund for Cancer Research, the John Templeton Foundation and a grant from B. Wu and Eric Larson. Benchmarking was performed on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

1.8 Author contributions

C.I.-D. and A.M.-M. performed autopsies and experiments; all authors analyzed data; J.G.R., J.M.G., and K.C. performed mathematical analyses; J.G.R. and J.M.G. developed algorithms, performed benchmarking and implemented the tool; all authors contributed to the writing of the manuscript.

1.9 Competing financial interests

The authors declare no competing financial interests.

2

Precancerous neoplastic cells can move through the pancreatic ductal system

2.1 Forward

This work considers the relationship between *matched* precancerous and cancerous lesions in the pancreas. Most people will develop precancerous lesions in the pancreas (and many other organs),

but the vast majority of these lesions will remain benign–at least over the course of a normal lifetime. Informed by the phylogenetic inference tool in Chapter 1, we provide the first characterization of the similarities and differences between these lesions. The work is the product of a longstanding collaboration between Martin Nowak, Christine Iacobuzio-Donahue, and Bert Vogelstein, and the experimental work described here began long before my involvement. I am indebted to Bert and Christine's vision and planning, which made the collection of such remarkable data possible.

Hannes Reiter, Alvin Makohon-Moore, and I collaborated to analyze the data in this study. We were all involved in reviewing data quality and the generation and analysis of phylogenies for each patient. I contributed most heavily to the analysis of mutational signatures in these patients and the inference of event times at the end of the chapter.

This work was first published in Ref.[159]:

Makohon-Moore, A. P.*, Matsukuma, K.*, Zhang, M.*, Reiter, J. G.*, Gerold, J.M.*, Jiao, Y., Sikkema, L., Attiyeh, M. A., Yachida, S., Sandone, C., Hruban, R. H., Klimstra, D. S., Papadopoulos, N., Nowak, M. A., Kinzler, K. W., Vogelstein, B., and Iacobuzio-Donahue, C. A. (2018). Precancerous neoplastic cells can move through the pancreatic ductal system. Nature, 561(7722), 201. (*equal contribution)

Supplemental materials can be found online at DOI 10.1038/s41586-018-0481-8

2.2 Abstract

Most adult carcinomas develop from noninvasive precursor lesions, a progression that is supported by genetic analysis. We analyzed the somatic variants of co-existing pancreatic cancers and precursor lesions sampled from distinct regions of the same pancreas. After inferring evolutionary relationships, we found that the ancestral cell had initiated and clonally expanded to form one or more lesions, and that subsequent driver gene mutations eventually led to an invasive pancreatic cancer. We

estimate that this multi-step progression generally spans many years. These new data reframe the step-wise progression model of pancreatic cancer by illustrating independent, high-grade pancreatic precursor lesions observed in a single pancreata often represent a single neoplasm that has colonized the ductal system, accumulating spatial and genetic divergence over time.

2.3 Introduction

The transformation of a normal cell to invasive cancer occurs through the accumulation of genetic and epigenetic changes[265]. Many invasive carcinomas of adults develop from morphologically recognizable noninvasive precursor lesions[264]. The most common precursor lesion associated with pancreatic ductal adenocarcinoma (PDAC) is pancreatic intraepithelial neoplasia (PanIN)[17]. At the morphologic level, low-grade PanINs (LG-PanIN, PanIN-1 and PanIN-2) have minimal to moderate cytologic atypia and higher-grade PanINs (HG-PanIN, PanIN-3) have severe cytologic atypia. HG-PanINs exhibit morphological features that are thought to facilitate progression to an infiltrating carcinoma[114].

Aspects of this progression are supported by genetic studies[114,262,123], yet fundamental questions about the development of PDAC remain[158]. The majority of PanINs (regardless of grade) harbor KRAS mutations; increasing grade of PanINs and invasive carcinomas are more likely to contain additional driver gene alterations such as those in TP53, CDKN2A, and SMAD4. Moreover, PanINs adjacent to PDACs often share many genetic alterations in both passenger and driver genes[177,113]. Collectively, these observations suggest a subset of PDACs arise from adjacent PanINs, just as a colorectal carcinoma can arise from an underlying adenoma[83]. However, in individuals with multiple anatomically distinct PanINs[168], the biologic and genetic relationships among these lesions and their clinical significance are not fully understood[280]. For instance, cancerization of the pancreatic ducts by an established PDAC recapitulates lesions with histopathologic features that are difficult

to distinguish from those of bona fide PanIN precursor lesions[134]. Further, the importance of non-invasive precursor lesions was recently challenged by a whole genomic sequence analysis of pancreatic cancers which proposed that pancreatic cancer tumorigenesis is neither gradual nor slow[189]. We posited that a genomic evaluation of PDAC and matched co-evolving PanINs would provide additional insights into the biology of pancreatic cancer precursors and the dynamics of step-wise progression.

2.3.1 Evolutionary scenarios

Figure 2.1 presents the conceptual framework underlying the interpretation of sequencing data generated from one PanIN and PDAC in the same patient, outlining three possible scenarios that in theory might be found. In the first scenario, the PanIN and the PDAC do not share any somatic mutations and arose independently. In the second scenario, the PanIN shares a subset of the somatic passenger and driver gene mutations with the PDAC, but the PDAC contains additional driver or passenger gene alterations not present in the PanIN. Scenario 2 presumes that a common ancestral cell underwent initiation and clonal expansion prior to seeding the PanIN and PDAC, but neither the common ancestral cell nor the founding PanIN cell had yet acquired all the genetic events required to generate an invasive neoplasm. In the third scenario, the PanIN and the PDAC share some passenger mutations and all driver gene alterations, and the ancestral cell that seeded both the PDAC and PanIN already acquired all alterations required to form a malignant cancer.

2.4 Results

To investigate the progression patterns of pancreatic carcinogenesis, >100 resected pancreata from over a three-year interval were prospectively screened to identify those samples in which at least one LG-PanIN (PanIN-2) or HG-PanIN (PanIN-3) was present in a region that was anatomically dis-

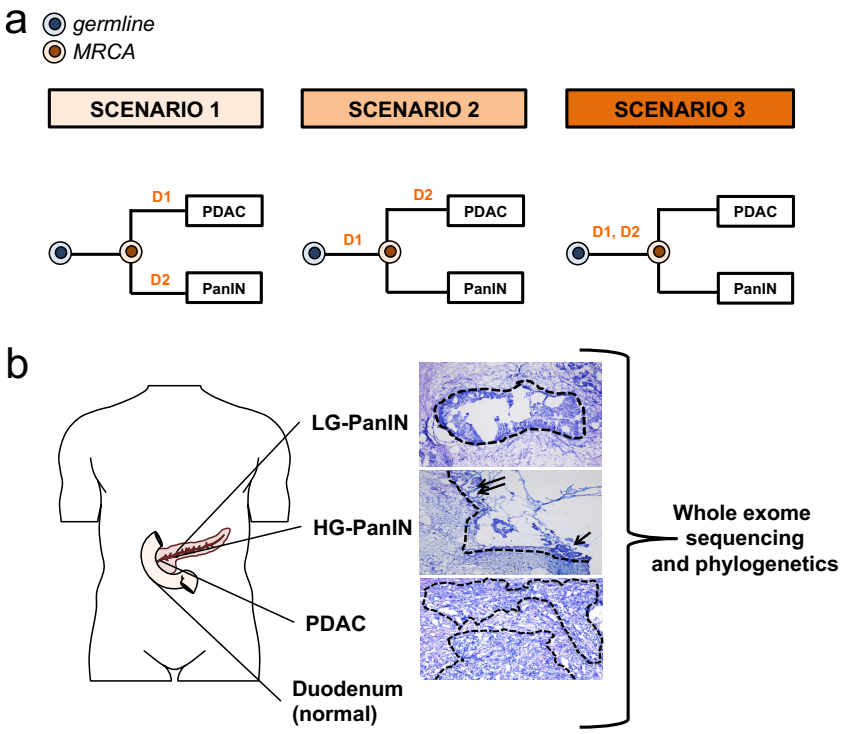


Figure 2.1: a. Evolutionary scenarios of coexistent PanIN(s) and PDAC. For each of the three evolutionary scenarios, D1 and D2 indicate two hypothetical driver gene alterations whereas the colored cells represent the germline (matched normal sample) in blue and the most recent common ancestor (MRCA) in orange for each PanIN/PDAC pair. The primary tumor is labeled "PDAC" while the PanIN is labeled by a letter. In scenario 1, none of the somatic gene alterations are shared by the PanIN and PDAC. Mutation D1 is private to PDAC and mutation D2 is private to the PanIN. In scenario 2, only D1 is shared by the PanIN and PDAC. The mutation in D2 is private to the PDAC. In scenario 3, both D1 and D2 driver gene alterations are shared by the PanIN and PDAC. b. Tissue collection, histological review and microdissection, whole exome sequencing (WES), and phylogenetic analysis of human patients. Body diagram was adapted from the Motifolio toolkit. Example of PanINs and matched PDAC. The dashed outlines indicate regions that underwent laser capture microdissection of DNA extraction followed by whole exome sequencing (WES). The low-grade PanIN (LG-PanIN) shows well formed papillary structures with nuclear crowding and cytologic atypia. The high-grade PanIN (HG-PanIN) has regions of pseudopapillary formation (arrows) with high nuclear to cytoplasmic ratio. The matched PDAC shows features of poorly differentiated carcinoma with desmoplasia.

tinct and far removed from that of the PDAC (Methods). We excluded any patient with a personal or family history of PDAC from our study, as the dynamics of initiation in patients with germline alterations may be different from that in sporadic pancreatic carcinogenesis[223]. Eight patients were identified, from which 12 PanINs and eight PDACs were sampled for the current study (Supplementary Table 1). All 20 tissue samples were laser-capture microdissected to ensure that a high fraction of the cells within each lesion were neoplastic (Figure 2.1b). Despite the microscopic size of the PanINs, we were able to obtain sufficient amounts of DNA to generate high quality libraries for whole exome sequencing (WES). Importantly, the generation of these libraries did not require whole genome amplification prior to WES, thus reducing potential errors in downstream analyses.

Sequencing libraries were prepared from each of the lesions as well as from normal tissues of each patient and used for massively parallel sequencing on an Illumina HiSeq instrument. We obtained a median canonical exon coverage of 253x across all samples. By comparison of each lesion with its matched normal DNA, a total of 2,886 somatic single base substitutions (SNVs) and small insertions or deletions (INDELs) were identified (Extended Data Figure 1, Supplementary Table 2). As a group, the PanINs harbored as many SNVs/INDELs as the PDACs (average of 75 vs. 80, Extended Data Figure 1b). We also analyzed somatic copy number alterations (CNAs) and structural variants (SVs) from the exomic sequencing data (Supplementary Tables 3 and 4, Extended Data Figure 1c and 2). The number of CNAs, unlike the number of SNVs/INDELs, was higher in PDACs compared to PanINs (average of 90 vs 68). Computational analysis (Methods) revealed somatic mutations in many well-known driver genes, such as KRAS, CDKN2A, TP53, SMAD4, U2AF1, and KMT2D (Supplementary Table 5). Collectively, the genetic features of this set of PanINs and PDACs were consistent with previous sequencing studies of these tumors[121,24,268,274,12]. To infer evolutionary relationships among the PanINs and PDACs for each patient based on the SNVs/INDELs, we employed Treeomics[218], a recently developed phylogenetic method designed specifically for analyzing sequencing data from spatially distinct tumors in the same individual[160] (Methods). Treeomics

identified high confidence phylogenies for the matched samples from each of the eight patients (Figure 2.2a-c, Extended Data Figures 3-5). These analyses allowed us to derive the evolutionary relationships between the coexisting PanINs and the PDAC in each patient.

2.4.1 Evolutionary patterns in pancreatic cancer and precursor lesions

In our cohort, we found two cases (PIN102 and PIN105) in which no passenger gene mutations were shared by the PDAC and PanIN (Figure 2.2a, Extended Data Figure 3). For example, in patient PIN105 both the PanIN and PDAC had a KRAS p.G12D missense mutation. The PDAC exhibited 80 additional point mutations, including a one basepair frameshift deletion in TP53, a missense mutation in ACVR1B p.C34Y, and a 15 basepair in frame deletion in SMAD4. Additionally, the PDAC acquired CNA losses affecting CDKN2A, MAP2K4, TP53, and SMAD4 (Extended Data Figure 3b). The PDAC and PanIN may have arisen independently and by chance accumulated the same KRAS mutation (scenario 1), or they may have been initiated by a single KRAS p.G12D mutant clone and subsequently diverged (i.e. scenario 2). Scenario 1 may be more likely given the high frequency of KRAS variants in PDAC (>90%)13 and the absence of any other shared somatic variants among the matched PanIN and PDAC samples in both of these patients. Moreover, the PanINs in both of these cases exhibited PanIN-2 histology, and a previous study indicated that low grade PanINs often harbor genetic features that support independent evolution. We note the previous observation included distinct KRAS variants in matched PanINs, contrary to the two cases presented here9.

Four of the eight cases showed unequivocal evidence for scenario 2, that is a common ancestral cell underwent initiation and clonal expansion to form one or more PanINs. Further clonal expansions driven by additional driver gene mutations in a PanIN cell eventually led to a PDAC (Figure 2.2b, Extended Data Figure 4). For example, in patient PIN101, the common ancestor of PanIN lesion A and the PDAC acquired 14 somatic passenger mutations, including a KRAS p.G12D, as well as losses affecting ACVR1B, MAP2K4, TP53, and SMAD4 (Figure 2.2b, Extended Data Figure 4a).

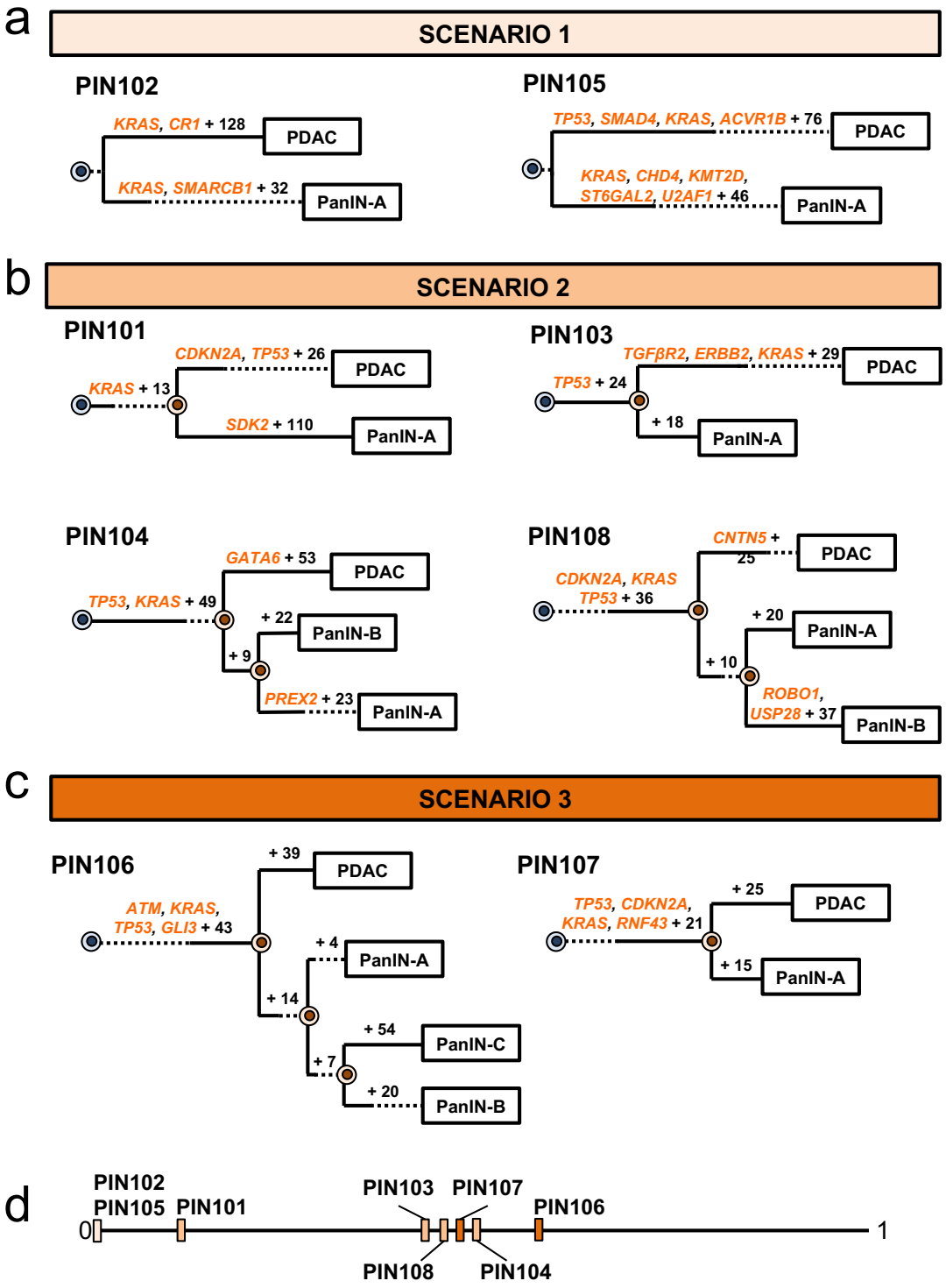


Figure 2.2: a-c. Phylogenetic trees from SNVs/INDELs. See Supplementary Table 1 for sample identities. For each phylogeny, gene names in orange text are SNVs/INDELs and the number of additionally acquired mutations are in black font. The branch lengths approximate the number of SNVs/INDELs. The dashed lines indicate branches that have been extended to accommodate gene annotation and variant numbers. The sequencing data for each driver gene variant was manually reviewed to verify phylogenetic position. a. PIN102 and PIN105 are both scenario 1. b. PIN101, PIN103, PIN104, and PIN108 are scenario 2. In manual review of the PIN101 sequencing data, a read supporting the presence of the KRAS p.G12D variant was detected in both the PDAC and PanIN-A samples and was thus moved to the trunk of the phylogeny despite the overall low coverage of KRAS in PanIN-A. c. PIN106 and PIN107 are scenario 3. d. Jaccard indices from SNVs/INDELs. For each evolutionary scenario, the average Jaccard index for each patient was calculated from all driver and passenger variants (see Supplementary Table 6 for values) and plotted on a range from 0 to 1, with values closer to 1 denoting higher genetic similarity.

The PDAC accumulated 28 point mutations including a p.A21D missense mutation in CDKN2A and a missense mutation in TP53 for p.R273H, as well as a loss affecting CDKN2A and a gain affecting MYC. The PanIN lesion A accumulated 111 point mutations, including a nonsense mutation in SDK2. Similar patterns were found in PIN103, PIN104 and PIN108, i.e. driver gene mutations common to all lesions as well as additional driver gene mutations specific to the PDAC (i.e., scenario 2, Figure 2.2b).

Finally, we observed two cases with phylogenetic patterns consistent with scenario 3 (PIN106 and PIN107) in which all lesions in a single pancreata shared all of the driver gene mutations identified (Figure 2.2c, Extended Data Figure 5). In patient PIN106, the common ancestor of all four samples harbored 47 somatic point mutations, including a p.G12D missense mutation in KRAS, a p.G266E missense mutation in TP53, a mutation affecting the splice region in ATM, and a p.Q597* in GLI3 (Figure 2.2c). The PDAC subsequently acquired 39 passenger mutations and losses affecting CDKN2A and SMAD4.

In summary, the lesions in four of these eight patients were unequivocally derived from the same precursor clone, as they shared multiple passenger genes and a subset of driver genes (scenario 2). The presence of these additional driver gene alterations, coupled with phylogenetic analysis, provides persuasive evidence that the PDAC was derived from a PanIN in each case. These results highlight the value of genetic evaluation of morphologically distinct legions in revealing the evolutionary dynamics of pancreatic carcinogenesis. Because the PanINs were all anatomically distinct and far removed from the PDAC (Methods), the data indicate that a single mutant clone had spread through the pancreatic ductal system to generate coexisting neoplastic lesions (Figure 2.3a). This situation is similar to what occurs in the bladder, wherein a single clone can form multiple anatomically distinct neoplasms[235]. Though it would seem much more challenging for a neoplastic cell to journey through the fluid in the pancreatic ductal system than to journey through the urine, this journey has been described in intraductal papillary mucinous neoplasms of the pancreas, and clearly occurred in

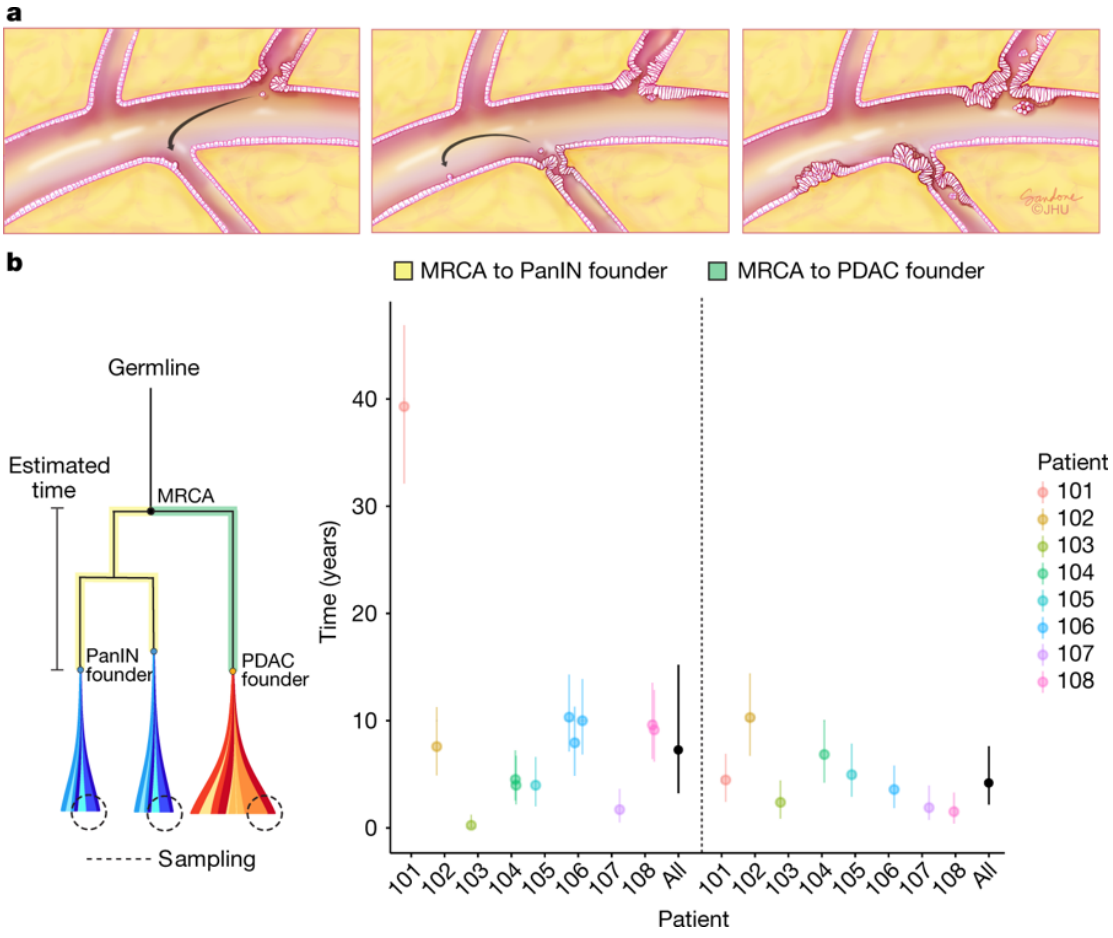


Figure 2.3: a. Spatial evolution and PanIN progression in intralobular ducts. Low grade PanIN (LG-PanIN) and high grade PanIN (HG-PanIN) lesions represent precursors with differing degrees of nuclear and cytologic atypia. A LG-PanIN develops and seeds a cell that travels to a second duct (arrow, left panel). The first LG-PanIN matures into a HG-PanIN, while a LG-PanIN develops at the second site and a cell subsequently travels to a third duct (arrow, center panel). The second site LG-PanIN matures into a HG-PanIN while a LG-PanIN develops at the second site (right panel). b. Estimated progression times. The lineage leading from the MRCA to the PanINs is illustrated in yellow, while the lineage leading from the MRCA to the PDAC is in green. Clonal passenger mutations were used to estimate progression times, shown for each patient with 90% CIs. Overall (black), the inferred median time elapsed between the common ancestral cell and the birth of the founder clone of a PanIN was 7.1 years (90% CI 3.3-12.2; MRCA to PanIN, n = 12). The median time elapsed between the common ancestral cell and the PDAC was 4.3 years (90% CI 2.3-7.2; MRCA to PDAC, n = 8). These estimates assume a mutation rate of 0.0224 per generation and a time per generation of 4 days (Online Methods).

these four patients as well[201].

To assess genetic relatedness using all somatic variants, we quantified Jaccard similarity coefficients between pairs of lesions within each scenario (Figure 2.2d, Supplementary Table 6). Interestingly, scenario 2 PanIN lesions tended to share fewer somatic variants with the matched PDAC as compared to PanIN lesions in scenario 3 (average Jaccard similarity coefficient of 0.39 vs. 0.50, respectively), although the range of Jaccard similarity coefficients overlapped between the two scenarios (scenario 2 range = 0.10 - 0.57, scenario 3 range = 0.44 – 0.70).

Our phylogenetic analysis also enabled us to estimate the mutational signatures operating in different tumor lineages that led to the PDAC or a coexisting PanIN (Extended Data Figures 6-8). Some signatures were shared between a PDAC and PanIN, while others operated only on a subset of different branches[226].

In PIN106 and PIN107, the PDACs and corresponding PanINs contained the same driver gene SNVs/INDELs (scenario 3, Figure 2.2 and Extended Data Figure 5). In addition to the lost copies of CDKN2A and SMAD4, several unobserved factors might contribute to their morphological differences. First, the PDAC may have accumulated additional genetic events of significance in regions of the genome not assessed by whole exome sequencing. Second, the PDACs may have acquired epigenetic alterations that were not detectable by the approach we used. Third, the microenvironment may have influenced the progression from a PanIN to a PDAC[134]. Finally, the PanIN lesions in PIN106 and PIN107 may represent cancerization of the ducts (invasive cancer growing back into the duct system and simulating PanINs). We note the PDACs in these two patients showed moderate to poorly differentiated histology, thereby decreasing but not fully eliminating this possibility[280].

2.4.2 Modeling progression time of pancreatic cancer evolution

The WES data allow us to estimate the time required for a cell to progress from a non-invasive, neoplastic clone to an invasive pancreatic cancer[279] (Methods). We used the number of acquired genetic

passenger alterations from a common ancestor to the PanINs and the PDACs, after removing mutations suspected to be drivers or subclonal, to infer the amount of passed time. Because the great majority of the mutations present in any of these lesions are passengers and are not associated with positive or negative growth advantage, these mutations can serve as a molecular clock. Based on previously estimated mutation rates[257] and cell division times[135] measured in PanINs, we found that the median time elapsed between the common ancestral cell (Figure 2.3b) and the birth of the founder clone of a PanIN was 7.1 years (90% CI of the median: 3.3 to 12.2 years). Similarly, the median time elapsed between the common ancestral cell and the founder cell of the PDAC itself was 4.3 years (90% CI 2.3 to 7.2 years). Because the PanIN samples are monophyletic in all patients, we cannot estimate how long the primary tumor lineage might have existed as a PanIN. Nonetheless, these intervals are conservative underestimates of the times required to develop neoplasia and radiographically detectable cancer because they do not include any clonal steps prior to the birth of the common ancestral cell nor the time between the birth of the PDAC founder cell and the multiplication of this cell to form a clinically evident mass. A larger patient cohort is required to assess whether or not this length of time is characteristic of the population of individuals with PDAC. When the time required for mass development is taken into account, the data suggest that it takes an average of at least 8.1 years elapsed between the birth of the common ancestral cell and the presence of a clinically evident mass (Methods).

2.5 Discussion

Comparison of our results with three recent studies is informative. First, Matsuda et al. found that 77% of patients without clinically evident pancreatic neoplasia actually harbored PanIN-1 lesions when autopsied[168]. Moreover, Wood et al. found that low grade PanINs (PanIN-1 and PanIN-2) from the same patient generally do not share the same genetic alterations, in contrast to our data

which show genetic relationships among high grade PanINs (PanIN-3)[113]. When taken together with our results, the data suggest that early neoplastic lesions in the pancreas may represent independent events, and that the success of the neoplastic cells in colonizing the ductal system is only achieved with histologic progression and the accrual of additional genetic alterations. Of interest in this regard, the budding off of small clusters of neoplastic epithelial cells into the lumen is one of the pathognomonic morphological features of a high-grade PanIN (PanIN-3)[115].

Our data are apparently at odds with the interpretation of a recent study that concluded PDACs do not arise in a gradual fashion[189]. This conclusion was based on genetic analyses of microdissected PDACs and did not include an analysis of PanINs, nor were models applied to the data to support such a conclusion. As such, it relied on assumptions about the timing of transition from precursor lesion to invasive carcinoma. By contrast, our data are directly based on genomic analyses of the precursor lesions and their corresponding PDACs. Our step-wise model is supported not only by the current data but also by a body of scientific literature[24,268,274,12,160,279,120,217] that suggests single/short base substitutions that gradually accumulate over many years form the great majority of the genetic alterations responsible for this tumor type. Our findings in no way contradict the observation that multiple chromosome translocations can occur simultaneously (chromothripsis) in a small subset of pancreatic tumors[189,217]. However, they do buttress the model that PDAC development is a multistep progression caused by the accumulation of somatic alterations in driver genes, a process that generally spans many years.

It could be argued that the cases we analyzed were unusual in that more than one advanced PanIN was found in each pancreas, and our selection of eight out of 100 patients potentially introduced an unintended bias in our cohort. However, Matsuda et al. have shown that multiple advanced PanIN lesions are the norm rather than the exception when the entire pancreas is methodically dissected[168]. Further, the mutations in driver genes and distribution of mutational signatures in this cohort are similar to those previously observed in pancreatic cancers. Finally, genomic anal-

ysis of a PDAC arising directly from an adjacent high grade PanIN lesion revealed a gradual genetic progression from PanIN to PDAC8 – similar to our findings for anatomically separate high grade PanIN lesions and their corresponding PDACs.

In summary, we have discovered that pancreatic intraepithelial neoplasia (i.e., PanIN-2 and PanIN-3) need not be a spatially localized lesion; rather, it is a disease that can spread through the entire ductal system. Additional studies—with more patients and a higher density of samples—will be required to determine the frequencies of the evolutionary scenarios we identified and to clarify which features of precursor lesions put them at substantial risk of transformation. Nonetheless, our data suggest that the multiple, apparently discrete PanIN lesions observed in an individual patient often represent a single neoplasm that can spread (contiguously or discontiguously) along the ductal system. This finding provides an explanation for the observation that patients who have had a high grade PanIN or PDAC removed by subtotal pancreatectomy are at high risk for the development of recurrent disease.

2.6 Methods

2.6.1 Patient selection

Human tissues were collected with the approval of the Johns Hopkins Hospital Institutional Review Board (protocols NA_00001584 and NA_00017879) after informed and written consent was obtained, following all relevant ethical regulations. Fresh-frozen samples from eight patients who underwent surgical resection of pancreatic cancer at Johns Hopkins Hospital (Jan 2009-Dec 2011) with pathologic confirmation of pancreatic ductal adenocarcinoma and geographically distinct PanIN-2 or PanIN-3 lesions were selected for study. For inclusion in the study, PDAC, PanIN lesion(s), and normal duodenum tissue were required for each patient. To minimize the possibility of studying cancerization of normal ducts, we only included PanINs in which at least 1.0 cm of un-

involved lobular parenchyma was present between the PanIN and the cancer, or the PanINs were present in a block that contained no cancer.

2.6.2 Processing of tissue samples

For each tissue sample, multiple sequential 5 μm thick cryosections were mounted on polyethylene naphthalate (PEN) membrane slides and stained with cresyl violet for visualization of histologic features and confirmation of adequate cellularity. Neoplastic epithelium was laser-microdissected using the Leica LMD7 laser microdissection system.

2.6.3 DNA extraction and quantification

Genomic DNA (gDNA) was extracted from each normal, PanIN, or tumor piece using a standard phenol and chloroform extraction followed by precipitation in ethanol. The gDNA was quantified by LINE assay (i.e. counting long interspersed elements (LINE) using real-time PCR. The LINE forward primer was 5'-AAAGCCGCTCAACTACATGG-3' and the reverse primer was 5'-TGCTTTGAATGCGTCCCAGAG-3'. The real-time PCR protocol was 50°C for 2 min, 95°C for 2 min, 40 cycles of 94°C for 10 s, 58°C for 15 s, and 70°C for 30 s, 95°C for 15 s, and 60°C for 30 s. The PCR reactions were carried out using Platinum SYBR Green qPCR mastermix (Invitrogen).

2.6.4 Whole exome sequencing and alignment

Whole exome sequencing (WES) was performed on an Illumina HiSeq 2000 platform for a target coverage of 150X. Upon the completion of WES, the data were analyzed in silico to determine overall quality and coverage. Sequencing reads were aligned to the hg19 human reference genome using BWA[146]. Read de-duplication, base quality recalibration, and multiple sequence realignment were performed using the Picard Suite and GATK version 3.1[71,175]. SNVs were called using Mutect version

1.1.6 and INDELs were detected using HaplotypeCaller version 2.4[71,54].

2.6.5 Filtering of whole exome sequencing data

WES generated a large list of potential mutations, and we evaluated these data to identify high quality mutations while removing sequence artifacts. Each mutant must have been observed with at least 5% variant allele frequency with 20x coverage in at least one neoplastic sample; each mutant must have been observed in less than 2% of the reads (or 3 reads total) of the matched normal sample with 10x coverage. This filtering yielded a total of 2,886 mutations for subsequent analysis (Supplementary Table 2).

2.6.6 Driver gene and mutation analysis

All somatic variants causing a frameshift deletion, frameshift insertion, in-frame deletion, in-frame insertion, missense, nonsense, nonstop, splice site/region, or a translation start site were considered. If a variant was a missense or nonsense mutation, we required the variant to have a CHASM p-value of ≤ 0.05 and an FDR of ≤ 0.25. In combination with manual review, driver gene mutations were identified if the gene was supported by at least three of the following four methods: 20/20+36, TUSON[65], MutSigCV[142] (see Table S1 in Ref.[255] for gene list), and a hotspot analysis[48]. In addition, we also considered genes significantly mutated in large PDAC sequencing studies[24,268,12,213]. Further, we required that each somatic variant have a variant allele frequency of $< 2\%$ in the patient-matched normal tissue as well as any normal tissue from another patient. If a deleterious variant was detected in a driver gene as described above, and was not detected abundantly in any normal tissue, it was considered a driver gene variant.

2.6.7 CNAs

Allele-specific copy number analysis was performed using FACETS[241]. Briefly, FACETS performs a complete analysis that includes library size and GC-normalization, and segmentation of total and allele-specific signals, using coverage and genotypes of single nucleotide polymorphisms simultaneously across the exome[241]. The resulting segments accurately identify points of change in the exome, accounting for diploidy, purity, and average ploidy for each sample. A maximum likelihood approach then assigns each segment with a major and minor integer copy number.

2.6.8 Evolutionary analysis

We derived phylogenies for each set of samples by using Treeomics 1.7.9[218]. Each phylogeny was rooted at the matched patient's normal sample and the leaves represented the PanIN or tumor samples. Treeomics employs a Bayesian inference model to account for error-prone sequencing and varying neoplastic cell content to calculate the probability that a specific variant is present or absent. Treeomics infers the global optimal tree based on Mixed Integer Linear programming. For Extended Data Figures 3-5, the CNAs were not directly used to infer phylogenies in order to prevent bias from potential false-negatives or false-positives, given that CNA calls from multiple samples within a patient are particularly sensitive to varying neoplastic cell content and depth of sequencing. Moreover, WES data usually does not capture the exact breakpoints of CNAs, further complicating phylogenetic analysis. Nevertheless, common PDAC driver genes KRAS, MYC, GATA6, and CDK6 were manually reviewed in the CNA data for evidence of gains, while CDKN2A, SMAD4, TP53, MAP2K4, TGFβR2, and ACVR1B were queried for losses. Allelic losses were defined as total copy number (tcn) = 1 or 0, and gains were defined as tcn $\geq$ 4. Given the CNA status of a given driver gene in each sample, the driver gene with the CNA status was manually placed on the corresponding position edge in the phylogeny (previously derived using SNVs/INDELs). This approach was used

with each PDAC driver gene affected by a CNA as defined above.

Our classification of each patient into one of three evolutionary scenarios was based on SNVs/INDELs that affect key driver genes in PDAC (e.g. KRAS G12D). Such alterations represent driver gene variants that are readily interpretable with respect to function as well as position on the phylogenetic tree. Nonetheless, CNAs can also affect driver genes involved in pancreatic cancer (e.g. CDKN2A deletion). If we reclassify the eight patients using both SNVs/INDELs as well as CNAs affecting driver genes (Extended Data Figures 3-5), we find that the evolutionary scenario does not change for six patients. For two patients (PIN106 and PIN107), the scenario changes from scenario 3 to scenario 2, indicating a step-wise progression of PanINs and PDACs for all eight patients. As noted above, the identification and placement of CNAs on a phylogenetic tree remains challenging. Nonetheless, we note that the SNV/INDEL phylogenies represent a minimum number of evolutionary steps: including additional CNAs would either confirm or increase the total number of steps in the evolution of the PDAC.

2.6.9 Structural variant analysis

We inferred structural variants (SV) using DELLY2 (v.0.7.5) to verify the reconstructed phylogenies[214]. Since the SVs were called for each sample independently, we merged SVs for which DELLY determined breakpoints differing by at most 250 base-pairs among the samples of each patient. In total, we found 154 distinct SVs in the eight subjects. After a comprehensive manual review of the called SVs, we developed additional criteria to minimize the number of false positives. We required that each SV has to pass one of the following two filters in at least one sample: 1) (a) SV is supported by at least 3 distinct split reads, (b) the ratio of split reads that support the SV to the total number of split reads at the position of the SV is greater or equal to 0.75, and (c) the number of the SV supporting split reads is greater than the number of split reads in the normal sample; or 2) (d) SV is supported by at least 5 discordantly paired (DP) reads, (e) the ratio of DP reads that support the SV

to the total number of DP reads at the position of the SV is greater or equal to 0.25, and (f) the number of the SV supporting DP reads is greater than the number of DP reads in the normal sample. After applying these filters, we obtained 40 SVs (Supplementary Table 4).

To create input files for Treeomics, we used the number of SV supporting split and DP reads as the number of variant reads. We normalized the coverage of SVs such that on average it approximately matched the median coverage of the SNVs (single nucleotide variants). Generally, the inferred phylogenies based on the SVs agreed well with the ones based on SNVs. However, since the significantly lower number of SVs per subject (median 4; range 0-14; Supplementary Table 4), the confidence in the inferred branches was significantly lower than in the phylogenies based on SNVs. For PIN106 (coverage in sample of PanIN-A was extremely low), we inferred a slightly different phylogeny as PanIN-A diverged before the PDAC, likely due to many false negatives resulting from the extremely low coverage and the therefore difficult detection of SVs in this sample. For PIN108, no SVs were shared across multiple samples and hence there were no parsimony-informative SVs such that a phylogeny could be inferred.

2.6.10 Mutation signatures

We assessed the presence of previously identified mutational signatures[4] in each patient. Our phylogenetic analysis enabled us to estimate the signatures operating at different stages of cancer evolution[226]. For SNVs acquired along each phylogenetic branch, we estimated the maximum likelihood signature proportions among 30 previously identified trinucleotide signatures[3] (see https://github.com/mskcc/mutation-signatures). We quantified the uncertainty in these estimates by performing 100 iterations of bootstrap resampling within each branch followed by signature re-estimation. We ignored branches with 5 or fewer mutations and removed signature 24 because of its similarity to smoking. The maximum likelihood signature estimates and 90% bootstrap confidence intervals for each branch are shown in Extended Data Figures 6-8. We detect signatures 1, 2, 3, and 6, consistent with previous studies[4].

Additionally, we find evidence for signatures 4 (associated with smoking) and 29 (associated with chewing tobacco). Signatures operating on different branches within a patient were not significantly more similar than those across patients (mean cosine distance similarity 0.62 vs 0.59, p=0.21, one-sided permutation test). We note that signature estimates had large bootstrap uncertainty and the number of patients as well as the number of mutations is limited.

2.6.11 Progression time inference

We assume that the number of passenger mutations n acquired along a lineage during time T (in cell generations) is Poisson-distributed with rate equal to T times the mutation rate per cell division[121]:

$$n_\mu | T \sim Poisson(\mu T) \tag{2.1}$$

We assume that a random sample from the population of PanINs or PDACs takes T generations to progress from a previous stage (either most recent common ancestor (MRCA) of all sampled PanINs and PDAC in a patient or the MRCA of the most closely related PanIN to the PDAC) to the founder of a particular PanIN or PDAC, and that the mutational clock time μT is gamma-distributed with hyperparameters shape k and scale θ ($k, \theta > 0$) uniform a priori:

$$\mu T \sim Gamma(k, \theta) \tag{2.2}$$

In order to infer the joint distribution of (T, k, θ), we use the following sampling strategy. For each sample i, we update T by sampling directly from the gamma posterior:

$$T_i \mu | n, k, \theta \sim Gamma(k + n, \theta/(1 + \theta)) \tag{2.3}$$

Using the updated values, we jointly update k, θ by Metropolis-Hastings sampling from the pos-

terior:

$$L(k,\theta) \propto \pi(k,\theta) \prod_{T_i} dGamma(k,\theta,\mu T_i) \tag{2.4}$$

where dgamma is the density function for the gamma distribution and $\pi(k,\theta)$ is the prior over the hyperparameters (uniform). This setup pools information about the time to progression for each sample toward the population of progression time estimates, with a flexible structure for the overall distribution of times provided by the gamma distribution.

In order to convert the inferred number of generations to absolute time, we follow a previous method[279] by multiplying by the average time for cell division. To estimate the division time, we again follow the previous method but instead note that 14% of Stage II PanINs stain positively for Ki-67[135]. We therefore estimate the generation time of PanIN Stage II cells to be 4 days. The mutation rate μ per generation is 0.0224, calculated for 35 Mb of exome sequencing multiplied by a point mutation rate of $6.4 * 10^{-10}$ per generation[257].

To calculate the expected time it takes that the PDAC founding cell grows to a detectable lesion of 1 cm^3 ($\approx 10^9$ cells), we used previously measured PDAC metastasis doubling times of 56 days[6] leading to an exponential growth rate of r=0.012 per day. The probability density function for the time an exponential branching process conditioned on survival takes to reach size $M = 10^9$ is approximately given by:

$$f_(t_M)(t) = exp(-\frac{r}{b} Mexp(-rt)) \frac{r^2 M}{b} exp(-rt) \tag{2.5}$$

where $b = 1/2.3$ per day is the assumed PDAC cell division rate[279,77].

2.7 Data availability

Sequence data have been deposited at the European Genomephenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG), under accession number EGAS00001002778. Further information about EGA can be found at https://ega-archive.org and "The European Genome-phenome Archive of human data consented for biomedical research" (http://www.nature.com/ ng/journal/v47/n7/full/ng.3312.html). Source data are provided for Figure 2.3, panel b, and Extended Data Figures 1, 7 and 8. All other relevant data are included within the manuscript or are available upon request from the corresponding author (C.I-D.).

3

Minimal functional driver gene heterogeneity among untreated metastases

3.1 Forward

This work grew out of a review of driver gene heterogeneity among untreated metastases. Hannes, Alvin, and I realized that no such analysis had been performed previously since most cancer studies

had only one or a few patients who met the treatment criteria. After Hannes completed the initial and painstaking work compiling the data from several sources and running Treeomics (see Chapter 1), we all came together with Chris, Bert, and Martin to analyze the results and decide how to proceed. The results of this were additional data and insight into targeted therapy outcome heterogeneity, a more stringent definition of driver genes, and a model of tumor progression and metastastic seeding which helped make sense of the data. I contributed most to the development and coding of the model, which was analyzed in parallel by Alex Heyde, who contributed several excellent insights. I also helped to check his analysis and contributed heavily to the manuscript revisions.

This work was first published in Ref.[219]:

Reiter, J. G.*, Makohon-Moore, A. P.*, Gerold, J. M.*, Heyde, A., Attiyeh, M. A., Kohutek, Z. A., Tokheim, C. J., Brown, A., DeBlasio, R. M., Niyazov, J., Zucker, A., Karchin, R., Kinzler, K. W., Iacobuzio-Donahue, C. A., Vogelstein, B., and Nowak, M. A. (2018). Minimal functional driver gene heterogeneity among untreated metastases. Science, 361(6406), 1033–1037. (*equal contribution)

Supplemental materials can be found online at DOI 10.1126/science.aat7171

3.2 Abstract

Metastases are responsible for the majority of cancer-related deaths. While genomic heterogeneity within primary tumors is associated with relapse, heterogeneity among treatment-naïve metastases has not been comprehensively assessed. We analyzed sequencing data for 76 untreated metastases from 20 patients and inferred cancer phylogenies for breast, colorectal, endometrial, gastric, lung, melanoma, pancreatic, and prostate cancers. We found that within individual patients a large majority of driver gene mutations are common to all metastases. Further analysis revealed that the driver gene mutations that were not shared by all metastases are unlikely to have functional consequences.

A mathematical model of tumor evolution and metastasis formation provides an explanation for the observed driver gene homogeneity. Thus, single biopsies capture most of the functionally important mutations in metastases and therefore provide essential information for therapeutic decision making.

3.3 Main

The clonal evolution model of cancer proposes that cells accrue advantageous mutations and clonally expand such that these mutations are eventually present in all tumor cells[196,83,114,97]. Recent studies reported mutations in putative driver genes that were only present in subpopulations of tumor cells[118,92]. The extent to which the acquisition of advantageous mutations continues after the initiation of the primary tumor[244] or during metastasis formation is unknown[259,167]. The growing list of putative driver genes and the increased sensitivity of next-generation sequencing have facilitated the discovery of subclonal driver gene mutations within a tumor[118,11]. Nevertheless, the evolutionary dynamics and the clinical significance of driver gene mutation heterogeneity in solid tumors are not fully understood.

Cells acquire a few mutations during each division due to imperfect DNA replication; hence, any population of cells is genetically heterogeneous[26]. Because cancer cells continue to divide after cancer initiation, many new mutations are expected to be present in tumor subpopulations. However, to assess functional heterogeneity, advantageous mutations in putative driver genes must be distinguished from neutral replication errors in those genes. For example, within oncogenes only few recurrently mutated positions are functional and therefore many mutations—even in driver genes—may not have important functional consequences. Moreover, although metastatic disease is responsible for most cancer-related deaths, the heterogeneity of driver gene mutations has predominantly been evaluated in primary tumors. Biopsies of metastatic lesions are not readily available and

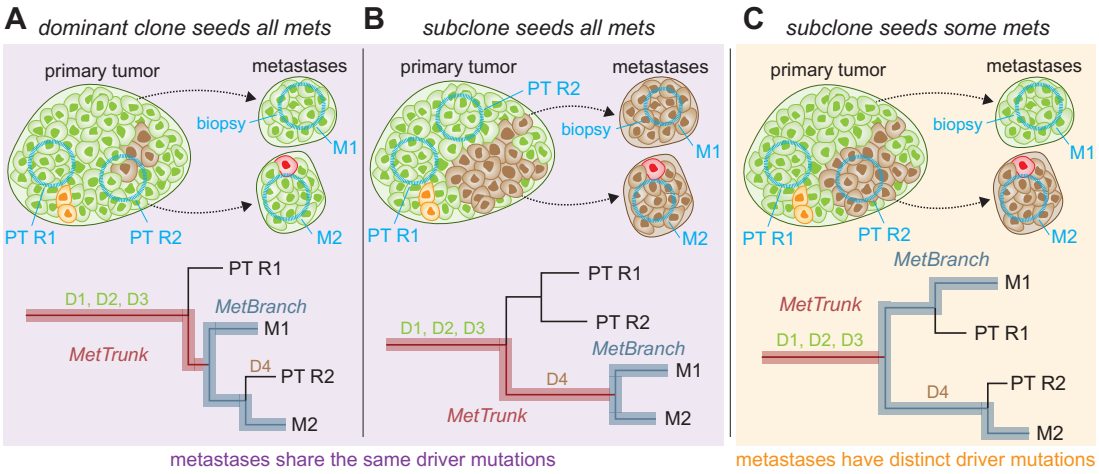
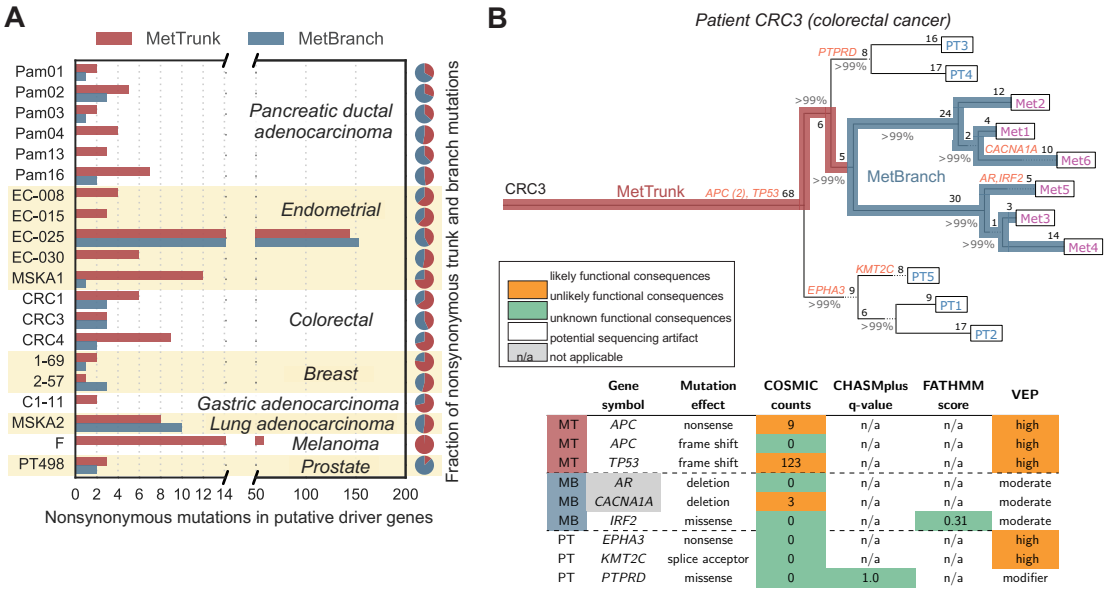


Figure 3.1: The original clone (green cells) contains three driver gene mutations (D1, D2, D3). Brown, yellow, and red cells acquired additional driver mutations during the growth of the primary tumor (PT) and may expand to form detectable subpopulations (brown) which can seed metastases. Top panels illustrate seeding subpopulations and biopsies (blue circles) of different regions (R1, R2) of the PT and of distinct metastases (M1, M2). Bottom panels illustrate reconstructed cancer phylogenies from those biopsies. (A) Original clone seeds all metastases. All metastases share same founding driver mutations. Subclones with additional driver mutations (D4) evolve too late to seed metastases, but might be detectable in the PT. (B) A single highly metastatic subclone evolves and gives rise to all metastases. All metastases share same founding driver mutations. (C) A new subclone with an additional driver mutation (D4) evolves and independently seeds metastases. PT regions and metastases exhibit driver mutation heterogeneity.

typically are acquired after exposure to toxic and mutagenic chemotherapies. These treatments can induce selective bottlenecks and confound the interpretation of genetic alterations.

Because driver gene mutations increasingly inform clinical treatment decisions, undetected driver heterogeneity among metastases poses a barrier to the success of this precision medicine approach [265]. If the founding cells of different metastases carry distinct driver gene mutations, disease progression and treatment could be fundamentally more complex than expected from a primary tumor biopsy alone. Additional driver gene mutations might be present in all or in a subset of metastases (Figure 3.1). In both scenarios, more biopsies would be necessary for accurate diagnosis and optimal treatment. Here, we comprehensively analyzed the evidence for driver gene mutation heterogeneity among untreated metastases across cancer types. We also developed a mathematical model to determine the evolutionary mechanisms that give rise to inter-metastatic driver mutation heterogeneity.



	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	VEP
MT	*APC*	nonsense	9	n/a	n/a	high
MT	*APC*	frame shift	0	n/a	n/a	high
MT	*TP53*	frame shift	123	n/a	n/a	high
MB	*AR*	deletion	0	n/a	n/a	moderate
MB	*CACNA1A*	deletion	3	n/a	n/a	moderate
MB	*IRF2*	missense	0	n/a	0.31	moderate
PT	*EPHA3*	nonsense	0	n/a	n/a	high
PT	*KMT2C*	splice acceptor	0	n/a	n/a	high
PT	*PTPRD*	missense	0	1.0	n/a	modifier

Figure 3.2: (A) Twenty patients with 76 untreated metastases. Thirteen patients acquired mutations in putative driver genes along the MetBranch (MB) while seven did not. (B) Inferred phylogeny of a colorectal cancer exhibits inter-metastatic driver mutation heterogeneity. Nonsynonymous mutations in driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Table shows predicted functional effects of mutations in driver genes. Heterogeneous driver mutations were predicted to have no functional effect or were likely sequencing artifacts (low coverage and low VAF across all sites). MetTrunk (MT) denotes that variant was acquired on the trunk of all metastases. Sample origin: rectum: PT1-5; liver: Met1-6.

We analyzed data from 20 cancer patients for whom genome- or exome-wide sequencing was performed for at least two distinct treatment-naïve metastases[234,111,131,94,160,41,202]. In total, we studied 115 samples including 76 untreated metastases samples from diverse tissues (mean of 3.8 and median of 3 metastases per patient) (Supplemental Figure 1; Supplemental Table 1). We assessed somatic mutations of patients with pancreatic, endometrial, colorectal, breast, gastric, lung, melanoma, and prostate cancer (Figure 3.2A). We classified nonsynonymous variants into putative driver and passengers mutations according to the TCGA consensus list of 299 putative driver genes (10). To allow for a consistent interpretation of driver gene mutation heterogeneity, we excluded two hypermutated subjects with more than 1000 nonsynonymous mutations and focused on the remaining eighteen subjects. In these subjects, we found a median of 4.5 mutated driver genes (range 2-18) (Figure 3.2A).

To determine the evolutionary timing of somatic mutations, we inferred cancer phylogenies and mapped all variants onto evolutionary trees[218] (supplementary materials; Supplemental Figure 2). We classified mutations into those present in all metastases (MetTrunk, hereafter referred to as trunk) and those present in a subset of metastases (MetBranch, hereafter referred to as branch) (Figure 3.2B). We observed similar numbers of nonsynonymous or splice-site variants (hereafter referred to as nonsynonymous) in both categories (Figure 3.2A). In contrast, trunks exhibited a 2 fold enrichment of the ratio of driver gene mutations to nonsynonymous mutations compared to branches (9.1% vs. 4.0%, two-sided paired t-test P=0.004; Figure 3.3A). Nevertheless, we observed mutations in driver genes that were heterogeneous among metastases for 12 of 18 subjects.

To investigate whether heterogeneous mutations in putative driver genes were likely to be functional, we employed a variety of approaches. We found that a large proportion of nonsynonymous variants in driver genes along trunks were previously detected at least once in other cancers (COSMIC, Catalogue Of Somatic Mutations In Cancer; 37.8%, 31/82) whereas a much smaller proportion along branches was present in COSMIC (15.6%, 5/32; two-sided Fisher's exact test P = 0.025; Figure 3.3B). The fraction of driver gene mutations in branches in COSMIC was in fact similar to that of

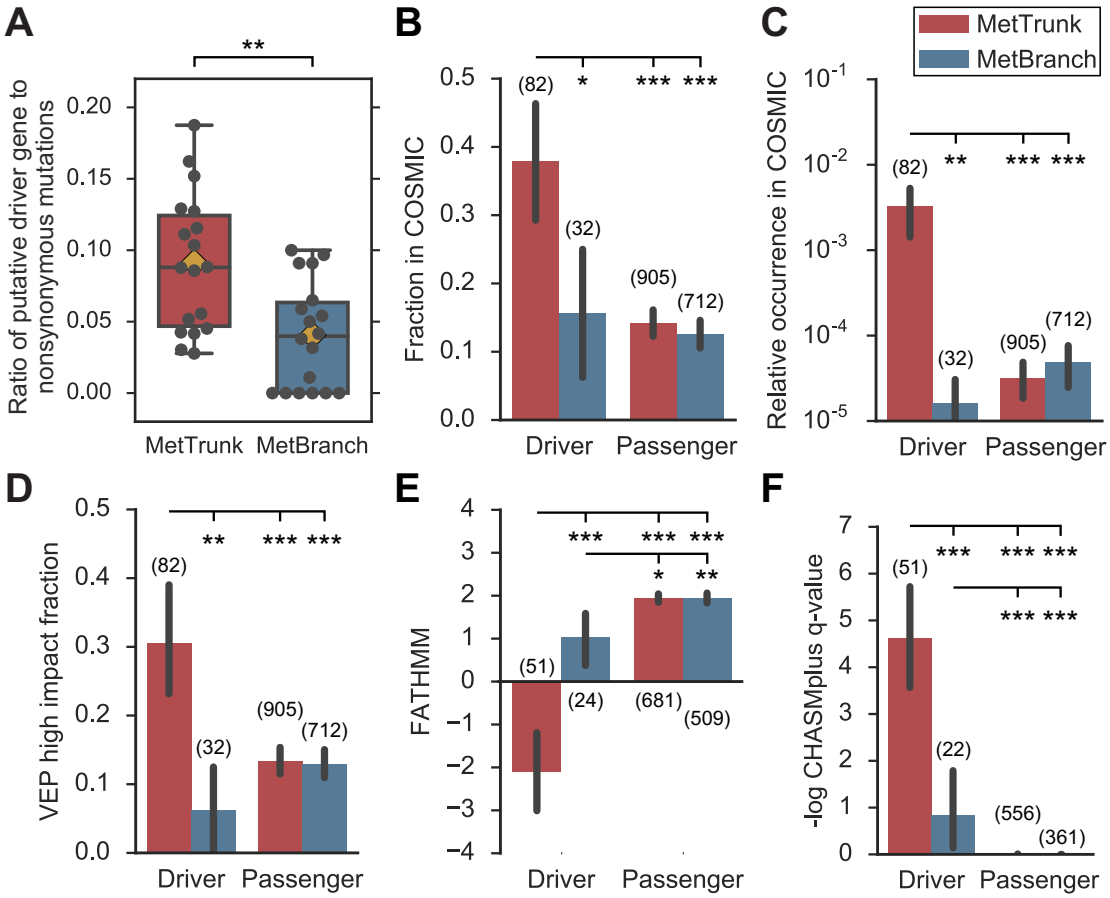


Figure 3.3: (A) Ratio of driver gene mutations to nonsynonymous mutations is enriched by 2-fold along trunks compared to branches. Orange diamond denotes mean, black bar denotes median (two-sided paired t-test P=0.004). (B) Fraction of nonsynonymous variants in driver genes along MetTrunk in COSMIC was 38% compared to 16% along MetBranch (two-sided Fisher's exact test P=0.025). (C) Relative occurrence of variants in driver genes along MetTrunk in individual COSMIC samples was 0.32% compared to 0.0016% along MetBranch (two-sided Wilcoxon rank-sum test P=0.008). (D) VEP inferred that 30% and 6% of driver gene mutations were of high impact along MetTrunk and MetBranch, respectively (two-sided Fisher's exact test P=0.006). (E-F) FATHMM (value below −0.75 indicates likely driver mutation) and CHASMplus predicted increased functional consequences for variants in driver genes in MetTrunk. Two-sided Wilcoxon rank-sum tests were used. Thick black bars denote 90% confidence interval. No other statistically significant differences were observed. Numbers in brackets denote number of variants in each group. * indicates P<0.05, ** P<0.01, *** P<0.001.

passenger gene mutations in either trunks or branches (14.1%, 128/905 and 12.5%, 89/712). Because mutations that are true drivers are often recurrent, we investigated how frequently identical nonsynonymous variants were found in COSMIC. While variants in driver genes along trunks on average occurred in 0.32% COSMIC samples (occurrence mean of 82.0 in 25,516 COSMIC samples), driver gene mutations acquired along branches occurred more than 100-fold less frequently (0.0016%; Figure 3.3C; Wilcoxon rank-sum test P=0.008).

We then utilized several methods to predict the functional impact of 1,755 nonsynonymous variants along trunks and branches. We found that driver gene mutations acquired along trunks were more likely to have predicted functional consequences (Figure 3.3, D-F; Supplemental Figure 3). Variants with the most likely protein changing effects (mutation consequences with high impact, e.g., frameshift or nonsense mutations) were frequently observed in driver genes along trunks but rarely observed along branches (30.5% vs 6.3%; Fisher's exact test P = 0.006; Figure 3.3D). The frequency of high impact variants in driver genes along branches was no higher than that in passenger genes. FATHMM [242] predicted significantly stronger functional effects for driver gene mutations along trunks than along branches (mean scores of -2.1 vs. 1.0; scores below -0.75 indicate likely driver mutation; Wilcoxon rank-sum test P<0.001; Figure 3.3E). Similarly, CHASMplus [254] predicted significantly higher gene-weighted scores for driver gene mutations along trunks than along branches (mean scores 0.47 vs. 0.16; higher values indicate likely functional effects; Wilcoxon rank-sum test P<0.001; Figure 3.3F).

To identify the evolutionary determinants of inter-metastatic heterogeneity, we developed a mathematical framework to assess how rates of growth, mutation, and dissemination give rise to driver gene mutation heterogeneity [102,77] (supplementary materials). The original clone in the primary tumor grows with a rate of $r_0 = b_0 - d_0$ per day (birth rate b_i, death rate d_i for each clone i) and disseminates cells to distant sites with rate q_0 per day (Figure 3.4A). When a cell divides, a daughter cell can acquire an additional driver mutation with probability u. This model produces

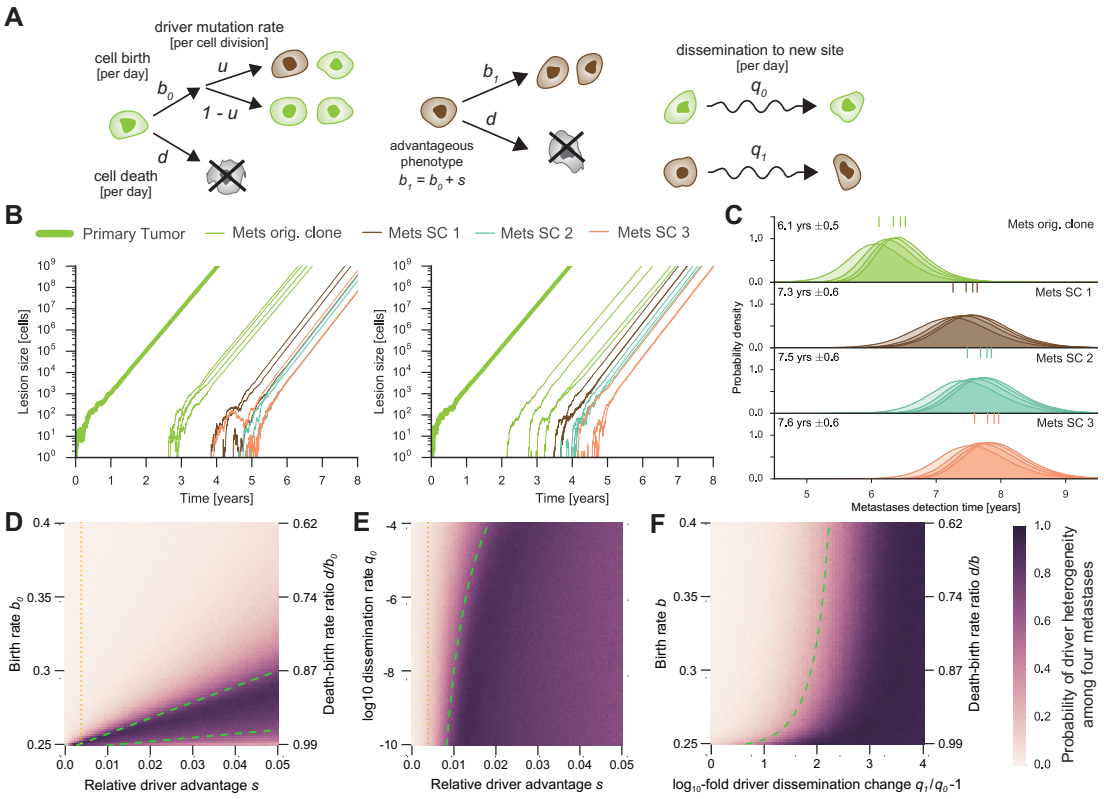


Figure 3.4: . (A) Primary tumor expands stochastically from a single advanced cancer cell and seeds metastases. Cells of original clone (green) divide at rate b0 and die at rate d per day. Additional driver mutations increase the birth rate to $b_1 = b_0(1+s)$, where s denotes the relative driver advantage ($b_1\, b_0$, $q_0 = q_1$; B-E), or increase the dissemination rate ($q_1\, q_0$, $b_1 = b_0$; F). (B) Representative model realizations for typical parameter values. Growth rate $r_0 = 1.24\%$ per day, $s = 0.4\%$, dissemination rate $q_0 = 10^{-7}$ per cell per day. (C) Distribution of metastases detection times for parameter values in B. Numbers denote mean ± standard deviation. Colored marks show mean detection times of first, second, third, and fourth metastases seeded by the corresponding subclone (SC). (D-F) Probability of distinct driver mutations among four metastases. Green dashed lines depict bounds separating parameter regions of likely inter-metastatic driver homogeneity from heterogeneity. Orange dotted lines denote $s = 0.4\%$. (D) Fixed $q_0 = 10^{-7}$. (E) Fixed death-birth rate ratio $db_0 = 0.95$. (F) Fixed $q_0 = 10^{-7}$. Other parameter values: $d = 0.2475$, driver mutation rate $u = 3.4 * 10^{-5}$ per cell division.

inter-metastatic heterogeneity if not all detectable metastases were seeded from the same subclone in the primary tumor.

Following previously measured growth and selection parameters, we assume a growth rate of $r_0 = 1.24\%$ per day and a relative growth advantage of a driver gene mutation of $s = 0.4\%$ ($s = b_i(b_0 - 1)$)[32,88]. To mimic the composition of our cohort, we consider the first four metastases that reach a detectable size of 10^8 cells (1 cm^3). We find that the probability of inter-metastatic driver heterogeneity is 10.5% (Figure 3.4; $d = 0.2475$, $q = 10^{-7}$). The original founding clone of the primary tumor most likely seeds all detectable metastases (green cells; Figure 3.1A). The increased growth rate conferred by a new driver mutation is insufficient to compensate for the time spent waiting for the driver mutation to occur (Supplemental Figures 4, 5).

The model reveals that the probability of observing inter-metastatic driver heterogeneity increases when the primary tumor grows very slowly before metastases are seeded, the average growth advantage of additional driver mutations is very large, and the driver gene mutation rate is high (Supplemental Figure 6C). In contrast, a high dissemination rate produces less inter-metastatic heterogeneity because metastases are established before driver subclones grow large (Figure 3.4E, Supplemental Figure 7C). For very high driver growth advantages but slowly growing cancers, another scenario is possible: all metastases are seeded from the same highly advantageous subclone (Figure 3.1B). Finally, if driver mutations instead increase the dissemination rate, an almost ten-fold increase is required to produce inter-metastatic driver heterogeneity (Figure 3.4F; Supplemental Figure 8).

In real patients, we expect less inter-metastatic heterogeneity for several reasons. First, driver gene mutations may not confer the same advantage in the microenvironment of the primary tumor and of a distant site, reducing the probability of heterogeneity (Supplemental Figure 9). Second, primary tumor growth may slow down due to space or nutrient constraints or surgical removal, also reducing the expected inter-metastatic heterogeneity (Supplemental Figure 10). Third, advanced cancer cells have already acquired multiple driver gene mutations in various pathways, possibly reducing

the number of additionally available driver gene mutations that confer a significant selective advantage (Supplemental Figure 6B).

Overall, we observed a depletion of heterogeneous mutations in putative driver genes among metastases (Figure 3.3). Moreover, the majority of those that were observed had only weak or no predicted functional effects. These results are compatible with multiple recent studies on neutrally evolving cancers after transformation[244,34,273]. However, the mathematical framework demonstrates that a lack of inter-metastatic driver heterogeneity does not imply neutral evolution but can also be explained by various other factors, including primary tumor growth dynamics (Figure 3.4). Furthermore, growth rates may saturate and fitness gains of additional driver gene mutations become smaller because available resources (nutrients, oxygen, etc.) are already almost optimally utilized; a phenomenon that is observed in bacterial evolution[96].

Several limitations of this study should be noted. First, we exclusively focused on single nucleotide variants and small insertions/deletions because their functionality can be predicted by multiple methods and their heterogeneity has immediate clinical consequences for therapy selection[265]. We did not assess recurrent noncoding, copy number, or epigenetic alterations since functional prediction methods for them are not yet available. Second, we cannot exclude the possibility that mutations in yet undiscovered driver genes of metastases are heterogeneous. Third, we could not evaluate micro metastases that are not visible clinically.

Because therapy selection and treatment success of previously untreated patients increasingly depends on the identification of genetic alterations, it will be critical to extend this analysis to larger cohorts and more cancer types to investigate whether minimal driver gene mutation heterogeneity is a general phenomenon of advanced disease. This pan-cancer analysis of untreated metastases suggests that a single biopsy accurately represents the driver gene mutations of a patient's metastases.

3.4 Funding and support

This work was supported by the National Institutes of Health grants K99CA229991 (J.G.R.), CA179991 (C.A.I. D.), F31CA180682 (A.P.M.-M.), T32 CA160001-06 (A.P.M.-M.), F31CA200266 (C.J.T.), U24CA204817 (R.K.), CA43460 (B.V.), as well as by the Lustgarten Foundation for Pancreatic Cancer Research, The Sol Goldman Center for Pancreatic Cancer Research, The Virginia and D.K. Ludwig Fund for Cancer Research, an Erwin Schrödinger fellowship (J.G.R.; Austrian Science Fund FWF J-3996), a Landry Cancer Biology fellowship (J.M.G.), and the Office of Naval Research grant N00014-16-1-2914.

3.5 Author contributions

J.G.R., A.P.M.-M., C.A.I.-D., B.V., and M.A.N. conceived and designed the study. A.P.M. M., M.A., Z.A.K., A.B., R.D., J.N., A.Z., and C.A.I.-D. performed autopsies. A.P.M.-M., M.A., Z.A.K., K.K., K.W.K., C.A.I.-D., and B.V. generated sequencing data. J.G.R. performed computational analysis. J.G.R., J.M.G., A.H., and M.A.N. performed mathematical modeling. C.J.T. and R.K. performed CHASMplus analysis. C.A.I.-D., B.V., and M.A.N. supervised the study. J.G.R., A.P.M.-M., J.M.G., A.H., C.A.I. D., B.V., and M.A.N. wrote the manuscript. All authors read and approved the manuscript.

3.6 Competing interests

K.W.K. and B.V. are founders of Personal Genome Diagnostics. B.V. and K.W.K. are on the Scientific Advisory Board of Sysmex-Inostics. B.V. is also on the Scientific Advisory Boards of Exelixis GP. These companies and others have licensed technologies from Johns Hopkins, and K.W.K. and B.V. receive equity or royalties from these licenses. The terms of these arrangements are being managed

by Johns Hopkins University in accordance with its conflict of interest policies.

3.7 Data and materials availability

Accession numbers for the raw sequencing data are available in the original publications (13–18). Data of Brown et al., Hong et al., and Makohon Moore et al. as well as of subjects MSKA1 and MSKA2 are deposited at the European Genome-Phenome Archive (http://www.ebi.ac.uk/ega) and are available under accession numbers EGAS00001000760, EGAS00001000942, EGAS00001002186, and EGAS00001002777, respectively. Data of Gibson et al. and Sanborn et al. are deposited to the database of Genotypes and Phenotypes (dbGaP) at the National Center for Biotechnology Information (NCBI) under accession codes phs001127.v1.p1 and phs000941.v1.p1, respectively. Data of Kim et al. are deposited to the Sequence Read Archive (SRA) at the NCBI under the project ID of PRJNA271316.

4

Quantifying clonal and subclonal passenger mutations in cancer evolution

4.1 Forward

Ivana Bozic invited me to contribute to this originally purely theoretical project after deciding that it would benefit from comparison to allele frequency spectra from TCGA sequencing data. My main

contributions were principly in bioinformatics: running variant calling software, filtering results, implementing fitting approaching and helping to produce figures. I benefited greatly from exposure to Ivana's incisive analysis of the stochastic model described here. The critical insight about the balance between birth and death rates in a branching process framework helped shape the direction of Chapter 5 as well.

This work was first published in Ref.[34]:

Bozic, I., Gerold, J. M., and Nowak, M. A. (2016a). Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. PLOS Computational Biology 12(2), e1004731.

Supplemental materials can be found online at DOI 10.1371/journal.pcbi.1004731

4.2 Abstract

The vast majority of mutations in the exome of cancer cells are passengers, which do not affect the reproductive rate of the cell. Passengers can provide important information about the evolutionary history of an individual cancer, and serve as a molecular clock. Passengers can also become targets for immunotherapy or confer resistance to treatment. We study the stochastic expansion of a population of cancer cells describing the growth of primary tumors or metastatic lesions. We first analyze the process by looking forward in time and calculate the fixation probabilities and frequencies of successive passenger mutations ordered by their time of appearance. We compute the likelihood of specific evolutionary trees thereby informing the phylogenetic reconstruction of cancer evolution in individual patients. Next, we derive results looking backward in time: for a given subclonal mutation we estimate the number of cancer cells that were present at the time when that mutation arose. We derive exact formulas for the expected numbers of subclonal mutations of any frequency. Fitting this formula to cancer sequencing data leads to an estimate for the ratio of birth and death rates of cancer cells during the early stages of clonal expansion.

4.3 Introduction

In healthy tissues, cell division and cell death are tightly controlled processes, which enable a precise balance assuring that the number of cells in the body remains approximately constant. However, during each cell division mistakes in DNA replication can occur, leading to accumulation of mutations in individual cells [257,86]. The majority of such mutations are effectively neutral (passengers), but some of them (drivers) can provide selective advantage to the cell, by tipping the balance of cell division and death slightly in favor of increased proliferation [162,18,32]. This unwanted evolution [35,195,191] of somatic cells can lead to a clonal expansion of cells with driver mutations, which can ultimately result in the formation of tumors and seeding of new lesions in distant tissues [265,224,276].

Sequencing efforts over the past decades have resulted in a compendium of genetic alterations in the exomes of common human cancers and revealed that adult cancers harbor dozens (leukemias) to hundreds (lung cancer and melanoma) somatic mutations. A typical tumor is thought to contain 2-8 driver gene mutations, with the rest being neutral passengers [265]. Unlike driver mutations, passengers cannot be attacked by conventional targeted therapy, but some of them can become targets for immunotherapy or induce resistance to treatment [211,227,36]. In addition, passenger mutations can provide information about the timing of cancer evolution in individual patients by acting as a molecular clock [279].

Recent studies found that the evolution of metastases in colorectal [121] and pancreatic cancer [279], and even the evolution of primary colorectal cancer follows largely neutral evolution [244], in which the founding cell starts a clonal expansion during which cells accumulate neutral mutations (passengers) rather than drivers. Here we study neutral evolution during clonal expansion and show that passenger mutations can be used to infer the parameters of the tumorigenic process. Our model is a generalization of the famous Luria-Delbrück model for studying resistance mutations in bacteria [153,74]. In contrast to the original Luria-Delbrück model, in which wild type and mu-

tant populations grew deterministically and mutation occurred stochastically, in our model all cell types grow stochastically; our model also includes cell death. Many authors have studied the fully stochastic version of the Luria-Delbrück model in which all mutants are treated as a single population[57,117,137,128,129,126]. In these models there are only two cell populations: wild type and mutant. In contrast, here we study populations started by individual passenger mutations separately.

4.4 Results

We model the accumulation of passenger mutations during clonal expansion of cancer using a multi-type branching process[10,132] that starts with a single type-0 cell. All cells in the process divide with rate b and die with rate d. At each division, one of the daughter cells receives a new passenger mutation with probability u, which starts a new type. We are interested in the mutations accumulating in the exome during tumor evolution, so we are mostly interested in the mutation rate $u = 0.015$, which is the product of the normal point mutation rate per cell division ($\sim 5 \cdot 10^{-10}$) and the number of base pairs in the exome ($\sim 3 \cdot 10^{7}$)[32].

Any new mutation that appears in the population can be lost due to stochastic fluctuations. The probability that its lineage will not survive is $\delta = d/b$, the ratio of the death and the birth rates[10], and we will see later that the limiting behavior of the process is strongly dependent on δ and not the individual values of b and d. We label the mutations with surviving lineages ("successful" mutations) according to their order of appearance. We are interested in the fraction of cells harboring mutation k, for $k \geq 1$, and the phylogenetic relationships[19] between first appearing successful mutations (Figure 4.1).

Throughout the paper we mostly assume that the birth rate is $b = 0.25$ per day, a typical value for colorectal cancer[32], but all results scale accordingly for other values of b. The ratio of death and birth rates in cancer has been estimated to be on the order of $\delta = 0.72$ in fast-growing colorectal cancer

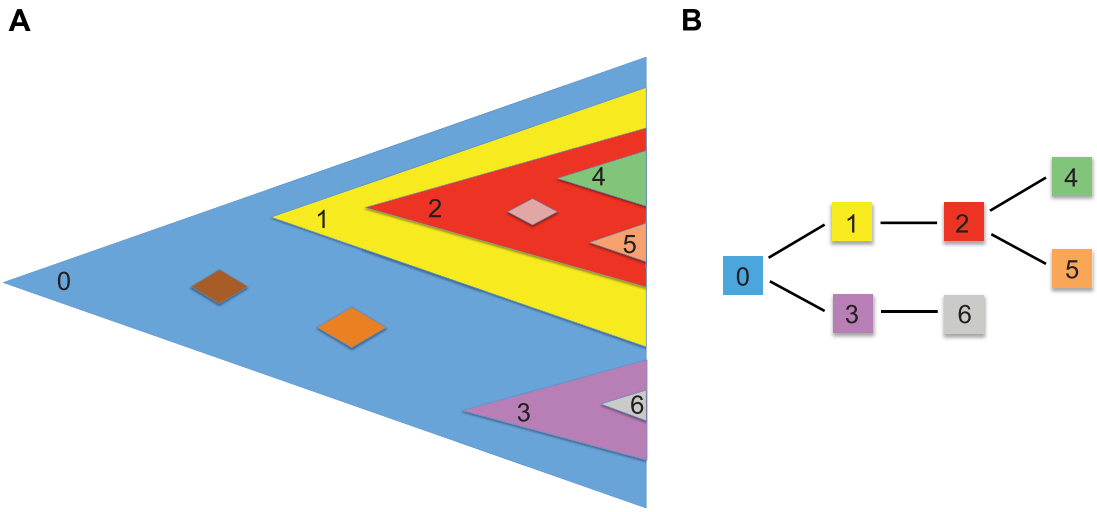


Figure 4.1: (A) New passenger mutations can be lost due to stochastic drift (diamonds). Successful mutations form surviving lineages. We order successful mutations by their time of appearance. Individual cells can harbor many passenger mutations and various different phylogenies can arise (B). In the example shown, mutation 2 appears in a cell that already harbors mutation 1. Thus all cells that have mutation 2 also have mutation 1. Similarly, all cells cells that have mutations 4 or 5 also harbor mutations 1 and 2. Mutation 3 forms an independent clone. We calculate the likelihood of different phylogenies and the expected number of subclonal mutations of any frequency.

metastases[75] to $\delta = 0.99$ in early tumors[32], and we focus on δ values in this range.

The average number of passenger mutations with surviving lineage that are present in a population of M cells is Mu[36]. For a tumor containing $M = 10^9$ cells (~ 1 cm in diameter) the number of passenger mutations is $\sim 1.5 \cdot 10^7$. The number of mutations is thus almost half of the length of the exome, but the vast majority of those mutations are present in a small number of cells, far below the detection limit of current sequencing technologies[93]. We are mostly interested in the mutations present in a sizable fraction of the population (above 0.1% of all cells). We assume an infinite allele model, in which each mutation can appear only once.

We perform Monte Carlo simulations of the multitype branching process using the Gillespie algorithm[95]. Between 5,000 and 10,000 surviving runs are used for each parameter combination.

4.4.1 Probability of fixation of new mutations.

In a pure birth process, $d = 0$, the founding cell (type-0 and no other mutations) is always present in the population, and thus all new mutations are subclonal; they are present in less than 100% of tumor cells. However, with death rate $d > 0$, new mutations appearing during clonal expansion can reach fixation in the population. We show in Methods that the probability that the k-th mutation with surviving lineage eventually fixates and becomes present in all cells is given by

$$\rho_k \approx \left(\frac{u}{u - \log \delta} \right)^k . \tag{4.1}$$

If the k-th surviving mutation reaches fixation, it is implied that all preceding $k-1$ surviving mutations (labeled 1 to $k-1$) also reach fixation. Therefore each cell in the lesion has the first k mutations.

From formula (4.1) we see that the probability of fixation increases with both the mutation rate, u, and the death-birth ratio, δ. Assuming normal mutation rate in the exome we have $u = 0.015$. Thus, in a fast-growing population, in which δ is significantly smaller than 1, it is unlikely that any

new mutation reaches fixation. For example, for $\delta = 0.72$, the probability that the first appearing mutation with surviving lineage reaches fixation is $\rho_1 \approx 0.04$. For $\delta = 0.96$ this probability is $\rho_1 \approx 0.27$. When growth is particularly slow, for example $\delta = 0.99$, the first mutation with surviving lineage has a $\rho_1 \approx 0.6$ chance of fixation, while the second has a $\rho_2 \approx 0.36$ chance and even the fifth can fixate with probability $\rho_5 \approx 0.08$.

4.4.2 Frequency and phylogenies

We show in Methods that the cumulative distribution function for the frequency of cells with the k-th mutation is given by

$$F_k(\alpha) \approx 1 - \left(\frac{u}{u - \log[1 - \alpha(1-\delta)]} \right)^k \tag{4.2}$$

for $0 < \alpha < 1$. Formula (4.2) is the probability that the frequency of cells carrying the k-th mutation is less than α. Note that $F_k(\alpha)$ does not approach 1 as $\alpha = 1$ due to a non-zero fixation probability. The excellent agreement between Formula (4.2) and exact computer simulations of the stochastic process is shown in Figure 2A,B. The fixation probability of the k-th mutation is precisely $\rho_k = 1 - F_k(1)$. For fast growing tumors, $\delta = 0.72$, the median frequencies of the first three surviving mutations are all smaller than 5% (Figure 4.2A). In contrast, for slow growing tumors, $\delta = 0.99$, the median frequencies of the first three mutations are all greater than 40% (Figure 4.2B).

Another significant difference between slow growing and fast growing tumors is exhibited by the phylogenetic relationships among the first surviving mutations. When $\delta = 0.72$, the most likely phylogeny including the founding population (type-0) and the first surviving mutations is star-like (first tree in Figure 4.3A). In contrast, when $\delta = 0.99$, the most likely phylogeny is linear (last tree in Figure 4.3A).

Formulas for the probabilities of all six phylogenetic trees involving the founding population

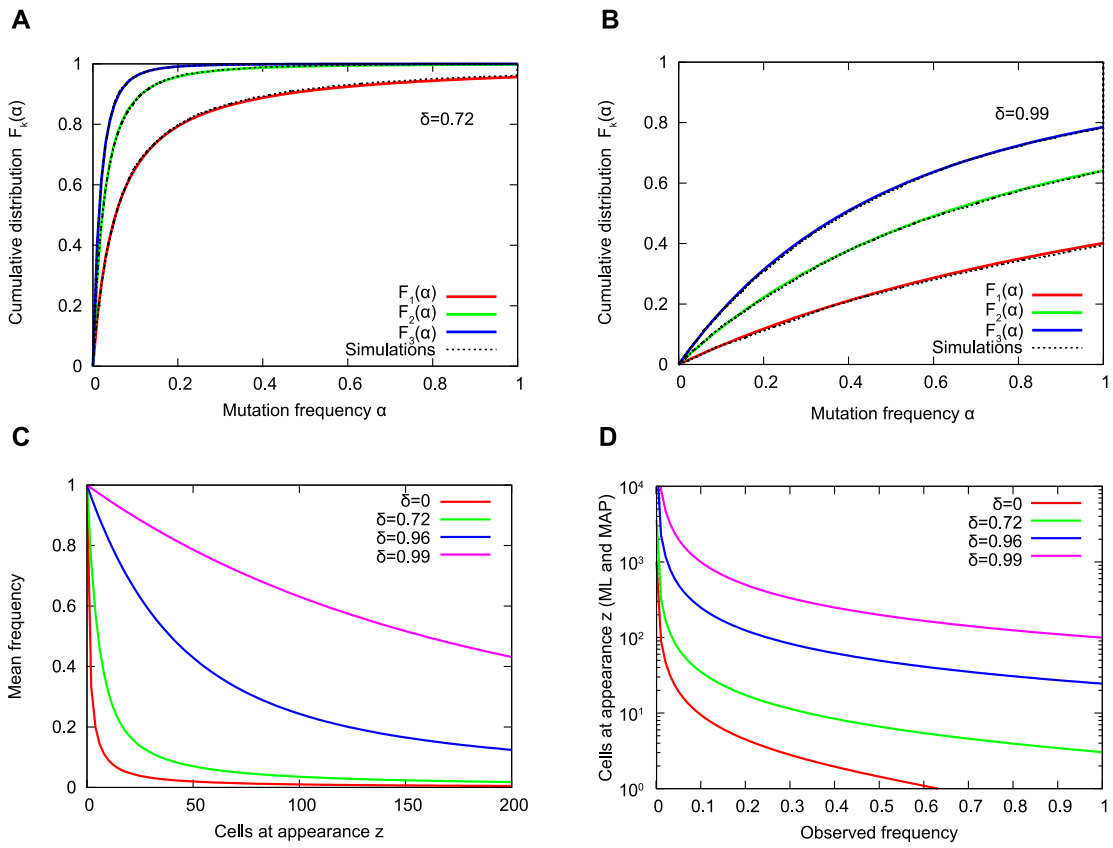
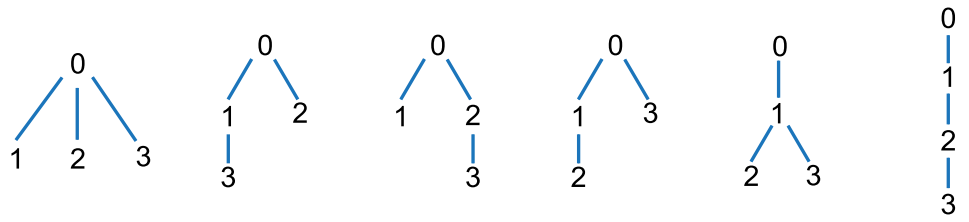


Figure 4.2: (A-B) Cumulative distribution function for the first three successful mutations. The y-axis shows the probability that the mutation has a frequency of less than α. Comparison between formula (4.2) and exact computer simulations of the stochastic process with death-birth ratios $\delta = 0.72$ (A) and $\delta = 0.99$ (B). For $\delta = 0.72$, the median frequencies of the first three successful mutations are below 5%. For $\delta = 0.99$, they are all above 40%. (C-D) Mutation frequency versus time of appearance. (C) Mean frequency attained by a mutation which arose when there were z other cells in the population, for different values of the death-birth ratio, δ. (D) Maximum likelihood and maximum a posteriori estimate (which are approximately equal) for the number of cells in the population when the mutation with frequency α arose. Passenger mutation rate $u = 0.015$ (product of the number of basepairs in the exome, $L \sim 3 \cdot 10^7$, and the normal point mutation rate during cell division, $\mu \sim 5 \cdot 10^{-10}$).

A



B

$\delta = d/b$						
0.7	71%	10%	4%	10%	4%	1%
0.97	20%	15%	10%	15%	17%	23%
0.99	5%	10%	8%	10%	15%	52%

Figure 4.3: (A) All six phylogenetic trees containing the first three surviving passenger mutations are shown. (B) Probabilities of each tree for different values of the death-birth ratio, δ (formulas shown in Methods). For $\delta = 0.72$, the first tree is the most likely. For $\delta = 0.99$, the sixth tree is the most likely. For intermediate $\delta = 0.97$, the most likely tree shape is that of trees 2-4. Passenger mutation rate $u = 0.015$.

and the first three surviving mutations are given in Methods. In Figure 4.3 we plot all six trees and their probabilities for various values of δ. For all values of δ, either the first or the last tree are the most likely. However, if we do not possess the knowledge of the order in which the first mutations appeared, then trees 2-4 (Figure 4.3A) are indistinguishable, and for intermediate δ (i.e. $\delta = 0.97$), the shape of trees 2-4 will be the most likely.

4.4.3 Frequency and time of appearance

Let us now assume that there were z other cells in the population when a certain mutation (with surviving lineage) appeared. In Methods we calculate the probability distribution for the eventual frequency of that mutation. We show that the expected frequency that the mutant eventually achieves is

$$\mathrm{E}(x) = \frac{1-\delta^{z+1}}{(1-\delta)(z+1)}. \tag{4.3}$$

If $\delta = 0$, which means no cell death, $d = 0$, the expected frequency is $1/(z+1)$. However, for $\delta > 0$ and especially when δ is close to 1, which is the most relevant case for cancer growth, the

expected frequency is much higher than $1/(z+1)$. For example, consider a mutation (with surviving lineage) which appears when there are 100 other cells present in the tumor; for $\delta = 0$ this mutations reaches an expected frequency of $E(x) = 0.01$; for $\delta = 0.99$ it reaches an expected frequency of $E(x) = 0.63$ (Figure 4.2C).

So far we have studied the branching process by looking forward in time. We now derive several results that are useful for inferring knowledge about the early evolutionary history of the tumor obtained from data at late stages of its evolution. To start, we ask an inverse question: if a mutation is present at frequency α, when did it first appear? We show in Methods that the maximum likelihood and maximum a posteriori estimates for the number of cancer cells that were present at the time when that mutation arose are approximately the same and given by

$$\hat{z}_{\text{MAP}} \approx \hat{z}_{\text{ML}} = -\frac{1}{\log[1-\alpha(1-\delta)]}. \tag{4.4}$$

Note that the estimated number of cells at appearance increases with δ (Figure 4.2D).

We see from formula (4.4) that a mutation that is present in 10% of the population has most likely appeared when there were as few as 10 cells (if $\delta = 0$ which is unlikely in cancer) to as many as 1000 cells (if $\delta = 0.99$). Similarly, a mutation that is present in 50% of cells most likely appeared when there were as few as 1 other cell (if $\delta = 0$) to as many as 200 cells (if $\delta = 0.99$).

4.4.4 Expected number of clonal and subclonal mutations

We prove in Methods that the expected number of subclonal mutations present at a frequency larger than α is

$$\overline{m}_s = \frac{u(1-\alpha)}{(1-\delta)\alpha} \tag{4.5}$$

Similarly, the expected number of clonal passenger mutations is given by

$$\overline{m}_c = \frac{\delta u}{1-\delta} \tag{4.6}$$

The number of sublclonal mutations is highly dependent on the ratio δ. When there is no cell death, $\delta = 0$, there is on average only a single passenger mutation that is present at a frequency higher than 1%. On the other hand, if $\delta = 0.99$, there will be about 150 passenger mutations present in more than 1% of cancer cells (Table 4.1).

Table 4.1: Expected number of subclonal and clonal mutations for different values of $\delta = d/b$.

δ	$> 0.1\%$	$> 1\%$	$> 10\%$	$> 50\%$	Clonal
0	15.0	1.5	0.14	0.015	0
0.72	53.5	5.3	0.48	0.05	0.04
0.96	374.6	37.1	3.37	0.38	0.36
0.99	1498.5	148.5	13.5	1.5	1.48
0.999	14985	1485	135	15	15

Values calculated using formulas (4.5) and (4.6).We assumed normal point mutation rate in the exome $u = 0.015$.

Formula (4.5) can be fitted to cancer sequencing data to determine how well the branching process model of neutral evolution describes the observed mutation frequencies and to extract the most likely parameters of the process. We fit our formula to the TCGA (http://cancergenome.nih.gov/) colorectal cancer dataset, publicly available at https://dcc.icgc.org/releases/current/Projects/COAD-US. All samples were classified as either microsatellite-stable (MSS) or instable (MSI) based on the sample's total number of mutations [252], and their purity and ploidy have been assessed [284]. We required samples with ploidy between 1.8 and 2.2 so that the cancer was not too far from diploid and chromosomal instability and LOH did not significantly alter the distribution of allele frequencies. We further required a purity estimate of at least 70%. A total of 42 samples passed both of these cri-

teria and were fit to our formula, after adjusting their allele frequency to account for sample purity.

In Figure 4.4 we plot the number of mutations with allele fraction between 0.12 and 0.25, found in two colorectal cancer samples from the TCGA dataset. Mutations with allele frequency of 25% or more may be clonal (for example, a heterozygous mutation present in one copy of a tetraploid chromosome). On the other hand, mutations with allele frequency below 10% can be difficult to detect (which translated to 12% corrected frequency as the average purity of our samples is 85%). In [272], the authors fit the same data to a formula they derived using a deterministic approximation with no cell death.

Assuming that there is no loss of heterozygosity and that all mutations are present in a single allele of a diploid cell, the allele frequency of a mutation is 1/2 of its cancer cell frequency. It follows from formula (4.5) that the number of mutations with allele frequency larger than α but smaller than 0.25 is given by

$$\frac{u}{2(1-\delta)}\left(\frac{1}{\alpha}-\frac{1}{0.25}\right). \tag{4.7}$$

Out of the 42 colorectal cancer samples that passed our filtering criteria, 16 had fits with with $R^2 \geq 0.9$, and we show them in Supplementary Figure 1. For the two cancers in Figure 4.4, the best fit is obtained for $a = u/(1-\delta) = 1.23$ (Figure 4.4A) and $a = 62.1$ (Figure 4.4B). More generally, the median value for a in MSS cancers is 2.86, and 27.61 for MSI cancers. For MSS cancers, assuming a normal passenger mutation rate in the exome $u = 0.015$ leads to birth-death ratio $\delta = 0.997$. This value is between the estimates of net proliferation rates in premalignant colorectal tumors ($\delta \sim 0.999$) and colorectal cancers ($\delta \sim 0.99$) obtained from cancer incidence data[151]. If the mutation rate is elevated 10-fold, then the best fit is obtained with death-birth ratio $\delta = 0.97$. Note that fitting formula (4.7) and assuming death rate $d = 0$ (which implies $\delta = 0$) to the data means that the mutation rate during tumor evolution needs to be ~ 400 fold higher than the normal mutation rate, which is unlikely.

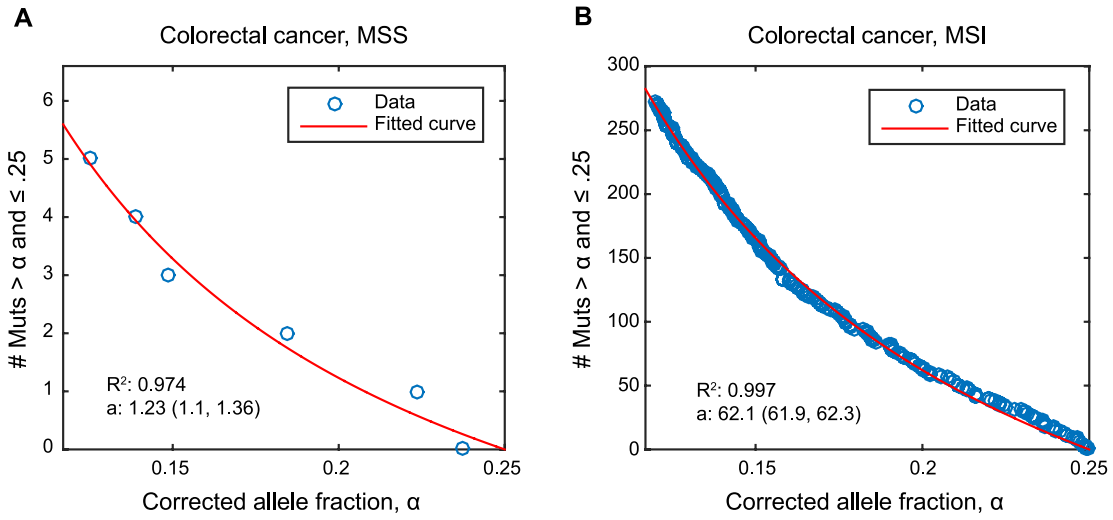


Figure 4.4: Exome sequencing data for two colorectal cancers from the TCGA dataset, (A) microsatellite stable (MSS) and (B) microsatellite instable (MSI), show the corrected allele fraction of each detected mutation (observed allele fraction divided by purity). Mutations with allele frequency of 25% or more may be clonal [272] and mutations with corrected allele frequency below 12% can be difficult to detect reliably. Thus we focus on mutations with fractions between 0.12 and 0.25, and plot the number of mutations with fraction between α and 0.25 as a function of α. The data are fit to the formula for the number of mutations with the corresponding allele frequency (4.7). The best fit for $a = \frac{u}{2(1-\delta)}$ and its corresponding 95% confidence interval is shown for each sample.

Interestingly, for δ close to 1, the number of subclonal mutations with frequency above 50% is approximately equal to the number of clonal passengers collected during tumor progression. Thus subtracting the number of subclonal mutations with frequency above 50% from the number of all clonal mutations in the cancer will be an estimate for the number of clonal mutations present in the first malignant cell.

4.5 Discussion

In summary, we have shown that the frequencies of the first and thus most abundant passenger mutations are influenced not only by the mutation rate, but also by the death-birth ratio, $\delta = d/b$, of the cancer cells. If δ is close to 1, which is the relevant case for slow overall growth, then several clonal passengers may not have been present in the first tumor cell, but were collected during clonal

expansion. The ratio δ also determines the shape of the cancer's phylogenetic tree, which is star-shaped for fast growth and linear for slow growth (Figure 3). Additionally, if we consider a mutation that has a certain observed frequency in the population of cancer cells, we can ask: how many cells were present when that mutation arose? The answer varies hundred-fold as the death-birth ratio, δ, changes from 0 to 0.99. This is particularly relevant as high levels of cell death are reported in both malignant and premalignant tissues[63,127].

In this work, we derive a simple form (4.7) for the cumulative distribution of mutant allele frequencies under a neutral model of cancer evolution. Fitting the cumulative distribution of allele fractions in a sample to this functional form and analyzing the goodness of fit provides information about the nature of the process underlying the generation of mutant alleles. We computed this distribution with data from each patient by first dividing all allele fractions in a single patient by the sample purity and then restricting our view to allele fractions in the interval [.12, .25]. The high end of interval was chosen to minimize the chance clonal mutations appeared in it and the low end was chosen to ensure that sequencing was powered to detect such mutations. The cumulative distribution of allele fractions in this interval was computed at each mutant allele fraction as the number of mutations greater than that fraction but less than or equal to the cutoff at 0.25. Finally, we performed a least-squares fit to the functional form described using MATLAB and report fits with R^2 values at least 0.9.

Our model is applicable to individual cancers in which there are no subclonal drivers at observable cell frequencies. This includes both liquid and solid tumors. It has recently been shown that colorectal tumors fit this model often[244,272]. In contrast, some liquid cancers such as chronic lymphocytic leukemia usually harbor subclonal drivers, and are thus not good candidates for the application of our model[140].

In a previous paper[36], we have studied the accumulation of individual resistance mutations in cancer using a fully stochastic Luria-Delbrück model. For targeted therapies, resistance mutations

are estimated to be rare: only about a hundred positions in the genome can give rise to resistance if mutated[75,144,37]. Therefore all results in Ref.[36] were derived in the limit of very small mutation rates, about $\sim 10^{-7}$ per cell division. In that scenario, two resistance mutations are virtually never present in the same cell, and the fixation of mutations in the population is extremely unlikely. In contrast, here we are interested in the accumulation of passenger mutations in the whole exome, which leads to a large number of passenger mutations in the population. Even individual cells contain many different mutations. Therefore, new questions arise and a new mathematical approach is needed.

Our model is a continuous time version of the infinite alleles branching process introduced by Griffiths and Pakes[98], and is a special case (birth-death process) of the model studied recently by Wu and Kimmel[278]. These works were mostly interested in the limiting frequency spectrum of the process, namely the number of mutations present in j individuals as time $t \to \infty$; for example, Wu and Kimmel provide an explicit expression for the mean limiting frequency spectrum for the birth-death process in terms of the hypergeometric function. We study the same process with respect to tumor size and derive explicit expressions for the frequencies of mutations according to their order of appearance. We also study the expected number of mutations above a certain frequency, which has connections to the frequency spectrum.

Sottoriva and Graham[244] estimate the number of cells, including the new cell, that were present when a mutation with observed frequency α appeared by $1/\alpha$, using a deterministic model with no cell death. A deterministic model is always useful as it provides the simplest approach to study evolutionary dynamics. Our formula (4.4) provides the stochastic correction to their prediction.

Recently, Durrett[76] derived formulas for the expected number of passenger mutations present at a frequency larger than α. His process differs from ours as in his model mutations occur independent of cell division. Consequently, the founding cell can collect mutations prior to the first cell division, and even for $d = 0$ there could be passenger mutations that are clonal. In contrast, in our model for $d = 0$ all mutations are strictly subclonal.

In this paper we do not consider loss of heterozygosity (LOH), which implies that mutations can be lost from the cell during cell division. The rate of LOH is on the order of 10^{-6} or lower per cell division in tumors that do not have chromosomal instability (CIN)[105]. For such small LOH rate, if LOH events are neutral or deleterious, the fraction of cells that gained a mutation but then lost it will be very small, and our results would still hold; our results would also hold for LOH events that occur at higher rates but are deleterious. In a subsequent paper we will study the effect of LOH events on the evolutionary dynamics of passenger mutations, and focus on situations where they occur with high frequency (CIN)[193,136].

Similarly, here we do not consider the effect of new driver mutations that may appear during clonal expansion of cancer[32,78,215,79]. The addition of drivers may change both the observed mutational frequencies and the phylogenies of the occurring mutations. Instead we focus only on the accumulation of neutral mutations, which is the relevant case for studying the growth of metastases and even some primary tumors[279,121,272].

While cancer spends much time in clonal expansion, plateau stages are also common. The effects of plateau stages after clonal expansion on the frequency of passenger mutations can be studied using a density-dependent branching process used previously in the context of resistance to cancer therapy[31]. This density-dependent model can be approximated analytically with a two phase model: a branching process with constant birth and death rates b and d corresponding to the growth phase, and a plateau phase of length T in which birth rate of all cells is approximately equal to d. In this model, mutations present at the end of the plateau phase can be either "old" mutations that were present at the end of the growth phase, or "new" mutations that appeared during the plateau phase and were not lost. In the large time T limit, all old mutations will either reach fixation or be lost in the population, but for shorter times T frequencies of old mutations can be approximated by their frequencies at the end of the exponential growth phase. On the other hand, the number of new mutations that are present above a certain frequency can be studied analytically using techniques

from[31] to show that new mutations will in general not be present at observable frequencies if the population size at the plateau is on the order of millions of cells or higher. Rodrigues-Brenes et al. recently studied inhibited cancer growth in the context of stem-cell driven cancers[225].

In our model we do not explicitly take into account the possible existence of a differentiation hierarchy within the cancer population: namely, the existence of cancer stem cells, which are able to propagate the tumor population indefinitely, and differentiated cells, which have limited lifespans. However, our results can also inform the study of this more complicated situation, and our analysis in fact applies to cancer stem cells. To add differentiated cells, we can consider the model in which, in addition to our basic assumptions, stem cells can also divide to produce one stem cell and one differentiated cell of generation 1, differentiated cells of generation i divide to produce two differentiated cells of generation $i+1$ and the lifespan of differentiated cells is n divisions (i.e. differentiated cells of generation n are lost from the population). A reasonable estimate for the number of divisions before mitotic arrest is $n = 10$ (e.g. 4-6 divisions in colon[208] and 15-20 in hematopoietic system[22]), which means that each differentiated cell of generation 1 produces ∼1000 cells before they are lost from the population. Typical detectable tumors contain billions of cells; mutations that occur in the lineage of a single differentiated cell will remain confined to that lineage, which will contain no more than ∼1000 cells (or no more than $\sim 10^6$ cells if $n = 20$), and will not reach a significant fraction in the population (less than $1/10^6$ for $n = 10$ or less than $1/10^3$ for n as high as 20). Hence mutations appearing in the lineages of differentiated cells will not be present at frequencies above 0.1% or 1% that we are interested in - mutations above these frequencies will be only those in the stem cell population, which will behave as described in our model. The only adjustment that needs to be made to our results when referring to cancer stem cells is the adjustment of the mutation rate to account also for mutations that occur to stem cells during asymmetric divisions.

We recently developed a spatial version of the model studied in this paper, which we mostly analyzed through computer simulations[267]. In this spatial model, tumor growth occurs on a 3-

dimensional lattice and birth rate is reduced in the presence of many neighboring cancer cells. This results in the inside of the tumor being in the state of equilibrium between birth and death, while the surface of the tumor is able to expand. If the effective birth rate of cells in the surface of the spatial model is comparable to the birth rate in the well-mixed model, there will be more mutations present above a certain frequency in the spatial case, as the spatial tumor experienced more divisions to reach the same size. However, the addition of migration of tumor cells allows cells to explore less crowded spatial positions and in turn reduces the number of divisions needed to reach the same size as well as the number of mutations above a certain frequency, bringing this model closer to the model without space[267].

4.6 Methods

4.6.1 Eventual fraction and time of appearance

We are interested in the eventual fraction of cells carrying a successful mutation, which appeared when there were z other cells in the population. Let Y be the population started by these z cells (i.e. cells without the mutation) and X the population carrying the mutation. The probability that exactly i out of z non-mutant cells have surviving lineage is

$$\pi_i = \binom{z}{i}(1-\delta)^i \delta^{z-i}. \tag{4.8}$$

When $i = 0$, the non-mutant fraction dies out and the eventual fraction of the mutant is 1. For $i \geq 1$, the number of non-mutant cells $Y \approx e^{(b-d)t}(V_1 + \cdots + V_i)$, where $V_1, \ldots, V_i$ are independent exponentially distributed random variables with mean $b/(b-d)$[36,79] and time t is large and measured from the time of appearance of the mutant. In other words, the number of cells

without the mutation is given by

$$Y = \begin{cases} e^{(b-d)t}(V_1 + \cdots + V_i) & \text{with probability } \pi_i,\ i \geq 1 \\ 0 & \text{with probability } \pi_0 \end{cases} \tag{4.9}$$

Similarly, $X \approx e^{(b-d)t}V$, where V is again exponentially distributed random variable with mean $b/(b-d)$. Thus when the number of non-mutant cells with surviving lineage $i > 0$, the eventual fraction of cells with the mutation is

$$x = \frac{X}{X+Y} = \frac{V}{V+V_1+\cdots+V_i} = \beta[1, i], \tag{4.10}$$

where $\beta[1, i]$ is a beta-distributed random variable with probability density function $i(1-w)^{i-1}$ [36]. This allows us to calculate the probability that the fraction of the population carrying the mutation is smaller than α, for $0 < \alpha < 1$:

$$\text{Prob}[x \leq \alpha | Y(0) = z] \approx \sum_{i=1}^{z} \binom{z}{i} (1-\delta)^i \delta^{z-i} (1-(1-\alpha)^i) \tag{4.11}$$

$$= 1 - (1 - \alpha + \delta\alpha)^z \tag{4.12}$$

Probability density function for the fraction of mutants, the first of which appeared when there were z other cells in the population is

$$f_z(\alpha) = (\text{Prob}[x \leq \alpha | Y(0) = z])' \tag{4.13}$$

$$= (1-\delta)z(1-\alpha+\delta\alpha)^{z-1} \tag{4.14}$$

Then the mean fraction that the mutant will eventually achieve in the population is

$$\mathrm{E}(x) = \delta^z + \int_0^1 \alpha f_z(\alpha)\, d\alpha = \frac{1-\delta^{z+1}}{(1-\delta)(1+z)} \tag{4.15}$$

We can obtain the maximum likelihood (ML) estimate for the number of cells that were present in the population when the mutation that is present at a fraction α (for $\alpha < 1$) appeared, by maximizing the probability distribution for the mutant fraction $f_z(\alpha)$ (4.14):

$$\hat{z}_{\mathrm{ML}} = -\frac{1}{\log[1-\alpha+\delta\alpha]} \tag{4.16}$$

To calculate the maximum a posteriori (MAP) estimate for the number of cells that were present in the population when the mutation that is present at a fraction α appeared, we let v be the probability that a particular single mutation appears during cell division. Then v is also the probability that this mutation apperas in the population when there are z cells and forms a surviving lineage, for all $z \geq 1$. We note that v is very small and on the order of 10^{-9}. The probability that the successful mutation first appeared when there were z cells is $p(z) = v(1-v)^{z-1}$ so to get the MAP estimate we will maximize $p(z)f_z(\alpha)$:

$$\hat{z}_{\mathrm{MAP}} = -\frac{1}{\log[1-v]+\log[1-\alpha+\delta\alpha]} \approx -\frac{1}{\log[1-\alpha+\delta\alpha]} \tag{4.17}$$

since v is very small.

Probability of fixation of k-th mutation

In a pure birth process (with $d = 0$) the founding population (type-o and no other mutations) will always be present in the population, and thus all mutations will be present in less than 100% of tumor cells. However, when death rate $d > 0$, new mutations appearing during clonal expan-

sion can reach fixation in the population. Probability that the k-th mutation with surviving lineage eventually fixates and becomes present in all cells is approximately given by

$$\rho_k \approx \int_0^\infty \frac{(zu)^{k-1} e^{-zu} u}{(k-1)!} \delta^z dz. \tag{4.18}$$

Here we use the fact that the population sizes at which mutations with surviving lineage appear can be approximated by a Poisson process on $[0, M]$ with rate u[36], that final population size M is large and that if k-th mutation is produced when there are z other cells in the population, it will reach fixation if and only if the lineages of the other z cells die out. Evaluating the integral above we obtain

$$\rho_k \approx \left(\frac{u}{u - \log \delta} \right)^k. \tag{4.19}$$

We can obtain the maximum likelihood (ML) estimate for the number of cells that were present in the population when the mutation that is present at a fraction α (for $\alpha < 1$) appeared, by maximizing the probability distribution for the mutant fraction $f_z(\alpha)$ (4.14):

$$\hat{z}_{\mathrm{ML}} = -\frac{1}{\log[1 - \alpha + \delta\alpha]} \tag{4.20}$$

To calculate the maximum a posteriori (MAP) estimate for the number of cells that were present in the population when the mutation that is present at a fraction α appeared, we let v be the probability that a particular single mutation appears during cell division. Then v is also the probability that this mutation apperas in the population when there are z cells and forms a surviving lineage, for all $z \geq 1$. We note that v is very small and on the order of 10^{-9}. The probability that the successful mutation first appeared when there were z cells is $p(z) = v(1-v)^{z-1}$ so to get the MAP estimate we

will maximize $p(z)f_z(\alpha)$:

$$\hat{z}_{\text{MAP}} = -\frac{1}{\log[1-v] + \log[1-\alpha+\delta\alpha]} \approx -\frac{1}{\log[1-\alpha+\delta\alpha]} \tag{4.21}$$

since v is very small.

Probability of fixation of k-th mutation

In a pure birth process (with $d = 0$) the founding population (type-0 and no other mutations) will always be present in the population, and thus all mutations will be present in less than 100% of tumor cells. However, when death rate $d > 0$, new mutations appearing during clonal expansion can reach fixation in the population. Probability that the k-th mutation with surviving lineage eventually fixates and becomes present in all cells is approximately given by

$$\rho_k \approx \int_0^\infty \frac{(zu)^{k-1}e^{-zu}u}{(k-1)!}\delta^z dz. \tag{4.22}$$

Here we use the fact that the population sizes at which mutations with surviving lineage appear can be approximated by a Poisson process on $[0, M]$ with rate u[36], that final population size M is large and that if k-th mutation is produced when there are z other cells in the population, it will reach fixation if and only if the lineages of the other z cells die out. Evaluating the integral above we obtain

$$\rho_k \approx \left(\frac{u}{u - \log\delta}\right)^k. \tag{4.23}$$

Fraction of cells with k-th mutation

Having already characterized the probability that the k-th passenger mutation reaches fixation in the population, we will now investigate the size of the population with the k-th mutation when it is

subclonal.

We can derive the cumulative distribution function for the fraction of cells with the k-th mutation by again using the fact that the population sizes at which mutations with surviving lineage appear can be approximated via a Poisson process on $[0, M]$ with rate u, where M is the final population size[36], together with result (4.12).

$$\text{Prob}[x_k \leq \alpha] \approx \int_0^\infty \frac{(zu)^{k-1} e^{-zu} u}{(k-1)!} [1 - (1 - \alpha + \delta\alpha)^z] dz \tag{4.24}$$

$$= 1 - \left(1 - \frac{\log(1 + (-1 + \delta)\alpha)}{u}\right)^{-k} \tag{4.25}$$

From here we can derive the median fraction of cells with the k-th mutation

$$\text{Med}(x_k) = \min \left\{ 1, \frac{1 - e^{u(1-2^{\frac{1}{k}})}}{1 - \delta} \right\} \tag{4.26}$$

Trees

In addition to the numbers of cells carrying specific mutations, we will also investigate the phylogenetic relationships between neutral mutations in tumors. We will show that the likelihood of a particular configuration depends on the parameters on the process.

We first calculate the probability that mutation 2 appears in the lineage of mutation 1 (and not 0). We will use the approximation that the probability that mutation 2 is offspring of mutation 1 is equal to the eventual fraction of cells with mutation 1 in the population. Then the probability that mutation 2 appears in the lineage of mutation 1 is

$$p_{1\to 2} \approx \text{E}(x_1) = \rho_1 + \int_0^1 \alpha g_1(\alpha) d\alpha, \tag{4.27}$$

where g_1 is the probability distribution function for the fraction of cells with the first mutation. In other words, g_1 is the derivative of the cumulative distribution function given in (4.25) for $k = 1$. For $b = 0.25$, $d = 0.18$ and $u = 0.015$, $p_{1\to 2} = 0.15$, while for $d = 0.2475$ and b and u same as before, $p_{1\to 2} = 0.77$, in excellent agreement with simulations results.

Next we want to estimate the probabilities of each of the six trees (Figure 3) involving the first three (successful) passenger mutations. For these calculations we will need the following two quantities, sub_1 and sub_2, mean fractions of cells with the first and the second mutation, conditioned on their subclonality. Mean fraction of cells with the first mutation, conditioned on that mutation being subclonal, is simply

$$\mathrm{sub}_1 = \frac{\int_0^1 \alpha g_1(\alpha)\, d\alpha}{1-\rho_1}. \tag{4.28}$$

On the other hand, mean fraction of cells with the second mutation, conditioned on that mutation being subclonal, is

$$\mathrm{sub}_2 = \frac{\int_0^1 \alpha g_2(\alpha)\, d\alpha}{1-\rho_2}, \tag{4.29}$$

where g_2 is the derivative of the cumulative distribution function given in (4.25) for $k = 2$.

Table 4.2: Likelihood of phylogenetic trees

	δ	1	2	3	4	5	6
Formulas	0.72	71.1%	9.8%	3.8%	9.8%	4.5%	1.2%
	0.96	26.4%	15.2%	9.9%	15.2%	16.1%	17.2%
	0.99	5.5%	9.7%	8.2%	9.7%	15.1%	51.7%
Simulations	0.72	73.9%	7.7%	3.5%	7.5%	6.2%	1.1%
	0.96	30.7%	12.7%	8.8%	12.7%	18.1%	16.9%
	0.99	7.4%	8.9%	7.6%	9.2%	15.2%	51.2%

Probability of each of the six trees for different values of death-birth ratio δ. Probabilities obtained using formulas from this section and results from 10,000 (surviving) runs of the computer simulation. Parameters: birth rate $b = 0.25$ and passenger mutation rate $u = 0.015$.

We will also need sub_2^{1c}, mean fraction of cells with mutation 2, conditioned on 1 being clonal and 2 not being clonal, and sub_2^{1nc}, mean fraction of cells with mutation 2, conditioned on 1 not being clonal and 2 not being clonal. $\mathrm{sub}_2^{1c} = \mathrm{sub}_1$ and since

$$\mathrm{E}(x_2) = \rho_1^2 + (1-\rho_1^2)\mathrm{sub}_2 = \rho_1^2 + \rho_1(1-\rho_1)\mathrm{sub}_2^{1c} + (1-\rho_1)\mathrm{sub}_2^{1nc} \tag{4.30}$$

we have

$$\mathrm{sub}_2^{1nc} = (1+\rho_1)\mathrm{sub}_2 - \rho_1\mathrm{sub}_1. \tag{4.31}$$

We will start with calculating the probability of tree 2, p_2, in which mutation 2 is not offspring of 1 (event A) and mutation 3 is offspring of 1 (event B). We will denote the event that mutation 1 does not fix as C. Then $A \subset C$ and

$$\begin{aligned} p_2 &= P(A \cap B) = P(A \cap B \cap C) \\ &= P(C)P(A|C)P(B|A \cap C) \\ &\approx P(C)P(A|C)P(B|C) \end{aligned} \tag{4.32}$$

In other words, we approximate $P(B|A \cap C)$ with $P(B|C)$ and obtain

$$p_2 \approx (1-\rho_1)(1-\mathrm{sub}_1)\mathrm{sub}_1. \tag{4.33}$$

Thus tree 2 occurs only when mutation 1 is subclonal (which occurs with probability $1-\rho_1$), mutation 2 is offspring of mutation 0 (which occurs with probability $1-\mathrm{sub}_1$) and mutation 3 is offspring of mutation 1 (which occurs with probability $\approx \mathrm{sub}_1$). When calculating the probabilities of individual trees, we again use the approximation that the probability that e.g. mutation 2 is offspring of mutation 1 is equal to the eventual fraction of cells with mutation 1 in the population.

Similarly, the probability of tree 4, p_4, in which mutation 2 is offspring of 1 and mutation 3 is offspring of 0 is given by

$$p_4 \approx (1-\rho_1)\mathrm{sub}_1(1-\mathrm{sub}_1) \approx p_2. \tag{4.34}$$

We will calculate the probability of tree 1, p_1, in a similar manner. Let A and C be the same events as above and let event B now be that mutation 3 is not an offspring of either 1 or 2. Then

$$\begin{aligned} p_1 &= P(A \cap B) = P(A \cap B \cap C) \\ &= P(C)P(A|C)P(B|A \cap C) \\ &\approx (1-\rho_1)(1-\mathrm{sub}_1)(1 - E(x_1|A \cap C) - E(x_2|A \cap C)) \\ &\approx (1-\rho_1)(1-\mathrm{sub}_1)(1-\mathrm{sub}_1 - \mathrm{sub}_2^{1nc}) \end{aligned} \tag{4.35}$$

Thus we have

$$p_1 \approx (1-\rho_1)(1-\mathrm{sub}_1)(1-\mathrm{sub}_1 - (1+\rho_1)\mathrm{sub}_2 + \rho_1\mathrm{sub}_1) \tag{4.36}$$

Similarly, the probability of tree 3, p_3, is

$$\begin{aligned} p_3 &\approx (1-\rho_1)(1-\mathrm{sub}_1)\mathrm{sub}_2^{1nc} \\ &= (1-\rho_1)(1-\mathrm{sub}_1)((1+\rho_1)\mathrm{sub}_2 - \rho_1\mathrm{sub}_1) \end{aligned} \tag{4.37}$$

We now turn to calculating the probability of tree 5, p_5. Tree 5 can occur when mutation 1 either fixes or does not fix. Mutation 1 fixes with probability ρ_1, and then mutation 2 must not fix in the population with mutation 1 (which occurs with probability $1-\rho_1$) and mutation 3 must not be offspring of mutation 2 (which occurs with probability $1-\mathrm{sub}_2^{1c} = 1-\mathrm{sub}_1$). If mutation 1 does not fix (which occurs with probability $1-\rho_1$), then mutation 2 is offspring of mutation 1 with

probability sub_1, mutation 2 does not fix in the population with mutation 1 with probability $1-\rho_1$ and mutation 3 is offspring of mutation 1 but not 2 with probability $\approx \mathrm{sub}_1 - \mathrm{sub}_2^{1nc2nc1}$. Here $\mathrm{sub}_2^{1nc2nc1}$ is the mean fraction of cells with mutation 2, conditioned on mutation 1 not being clonal and mutation 2 being offspring but not clonal in 1. We have

$$\begin{aligned} p_5 &\approx \rho_1(1-\rho_1)(1-\mathrm{sub}_2^{1c}) + (1-\rho_1)\mathrm{sub}_1(1-\rho_1)(\mathrm{sub}_1 - \mathrm{sub}_2^{1nc2nc1}) \\ &= \rho_1(1-\rho_1)(1-\mathrm{sub}_1) + (1-\rho_1)^2\mathrm{sub}_1(\mathrm{sub}_1 - \mathrm{sub}_2^{1nc2nc1}) \\ &\approx (1-\rho_1)[\rho_1(1-\mathrm{sub}_1) + (1+\rho_1)\mathrm{sub}_1(\mathrm{sub}_1 - \mathrm{sub}_2)] \end{aligned} \tag{4.38}$$

In the last equality we used the fact that

$$\mathrm{sub}_2^{1nc2nc1} \approx \frac{\mathrm{sub}_2^{1nc} - \rho_1\mathrm{sub}_1}{1-\rho_1}. \tag{4.39}$$

Using similar reasoning and 4 scenarios: 1) mutation 2 fixes, 2) mutation 1 fixes, but mutation 2 does not, 3) mutation 1 does not fix but mutation 2 fixes in the population with mutation 1 and 4) mutation 1 does not fix, mutation 2 does not fix in the population with mutation 1 we obtain

$$\begin{aligned} p_6 &\approx \rho_2 + \rho_1(1-\rho_1)\mathrm{sub}_1 + (1-\rho_1)\rho_1(\mathrm{sub}_1)^2 + (1-\rho_1)^2\mathrm{sub}_1\mathrm{sub}_2^{1nc2nc1} \\ &\approx (\rho_1)^2 + \rho_1(1-\rho_1)\mathrm{sub}_1(1-\mathrm{sub}_1) + (1-\rho_1^2)\mathrm{sub}_1\mathrm{sub}_2 \end{aligned} \tag{4.40}$$

Expected number of subclonal mutations

Let u_z be the probability that, when there are z total cells in the population, a new mutation is produced that will become subclonal and present in a fraction larger than α. The probability that a new mutation with surviving linage is produced before going to $z-1$ or $z+1$ cells is $bu/(b+d)\cdot(1-d/b) = u(1-\delta)/(1+\delta)$ for $z>1$. When $z=1$, the probability that a new mutation with surviv-

ing lineage is produced before going to 2 cells (the only option in the process without extinction) is $u \cdot (1 - d/b) = u(1 - \delta)$. For all $z > 0$, probability that the newly produced mutation is subclonal is $1 - (d/b)^z$.

Using formula (4.12), probability that a subclonal mutation with surviving lineage, that appeared when there are z other cells, is present in a fraction larger than α is

$$\text{Prob}[x > \alpha | x < 1] = \frac{(1 - \alpha + \delta\alpha)^z - \delta^z}{1 - \delta^z} \tag{4.41}$$

This leads to

$$u_z = u\frac{1 - \delta}{1 + \delta}((1 - \alpha + \delta\alpha)^z - \delta^z) \tag{4.42}$$

for $z > 1$ and

$$u_1 = u(1 - \delta)((1 - \alpha + \delta\alpha)^1 - \delta^1) \tag{4.43}$$

We are interested in the expected value for the number of subclonal mutations with fraction larger than α. Let $\overline{m}_k$ be the expected value for the number of such mutations when the process starts with k cells. We are again only interested in the process that does not go extinct. Thus we have

$$\overline{m}_1 = u_1 + \overline{m}_2 \tag{4.44}$$

$$\overline{m}_k = u_k + p\overline{m}_{k+1} + (1 - p)\overline{m}_{k-1}, \tag{4.45}$$

for $k > 1$ and $p = 1/(1 + \delta)$. Expressing $\overline{m}_1$ in terms of $\overline{m}_k$ we get

$$\overline{m}_1 = u_1 \sum_{i=0}^{k-2} \left(\frac{1 - p}{p}\right)^i + \frac{u_2}{p} \sum_{i=0}^{k-3} \left(\frac{1 - p}{p}\right)^i + \cdots + \frac{1}{p} u_{k-1} + \overline{m}_k \tag{4.46}$$

Taking the limit of the right-hand side when $k \to \infty$ and noting that $\overline{m}_k \to 0$ when $k \to \infty$ we get

$$\begin{aligned}
\overline{m}_1 &= u_1 \sum_{i=0}^{\infty} \left(\frac{1-p}{p}\right)^i + \sum_{z=2}^{\infty} \frac{u_z}{p} \sum_{i=0}^{\infty} \left(\frac{1-p}{p}\right)^i \\
&= \left(u_1 + \sum_{z=2}^{\infty} \frac{u_z}{p}\right) \sum_{i=0}^{\infty} \left(\frac{1-p}{p}\right)^i \\
&= \left(u_1 + \sum_{z=2}^{\infty} \frac{u_z}{p}\right) \frac{p}{2p-1}
\end{aligned} \tag{4.47}$$

Plugging in the expressions for u_z and p we get that the expected value for the number of subclonal mutations with fraction larger than α is

$$\begin{aligned}
\overline{m}_1 &= \left(\sum_{z=1}^{\infty} u(1-\delta)((1-\alpha+\delta\alpha)^z - \delta^z)\right) \frac{1}{1-\delta} \\
&= \sum_{z=1}^{\infty} u((1-\alpha+\delta\alpha)^z - \delta^z)
\end{aligned} \tag{4.48}$$

Evaluating the last sum we obtain the expected number of subclonal mutations present in a fraction larger than α

$$\overline{m}_s = \overline{m}_1 = \frac{u(1-\alpha)}{(1-\delta)\alpha} \tag{4.49}$$

Expected number of clonal mutations

Using the same reasoning as in the previous section we can calculate the expected number of clonal mutations.

$$\overline{m}_c = \sum_{z=1}^{\infty} u\delta^z = \frac{\delta u}{1-\delta} \tag{4.50}$$

5

Evolutionary dynamics of neoantigens under immune surveillance

5.1 Forward

This work grew out of an interest in cancer immunotherapy responses which began after discussions with Luis Diaz. Hannes Reiter and I realized that we could not answer questions about how many

mutations would be sufficient for a robust immunotherapy response without working back to the founding cell of the tumor. In doing so, we developed a model that focuses on the interaction of immunity with an evolving tumor population and is less focused on the particulars of therapy response. I contributed to several aspects of this work, including the original model descriptions, code, and analysis. The simulations and analysis were improved greatly by Alex Whatley and Michael Nicholson, who contributed especially to the weak immunity model. The work benefited greatly from input and guidance from Martin Nowak and comments from Bert Vogelstein. We are preparing this work for publication.

5.2 Introduction

Genetic mutations inflict changes in all human cells over time[154]. While most mutations are passengers of no phenotypic consequence, some confer a growth advantage eventually sufficient to initiate the development of a tumor[266]. The expansion of a tumor reveals the mutations acquired during its development. When these mutations give rise to altered peptide fragments, their display on the cell surface can trigger an adaptive autoimmune response. Accordingly, the adaptive immune system has long been suspected to surveil the body for expanded populations of somatic cells and eliminate them[130], thereby reducing the incidence of cancer. Recent results in checkpoint-blockade immunotherapy confirm the presence and efficacy of endogenous, tumor-reactive cytotoxic T cells in some tumors[143,170,99]. Further, durable and complete responses to immunotherapy have been observed even when the checkpoint-blockade inhibitor is withdrawn, raising the possibility that novel immune responses continue to mount after immunotherapy is withdrawn. However, the success of checkpoint-blockade immunotherapy has so far been confined to a small subset of tumors, including mismatch repair-deficient (MMRD) colorectal cancers[143], smoking-associated lung cancers[221], and UV exposure-driven melanomas[169,220], all of which harbor relatively high numbers of detectable

mutations.

These observations have inspired the hypothesis that a sufficient number of mutation-associated neoantigens are required for robust response to immunotherapy and eventual clearance or control of a tumor[221,233]. Still, which tumors might be good candidates for immunotherapeutic intervention remains an important open question, and the landscape of neoantigens in a clinically relevant tumor remains unclear. Because many mechanisms exist to thwart immune elimination of tumor cells[200], a lack of response in other tumor types could in principle reflect the need for different systems-immunologic perturbations in order to unleash endogenous immunity. Here, we explore the effectiveness of immune surveillance and immunotherapy with a simple model of tumor immunity. The model incorporates the accumulation and loss of neoantigens as well as the growth dynamics of a tumor. The model establishes that, for most tumors, endogenous immunity is insufficient for tumor control even in the absence of barriers to immunity like resistance or HLA loss.

5.3 Model

5.3.1 Background

Human cells display endogenous protein fragments to antigen-specific T cells via the MHC class I pathway. T cell responses against unmutated protein fragments are rare because T cells must survive negative selection in the thymus, where autoreactive cells are removed. However, a genetic change in a cell can lead to production of a mutation-assocaited neoantigen, a mutant protein fragment capable of eliciting an immune response from an extant T cell clone. An expanding tumor clone harboring a neoantigen will eventually encounter antigen-specific T cells. In a successful anti-tumor immune response, these T cells expand and specifically eliminate the tumor clone harboring the neoantigen.

Tumors arise as a result of a stepwise process of accumulated genetic change. The genetics of large

tumors are consistent with a "big-bang" model of initiation in which the cells in a tumor descend from a single founder with a selective advantage and whose early growth is exponential[24,5]. Differences between an individual's germline and tumor can be partitioned according to whether they arose before this initiation event or after. Mutations acquired before initiation might also be present in other, presumably healthy, somatic cells. Mutations acquired after may be also be clonal, though the vast majority of tumor mutations will be subclonal[33]. Immune surveillance, or the sensing and elimination of tumor clones which harbor immunogenic peptides, might act on mutations which occur before or after initiation.

5.3.2 Model Description

We consider three possible models of tumor interaction with the immune system. In all the models, the initiating cell of a tumor can begin with or without neoantigens. Let the number of neoantigens in the initiating cell of a tumor be k. The tumor population grows as a branching process with cell division rate b and death rate $d (b > d)$. Upon cell division, a new neoantigen is produced in one daughter with probability u, and every neoantigen is unique. In the first model, the immune system purges neoantigens from a growing tumor immediately after they arise. This situation is illustrated in figure 5.1a. In the second case, a clone of cells harboring a particular neoantigen does not stimulate an immune response until it grows to some threshold size T. After reaching this size, the clone is quickly eliminated. Figure 5.2a illustrates this scenario. Finally, an immune response might emerge against a neoantigen which is constant but very weak compared to the growth advantage of the tumor. In this case, the death rate of cells harboring k neoantigens is increased by $k * s$. Figure 5.3a illustrates this scenario.

5.4 Results

5.4.1 Analysis

We first consider the case when a tumor is initiated in a cell without any neoantigens and no loss of neoantigens is possible. The fate of a tumor population depends on the birth and death rates of the initiated cell and the rate at which cells gain neoantigenic mutations. In this case the original cell has some chance to survive and reach detectable size if the growth rate of cells free of any neoantigens is positive, that is if

$$u < 1 - \frac{d}{b}$$

In this regime, it is possible for the population of cells without any neoantigens to sustain itself. Although cells continue to accumulate mutations, the whole population is expanding exponentially. However, when the rate of acquiring neoantigens is higher than this threshold, all cells will acquire neoantigens and immune responses will eventually drive the population extinct. Figure 5.1b shows an example simulation trajectory for this process. The elimination of neoantigenic clones slows the growth of the tumor population by exactly the rate that neoantigens are produced. Neoantigenic mutations will never be found in such a population, but the effect of their elimination will be implicitly reflected in the apparent rate of tumor growth. Using this insight and previous results for the variant allele frequency spectrum in exponentially growing populations[34], the variant allele frequency spectrum for neutral mutations under this model of immunity can be calculated. Figure 5.1c shows the results of this calculation. Both the intrinsic tumor growth rate and the rate of gaining neoantigens influence the proportion of initiated tumors which will be eliminated by the immune system, as shown in Figure 5.1d.

The immediate killing model is an optimistic possible scenario for the effectiveness of immune surveillance. However, the immune system might instead require some time or signals to mount an

effective immune response, resulting in immune pressure which is very effective but only mounts after a tumor clone harboring a neoantigen reaches a certain size T. In this model, the whole tumor might be much larger than T but be mostly devoid of neoantigens; it is only the size of the subclone harboring the neoantigen which determines whether or not an immune response specific to that neoantigen will be mounted.

For this model (Figure 5.2a), immune surveillance is equally effective at removing initiated tumor cells because every cell which acquires a neoantigen will, eventually, go extinct. However, a surviving trajectory exhibits more interesting dynamics (Figure 5.2b): large fluctuations in the size are observed early on during moments of immune activation. Two phases of growth can be observed, corresponding to the times before and after immunological pressure reduces the overall growth rate of the tumor. The expected size of the tumor is shown for different thresholds in Figure 5.2c, highlighting the different regimes. The threshold also influences a simulated immunotherapy treatment response (Figure 5.2d) in which a tumor grows without any immune pressure and then is subject to immune pressure under the threshold model at a later time. Tighter thresholds produce a more dramatic response.

In the third model of immune surveillance, we consider the case that an immune response only weakly curtails the growth of a clone harboring a particular neoantigen (Figure 5.3a). In this regime, it is important to account for the accumulation of neoantigens, and subsequently increased immune killing, in a particular cell. A population of cells without any neoantigens will grow exponentially, giving rise to populations with subsequently more neoantigens over time, as shown in Figure 5.3b. In the long-time limit, the proportion of the population relative to the unmutated population which harbors a particular number of neoantigens will stabilize[184]. This proportion is shown as a function of the rate of gaining neoantigens in Figure 5.3c. Surviving neutral and neoantigenic mutations will have different variant allele frequency spectra as shown in Figure 5.3d.

5.4.2 Loss of neoantigens

We have until now only considered the case in which gained neoantigens could not be lost. However, consider the case in which a cancer population with some neoantigens present in the initiated founding cell grows to a clinically detectable size as a result of some dysfunction in the immune response. If this dysfunction is subsequently reversed, what should happen to the population of tumor cells? If the above condition is not satisfied, then the immune response will certainly (though perhaps slowly) drive the population extinct under any of the immune models considered. However, if survival of an unmutated cell is possible and many such cells exist at detection, then immunotherapy might not lead to tumor eradication. Figure shows the expected number of cells without neoantigens in a tumor that is not subject to immune pressure for different mutational scenarios. When a tumor with a normal mutation rate reaches detectable size, it is expected to have produced many cells without neoantigens. When $t * v$ is large, a tumor of size N is expected to produce an unmutated cell if

$$log(N) > \frac{u}{v}$$

5.4.3 Resistance to immunity

All of the models considered so far predict the elimination of tumors which begin with a very high mutation rate. However, such tumors are frequently observed clinically. One possible explanation for their appearance is the presence of a heritable mutation which confers resistance to immunity. A possible example of this mutation class is constitutive cell surface expression of PD-L1. In the context of the threshold immune response model, we derive an exact expression for the survival probability of such an initiated cell as a function of the resistance probability and the threshold size. Figure 5.5 illustrates the resistance model and survival probabilities. When the resistance probability

γ is greater than approximately $1/T$ per generation (Figure 5.5b, right panel $T = 10^4$), The survival probability of an initiated cell is not significantly altered by immune surveillance.

5.4.4 Applications

We establish bounds on the effectiveness of immune surveillance by estimating the chance that a newly arising mutation can stimulate an immune response. In order to produce a neoantigen, a point mutation must change the amino acid sequence of a protein, then be processed and displayed on the cell surface by the antigen presentation pathway, and finally recognized by a T-cell clone specific to that antigen. 75% of newly arising mutations are coding mutations, which can potentially generate several neopeptides depending on how the protein is digested. Only a small proportion of mutant peptide fragments can be successfully presented by a particular HLA allele. Both computational predictions of binding affinity[165,166] and direct measurement of peptide presentation by mass spectrometry[16] suggest that the proportion of peptides which can be presented by a cell is very small. Pyke et al, applying NetMHCPan, find approximately 1% percent of peptides can be presented by an individual with diverse HLA alleles, though this calculation requires the choice of a binding affinity cutoff. Direct measurements of presented peptides by mass spectrometry find similarly that slightly less than 0.1% of possible peptidome fragments can be detected on the surface of human cells, corresponding to an overall presentation probability of about 1% for an amino acid substitution. Accordingly, only a small proportion of mutated peptides have corresponding expanded T-cell clones[7]. We conservatively assume that any presented neoantigen can be recognized by some T-cell clone.

This estimate suggests that many tumors are initiated on a background lacking any neoantigens: fewer than 1 in 100 single nucleotide polymorphisms are neoantigenic. Figure 2a illustrates what point-mutational burden produces a 90-percent chance of harboring at least one pre-existing neoantigen if they are distributed in this way. We note frameshift In/Dels might be more likely to produce neoantigens: a frameshift mutation resulting in a chain of 90 new amino acids can in prin-

ciple produce 100 new peptide fragments, a factor of 10 larger than an amino acid substitution. If an individual has a lower germline diversity of HLA alleles, the proportion of mutations which generate neoantigens will be smaller. Finally, it might be the case that immune surveillance acts to remove neoantigens from populations of precancerous or even healthy cells. In this case, the number of founding neoantigens in a tumor will be lower than what is expected by chance.

Using a normal mutation rate μ of $5 * 10^{-10}$ per base per cell division[154] and assuming 75% of all exonic mutations are coding mutations, the contribution of normal point mutations to the rate u is not larger than $\mu \cdot 3 * 10^7 \cdot 0.75 \cdot 0.01 \cdot 2 \approx 2 * 10^{-4}$ per cell division. While death rates in cancer branching process models have been estimated to be only slightly smaller than birth rates, the growth advantage of primary tumor cells dwarfs the probability of gaining a neoantigen. With b normalized to 1, estimates for d range between $d = 0.999$ in premalignant lesions and $d = 0.99$ in early tumors[33]. This suggests that even immediate elimination of any cell harboring a neoantigenic peptide would be insufficient to drive extinct most clinically relevant tumors. Further, if the observed dynamics of tumor growth already reflect immune surveillance, the true growth advantage of transformed cells is even higher.

Nonetheless, certain cancers exhibit elevated mutation rates, and some mutational processes are more likely to lead to the creation of neoantigens. For example, in tumors with DNA repair deficiency or environmental mutagenic pressure from smoking or UV light, the underlying mutation rate can be increased by orders of magnitude, and u can exceed the growth advantage of cancerous cells. In this regime, a tumor must contend with adaptive immunity in some way. We conclude that while the adaptive immune system might play a role in the dynamics of premalignant tissue or malignant tissue with an elevated mutation rate, it cannot arrest the growth of malignant tissue when the mutation rate is normal.

The mutation rate in certain tumors changes over time. For example, in UV-light associated melanoma, the founding tumor cell has experienced a high mutational burden, but subsequent

tumor growth shields most cells from this mutagenic pressure. Therefore, in our model this process is best represented by a founding cell with a large number of neoantigens, but a low rate of gaining new neoantigens. When this tumor evolves in the context of immune suppression and grows to a clinically detectable size, it is expected to produce cells which harbor none of the original neoantigens and no new neoantigens, despite starting with such a high number. This situation is distinct from MMRD colorectal cancer, where neoantigens continue to be accumulated at a very high rate and our model predicts cells without any neoantigens are very unlikely.

5.5 Discussion

Many additions to our model of tumor immunity are possible. For example, defects in antigen presentation machinery and immune exhaustion are well-documented mechanisms of immune evasion. However, these additions would only make controlling a nascent tumor more challenging. Likewise, continued accumulation of driver mutations after cancer transformation would produce cells tolerant of an even higher rate of gaining neoantigens. On the other hand, environmental pressures besides immunity might arrest a tumor at a carrying capacity, precipitating the accumulation of mutations. Thus, late in the life of a tumor, cells at the carrying capacity might acquire many rare neoantigens, and the ability of these cells to seed new metastases might well depend on mechanisms to evade immune pressure. Finally, we have assumed that the drivers of tumorigenesis are not themselves neoantigens. We believe this is a reasonable approximation for the vast majority of tumors because antigenic driver mutations are probably as rare as antigenic passengers and, if the immune system is acting to eliminate neoantigens, even more rarely observed. However, there is suggestive evidence that the very small proportion of driver mutations surveilled by the immune system slightly reduce cancer incidence[166], though this observation may be confounded by other factors associated with high HLA diversity.

Our model clarifies the challenge in immunologic control of a tumor: the high fidelity of DNA replication which suppresses the accumulation of cancer drivers also prevents distinguishing between cancerous and normal tissue. Given that a tumor has emerged under the influence of a normal mutation rate, it is unlikely to harbor or acquire enough neoantigens to be driven extinct. We emphasize that this result holds in the best possible case for the immune system; no amount of systems-immunologic perturbation will improve it. This situation is analogous to the error threshold model[190], though that model was originally studied in a population with fixed size. The sensitivity of the immune system establishes an error threshold for cancer cells, but in most cases cancers exist well below this threshold because the expansion of the tumor population gives the population many chances to avoid neoantigens (see Figure 5.6).

However, even in this case, immunotherapies can provide clinically meaningful benefit, especially in combination with other therapies[37]. When tumors emerge with a high mutation rate, our results underscore the ability of the immune system to lead to durable control and point to genomic features which might distinguish tumors which have already been depleted of neoantigens from those which might respond to immunotherapy.

5.6 Derivations

5.6.1 Simple Model

Let $Z(t)$ be a continuous-time exponential branching process with division rate b and death rate d, $b > d$. The long-term extinction probability p_{ext} of this process can be derived by considering a one-step conditional recurrence for a single cell:

$$p_{ext} = P(Birth) * p_{ext}^2 + P(Death)$$

$$p_{ext} = \frac{b}{b+d} * p_{ext}^2 + \frac{d}{b+d}$$

The solution to this polynomial between 0 and 1 is

$$p_{ext} = \frac{d}{b}$$

The probability that the branching process survives in the long run is $\rho = 1 - \frac{d}{b}$.

5.6.2 Mutation

We introduce neoantigens into this system as follows. Upon division, each daughter will acquire some mutations which might be neoantigens. Let $Z_k(t)$ be the number of cells with k neoantigens. Since there are very many possible mutations, we model the acquisition of neoantigens in an infinite sites framework. If the number of new neoantigens in one daughter cell is Poisson distributed with mean $u/2$, then the probability that both daughter cells have no neoantigens P_{00} is e^{-u}. The probability that one daughter has at least one new neoantigen but the other does not have any P_{01} is $2e^{-u/2}(1 - e^{-u/2})$, and the probability that both daughters have new neoantigens P_{11} is $(1 - e^{-u/2})^2$.

5.6.3 Low rate of gaining neoantigens

The rate of gaining neoantigens is quite small (we estimate that the expected number of neoantigens gained in a healthy cell per division is about $u = 10^{-4}$), so for convenience we will assume $u << 1$. In this regime,

$$P_{00} = 1 - u + O(u^2)$$

$$P_{01} = u + O(u^2)$$

$$P_{11} = O(u^2)$$

We will neglect terms of order $O(u^2)$ and describe the rest of our results in terms of u, the approximate probability that one daughter gains a mutation.

5.6.4 Lethal neoantigens

In the simplest case, neoantigens immediately cause the death of a cell harboring them. In this case, the dynamics of the surviving process are equivalent to one in which the birth rate is reduced by a factor $1-u$. The process can survive in principle if $b(1-u) > d$, and when this is the case it survives with probability $1 - \frac{d}{b(1-u)}$.

5.6.5 Models without loss

In the threshold model without loss, any cell which acquires a neoantigen with *eventually* die. Therefore, the survival probability of this process is unaltered. Likewise, in the model of constant immune pressure, if this pressure is sufficient to guarantee extinction of any clone with a neoantigen (that is, if $d + w > b$, then the survival probability of this process is also unaltered.

5.6.6 Neoantigen loss

Upon division, each neoantigen can be lost with probability ν. Starting from a cell with k neoantigens, in the limit of small loss rates the probability of generating a daughter cell with $k-1$ neoantigens and another with k neoantigens is approximately $2k\nu$. Let $v = 2\nu$ as before. In our model extended to loss, a cell division produces two daughters with k neoantigens with probability $1-u-kv$, one daughter with k and one with $k+1$ neoantigens with probability u, and one with k and one with $k-1$ with probability kv.

5.6.7 Gain and loss approximation

Begin with a population of cells with k neoantigens. The expected proportion of cells $f(x, t)$ with x neoantigens at time t obeys the differential equation:

$$\frac{df(x,t)}{dt} = -b(u+vx)f(x,t) + buf(x-1,t) + bv(x+1)f(x+1,t)$$

with initial condition $f(k, 0) = 1$, where b is the birth rate, u is the probability that one daughter cell is produced with a neoantigen upon division, and v is the probability per neoantigen that a neoantigen is lost upon division.

The generating function $g(s, t)$ corresponding to this distribution, i.e.

$$g(s,t) = \sum_{x=0}^{\infty} s^x f(x,t)$$

obeys the differential equation:

$$\frac{dg}{dt} = (s-1)bug - (1-s)bv\frac{dg}{ds}$$

with initial condition $g(s, 0) = s^k$. It admits the following solution

$$g(s,t) = e^{\frac{(1-e^{-btv})(s-1)u}{v}}(1+e^{-btv}(s-1))^k$$

This generating function can be recognized as the generating function corresponding to the sum of two random variables, a Poisson with mean $(1 - e^{-bvt})u/v$ and a binomial with k trials and success probability e^{-bvt}.

The expected proportion of cells with no mutations, $f(0, t)$ is therefore:

$$f(0, t) = e^{(e^{-bvt}-1)u/v}(1 - e^{-btv})^k$$

We wish to calculate the expected number of cells with no neoantigens that exist when a single mutant cell grows until an immune response is triggered. First, the size of the neoantigenic clone grows in expectation according to

$$\mathbb{E}[Z_1(t)] = e^{(b(1-v)-d)t}$$

because accumulation of additional neoantigens does not change the size of the original clone. We require that $b(1 - v) > d$ so that the mutant clone grows in size. We approximate the time at which the immune response occurs against the founding clone as the time at which this expectation reaches the threshold:

$$t^* \approx \frac{log(T)}{b(1 - v) - d}$$

The total population grows in expectation according to

$$\mathbb{E}[Z] = e^{(b-d)t}$$

so the number of cells with no neoantigens produced by a clone with 1 neoantigen M_1 is approximately

$$M_1 \approx f_1(0, t^*)e^{(b-d)t^*}$$

Furthermore, the expected number cells with no neoantigens produced independently is approximately

$$\lambda \approx \int_0^{t*} bvf_1(1, \tau)e^{(b-d)\tau}\, d\tau$$

And for very small rates v this number is approximately Poisson distributed.

5.6.8 Threshold return criteria

We approximate the survival probability of a cell which begins with no neoantigens in a process with neoantigen gain rate u, loss rate v, threshold T and birth and death rates b, d.

$$p_{ext} = P(Birth)\big((1-u)p_{ext}^2 + up_{ext}p_{ext,u}\big) + P(Death)$$

Where $p_{ext,u}$ is the probability that a process beginning with one neoantigen goes extinct. We estimate $p_{ext,u}$ as follows. First, we assume that the threshold is sufficiently large that, in the absence of an immune response, if a population of cells reaches the threshold then it is certain to survive. Thus, a mutant cell goes extinct either by fluctuating extinct (this happens with the natural extinction probability of the process) or by reaching the threshold. All cells with the neoantigen will be driven extinct, but on the way the mutant clone might produce cells without the neoantigen. It produces N such cells (calculated above). Thus,

$$p_{ext,u} = d + (1-d)\sum_{n=0}^{\infty} p_{ext}^n P(N=n)$$

Using the definition above for λ,

$$p_{ext,u} = d + (1-d)\sum_{n=0}^{\infty} p_{ext}^n \frac{\lambda^n e^{-\lambda}}{n!}$$

$$p_{ext,u} = d + (1-d)e^{\lambda(p_{ext}-1)}$$

The full recurrence is as follows:

$$p_{ext} = \frac{b}{b+d}\big((1-u)p_{ext}^2 + up_{ext}(d + (1-d)e^{\lambda(p_{ext}-1)}\big) + \frac{d}{b+d}$$

and can be solved numerically.

5.6.9 Gain, loss, and a continuous immune response

Suppose the immune system does not eliminate sufficiently large populations of antigenic cells but rather it increases the death rate of a cell with k neoantigens by kw. In this case, the expected number of cells $Z(x, t)$ with x neoantigens at time t obeys the differential equation:

$$\frac{dZ(x,t)}{dt} = Z(x,t)\big(b(1-u-xv)-d-wx)\big) + Z(x+1,t)bv(x+1) + Z(x-1,t)bu$$

The corresponding generating function is

$$\frac{dg}{dt} = g\big(b(1-u(1-s)-d\big) + \frac{dg}{ds}\big(bv(1-s)-sw\big)$$

with initial condition $g(s,0) = Z_0 s^k$ for a population of size Z_0 with k neoantigens.

While exactly solvable, the solution to this PDE is quite complicated. However, the dependence on the initial condition decays with rate $bv+w$, and after this decay, growth of the population overall is controlled by a simple exponential,

$$Z(t) \approx C * exp\big(t * \frac{b(b-d)v + (b(1-u)-d)w}{bv+w}\big)$$

Thus, the population will grow in expectation iff the exponent is positive. After some algebra this is equivalent to

$$v + \frac{w}{b}(1-\frac{u}{\rho}) > 0$$

Further, relative to a population with mutation rate u, a population with mutation rate $\frac{u}{2}$ will expe-

rience a growth rate advantage s:

$$s = \frac{buw}{2(bv + w)}$$

.

5.6.10 Expected number of cells at time t

Let the number of cells at time t with no neoantigens be $Y_0(t)$ and set $Y_0(0) = 1$. A new mutant population is produced from Y_0 at rate $buY_0(t)$. The ith mutation initiates a population Y_i. Let K_t be the number of mutations occurred by time t at times $(u_t)_{t=1}^{K_t}$. The immune response is triggered when a particular neoantigen reaches size T. This population is then killed. Therefore the total population at time t is

$$Z(t) = Y_0(t) + \sum_{i=1}^{K_t} Y_i(t)$$

Let $(X_i(t))_{i=1}^{\infty}$ be iid linear birth death processes and $(\tau_i)_{i=1}^{K_t}$ the unorered mutation times. Then

$$Z(t) \stackrel{d}{=} X_0(t) + \sum_{i=1}^{K_t} X_t(t - \tau_i)\mathbb{I}_{\{0 < X_t(t-\tau_i) < T\}}$$

where equality is noted in distrubtion. Next, we approximate that each population grows deterministically and only mutations are stochastic. Thus, each $X_i(t)$ is approximated by $e^{\lambda t}$ where $\lambda = b(1-u) - d$ for unmutated cells and $\lambda = b - d$ for mutated cells.

5.6.11 Mean number of cells under approximation 1

Let $\lambda_0 = b(1-u) - d$, $\lambda_1 = b - d$ and $d = d/b$. Under approximation 1, K_t is Poisson with parameter $b(1-d)\,u(e^{\lambda_0 t} - 1)/\lambda_0$. From this and Wald's lemma,

$$\mathbb{E}[Z(t)] = e^{\lambda_0 t} + \frac{bue^{\lambda_0 t}}{\lambda_1 - \lambda_0} T^{(\lambda_1 - \lambda_0)/(\lambda_1) - 1}$$

$$= e^{b(1-u)-d)t} T^{u/(1-d)}$$

For $e^{\lambda_1 t} \geq T$, otherwise $\mathbb{E}[Z(t)] = e^{\lambda_1 t}$.

5.7 Time to resistance

Michael, not sure if we will use this. But going through it again I think it is a really great derivation. Let's discuss.

5.7.1 Yule model limit

As an aside, we note the special case when there is no death (a Yule process). Here, the expected number of unmutated offspring produced in one division is:

$$E(Z_0(t+dt)|Z_0(t)=1) = 2e^{-u/2} + 2e^{-u/2}(1-e^{-u/2}) \tag{5.1}$$

$$E(Z_0(t+1)|Z_0(t)=1) = 2e^{-u/2}$$

This is greater than 1 when

$$u < 2log(2)$$

This condition places an absolute upper bound on the rate at which a dividing population can gain mutations but still hope to maintain an ancestral sequence.

5.7.2 VAF for v = 0 in continuous model

we provide a heuristic argument for the distribution of the number of cells carrying a uniformly chosen neoantigen. Suppose we observe a tumor at time t, and let $L = \max\{\text{number of neoantigens in cell}\}$.

Thus we can partition the growing tumor population into $(Z_i(s))_{i=0,s=0}^{L,t}$, where $Z_i(s)$ is the number of cells with i neoantigens at time s. Let the growth rate of cell with i neoantigens be $\lambda(i)$, and we will assume this is a monotone decreasing function of i. To avoid clutter let $\lambda = \lambda(0)$ and $\lambda(i) = \lambda_i$. Further suppose at rate ν each cell is replaced by an identical copy of itself and a copy possessing a never seen before neoantigen. While L is a time-dependent random variable, if we consider fixed L, Theorem 3 in[184] informs us that conditional on type 0 survival, almost surely,

$$\lim_{t\to\infty}(e^{-\lambda t}Z_i(t))_{i=0}^{L} = W\left(\prod_{j=1}^{i}\frac{\nu}{\lambda-\lambda_j}\right)_{i=0}^{L} = W(c_i)_{i=1}^{L},$$

where W is exponential with parameter $\lambda/\alpha(0)$. If we suppose $\lambda - \lambda_i = is$ then $c_i = \frac{\nu^i}{s^i}\prod_{j=1}^{i} j^{-1}$. For the remainder of this discussion we will assume L is fixed and large and use the approximation $Z_i(s) = Wc_i e^{\lambda s}$. The number of new neoantigens acquired in the population by time t will be denoted K.

At time t let the number of new neoantigens generated by a cell with $i-1$ neoantigens be K_i, with $1 \le i \le L$. Such neoantigens are generated at rate $\nu Z_i(s) = Wc_{i-1}e^{\lambda s}$. Therefore the arrival density for any such neoantigen that arrives before t is approximately

$$f_{T_i}(s) = \frac{\lambda e^{\lambda s}}{e^{\lambda t}-1}.$$

Let $Y_i(t)$ be the number of cells possessing such a neoantigen that arrived at T_i, which we refer to as the clone initiated by that neoantigen. Cells of this clone will continue to accumulate neoantigens, and so we let $Y_{ij}(t)$ be the number of cells in the clone that have j neoantigens $i \le j \le L$. We may apply Theorem 3 in[184] to the growth of this clone which yields the approximation

$$Y_{ij}(s) = W_i c_{i,j} e^{\lambda_i(s-T_i)}$$

for $s \geq T_i$, and with $W_i \sim \text{Bern}(\lambda_i/\alpha_i)\text{Exp}(\lambda_i/\alpha_i)$, $c_{i,j} = \frac{\nu^{j-i}}{s^{j-i}} \prod_{k=1}^{j-i} k^{-1}$. Hence the total number of cells in this clone at t is

$$Y_i(t) = W_i e^{\lambda_i(t-T_i)} \sum_{j=i}^{L} c_{i,j} = W_i e^{\lambda_i(t-T_i)} C_i.$$

Thus the clone size distribution of neoantigens who arrive in a cell already possessing $i-1$ neoantigens is

$$\begin{aligned}
\mathbb{P}(Y_i(t) > y) &= \mathbb{E}[\mathbb{P}(W_i e^{\lambda_i(t-T_i)} C_i > y)] && (5.2)\\
&= \mathbb{E}\left[\int_0^{t-\lambda_i^{-1}\log\frac{y}{C_i W_i}} f_{T_i}(s)\, ds\right] && (5.3)\\
&= \mathbb{E}\left[\frac{e^{\lambda t}(C_i W_i/y)^{\lambda/\lambda_i} - 1}{e^{\lambda t} - 1}\right] && (5.4)
\end{aligned}$$

If W is fixed, then each of the K_i is Poisson with parameter $Wc_i(e^{\lambda t}-1)/\lambda$. Therefore if from the possible K neoantigens we select a particular neoantigen uniformly at random, this neoantigen was initially acquired in the division event to a cell already possessing $i-1$ neoantigens with probability $\frac{c_{i-1}}{\sum_{j=0}^{L-1} c_j}$. Let $\sum_{j=0}^{L-1} c_j = A$. Hence at time t, the distribution of the number of cells possessing a uniformly chosen neoantigen is

$$\mathbb{P}(Y(t) > y) = \sum_{i=1}^{L} \frac{c_{i-1}}{A} \mathbb{E}\left[\frac{e^{\lambda t}(C_i W_i/y)^{\lambda/\lambda_i} - 1}{e^{\lambda t} - 1}\right] \approx \sum_{i=1}^{L} \frac{c_{i-1} C_i^{\lambda/\lambda_i} \mathbb{E}[W_i^{\lambda/\lambda_i}]}{A y^{\lambda/\lambda_i}} \qquad (5.5)$$

where the last approximation is due to $e^{\lambda t} \gg 1$. Note that here L is random and an expectation should be taken upon the above inequality with respect to the distribution of L. In principle this

distribution may be obtained using Theorem 1 of[184]. However regardless of L, if $\nu/s < 1$ then

$$\frac{\left(\frac{\nu}{s}\right)^i \prod_{j=1}^{i} j^{-1}}{A} C_i^{\lambda/\lambda_i} = \frac{\left(\frac{\nu}{s}\right)^i \prod_{j=1}^{i} j^{-1}}{A} \left(\sum_{j=i}^{L} \frac{\nu^{j-i}}{s^{j-i}} \prod_{k=1}^{j-i} k^{-1} \right)^{\lambda/\lambda_i} \leq \frac{\left(\frac{\nu}{s}\right)^i \prod_{j=1}^{i} j^{-1}}{A}. \tag{5.6}$$

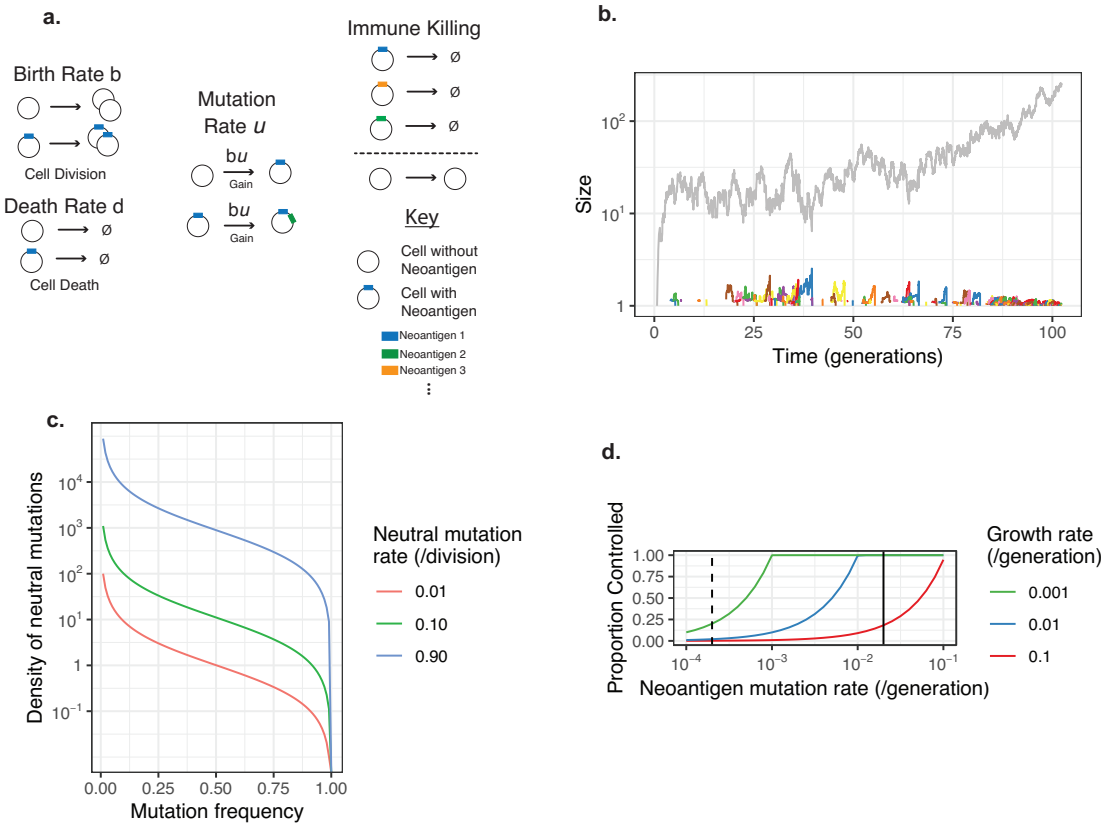


Figure 5.1: a, All cells grow according to a branching process with division rate b and death rate d. Mutations occur during division. A dividing cell can produce a daughter with an additional neoantigen with probability u. These neoantigenic mutations are neutral except when interacting with the immune system. In this model, the immune system eliminates a particular neoantigen essentially immediately. **b,** A simulation of a cell starting with no neoantigens (grey population). Some divisions produce cells with neoantigens (colored trajectories), which for illustrative purposes here are allowed to persist briefly. **c,** Variant allele frequency spectrum predictions for three different rates of mutation. **d,** The proportion of initiated tumors which are eventually controlled by this type of immune response as a function of the rate of gaining neoantigens for three different growth rates.

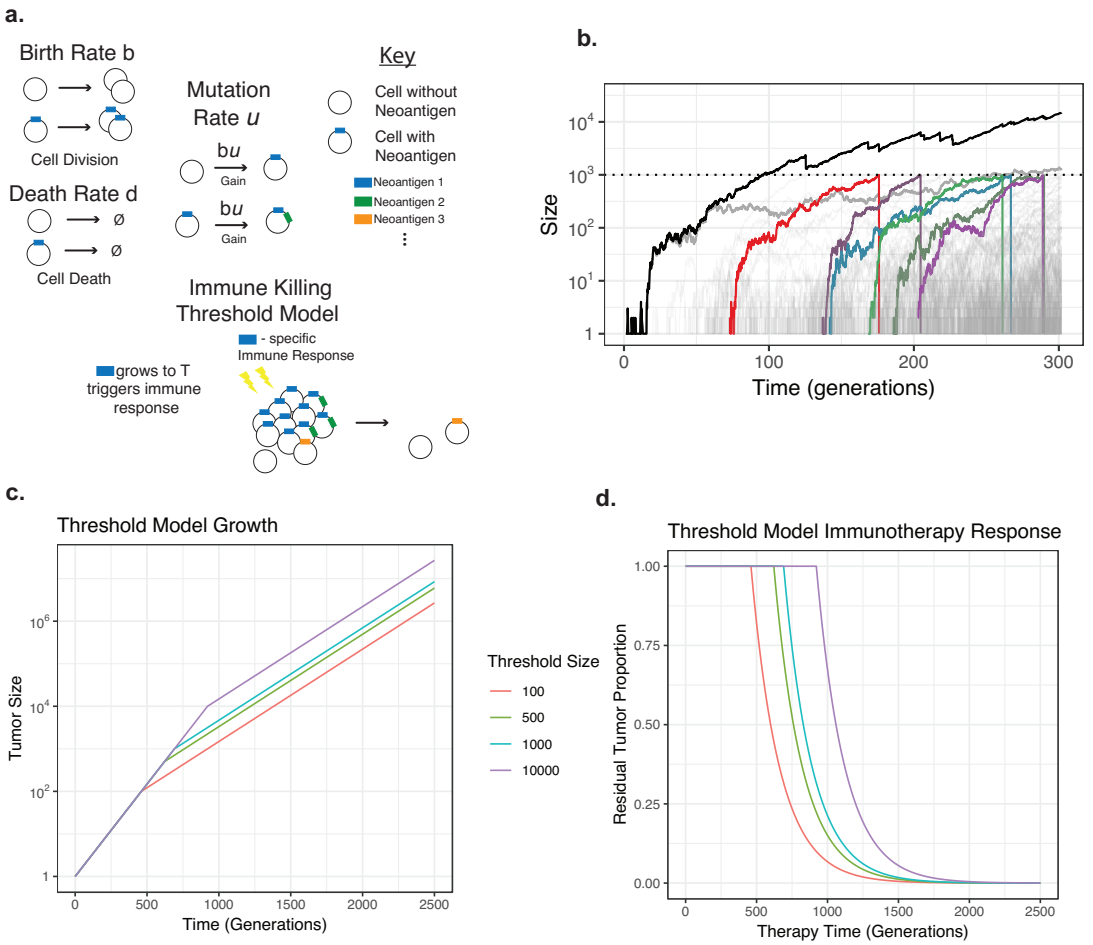


Figure 5.2: a, All cells grow and mutate as in the model from Figure 5.1. However, in this model the immune system does not elimiate a clone with a neoantigen until the clone reaches a threshold size T. **b,** A simulation starting with a cell that lack a neoantigen (thick grey line). The total population size is shown in black, and the dotted line indicates the level of the immune threshold. Example neoantigenic clone trajectories are illustrated in color from all the neoantigenic clones (thin grey lines). **c,** The expected size of the tumor population as a function of time for different thresholds. **d,** Simulated therapy response under this model for different thresholds, where the administration of an immunotherapy reduces an exponentially growing tumor to the size predicted in **c**.

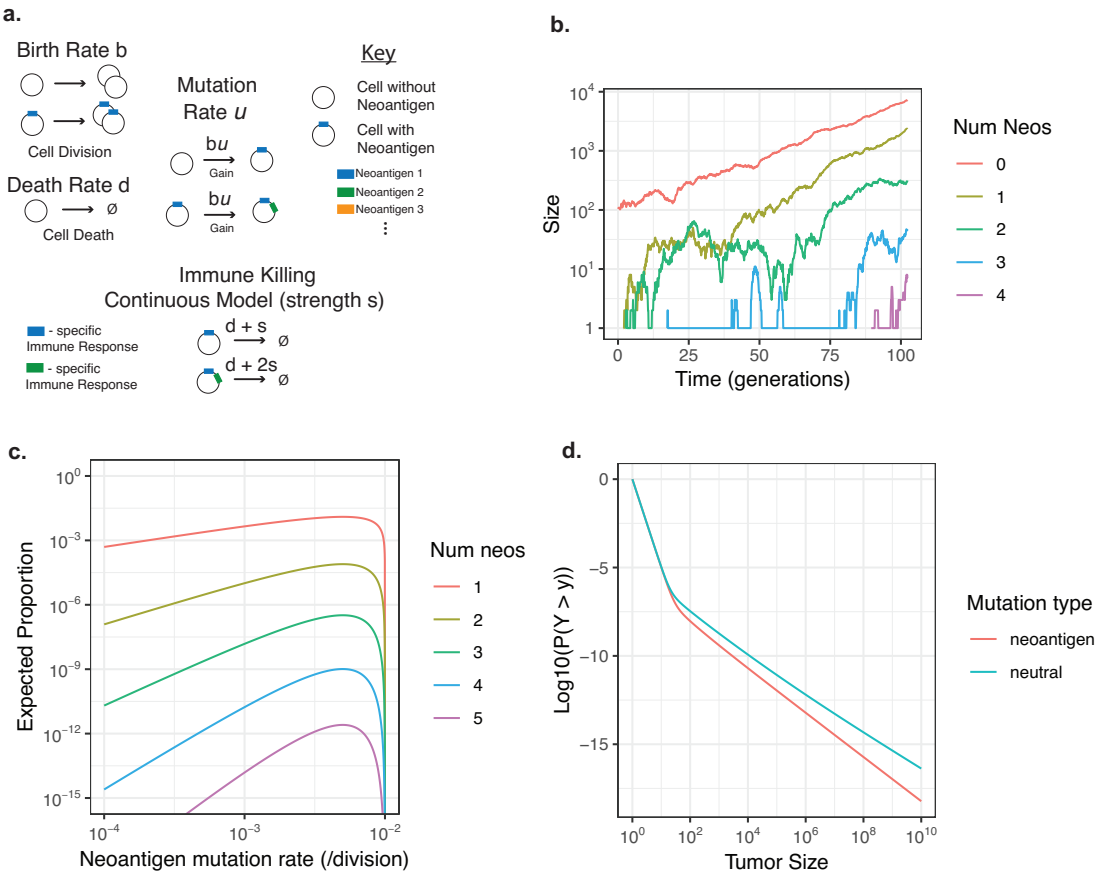


Figure 5.3: a, All cells grow and mutate as in the model from Figure 5.1. However, in this model the immune system exerts a constant rate of killing s for each neoantigen present in a cell. **b,** In simulations starting with cells that have no neoantigens (red line), cells are produced with progressively more neoantigens over time. **c,** The expected proportion of cells with 1 to 5 neoantigens is shown in the long-time limit of this process as a function of the rate of gaining neoantigens. When the rate of gaining neoantigens approaches the growth rate of the tumor, the mutated populations collapse. **d,** The variant allele frequency spectrum for neutral and neoantigenic mutations. At low frequencies, it is difficult to distinguish neutral and neoantigenic mutations, but neoantigenic mutations are severely depleted at high frequencies.

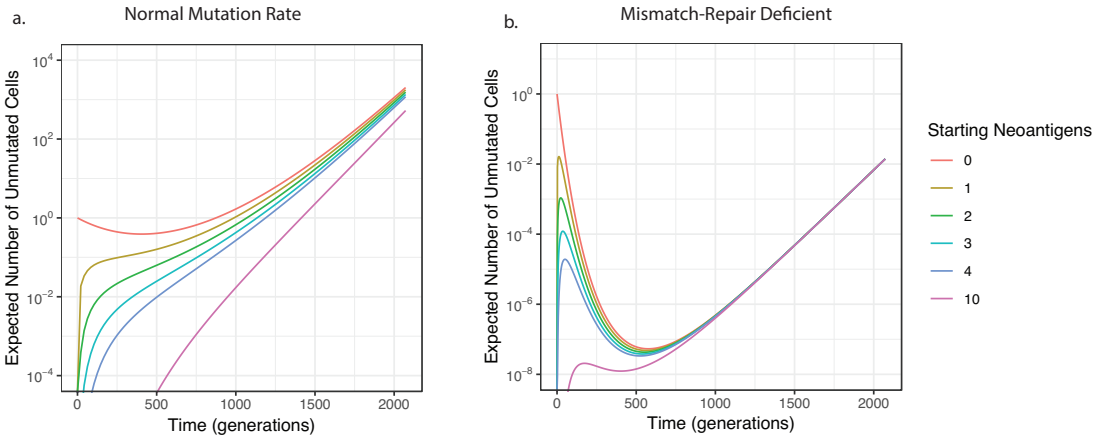


Figure 5.4: The expected number of unmutated cells in a population shielded from immunity is shown. The total size of the tumor population grows exponentially at rate 1% per generation, reaching size 10^9 at the last generation. The starting number of neoantigens has the largest effect on the early number of unmutated cells. **a.** Neoantigen gain and loss rates for normal cells ($u = 2 * 10^{-4}$, $v = 10^{-7}$). **b.** Neoantigen gain and loss rates for a MMRD and CIN tumor ($u = 0.1$, $v = 4 * 10^{-3}$).

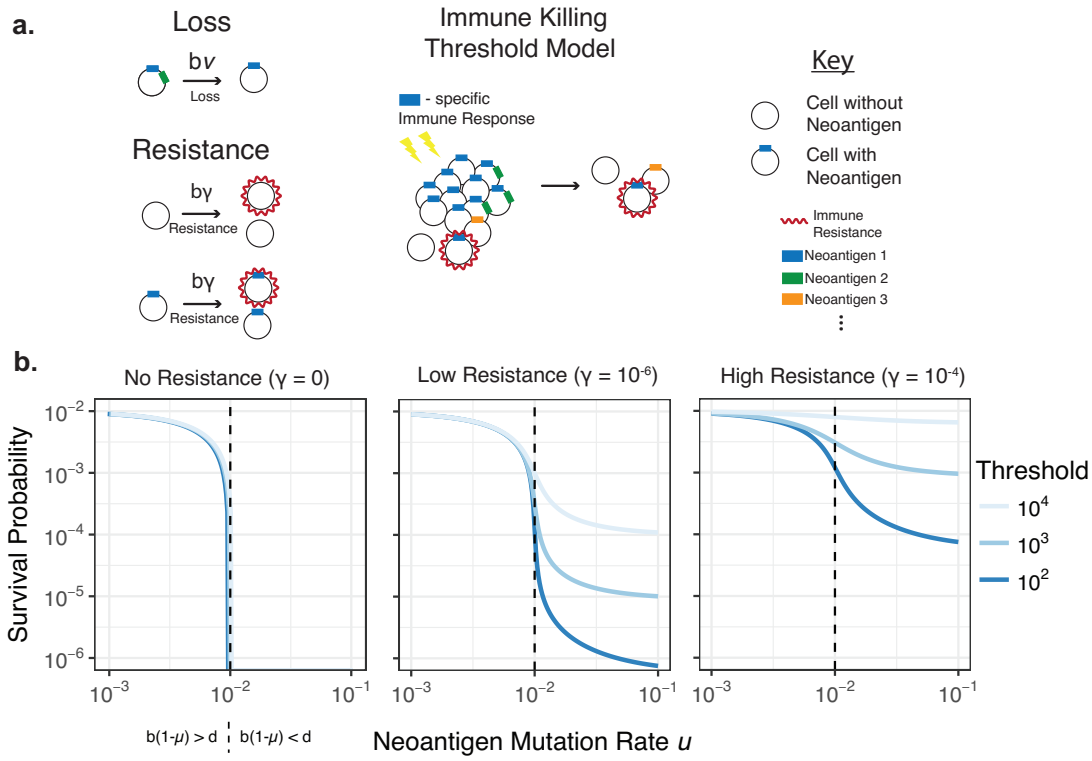


Figure 5.5: a. All cells grow, mutate, and are subject to immune pressure as in the threshold model (Figure 5.2a). However, in this model, an additional type of mutation is possible: with a probability per division γ, a cell is produced which is hidden from an immune response ("immune resistance" outlined in red). When an immune response is triggered against a particular neoantigen, cells with immune resistance are not eliminated. **b.** The survival probability of an initiated cell depends on the rate of immune resistance γ and on the threshold at which an immune response is triggered. higher rates of immune resistance and more relaxed thresholds produce a higher survival probability.

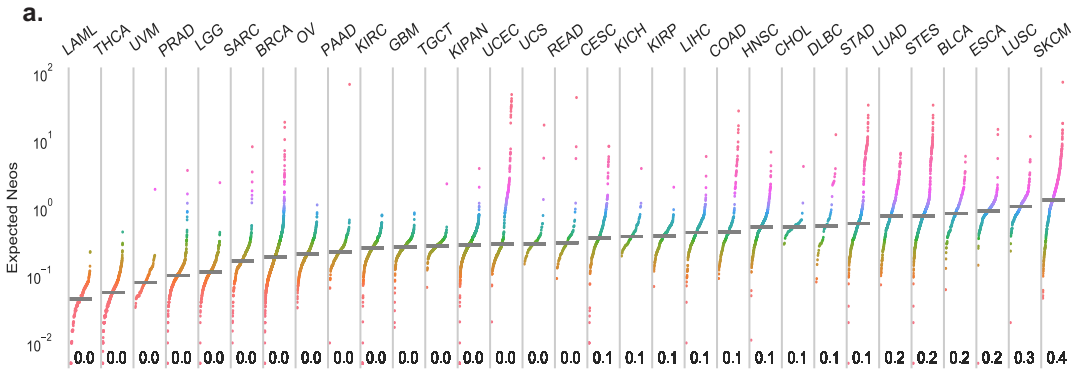


Figure 5.6: a, Many tumors arise on a background of no neoantigens. **b,** Eliminating tumors initiated without neoantigens is possible only when the rate of gaining neoantigens is high. Solid and dotted green lines indicate estimates of the probability of gaining a neoantigen per division in healthy cells and cells with microsatellite instability (100x healthy), respectively. As tumors progress, the growth advantage increases from low in pre-malignant lesions (blue) to high in metastases (red).

6

Viral rebound kinetics following single and combination immunotherapy for HIV/SIV

6.1 Forward

This work builds on previously published modeling efforts in viral dynamics[148,29], providing an improved model, better statistical fitting approaches, an expanded data set, and a data-driven con-

nection to human viral rebound. Many people have contributed to this work, including Melanie Prague, Chloe Pasin, Irene Balelli, James Whitney, Dan Barouch, and Jon Li. Alison Hill supervised this project and my work on it. I contributed to the development and analysis of the model, merging the stochastic and deterministic regimes in the model, and simulating rebound dynamics. We are preparing this work for publication.

6.2 ABSTRACT

HIV infection can be treated but not cured with combination antiretroviral therapy, and new therapies that instead target the host immune response to infection are now being developed. Three recent studies of such immunotherapies, conducted in an animal model (SIV or SHIV-infected rhesus macaques), have shown that agents which target the innate immune receptor TLR7 along with recombinant viral-vector vaccines or monoclonal antibodies can prevent or control the rebound in viremia that usually accompanies the discontinuation of antiretroviral drugs. However, the mechanism of action of these therapies remains unknown. In particular, it is unclear what relative role was played by reduction of the pool of latently infected cells versus boosting of anti-viral immune responses, and whether the therapies acted independently or synergistically. Here we conduct a detailed analysis of the kinetics of viral rebound in this collection of studies, and use mechanistic mathematical models combined with rigorous statistical methods for model fitting and selection to quantify the impact of these immunotherapies on viral dynamics. We find that the therapeutic vaccine reduced the effective reactivation rate from the latent reservoir by an average of 4-fold (95% CI [2,8]), and boosted the avidity of antiviral immune responses by 17-fold [5, 67] when alone and 210-fold [30, 1400] when combined with the TLR7-agonist. In the context of later initiation of antiretroviral therapy only (9 weeks vs 1 week after infection), the TLR7-agonist reduced the reservoir contribution to rebound by an average of 8-fold [4, 16], and also slightly increased target cell

availability (1.5-fold). The monoclonal antibody boosted immune response avidity by 8-fold [3,16] and displayed no detectable synergy with the TLR7 agonist. These results provide a framework for understanding the relative contributions of different mechanisms of preventing viral rebound and highlight the multi-faceted roles of TLR7-agonists as immunotherapy for HIV/SIV cure.

6.3 Introduction

Worldwide, over 39 million people are currently infected with HIV, and 2 million individuals are newly infected each year[260]. While combination antiretroviral therapy (ART) can suppress viral replication, preventing both transmission and progression to AIDS, it cannot completely clear the infection. A latent reservoir of integrated virus exists in long-lived lymphocytes and can re-initiate the infection ("rebound") whenever treatment is stopped[243,178]. Consequently, current therapy must be taken for life, and new research efforts are underway to find a permanent cure for HIV[69].

Two general approaches are being taken to prevent HIV rebound and hence allow therapy to be completely stopped ("cure"). One approach, often called a "sterilizing cure", aims to purge all remaining latent virus from the body[53], ideally re-capitulating the effects of case-studies involving bone marrow transplants (e.g.[283,106]) or extremely early treatment initiation (e.g.[204,107]). Another approach, often called a "functional cure", is to instead equip the immune system with the ability to control virus that reactivates from latency, perhaps mimicking what naturally occurs in so-called elite controllers[25] or post-treatment controllers[230]. Because of the difficulties in detecting latent virus and the lack of known immune correlates of HIV control, all current potentially-curative interventions must be evaluated by conducting treatment interruption studies, in which recipients eventually stop all therapy in a controlled manner and are monitored closely for viral rebound[82].

Here we analyze data from three recent studies[29,148,30] that are part of a larger effort to use therapies that perturb the immune response (known as "immunotherapy") to induce viral control either

by clearing latent virus, boosting anti-viral immune responses, or both. One component of the investigation therapy is a small-molecule agonist of the Toll-like receptor 7 (TLR7), part of the innate immune system involved in antiviral defense[149]. Another component is a "therapeutic vaccine" (administered *during* infection, as opposed to traditional *preventative* vaccines). The vaccine contains HIV (or SIV) DNA encoded in a viral vector (either Ad26, an adenovirus vector, or MVA, a modified vaccinia virus vector) and is administered in a prime-boost regimen (one vector, then the other)[13,14] (Figure 6.1). The third component is the PGT121 monoclonal antibody, which targets the V3 loop of the HIV envelope protein and has the highest potency of any antibody isolated to date, neutralizing $\sim$ 70% of HIV isolates with a median IC_{50} of around 50 ug/mL[122,269]. Innate immune stimulation as a strategy to treat chronic viral infections[199,87], supplement vaccination[248,116,125], or enhance the effect of monoclonal antibodies[237] has previously been shown to be promising.

Two of these pre-clinical studies were conducted in SIV-infected rhesus macaques, a well-validated animal model of HIV infection which recapitulates HIV pathogenesis and ART response[8]. For studies involving the monoclonal antibody, animals were instead infected with SHIV, a chimeric virus consisting of the HIV envelope gene in an SIV backbone. One or two of the immunotherapies was given to animals during long-term ART, and viral levels were monitored once all therapies were discontinued. The kinetics of viral rebound were altered in many treated animals, and a subset of animals showed unprecedented responses - some animals never rebounded and appeared to have a achieved a sterilizing cure, and another subset rebounded temporarily but then achieved complete suppression of virus (apparent functional cure) (Figure 6.1). The goal of this study was to combine mechanistic mathematical models with rigorous statistical methods to characterize in detail the changes in rebound kinetics in all study animals, compare hypotheses about the effects of each component of the immunotherapy intervention individually, evaluate their synergy, and provide recommendations about future trials.

Mathematical models have a long history of informing the dynamics of infections within individ-

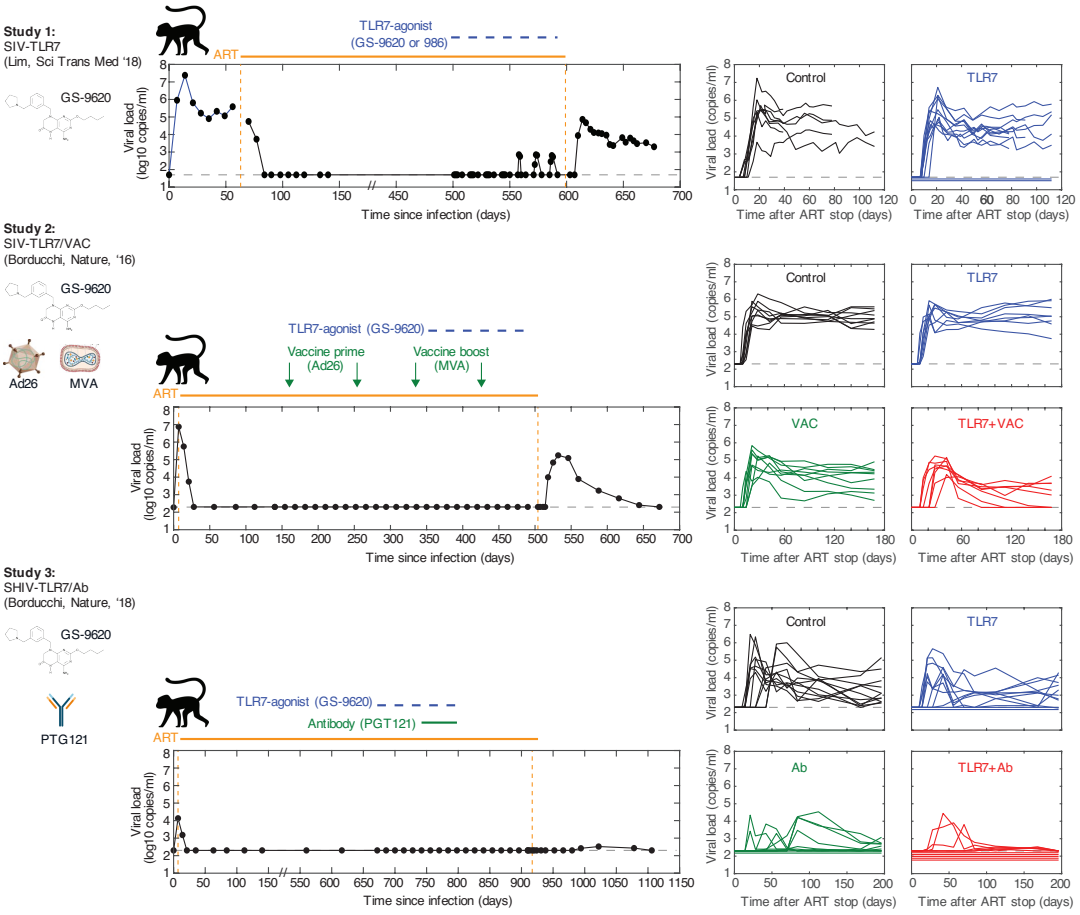


Figure 6.1: In **Study 1** [148], rhesus macaques were infected with SIV, and treated with ART after 9 weeks. During ART, one group of animals was administered repeated doses of a TLR7-agonist compound. ART was stopped after between 1.5-2.2 years. In **Study 2** [29], rhesus macaques were infected with SIV and treated with ART after 1 week. During ART, animals were divided into four groups, receiving a TLR7-agonist, an Ad26/MVA therapeutic vaccine in a prime-boost regimen, both, or neither. ART was stopped after 1.4 years. In **Study 3** [30], rhesus macaques were infected with SHIV (chimeric SIV with HIV envelope) and treated with ART after 1 week. During ART, animals were divided into four groups, receiving a TLR7-agonist, the PGT121 monoclonal antibody, both, or neither. ART was stopped after 2.5 years. Figures on the left show the full time-course of viral load for one example animal from each study (animal IDs: Study 1 - 156-08, Study 2 - 5888, Study 3 - 6377). Figures on the right show viral loads after ART cessation for all animals in each study. Viral load values for animals with no detectable virus are shown below the detection limit for visualization purposes only. More details of the experimental design is provided in the **Methods** and in the original study manuscripts.

ual hosts[194], and have been used to study viruses such as hepatitis B[192,203,64] and C[183,49], influenza[176], dengue[197,56], and herpes simplex virus[238]. For HIV, these viral dynamics models have been instrumental in understanding many important aspects of infection, such as the cause of viral load decline post-peak[205,247], the lifespan of infected cells[270,110], the rate of seeding of the latent reservoir[9], and the effects of treatment with antiretroviral therapy[164,277,228,46], and investigational immunotherapies such as antibodies[150] and IL-7[253]. These models are generally represented in the form of non-linear ordinary differential equations, and are often fit to data to estimate parameters that cannot be directly measured. Traditionally, such fitting was done on an individual-by-individual using least-squares (e.g.[43,150]) or maximum likelihood-based optimization, or fully Bayesian approaches (e.g.[277,152]), but these methods may suffer from identifiability issues when data is sparsely sampled or variables are unobserved, and do not provide a formal way of comparing dynamics between treatment groups. More recently, inference methods based on non-linear mixed-effects models have been developed to jointly infer parameters from groups of individuals and formally test for differences between treatment groups[209,46].

In this paper we describe the use of mathematical modeling to understand the effects of the single and combination immunotherapy on viral rebound kinetics following ART-cessation. First, we develop an augmented model of HIV/SIV dynamics which includes latent infection and an adaptive immune response. Then we analyze the dynamics of this model and investigate the theoretical identifiability of its parameters from longitudinal viral load data. We present a Bayesian non-linear mixed effects statistical inference framework to estimate model parameters from the data and use this to evaluate the most likely mechanism of action of each component of the treatment.

6.4 Results

6.4.1 Development of a viral dynamics model for rebound and control

Animals who received either ART alone or augmented with the TLR7-agonist, therapeutic vaccine, monoclonal antibody, or combination immunotherapy exhibited a wide range of viral rebound trajectories (Figure 6.1). The standard mathematical model of HIV viral dynamics, which describes interactions only between virus and target CD4 T cells, cannot explain prominent features of these kinetics, such as a large difference between peak and set-point viral load or eventual control of infection [43]. We hypothesized that these kinetics were influenced by the induction of an adaptive immune response before and during rebound. This idea is supported by the obsevation that immunotherapy led to perturbations in interferons and interferon-stimulated genes, activation of multiple lymphocyte subsets, and expansion of cellular immune responses to viral peptides [29,148,30].

To infer the mechanisms of immunotherapy action across the full range of observed viral rebound kinetics, we augmented the standard model of viral dynamics to account for an adaptive immune response. In addition to modeling uninfected and infected target cells and virus, we included population of effector immune cells which suppress infection and a longer-lived precursor population which produces effectors and provides immunological memory. Our model is general enough to represent either cellular or humoral responses. We also modeled the reactivation of latently infected cells, which provides the initial source for rebounding virus. The model was specifically developed to be flexible enough to capture rebound kinetics both in the regime where latent cells reactivate frequently and rebound occurs rapidly, and in the regime when reactivation from latency is rare and there are stochastic delays until the first fated-to-establish lineage exits the reservoir [152,108,206]. Figure 6.2 shows the model schematic and associated mathematical description. This augmented model is able to qualitatively reproduce the diverse rebound trajectories seen in data from the two studies, including rebound followed by a high set-point and rebound followed by immune control (Figure

6.3 and SI Figures TBA). The Methods section details the study designs, therapies, data collection, model structure, and fitting methods.

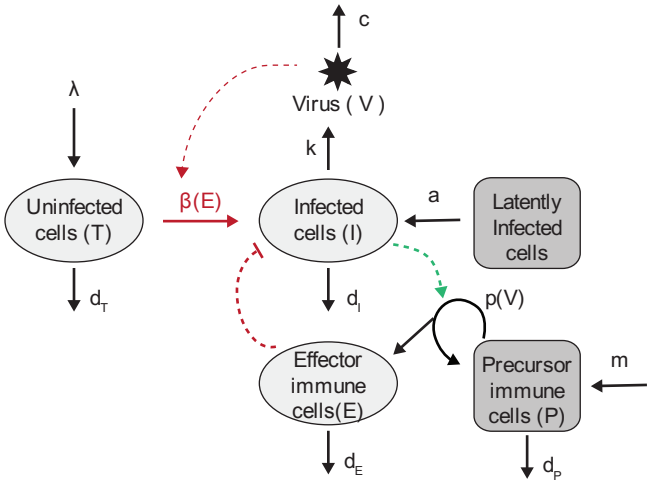


Figure 6.2: Briefly, free viruses V enter target cells T (with infection rate β), producing infected cells I. Infected cells in turn release free virus (rate k). Long-lived precursor immune cells P which encounter viral antigen proliferate ($p(V) = pV/(V + N_P)$) and produce short-lived effector immune cells E. Effector immune cells eliminate some infected cells before they contribute to ongoing infection by producing new virus ($\beta(E) = \beta/(1 + E/N_E)$). A fraction f of expanded precursor cells revert to the precursor state after encountering antigen, forming immunological memory. Both uninfected target cells and precursor immune cells are produced at a constant rate (λ and m respectively). While during acute infection m likely represents activation of naive cells, during viral rebound, it may be dominated by reactivation of memory cells. Latently infected cells reactivate with rate a (or equivalently, every t_a days) to become productively infected cells. Virus is cleared at a rate c and each cell type i dies with death rate d_i. More details are provided in the **Methods**.

6.4.2 Simulation analysis of determinants of viral rebound kinetics

The goal of this study was to understand the effects of immunotherapy by estimating the parameters of the model from the observed rebound trajectories and then comparing these parameters between treatment groups. To justify this approach, we first sought to determine which model parameters could be identified from viral load time-series alone, and to understand the expected influence of these parameters on rebound kinetics. A comprehensive identifiability analysis, detailed in the

Methods, found that the combination of reservoir reactivation rate a, target cell replenishment rate λ, and viral burst size k can be adjusted to produce equivalent observed trajectories (since we only have longitudinal measures of viral load V and not infected cells I). Likewise, we can only identify the ratio of precursor immune cell production rate m and half-maximal inhibitory concentration (IC_{50}) of effector cells N_E (since we don't have measures of effector immune cell levels E). Accordingly, we fixed the viral burst size k and the effector half-max N_E so the remaining parameters were identifiable. Furthermore, viral clearance rate, lifespans of each cell population, and the fraction of expanded immune cells that revert to memory can be estimated from other data sources, and so we fixed these parameters at experimentally-determined values (Table 6.1 and Methods).

The resulting model has six remaining parameters to be estimated: target cell replenishment rate λ, infection rate β, the latent cell reactivation rate a, precursor immune cell production rate m, maximal proliferation rate of immune cells p, and the half-maximally stimulating level of virus N_P. We verified that this combination of parameters was formally identifiable (see Methods), and then simulated the model under systematic variations in each parameter to understand how each affects rebound kinetics. If the immune response is not strong enough, then the kinetics of this model reduce to those of previous viral dynamics models without immune responses (Figure 6.3, top row). The availability of target cells (λ) and the baseline viral infectivity (β) control the early viral growth rate, while the timing of rebound depends on the rate at which latent cells reactivate (a). The density of target cells that the virus can access (λ) also influences the eventual setpoint viral load.

However, when viral antigen stimulates immune cells sufficiently, the immune response can curtail rebounding infection (Figure 6.3, bottom row). Increases in the rate at which effector immune cells expand when stimulated by antigen (p) and the rate of immune precursor cell production (m), which here determines the initial level of precursor cells, e.g. the size of the memory pool, at the time of treatment interruption, reduce the height and timing of peak viremia. The degree of control of the viral load setpoint is determined mainly by N_P, the viral load level at which antigen-stimulation

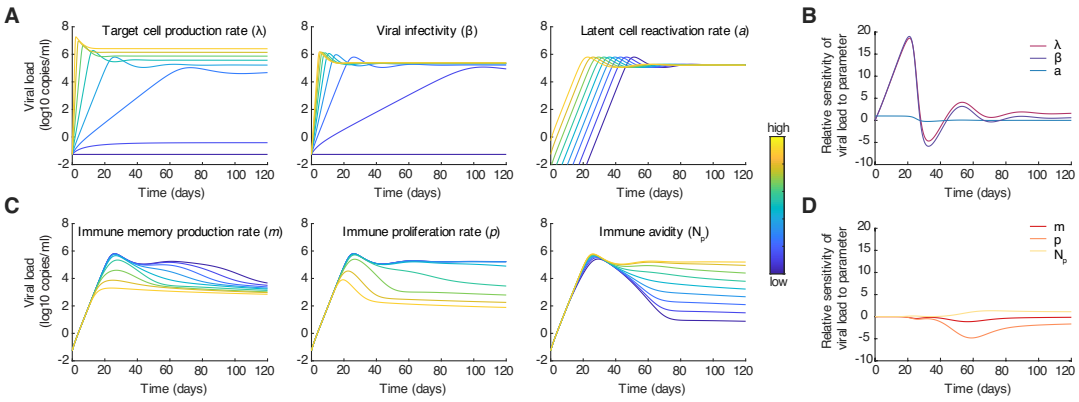


Figure 6.3: Top row: Weak immune response ($p = 0.1$). A) Viral load trajectories produced by the model for different values of either target cell production rate (λ), viral infectivity (β), or latent cell reactivation rate (a). B) Sensitivity of viral load to parameter values λ, β, or a over time. Relative sensitivity to parameter θ is defined as $\frac{\partial V}{\partial \theta}\frac{\theta}{V}$. Bottom row: Strong immune response ($p = 1$). C) Viral load trajectories produced by the model for different values of either immune memory production rate (p), immune proliferation rate (p), or immune repsonse avidity (N_P). D) Sensitivity of viral load to parameter values m, p, or N_P over time. Parameter values, when not varied and [min,max] when varied: $\lambda = 50$ [0, 500] cells mL^{-1} day^{-1}, $\beta = 5 \times 10^{-7}$ [0, 20] mL copies^{-1} day^{-1}, $N_E = 10^4$ cells mL^{-1}, $d_T = 0.05$ day^{-1}, $a = 10^{-5}$ [10^{-12}, 10^{-4}] cells mL^{-1} day^{-1}, $d_I = 0.4$ day^{-1}, $k = 5 \times 10^4$ virions cells^{-1} day^{-1}, $c = 23$ day^{-1}, $m = 1$ [0.01, 100] cells mL^{-1} day^{-1}, $f = 0.9$, $p = 0.1$ (A,B) or $p = 1$ [0.01, 10] (C,D) day^{-1}, $N_P = 10^4$ [10^2, 10^6] copies mL^{-1}, $d_E = 1$ day^{-1}, $d_P = 0.001$ day^{-1}.

is half-maximal, determines the degree of control of the viral load setpoint. When N_P is smaller, immune cells can still effectively proliferate even when viral load drops, thus maintaining control. The timing and rate of early viral growth - before the immune response has expanded to sufficient levels - is still controlled by λ, β, and a (see SI Figure TBA).

These results are corroborated by formal sensitivity analysis (Methods), which indicates that viral load values early on during rebound tend to provide the most information about the parameters related to target cells (λ), viral fitness (β), and reservoir reactivation (a), whereas values later on have more information about the immune response (m, p, N_P) (Figure 6.3, right).

6.4.3 Estimation of immunotherapeutic treatment effects from viral rebound data

After confirming that our model can qualitatively capture a wide range of viral rebound dynamics, we used a statistically rigorous group-level fitting approach to identify the model parameters from the observed viral rebound data. Our main goal was to compare parameters between groups receiving different combinations of immunotherapy. In inference framework, baseline parameters governing the dynamics in each individual are assumed to be drawn from a shared distribution which allows for heterogeneity between individuals, known as the *random* effects. Fixed *treatment* effects alter the parameters according to the treatment each individual received. Here, the treatments we consider include the TLR7-agonist, the Ad26/MVA therapeutic vaccine, and the PGT121 antibody. We considered animals infected with SIV and SHIV separately, as we expected that many viral dynamic paramters could differ between these virus strains. For the two SIV studies, we also included the study identity (Study 1/Whitney[148] or Study 2/Barouch[29]) as a "treatment" in order to search for systematic differences in rebound kinetics, which are most likely to be caused by the different timing of ART initiation between the studies (9 vs 1 week after infection, respectively). In addition to exploring a collection of biologically-motivated models (Methods, Tables 6.5 and 6.6), we refined

our model using an iterative model selection procedure based on the Bayesian Information Criterion (BIC). Inference was performed using an implementation of the Stochastic Approximation of Expectation Maximization (SAEM)[232]. The Methods section contains details about parameter inference and model selection.

We first examined the effects of the TRL7-agonist and therapeutic vaccine in the studies involving SIV-infected animals (Studies 1 and 2 in Figure 6.1). Our model-fitting procedure reliably identified effects of the immunotherapy on several model parameters, suggesting mechanisms for the efficacy of these treatments (Figure 6.4). We found that therapeutic vaccination reduced the rate of successful reactivations from the latent reservoir ($\uparrow\ t_a$) by 4-fold (95% CI 2-8) and increased the sensitivity of immune cells to antigen stimulation ($\downarrow\ N_P$) by 17-fold (95% CI 5-67). In addition, our fitting supported an interaction between the vaccine and the TLR7-agonist, resulting in an additional 12-fold increase (95% CI 3-47) in immune sensitivity to antigen ($\downarrow\ N_P$) when both therapies were administered together, for a total 210-fold boost. We hypothesize that vaccination, alone and in concert with TLR7-agonist treatment, establishes an adaptive immune response with wide breadth which reduces the fraction of viruses archived in the reservoir which can successfully reactivate, and primes adaptive immune response to expand in the event of reactivation.

Furthermore, we identified several study-specific effects. In the Whitney study (Study 1[148]), viral rebound kinetics supported a 10-fold elevation (95% CI 3-30) in the responsiveness of immune cells ($\downarrow\ N_P$) in all groups, and an 8-fold reduction (95% CI 4-16) in the latent reservoir reactivation rate in the presence of TLR7-agonist treatment (8-fold decrease in time between reactivations, t_a). We hypothesize that the first finding is due to the longer time after initial infection that ART was started in this study compared to Study 2 (8 weeks vs 1 week), which could have allowed for the formation of a more effective memory response. Indeed, previous investigation of immunological dynamics early in acute infection indicate that the timing of ART initiation determines the strength of HIV-specific immune responses and the kinetics of the subsequent rebound[271]. The inferred reduction

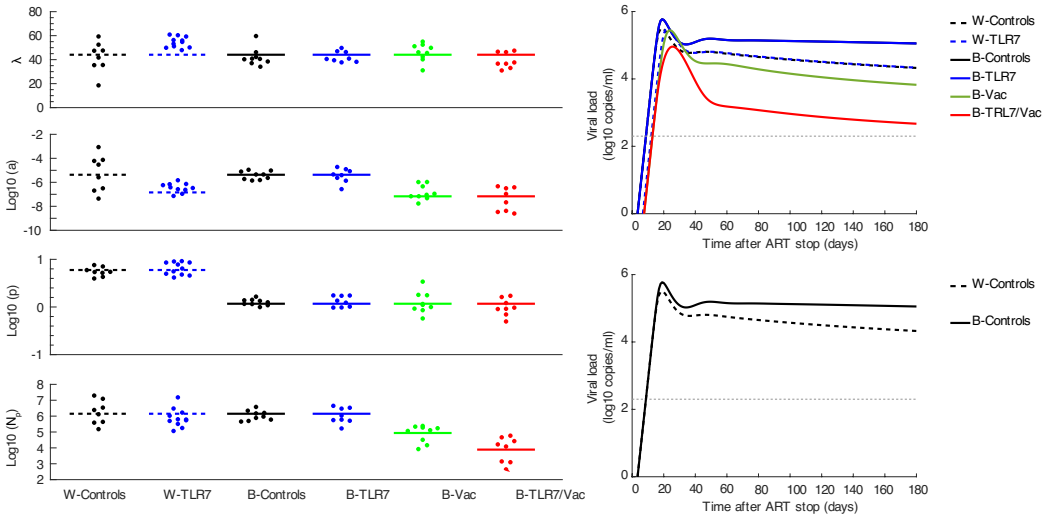


Figure 6.4: A) Individual and group-level mean parameters estimated for each model parameter for the studies conducted in SIV. B) Simulated viral load trajectories using the group-level mean value of each model parameter for all treatment groups in Studies 1 and 2. Note that since there are no group-level differences inferred between the Control and TLR7-agonist groups in Study 2, these curves are on top of each other. C) Comparing the control groups of Study 1 and 2 only. D) Individual and group-level mean parameters estimated for each model parameter for the study conducted in SHIV. Simulated viral load trajectories using the group-level mean value of each model parameter for all treatment groups in Study 3. Model parameters are: λ, target cell input rate, β, viral infectivity, t_a, time between latent cell reactivations, m, immune memory input rate, p, maximal immune proliferation rate, N_P, antigen threshold for immune stimulation.

in the latent reservoir exit rate is consistent with the observation of large viral blips despite ART during TLR7-agonist treatment in the Whitney study[148], which suggests reactivation and clearance of latently infected cells. This reduction was inferred even if we excluded the two animals who never rebounded (outliers with large values of t_a in Figure 6.4A).

We next examined the impact of the TLR7-agonist and the monoclonal antibody alone and in combination in SHIV-infected animals (Study 3 in Figure 6.1,[30]). The SHIV data exhibited greater variability than the SIV data overall: there was greater heterogeneity in the time to viral rebound and larger differences between peak and setpoint viral load. Furthermore, the residual error inferred by our fitting procedure was larger for the SHIV data compared to SIV, suggesting our model was better able to capture the dynamics in the SIV data. Nonetheless, we found that PGT121 antibody administration produced a 2.2-fold reduction (95% CI 1.7-2.9) in viral infectivity ($\downarrow \beta$) and a 7-fold improvement (95% CI 3-16) in the responsiveness of immune cells ($\downarrow N_P$) during rebound. While the mechanisms causing these effects remain unclear, PGT121 antibody administration might eliminate more fit viral strains from the latent reservoir, leaving behind less-infectious virus. Antibody administration might also have boosted endogenous antiviral immune responses, as has been previously observed[187].

To evaluate the robustness of our results, we used three different algorithms for selecting the optimal combination of treatment effects and for each of these we tested optimization based on both Bayesian Information Criteria (BIC) and log-likelihood (see Methods). We also varied the initial conditions for all parameter values fed to the model. In all cases, the models we report are robust to these variations. Despite the theoretical identifiability of all parameters we fit for, there was sometimes evidence of mutual information shared between the viral infectivity β and the target cell density λ. We therefore tested that an alternate model structure to our best-fit selection, which swapped the location of a treatment effect between β and λ, was indeed worse. The same procedure was conducted for treatment effects on N_P, p, and m, which all describe some aspect of the immune re-

sponse and hence could be hard to separate. We again found our selected model was optimal (Tables 6.5, 6.6). Before beginning the selection procedure, we defined a set of models based on biologically-motivated hypothesis about the potential effects of these treatments, and we later tested that none of these models were better than the selected one. All these results are reported in Tables 6.5 and 6.6. In Study 1 and Study 2 there was some variation in the doses of the TLR7-agonist given between animals (Table 6.7), and after including dose as a covariate in the model we could find no discernible influence on any kinetic parameters.

Finally, we carefully explored the relationship between the different treatment effects associated with each best-fit model. After finding the best-fit model in each study, we tested support for that model structure in data from the other study. We find that the model structure identified in the SHIV study is strongly disfavored by the SIV data. However, the model structure identified by the combination of SIV studies has nearly as much support in the SHIV data as the best-fit SHIV model. Accordingly, the specific parameters on which the combination of TLR7 and PGT121 treatment in SHIV are acting is less clear.

6.4.4 Predicting the effects of immunotherapeutic treatment in humans

Finally, we used these results to predict how TLR7-based immunotherapies would alter rebound kinetics in human trials. Our approach was to first develop a calibrated model of HIV rebound, and then simulate the model after adding in the treatment effects (for the antibody, vaccine, etc) that we identified in the SIV and SHIV studies. To characterize HIV rebound, we assembled data from a series of clinical trials that included treatment interruptions after long-term suppressive antiretroviral therapy initiated during chronic infection, totaling 69 individuals sampled at least weekly[147]. We fit our mathematical model to these viral rebound trajectories to determine the population-level distribution of the model parameters (see Methods). Comparing the rebound kinetics between SIV, SHIV, and HIV (Table 6.8), we found a large differences in the parameters estimated for differ-

ent viruses. The rate of reactivation from the latent reservoir was estimated to be highest for HIV (smallest time between reactivations, t_a), followed by SIV and then SHIV. This parameter is not scale-invariant and instead depends on the absolute number of latently infected cells, so the larger body size of humans as compared to macaques likely explains the apparent difference number of cells reactivating per day. The target cell density was inferred to be largest for HIV ($\uparrow \lambda$), whereas SHIV was found to have both the highest intrinsic viral infectivity ($\uparrow \beta$) and most sensitive immune response ($\downarrow N_P$). The inter-individual variation in rebound trajectories (only considering control animals) was low for SIV but higher for HIV and SHIV.

To predict how a human cohort might respond to immunotherapy treatment, We next conducted simulations where we altered the baseline parameters for HIV rebound by the immunotherapy effects identified in this study. Underlying this approach is the assumption that each therapy component will have the same relative effect in humans as in macaques (e.g. a five-fold reduction in reservoir reactivation). All single and combination immunotherapies involving the TLR7-agonist, the Ad26/MVA vaccine, and the PGT121 antibody were simulated in a hypothetical population of 200 individuals, and we calculated the distribution of peak viremia, setpoint viral load, and time to rebound for each case (Figure 6.5, 6.7-6.12). These simulations predict that the TLR7-agonist/vaccine combination could be effective in humans: 42% of simulated individuals rebounded above the detection threshold and then subsequently controlled viremia below it, and another 17% had no detectable viral rebound for a year after treatment interruption. The combination of TLR7-agonist/PGT121 Ab treatment is predicted to result in a dramatic suppression of viral rebound in humans, primarily because the antibody's reduction of viral infectivity (β) is sufficient to push the viral growth rate below the critical threshold $R_0 = 1$ in many individuals. Figures 6.7, 6.8, 6.9, 6.10, 6.11, and 6.12 show predictions for many other hypothetical treatment scenarios in humans and macaques.

6.5 Discussion

In this study, we developed and applied a joint stochastic and deterministic model of viral rebound with immune control to reveal treatment effects in several ATI immunotherapy trials. Using a statistically rigorous group-level fitting framework, we identified treatment effects on several immunological parameters and on the latent reservoir itself. Notably, we identify multiple synergistic effects between TLR7 agonist treatment and vaccination or antibody treatment. Beyond suggesting treatment effects, our results highlight the impact of ART initiation time on later rebound and suggest explanations for differences between dynamics in SIV, SHIV, and HIV. Finally, the work provides a framework with which to make predictions, however imperfect, for outcomes in human trials.

Our approach is not without limitations. Even this expanded model is unable to capture certain patterns in viral rebound observed in these studies. Features of viral rebound which reflect these processes might be erroneously attributed to parameter differences and treatment effects in our fitting. For example, the model fails to account for the continued increase in viral load, characteristic of the transition to immunodeficiency, in subject 13 at the end of the Whitney study. Our model also does not account for within-host evolution of the virus, and our model of immunity is very simplified. Future longitudinal sampling of immunological covariates will be critical for the development of more detailed dynamical models of systems immunity in the context of infection. The much larger unexplained variability associated with SHIV relative to SIV suggests these factors might be especially important for producing more accurate models of SHIV infection.

Because viral load was the only densely-sampled quantity in these studies, many model parameters were fixed in order to improve identifiability. Therefore, absolute values of the fit parameters are less reliable than trends observed between groups. This lack of identifiability also contributes to uncertainty about the location of treatment effects on particular parameters: a different choice of fixed parameters could lead to different treatment effect locations, especially among the parameter

groups [λ, t_a, k] and [m, N_E]. Additional data supporting the treatment effects we inferred and these alternative possibilities are discussed below.

Our statistical approach assumes similarities among the model parameters across individuals and assumes a particular form for treatment effects which influence the interpretation of our results. Though the population similarity assumption provides additional statistical power, if it is severely mis-specified, the results will not be reliable. We assumed multiplicative treatment effects for ease of interpretation, but our finding of treatment interactions must be interpreted in the context of this underlying assumption. Our model priors—log-normal—were designed to allow for a wide range of fitting results over all possible positive parameter values, but in reality, certain additional parameter regimes are implausible a priori. Finally, we adapted tools designed primarily for fitting deterministic models to the problem of joint stochastic-deterministic viral rebound (see methods). Future work might approach fitting fully stochastic simulations of viral rebound while maintaining this group structure among individuals.

Despite these limitations, this work makes several important contributions. First, we apply a minimal model of adaptive immunity which can give rise to a wide range of rebound dynamics to data from several studies. The model offers an alternative to previously proposed "bi-stability" explanations for low set-point viral load[60].

Additionally, our model merges the stochastic reactivation phase with the dynamics of viral rebound so that the combined effects of parameters in each regime can be assessed coherently. Long waiting times to viral rebound have been observed in individuals with very low latent reservoirs, suggesting that the stochastic reactivation regime will become only more relevant as therapies which reduce the size of the latent reservoir improve. In addition to stochastic exit from the latent reservoir compartment, our model also captures the probability that a reactivating cell gives rise to productive infection, which is itself a function of the model parameters. Because our stochastic model is integrated with later deterministic dynamics and accounts for this probability, our approach can

capture information from the whole course of viral rebound to more precisely estimate the rate of reactivation from the latent reservoir when compared to stochastic reactivation models alone[206,59].

Second, our model provides insights into mechanisms of drug action which are supported by additional data. In Study 1, TLR7 agonist administration was inferred to reduce the rate of latent reservoir reactivation. The observed reduction in integrated SIV DNA and two negative viral outgrowth assays in this group support the hypothesis that this treatment reduced the size of the latently infected population. Detectable viral "blips" concurrent with the administration of TLR7 agonist also suggest that TLR7 treatment has the capacity to reactivate, and thus presumably lead to the clearance of, a significant proportion of latently infected cells[148]. TLR7 agonist administration was also inferred to increase the rate of target cell production. Whether TLR7 treatment can effect a long-term change in the amount of target cells—a rare CD4 T cell subset which is difficult to measure directly—has not been directly assessed. However, TLR7 has been shown to transiently alter the proportion of activated cells in several immune subsets[148]. A change in the rate of target cell production might also be reflective of a more generally healthy immune system.

In the Barouch study, vaccination was inferred to reduce the rate of latent reservoir reactivation. One possible explanation for this effect is that the latent reservoir archives a large diversity of viral sequences[249,40,42,119], and after vaccination, many of them will not be able to give rise to productive infection. Vaccination might also have led to the creation of an immune response which began eliminating latently infected cells even before the end of antiretroviral therapy. Vaccination is also inferred to lead to the presence of an immune cell population which is very sensitive to virus. The creation of such a long-lived memory immune population is a hallmark of successful vaccination. The interaction effect between TLR7 agonist treatment and vaccination produces an even more sensitive immune population. This effect is supported by the role of TLR7 as a bridge between innate and adaptive immunity[149].

Our fitting results suggest a difference between studies 1 and 2. In study 1, the baseline level of NP

was reduced and TLR7 agonist treatment increased both the rate of target cell production ($\uparrow \lambda$) and immune sensitivity to antigen ($\downarrow N_P$). We ascribe these differences between the studies to the later ART start time in the Whitney study. Later ART start time has been shown to increase the size of the latent reservoir[271], impact the formation of memory immune responses[271], and might also affect immune exhaustion or the potential for viral evolution.

As a result of parameter unidentifiability, the effects of TLR7 and vaccine treatment on the target cell replenishment rate λ and time between latent cell reactivations ta might be explained by an effect on a different pair of parameters (t_a and k or λ and k) instead. However, we believe the effects on ta and lambda are more likely because they are supported by direct measurements of integrated viral DNA and known mechanisms of TLR7.

In the SHIV study, the data are generally more challenging to interpret. Overall, the amount of variability in these data are much larger; we obtained a best-fit error term more than twice as large for the SHIV data as compared to the SIV data. We find that the effect of antibody treatment is, first, to reduce the infectivity of virus. This is surprising because the antibody is no longer present when antiretroviral therapy is withdrawn. However, if antibody is eliminating certain viral strains from the reservoir, perhaps only the more mutated, and therefore possibly less fit, viral strains avoid elimination. Antibody treatment is also inferred to produce an immune response which is more sensitive to virus. This also might be the result of selection for mutated viral strains against which it is easier to mount an effective immune response. Finally, in this study TLR7 was inferred to increase the basal rate of immune precursor production, altering the initial size of the immune response against virus at the time of ART cessation. One possible explanation for this observation is that TLR7 leads to the permanent expansion of SHIV-specific immune responses just as it might have led to the expansion of the target cell population in the SIV context. Because SIV and SHIV viral particles have different exteriors, it is possible that the true target cell populations for these viruses are different[188], explaining the lack of an effect on the target cell replenishment rate λ in this study.

Another possible explanation is that the treatment effect on m is not identifiable with an effect on N_E, the per-cell effectiveness of immune effectors. Previous systems immunologic analysis suggests that outcomes in this study are connected to NK cell function, and TLR7 signaling plays a critical role in NK cell function, at least via cytokines produced by dendritic cells[1]. TLR7 treatment might lead to the creation of a mature dendritic cell population very efficient at antigen presentation, thereby reducing N_E. The fact that the best-fitting model did not support an effect of antibody treatment on the rate of latent reservoir reactivation is different from the mechanism of action hypothesized in the original study. However, we cannot exclude this mechanism because competing models of the effect of antibody and TLR7 treatment on the rate of latent reservoir reactivation had only slightly worse statistical support (see Table 6.5).

Third, by merging HIV rebound data with immunotherapy treatment effects inferred by our model, we generated hypotheses for the outcomes of different immunotherapeutic agents in humans. Treatments which move closer to viral control in humans are expected to produce higher variability in rebound dynamics and give rise to dynamics below the limit of quantification for the most common HIV RNA assays. The possibility of rebound dynamics with high peak viral load followed by control also complicates the design of human trials. Future trials in humans and also present an opportunity to build on the knowledge gained from the animal models analyzed here. Increasing the density of sampling after ART cessation and continuing post-peak viral load, expanding the collection of longitudinal immune covariates, precisely measuring drug washout kinetics in each individual would all contribute to an improved understanding of treatment effects. In macaques, using barcoded virus would also contribute to an improved estimate of latent reservoir reactivation rate. Overall, if the therapeutic effects of these treatments transfer to a human context, our results suggest they will dramatically alter the course of viral rebound.

6.6 Methods

6.6.1 Data

The design of the studies we considered is explained in detail elsewhere[29,148,30], but summarized here. All studies were conducted in rhesus macaques, which were infected intrarectally with the SIVmac251 virus and later treated with the combination antiretroviral therapy (ART) regime of tenofovir, emtricitabine, and dolutegravir (TFV/FTC/DTG). In the first study, all animals were given ART starting 65 days after infection, and were treated with ART for at least 400 days, before ART was stopped. During ART, some animals were additionally administered repeated doses of a TLR7-agonist (treatment group), while others received a placebo (control group). The study was divided into several phases and arms, in which the treatment groups received slightly different courses of the TLR7 therapy (see Figure 6.1A-B, Table 6.2). Overall, there were 8 control animals and 13 TLR7-agonist treated animals. In the second study, all animals started ART after only 1 week of infection, and were treated with ART for 500 days before stopping. Animals were divided into four groups of 8-9 individuals each - one control group who only received ART, one group who additionally received the TLR7-agonist (10 doses 2 weeks apart), another group who additionally received a prime-boost vaccine regimen (2 doses of Ad26 followed by 2 of MVA, each 12 weeks apart), and a fourth group who additionally received both the TLR7-agonist drug and the vaccine regimen. In all animals in both studies, there was at least 2 weeks between the time the last immunotherapy intervention was given and when ART was interrupted, which was chosen to insure that any non-ART treatment had washed out of the system by the time ART was stopped. This way, any change in viral rebound kinetics caused by the intervention must be due to a permanent perturbation made to the system, and not a direct inhibitor effect of the immunotherapy. In all studies, viral load was measured every 3-4 days after ART cessation. Additionally, viral load values were measured during acute infection and during ART administration. In all cases viral load values below the detection

limit of assays (50 or 200 copies/mL) are censored.

In the first study with TLR7-agonist only, all but two animals experienced persistent viral rebound. Two animals in the TLR7-agonist treatment group never experienced detectable viral load after ART was stopped. This represents the first observed case of a potential sterilizing cure in this SIV system. In the second study with TLR7-agonist and vaccine, all animals initially rebounded, but three animals treated with the combined immunotherapy eventually re-suppressed virus below the detection limit. The rebound trajectories of all animals are shown in Figures 6.1 (C-H).

6.6.2 Model development

The basic viral dynamics model used to describe HIV infection before and during antiretroviral therapy can also be used to describe the rebound of infection when treatment is stopped (which displays similar kinetics to acute infection). This often-used model reproduces many aspects of infection kinetics, such as exponential increase in viremia after initial infection or rebound, declining viral load after a peak is reached, and eventual stabilization at a "set point". However, it cannot describe the diversity of viral rebound trajectories seen in these studies - such as large declines in viral load from peak to setpoint or eventual post-rebound control - and does not explicitly consider viral latency nor antiviral immune responses. To address these issues, we developed an augmented model of HIV/SIV infection dynamics which incorporated ideas from multiple different existing models of various viral infections (Figure 6.2). We previously used this model in a preliminary analysis of a subset of this data[29].

The model we used is described by a system of ordinary differential equations that track changes in the levels of uninfected (T) and infected (I) target cells, free virus (V), and precursor (P) and effec-

tor (E) immune responses over time (Figure 6.2):

$$\begin{aligned}
\dot{T} &= \lambda - \beta TV - d_T T \\
\dot{I} &= a + \frac{\beta TV}{1 + (E/N_E)} - d_I I \\
\dot{V} &= kI - cV \\
\dot{P} &= m + p(1-f)\frac{V}{V+N_P}P - d_P P \\
\dot{E} &= pf\frac{V}{V+N_P}P - d_E E
\end{aligned} \tag{6.1}$$

This model assumes that infection is well-mixed throughout the blood and other lymph tissue, ignoring any spatial structure or compartmentalization. All variables are expressed as concentrations (per mL of plasma). Model variables and parameters are summarized in Table 6.1.

Susceptible, uninfected target cells (T) are produced at a constant rate λ and die at a per capita rate d_T. These cells are assumed to be CD4+ T cells, but may only be a subset of the total CD4+ population. Although the specific phenotype of CD4+ T cells that confers susceptibility is not completely clear, it is known that activated cells are more susceptible to infection than resting cells, and that only a small fraction of all CD4+ T cells are productively infected even at peak viremia (more may be abortively or latently infected). Here we ignore heterogeneity in infected cell subpopulations.

New infections occur proportionally to the density of free virus (V), target cells (T), and the infectivity rate β. Infected cells (I) release virus at rate k and die at a rate d_I. Free virus is cleared at rate c. We do not explicitly track latent infection, since it only significantly impacts infection levels when viral loads are very low, but instead use parameter a to describe the rate which latently infected cells reactivate to produce productive infection (which is necessary to kick-start rebound). This rate incorporates both the number of cells latently infected with intact virus as well as their per-

Table 6.1: Variables and parameters of the viral dynamics model (Eq. (6.43), Figure 6.2.

	Description	Units	Value	Source
Variables:				
T	(Uninfected) target cells	cells mL^{-1}		
I	(Productively) Infected cells	cells mL^{-1}		
V	Free virus	RNA copies mL^{-1}		
P	Precursor immune cells	cells mL^{-1}		
E	Effector immune cells	cells mL^{-1}		
Fixed parameters:				
d_T	Death rate of target cells	day^{-1}	0.05	
d_I	Death rate of infected cells	day^{-1}	0.4	
d_P	Death rate of precursor immune cells	day^{-1}	0.001	
d_E	Death rate of effector immune cells	day^{-1}	1	
c	Virus clearance rate	day^{-1}	23	
k	Virus production rate	virions $cells^{-1}$ day^{-1}	50 000	
m	Production rate of precursor immune cells	cells mL^{-1} day^{-1}		
f	Fraction of effector immune cells that don't revert to memory		0.9	
N_E	Effector concentration at which half-maximal inhibition occurs	cells mL^{-1}	10 000	
Estimated parameters:				
λ	Production rate of target cells	cells mL^{-1} day^{-1}		
β	Viral infectivity	mL $copies^{-1}$ day^{-1}		
a	Latent cell reactivation rate	cells day^{-1}		
p	Maximum proliferation rate of immune cells	mL $copies^{-1}$ day^{-1}		
N_P	Viral load at which half-maximal proliferation occurs	copies mL^{-1}		

capita rate of reactivation and the probability that they produce a lineage that escapes extinction and reaches detectable viremia.

We include a relatively general model of an antiviral immune response, which could represent cellular or humoral effects. Long-lived precursor immune cells (which includes both naïve and memory subsets) are produced at a baseline rate m. If this model were used during acute infection, m would be related to the frequency of naive precursors, whereas during viral rebound, m is dominated by the reactivation of memory cells formed during acute infection. In response to antigen (assumed here to be free virus, but could instead be infected cells), these cells are stimulated to proliferate at an antigen-dependent rate. The maximum proliferation rate is p and half-maximal proliferation occurs at viral load N_P. [Explain functional form of relationship, cite de Boer, Perelson work] A fraction f of all proliferating cells become short-lived effectors (E), while the remaining fraction will return to a long-lived memory state (P). Long-lived precursor immune cells die at rate d_P and short-lived effectors die at rate d_E. Effectors reduce the rate at which actively infected cells are produced (either by inactivating free virus or killing early-stage infected cells[89]), with half-maximal inhibition occurring at a concentration N_E. We could also have modeled effectors as killing infected cells, though previous work has shown that this mechanism is only consistent with existing data if the model explicitly includes an extra compartment for early stage infected cells that are not yet targeted by cytolytic immune responses[89], we have included non-lytic effects, and either model has similar qualitative behavior during rebound.

During ART, we assume infection is completely blocked ($\beta = 0$), and that treatment is given for long enough that virus and cells reach steady states, which we take as the initial conditions at the

time of ART interruption:

$$
\begin{aligned}
T_0 &= \frac{\lambda}{d_T} \\
I_0 &= \frac{a}{d_I} \\
V_0 &= \frac{a}{d_I}\frac{k}{c} \\
P_0 &= \frac{m}{(d_P - p(1-f)\,V_0/(\,V_0 + N_P))} \\
E_0 &= \frac{pfP_0\,V_0}{d_E(\,V_0 + N_P)}
\end{aligned}
\tag{6.2}
$$

6.6.3 Model parameters and identifiability

The data available for fitting this model consists only of longitudinal values of viral load (V), and therefore all parameters of the model are unlikely to be identifiable. While longitudinal values of total CD4+ T cells were collected, the relationship between this number and target cell density T is unclear for the reasons discussed in the previous section, and, this measurement is notoriously noisy in non-human primates sampled under anesthesia. Actively infected cells (I) are difficult to quantify separately from forms of latent or defective infection. Characterization of anti-viral immune responses (related to P, E) was only done once before and after rebound.

We conducted both analytic and numeric investigation of the model to determine principled ways to reduce the number of parameters to be estimated from the data. We simulated the model under a wide range of parameter conditions, systematically varying one parameter at a time, to understand the role that each parameter played in the viral rebound trajectories (Figure 6.2). In addition, we applied the differential algebra algorithm DAISY[20] to determine the formal identifiability of the model parameters, given perfect measurement of free virus levels.

When only viral load is observed, one of the parameters from each of the sets $\{\lambda, a, k\}$ and

$\{N_E, m\}$ is always non-identifiable. Therefore we fixed viral burst rate to $k = 5\times10^4$ virions/cell/day[51], and immune-response efficacy $N_E = 10^4$ cell/day. The latter value is chosen arbitrarily, since inhibition by immune cells is extremely difficult to measure quantitatively in experiments. However, the choice is inconsequential and only results in a scaling of the inferred m value.

While most other parameters are theoretically identifiable, many are practically impossible to infer from the available data. The fast time-scale of virus relative to infected cells makes the viral clearance rate c practically non-identifiable, and so we fixed this value to c=23/day based on plasma aphoresis studies[212]. Cell death rates are very hard to identify from viral loads during active infection, and didn't have a large influence on viral load kinetics within a certain range, so we fixed values from isotope-labeling studies in the literature (/day) as 0.05 for target cells, 0.4 for infected cells, 0.001 for precursor immune cells, and 1 for effector immune cells. Target death rate was estimated based on the estimated death rate of a fast subpopulation of CD4+ memory T cells in rhesus macaques (0.05 in uninfected animals, 0.1 in those with high SIV loads)[90] and of activated memory CD4+ T cells in humans (0.08)[156]. Infected cell death rates were taken from the rate of viral load decline during ART for SIV observed in previous analyses[271,29,148]. The death rate of precursor immune cells, which actually represents the net decay combining cell death and homeostatic (antigen-independent) proliferation, was roughly estimated from turnover rates of slow-proliferating memory CD8+ T cells in uninfected macaques (0.0025)[90], decay of human CD8 responses to yellow fever (0.006)[2] and smallpox (0.0002)[104] vaccination, and decay of murine responses to LCMV (<0.0005)[52]. The death rate of effector immune cells was estimated from LCMV infection in mice (0.4)[27,67] and acute mononucleosis infection in humans (0.8)[156]. We assumed the fraction of proliferating effectors that return to a long-lived memory state was (1-f)=0.1, consistent with experiments in multiple animals (reviewed in[62]), and results were insensitive to values other than those very near 0 or 1.

This left a model with six remaining unknown parameters: β, λ, a, m, p, N_P. We repeated the

formal identifiability analysis in DAISY with this reduced model. We confirmed this reduced model is locally identifiable, and when m is fixed the model is globally identifiable. To understand what features of the viral rebound kinetics would be most informative of each of these parameters, and to understand how the density of samples may impact identifiability, we assessed the sensitivity of viral load (V) to each parameter (generically denoted θ) over time by evaluating $\frac{\partial V}{\partial \theta}\frac{\theta}{V}$, taking into account the role of the parameter in both the initial condition of the system and the subsequent time evolution.

6.6.4 Model fitting

The same model fitting procedure was used for SIV-TLR7-VAC and the SHIV-TLR7-AB data. However, the two datasets were not combined as it was assumed that values for parameters will be highly different in the two population as the virus studied was different (SIV vs. SHIV).

The model in Equation (6.43) was fit to the longitudinal viral load data to estimate the values of the parameters. Briefly, in this framework, the parameter values for all individuals in the population are assumed to be drawn from a common distribution, and the goal is to estimate the mean and variance of this hyper-distribution. To ensure positivity, all parameters are estimated in log-transformed scale, denoted as $\log(\theta) \rightarrow \tilde{\theta}$. Individuals in the study may receive different treatments, and the mean parameter value may be shifted by each treatment in a different way. More explicitly, this statistical model assumes that the value of parameters $\theta_i = (\beta_i, \lambda_i, ta_i, m_i, p_i, N_{Pi})$ in individual i can be broken down into the following components

$$\begin{aligned} \tilde{\theta}_{ij} = log(\theta_{ij}) &= \mu_j + \sum_k G(i,k)\omega_{jk} + \sum_{k \neq k'} G(i,k)\, G(i,k')\omega_{jkk'} + \eta_{ij} \\ \eta_i &\sim \mathcal{N}(0, \Sigma) \end{aligned} \tag{6.3}$$

where $j \in \{\beta, \lambda, a, m, p, N_P\}$.

The first term μ_j, known as the *fixed effect*, is the mean value of parameter $\tilde{\theta}_j$ across the population in an individual who does not receive any of the treatments. The logical matrix $G(i,k)$ describes the 'treatments' each individual received. If individual i received treatment k, then $G(i,k) = 1$ whereas it is 0 otherwise. In this study $k = 1..4$, as we consider an effect of the TLR7-agonist, of the therapeutic vaccine, of the study identity (timing of ART initiation) and of the antibodies. The terms ω_{jk} describe the *treatment effects*: treatment k modifies the mean of parameter j by an amount ω_{jk}. We also consider the interactions between these treatments that are identifiable by design, i.e. therapeutic vaccine $\times$ TLR7-agonist, TLR7-agonist $\times$ study effects and antibodies $\times$ TLR7-agonist. The terms $G(i,k)\,G(i,k')\omega_{jkk'}$ describe the *interaction effects*: combination of treatment k and treatment k' modifies the mean of parameter j by an amount $\omega_{jkk'}$. The final term, η_{ij}, is known as the *random effect*, and describes the amount by which the observed parameter value θ_{ij} differs from the expected mean in the treatment group. The *random effects* of an individual i, η_i, are assumed to be normally distributed with diagonal variance-covariance matrix $\Sigma = diag(\epsilon_j)_{j=1...5}$. It means that is is supposed independent for each parameter that appears in Eq. (6.43).

For example, in the SIV-TLR7-VAC data, if we are considering the parameter for the time for reservoir reactivation ta ($ta_i = e^{\tilde{ta}_i}$), then this may be affected by TLR7-agonist, therapeutic vaccine administration, the study identity (timing of ART initiation), and interactions between these interventions. The equation for the components of this parameter becomes:

$$\begin{aligned} \tilde{ta}_i = \mu_{ta} &+ G(i,\, TLR7)\omega_{ta,\,TLR7} + G(i,\, Vac)\omega_{ta,\,Vac} \\ &+ G(i, Study)\omega_{ta,Study} + G(i,\, TLR7)\,G(i,\, Vac)\omega_{ta,\,TLR7\mathrm{xVac}} \\ &+ G(i,\, TLR7)\,G(i, Study)\omega_{ta,\,TLR7\mathrm{xStudy}} + \eta_{ta,i} \\ \eta_{ta,i} &\sim \mathcal{N}(0, \epsilon_t a^2) \end{aligned} \tag{6.4}$$

We assumed that the data are observed with measurement error, thus we defined a residual er-

ror model. In other words, the observed $\log_{10}$ viral load of patient i at time t, denoted $Y_i(t)$, is normally-distributed with a constant error ($\zeta_1 \rho_i(t)$) and a proportional ($\zeta_2\, Y_i^*(t)\rho_i(t)$) error term around the true viral load $Y_i^*(t)$ computed from the model derived from Eq. (6.43). This model referred as combined error model is classic when considering blood counts. It is expected that the magnitude of the error may depend on a constant measurement error and on the magnitude of the quantity observed[210] and writes:

$$\begin{aligned} Y_i(t) &= Y_i^*(t) + (\zeta_1 + \zeta_2\, Y_i^*(t))\rho_i(t) \\ Y_i^*(t) &= log10[\,V(t, \beta_i, \lambda_i, ta_i, m_i, p_i, N_{Pi})] \\ \rho_i &\sim \mathcal{N}(0, 1) \end{aligned} \tag{6.5}$$

Finally, many observed viral load values in these studies are left-censored, because they are below the detection limit of available assays (either 50 or 200 copies/mL, depending on the study). All these data points were included in our fitting. but considered as censored.

We used the Monolix software (MonolixSuite2019R2) to estimate the values of all the model parameters (μ, ω, ϵ and ζ) by maximizing the likelihood of the data given the model and parameters[236]. We assumed good pratical identifiability, thus we did not use prior knowledge about the values of the parameters we attempted to fit, and so therefore did not implement a penalized likelihood strategy. The software uses a frequentist version of the stochastic approximation expectation maximization (SAEM) algorithm[70]. SAEM is an iterative algorithm that essentially consists of constructing Markov chains that converge to the conditional distributions of the parameters given the data. The final parameters estimates are given by the mean parameters values over the iterations during the smoothing phase of the Markov chain. The standard errors of these parameters represent the uncertainty of the estimated population parameters. They are calculated via the estimation of the Fisher Information Matrix[138]. It is derived from the second derivative of the log-likelihood which

is evaluated by importance sampling[222]. All the statistical aspects of this fitting approach has been described elsewhere[47].

6.6.5 Model Selection

In order to focus on the strongest effects and control the number of hypotheses, we determined the location of treatment effects in each dataset by employing an iterative model selection approach. Starting from a model without treatment effects but with random effects on each parameter, we applied two independent variable selection methods.

First, we applied the method Stepwise Covariate Modeling (SCM). SCM has two phases: first, each covariate's potential effect on each parameter is examined, and the effect leading to the best improvement in the target criteria is incorporated in the following iteration. Next, when no further improvement in the target criteria is possible, each previously introduced effect is removed if it fails to contribute to an improvement in the target criteria. Second, we applied the method COnditional Sampling for Stepwise Approach based on Correlation tests (COSSAC). The COSSAC method also has two phases and is similar to SCM but less exhaustive. COSSAC introduces the covariate most correlated with the random effects of a paramater into the next iteration of model fitting. When this simpler incorporation procedure stops yielding an improvement in the target criteria, a backward elimination is performed as in SCM. Both methods are illustred in[50]. We used both likelihood and BIC as target criteria. After each selection procedure, we verified that the sign of a treatment effect was inferred unambiguously, so that it could be assigned a clear biological interpretation. To do so, we used a Wald test at level 5% to determine if treatment effects differed significantly from zero, $\omega_{jk} \neq 0$.

In the SIV-TLR7-VAC data, we ran the selection procedure on each of the five treatment covariates: TLR7-agonist, therapeutic vaccine, study identity (ART initiation time) and the two available interactions TLR7-agonist-therapeutic vaccine and TLR7-agonist-study identity. In the SHIV-

TLR7-AB data, we ran the selection procedure on each of the three treatment covariates: TLR7-agonist, antibodies, and the interaction bteween TLR7-agonist-antibodies.

In the two datasets, effects were tested on parameters ($beta$, N_P, ta, m). After the selection performed, the effects found on $beta$ were tested on $lambda$ and effects found on N_P were tested on p. When fitting abilities were similar according to the target criteria for each permutations, the effect was kept on the parameters which led to the most straightfoward biological explanation.

Repeated fitting suggests that the SAEM algorithm employed by Monolix converged to a global maximum. We ran 100 final estimations with high number of iteration in the burn-in, exploratory, and smoothing phase of the SAEM algorithm. Each fitting produced cosistent final estimates, and we selected the best in term of maximization of the log-likelihood for further interpretation. These highest-likelihood estimates of various effects were considered as final and presented in the article. We confirmed that this last model was the best in term of BIC compared to all other models tested. We also investigate over-parametrization and thus overfitting by checking the ratio between the largest and the smallest eigen value of the Fisher Information Matrix, which remained small, in both cases around 100 (smallest 0.033, largest 4.1 for SIV-TLR7-VAC and smallest 0.049, largest 3.8 for SHIV-TLR7-AB).

Misspecification in the structural model, the error model, and the covariate model can be detected by discrepancies between the observed percentiles and their prediction intervals. Visual predictive checks illustrate these intervals. We found that the model was even able to fit the two animals in the TLR7-agonist treatment group who never experienced detectable viral load after ART was stopped.

In order to characterize the robustness of our results, we investigated several perturbations to the model and fitting procedure. In the statistical model described in Eq. 6.3, we assumed an on-off effect of TLR7 agonist, i.e. $G(i,\ TLR7)$ is 0 or 1. We tested several model accounting for the design of TLR7 agonist administration (different number of doses and concentrations administered in

subsets of monkeys). The covariate $G(i, TLR7)$ was replaced by a variable representing the number of doses, the average dose, the cumulative dose and the maximal dose of drug administered. None of these modelings produced a better fit in term of log-likelihood or BIC.

In the SIV-TLR7-VAC data, we investigated if the effect of the interaction between study identity and TLR7 agonist could be replaced by study effect of TLR7 effect alone. However, such a model exhibits suboptimal log-likelihood and BIC values. We also verified that, when restricting the data to control group monkeys in each study, no strong study identity effects appear. Additioanally, we ran an analysis excluding the two nonrebounding monkeys. No major change in the type and magnitude of the treatment effects was observed.

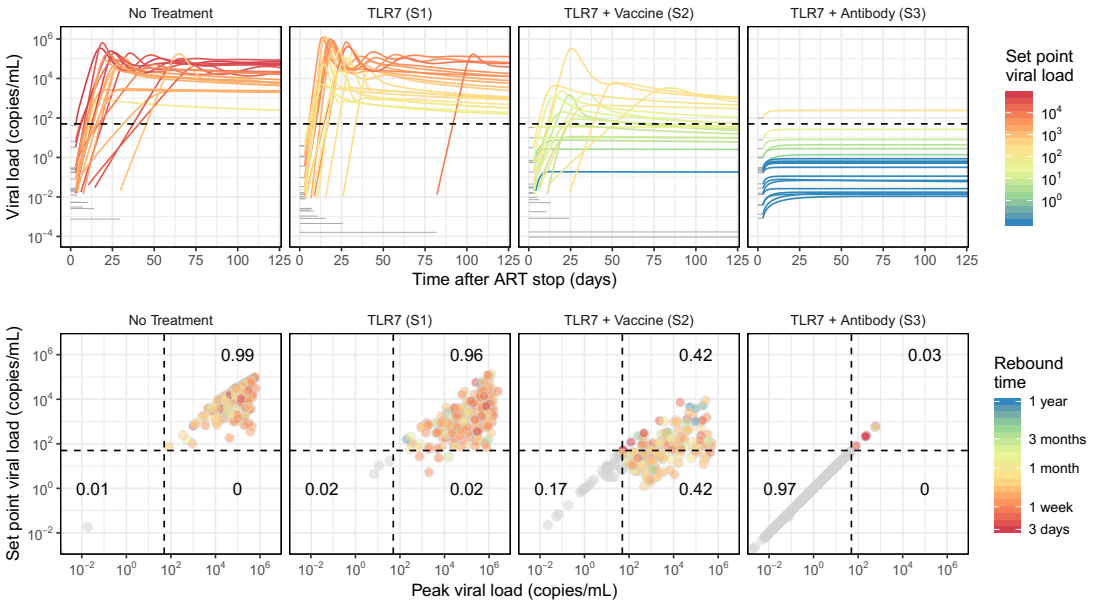


Figure 6.5: Viral rebound trajectories were simulated by combining baseline population heterogeneity from human HIV rebound data with treatment effects from the macaque studies. To simulate an individual rebound trajectory, a set of parameters was sampled from the human population fitting results and then the relevant treatment effects were applied. The top row shows 20 example rebound trajectories colored by viral load. Viral load prior to succesful rebound is illustrated by a grey horizontal line at the level expected according to the simulated reservoir size. Each trajectory is colored according to its viral load at one year. The bottom row summarizes 200 simulations for each treatment according to their peak and final viral loads. Each dot summarizes a particular individual and is colored according to the first time at which that individual rebounded (crossed 50 copies/mL). Individuals who never crossed this threshold are shown in gray. Dotted lines indicate the detection threshold for standard viral load assays (50 copies/mL) and dark black numbers show the proportion of individuals falling in the indicated quadrant. Viral rebound trajectories were simulated for one year.

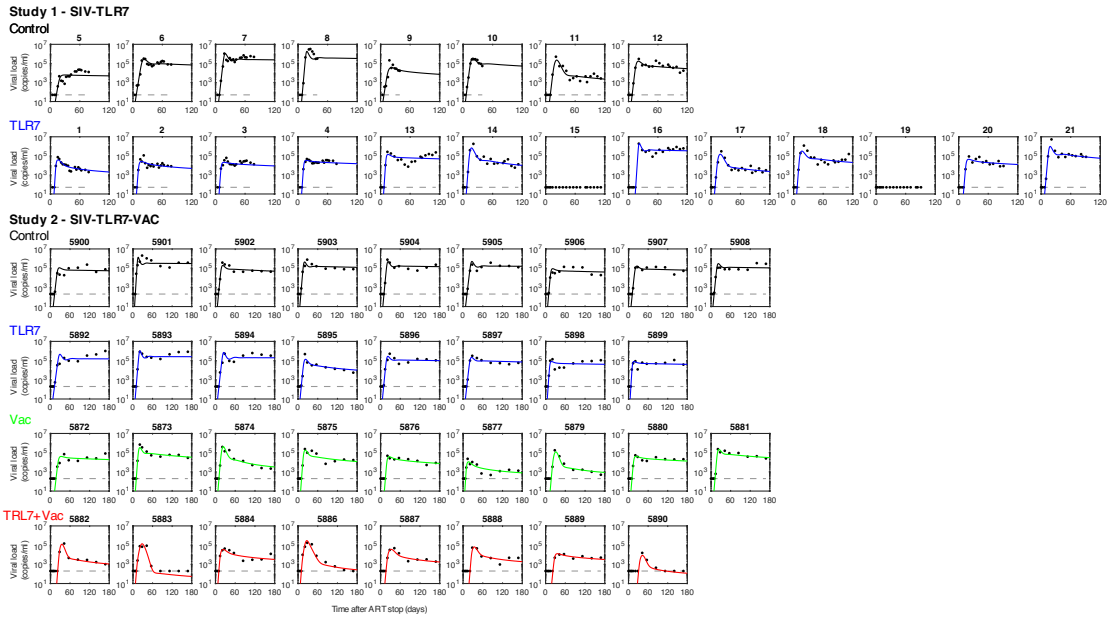

Study 1 - SIV-TLR7
Control
TLR7
Study 2 - SIV-TLR7-VAC
Control
TLR7
Vac
TRL7+Vac
Viral load (copies/ml)
Time after ART stop (days)

Figure 6.6

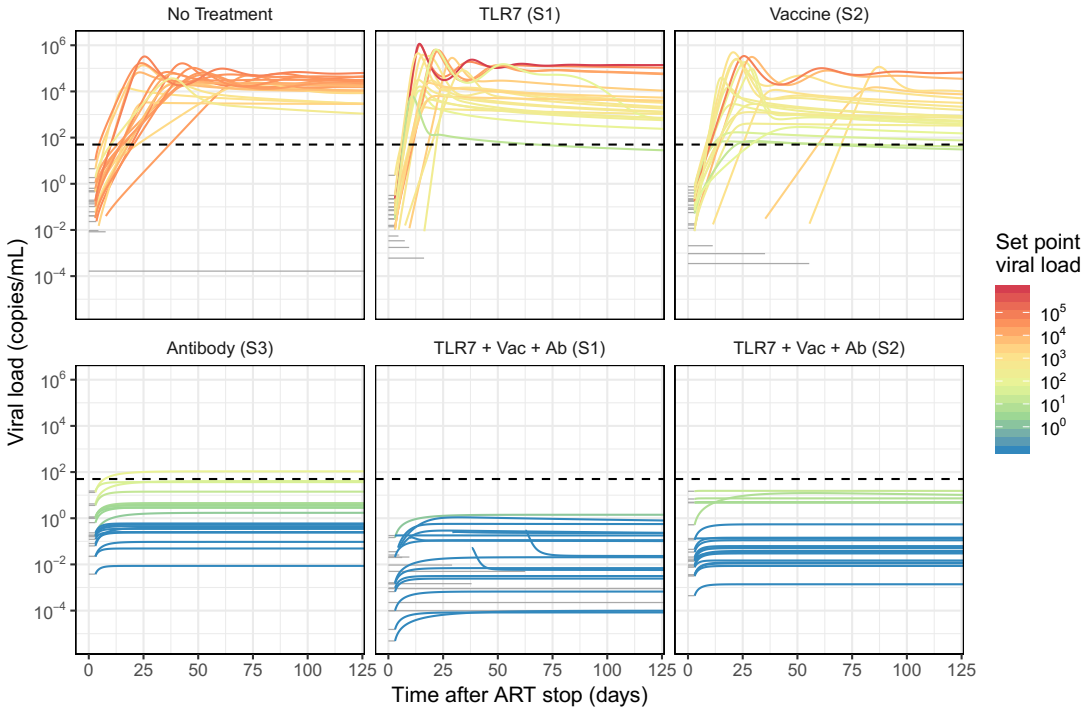


Figure 6.7: Viral rebound trajectories were simulated by combining baseline population heterogeneity from human HIV rebound data with treatment effects from each study. To simulate an individual rebound trajectory, a set of parameters was sampled from the SIV population fitting results and then the relevant treatment effects were applied. 20 example rebound trajectories are shown for each treatment. Viral load prior to succesful rebound is illustrated by a grey horizontal line at the level expected according to the simulated reservoir size. Each trajectory is colored according to its viral load at one year. Dotted lines indicate the detection threshold for standard viral load assays (50 copies/mL).

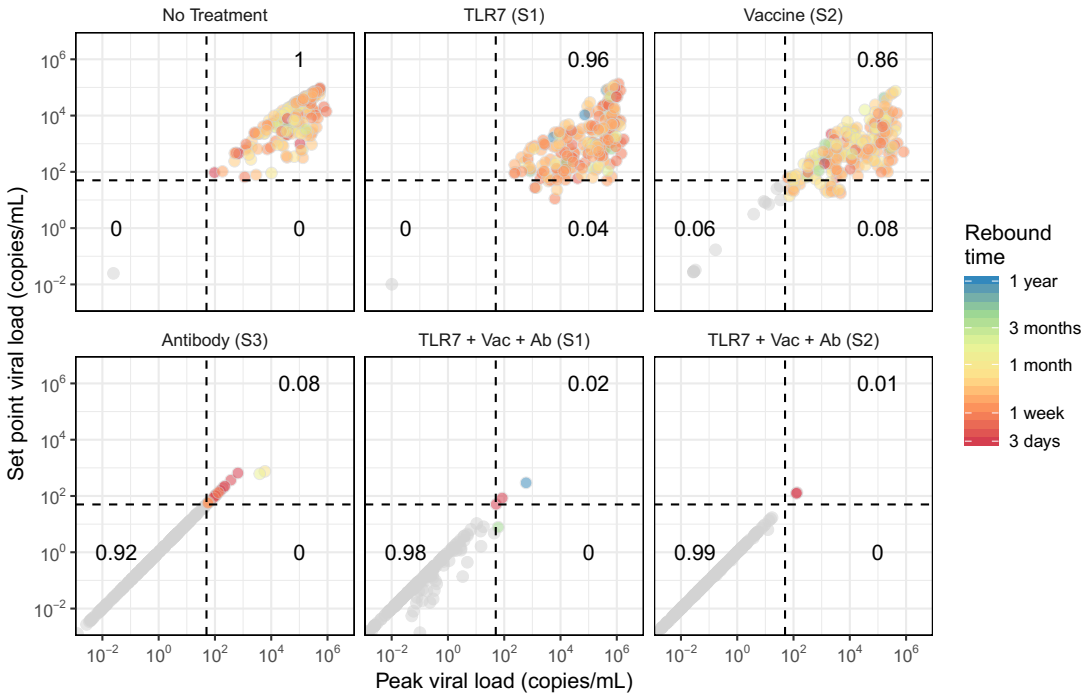


Figure 6.8: 200 simulations for each treatment summarized by their peak and final viral loads. Each dot represents a particular individual and is colored according to the first time at which that individual rebounded (crossed 50 copies/mL). Individuals who never crossed this threshold are shown in gray. Dotted lines indicate the detection threshold for standard viral load assays (50 copies/mL) and dark black numbers show the proportion of individuals falling in the indicated quadrant. Viral rebound trajectories were simulated for one year.

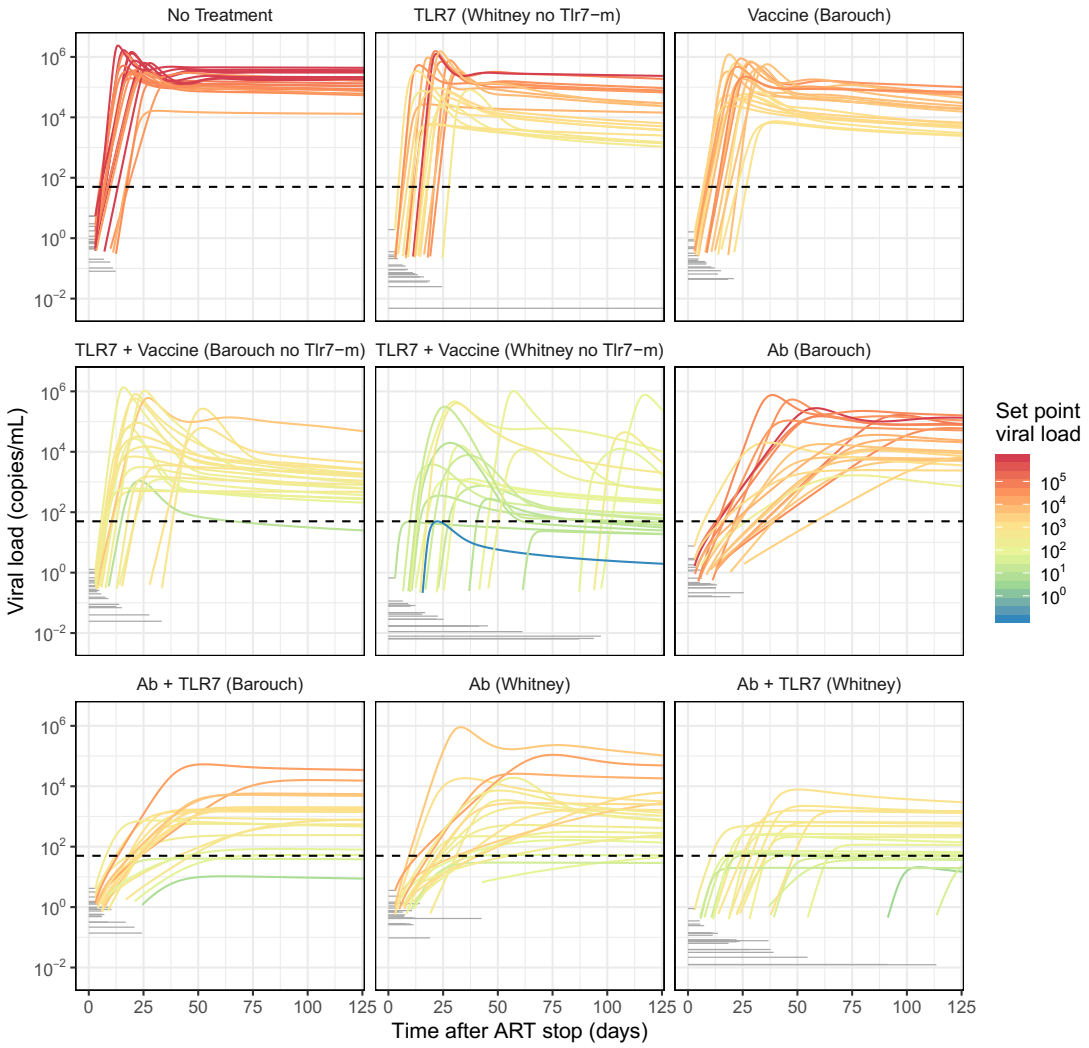


Figure 6.9: Viral rebound trajectories were simulated by combining baseline population heterogeneity from macaque SIV rebound data (studies 1 and 2) with treatment effects from each study. To simulate an individual rebound trajectory, a set of parameters was sampled from the SIV population fitting results and then the relevant treatment effects were applied. 20 example rebound trajectories are shown for each treatment. Viral load prior to succesful rebound is illustrated by a grey horizontal line at the level expected according to the simulated reservoir size. Each trajectory is colored according to its viral load at one year. Dotted lines indicate the detection threshold for standard viral load assays (50 copies/mL).

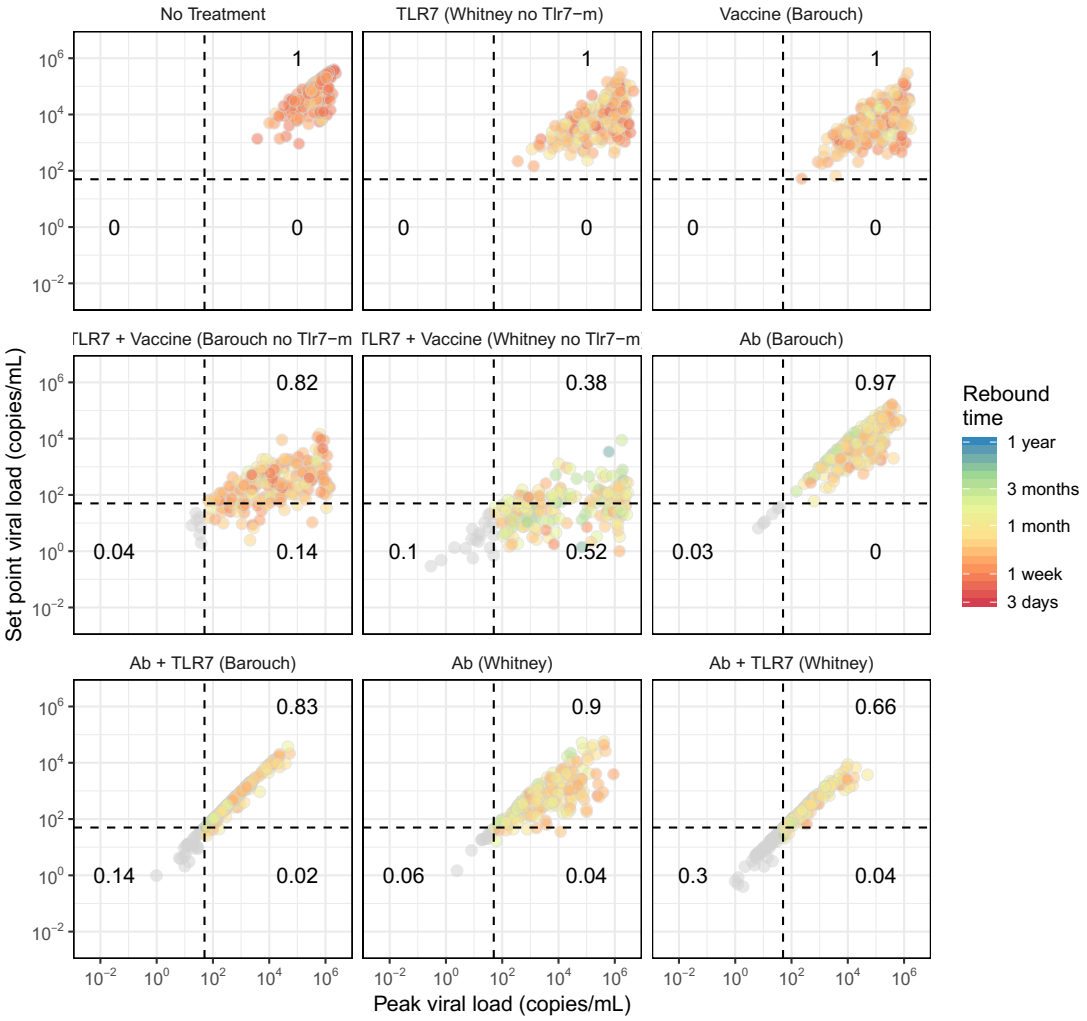


Figure 6.10: 200 simulations for each treatment summarized by their peak and final viral loads. Each dot represents a particular individual and is colored according to the first time at which that individual rebounded (crossed 50 copies/mL). Individuals who never crossed this threshold are shown in gray. Dotted lines indicate the detection threshold for standard viral load assays (50 copies/mL) and dark black numbers show the proportion of individuals falling in the indicated quadrant. Viral rebound trajectories were simulated for one year.

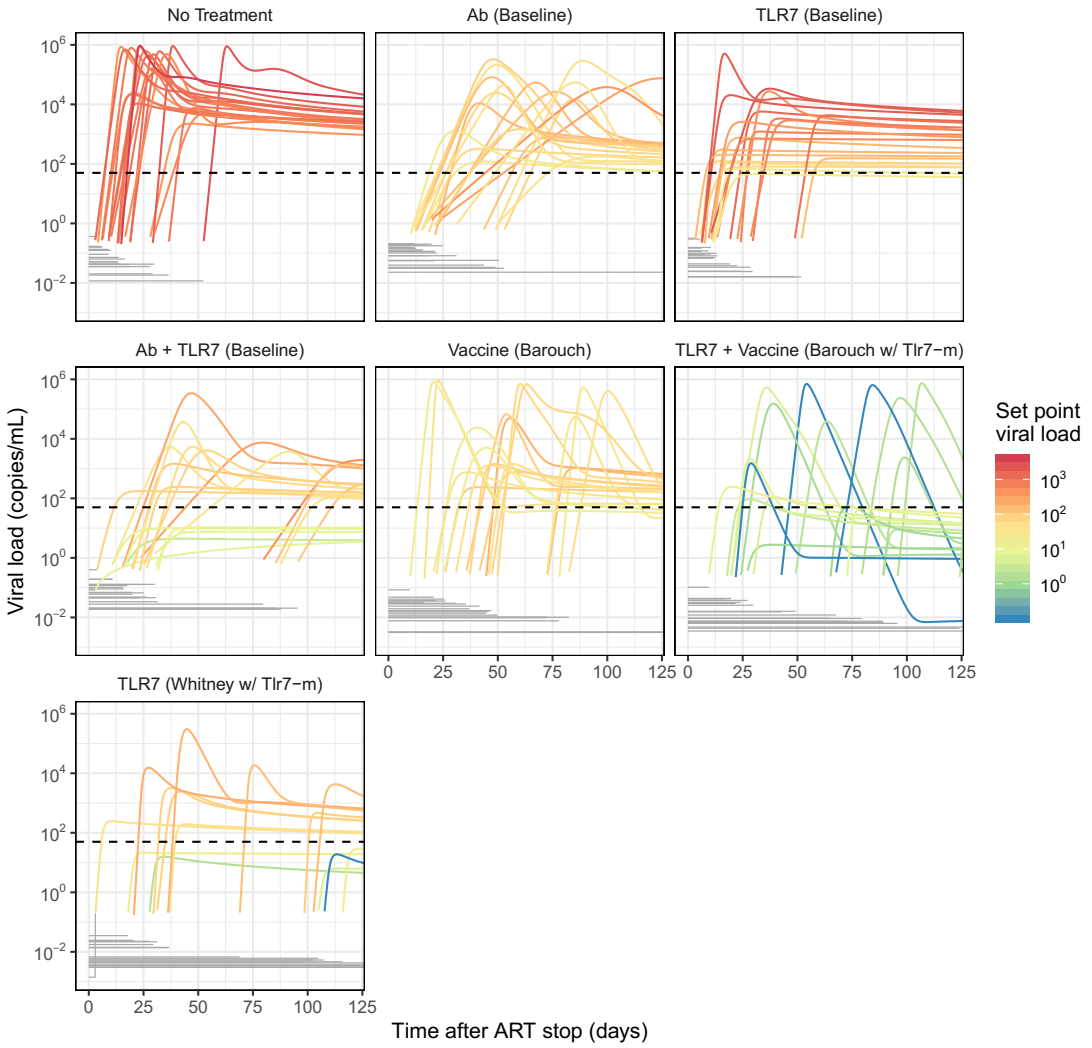


Figure 6.11: Viral rebound trajectories were simulated by combining baseline population heterogeneity from macaque SIV rebound data (study 3) with treatment effects from each study. To simulate an individual rebound trajectory, a set of parameters was sampled from the SIV population fitting results and then the relevant treatment effects were applied. 20 example rebound trajectories are shown for each treatment. Viral load prior to succesful rebound is illustrated by a grey horizontal line at the level expected according to the simulated reservoir size. Each trajectory is colored according to its viral load at one year. Dotted lines indicate the detection threshold for standard viral load assays (50 copies/mL).

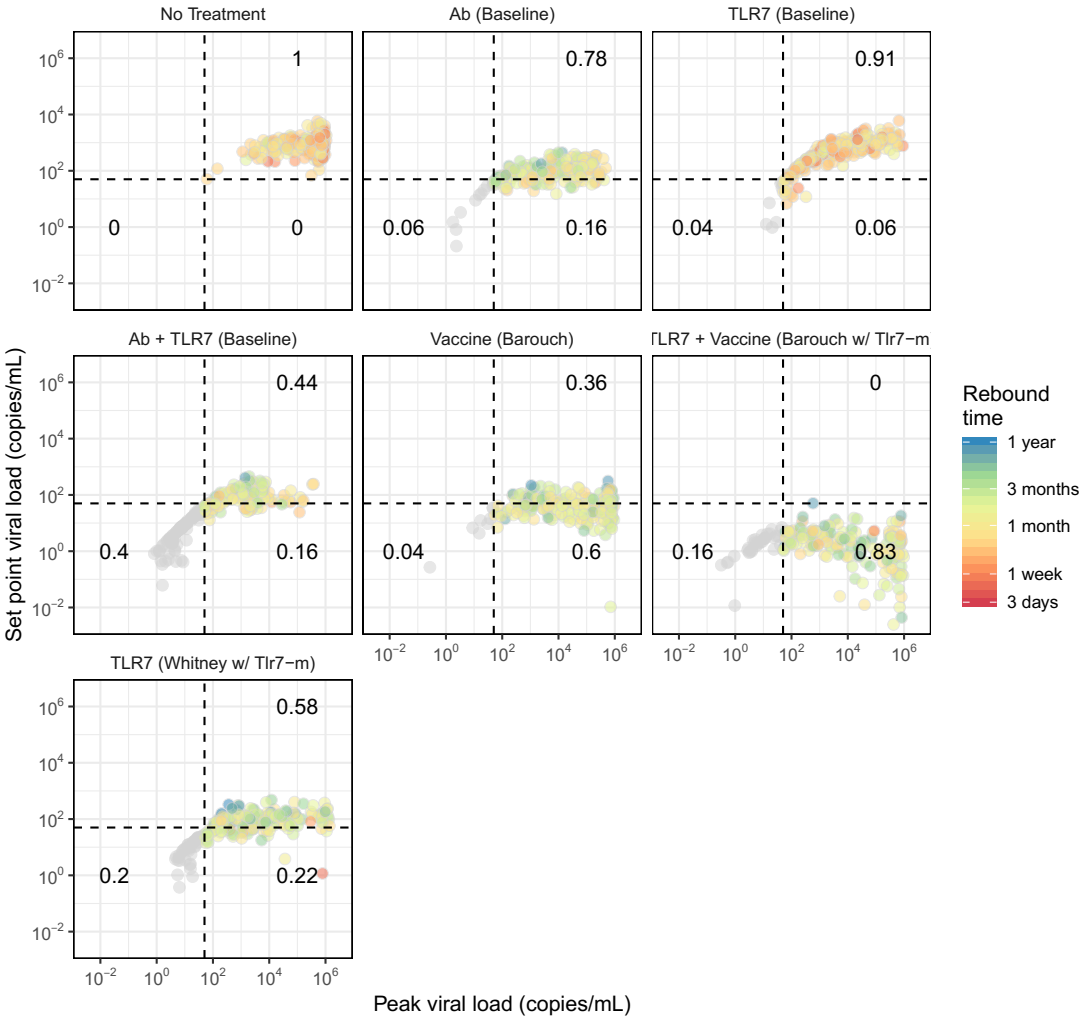


Figure 6.12: 200 simulations for each treatment summarized by their peak and final viral loads. Each dot represents a particular individual and is colored according to the first time at which that individual rebounded (crossed 50 copies/mL). Individuals who never crossed this threshold are shown in gray. Dotted lines indicate the detection threshold for standard viral load assays (50 copies/mL) and dark black numbers show the proportion of individuals falling in the indicated quadrant. Viral rebound trajectories were simulated for one year.

6.7 Supporting Tables

6.8 Model derivations

6.8.1 Re-parameterizing the model to account for stochastic reactivation from latency

We have expressed our model for the dynamics of active and latently infected cells, free virus, and anti-viral immune responses as a system of differential equations (Eqs. 1), which implicitly assumes that each variable and parameter is large enough that transitions happen continuously and fluctuations are small relative to the expected dynamics. This is in general a good assumption, especially in the regime we fit to where viral loads are above a detection limit of 50-200 copies/mL. However, there is one reaction for which we believe this assumption often fails: the reactivation of latently infected cells. In the model presented in the main text, we have assumed that reactivation occurs at a continual rate a. In reality, reactivations are discrete events occurring to single cells, and when the latent reservoir size is small enough or the per-cell reactivation rate low enough, there could be long waiting times between these events. Previous studies in HIV-infected humans and SIV-infected macaques have estimated these reactivations rates to be between 0.5-5 cells/day[108,85], and so in the presence of reservoir-reducing therapies, these rates could be much lower. In our data for SHIV-infected macaques especially (Figure 1, Study 3), we often see long delays to rebound followed by relatively rapid viral growth, which are suggestive of low rates of reservoir reactivation.

The differential equation model in Eqs. 1 can always still be fit to data in which reservoir reactivation happened after a delay, and would just result in a smaller effective a value. However, there would be two major problems in interpreting these fit values. One would be that it would not be possible to compare a values between two animals or treatment groups and claim that the differences were proportional to differences in reservoir size. As we will show below, when reactivation is

Table 6.2: Summary of study designs.

	Study	ART regimen	ART start	Immunotherapy			ART duration	N
1	SIV-TLR7 [148]	TDF+ FTC+ DTG (daily)	9 wks	Control			1.5 yrs (76 wks)	6
				Control			2.2 yrs (115 wks)	2
								8
				TLR7	GS-986	0.1 mg/kg wk 70, 0.2 mg/kg wk 72, 0.3 mg/kg wks 74:2:82	1.5 yrs (76 wks)	4
					GS-986	0.1 mg/kg wks 72:2:90, 108:2:124	2.2 yrs (115 wks)	3
					GS-9620	0.05 mg/kg wks 72:2:90, 108:2:124	2.2 yrs (115 wks)	3
					GS-9620	0.15mg/kg wks 72:2:90	2.2 yrs (115 wks)	3
								13
				ALL				21
2	SIV-TLR7-Vac [29]	TDF+ FTC+ DTG (daily)	1 wk	Control			1.4 yrs (71 wks)	9
				TLR7	GS-9620	0.15mg/kg wks 50:2:70		8
				Vac	Ad26/MVA	Ad26 wks 24, 36; MVA wks 48, 60		9
				TLR7-Vac	GS-9620+ Ad26/MVA	See above		8
				ALL				34
3	SHIV-TLR7-Ab [30]	TDF+ FTC DTG (daily)	1 wk	Control			2.5 yrs (129 wks)	11
				TLR7	GS-9620	0.15 mg/kg wks 96:2:114		11
				Ab	PGT121	10 mg/kg wks 106:2:114		11
				TLR7-Ab	GS-9620 + PGT121	See above		11
								44

Table 6.3: Population parameter values and treatment effects for best estimated model for TRL7-Vac.

	Mean	Random effect (fold-change)	Treatment effect(fold-change)				
			TLR7	Vac	TLR7*Vac	Study	TLR7*Study
λ	71	1.2	1.5 [1.3, 1.8]				
β	4.4×10^{-7}	1.1					
t_a	0.30	2.7		3.9 [1.9, 7.9]			7.6 [3.6, 16.2]
m	32	5.4					
p	0.85	1.8					
N_P	1.6×10^{6}	3.5		0.057 [0.015, 0.22]	0.083 [0.021, 0.32]	0.10 [0.033, 0.32]	

Table 6.4: Population parameter values and treatment effects for best estimated model for TRL7-Ab.

	Mean	Random effect (fold-change)	Treatment effect(fold-change)		
			TLR7	Ab	TLR7*Ab
λ	39				
β	1.0×10^{-6}	1.3		0.45 [0.34, 0.59]	
t_a	3.2	2.3			
m	1.9	23	100 [6.1, 1700]		
p	3.6				
N_P	7.7×10^{4}	1.9		0.14 [0.061,0.30]	

Table 6.5: Alternative models tested for TLR7-Vac. BIC = Bayesian Information Criterion

	Treatment effects						
	TLR7	Vac	Study	TLR7*Study	TLR7*Vac	Justification	BIC
(Best Fit)	–	t_a, N_P	N_P	λ, t_a	N_P	This was the best fit model to arise from the selection procedure	932.1
BH1	–	t_a, N_P	t_a	λ, t_a	N_P	Our initial hypothesis was that the ART start time should influence the size of the latent reservoir	947.9
BH2	–	t_a, N_P	N_P	λ, t_a	t_a	Our initial hypothesis was that TLR7 and Vac may act synergistically to reduce the latent reservoir size	942.0
BH12	–	t_a, N_P	t_a, N_P	λ, t_a	t_a, N_P	Combing both of the above hypotheses	940.6
AM1	–	t_a, p	p	λ, t_a	t_a, p	Same as best fit model but with immune effects on p instead of N_P	954.0
AM2	–	t_a, m	m	λ, t_a	m	Same as best fit model but with immune effects on m instead of p	954.8
AM3	–	t_a, N_P	N_P	t_a, λ	N_P	Same as best fit model but with-TLR7 effects on β instead of λ	940.7
RM1	–	β, N_P	–	m	–	Use the effects that were estimated for TLR7-Ab study	980.2
RM2	–	β, N_P	N_P	m	N_P	Use the effects that were estimated for TLR7-Ab study (except for ones that could not be be estimated there, for them use values from best fit model)	967.0

Table 6.6: Alternative models tested based for TLR7-Ab. BIC = Bayesian Information Criterion.

	Treatment effects				
	TLR7	Ab	TLR7*Ab	Justification	BIC
(Best Fit)	m	β, N_P	–	This was the best fit model to arise from the selection procedure	746.5
BH1	m	t_a, N_P	–	Our initial hypothesis was that the Ab would reduce the size of the latent reservoir	763.7
BH2	t_a	β, N_P	–	Our initial hypothesis was that TLR7 would reduce the size of the latent reservoir, and, this effect was selected in the TLR7-Vac study	766.0
BH3	m	β, N_P	N_P	Our initial hypothesis is that the TLR7 may increase the effect of the Ab, and this interaction was observed for the vaccine in the TLR7-Vac study	761.5
AM1	N_P	β, N_P	–	same as best fit model but including a TLR7 effect on N_P instead of on m	760.7
AM2	m	β, m	–	same as best fit model but including a vaccine effect on m instead of N_P	770.7
AM3	λ	β, N_P	–	same as best fit modeling but including a TLR7 effect on λ instead of m, since λ was selected for TLR7 in the TLR7-Vac study	769.4
AM4	m	λ, N_P	–	same as best fit model but including a vaccine effect on λ instead of on β	754.4
RM1	λ, t_a	t_a, N_P	–	Use the effects that were estimated for TLR7-Vac study	747.5
RM1b	λ, t_a	t_a, N_P	t_a	Use the effects that were estimated for TLR7-Vac study plus t_a interaction	764.5
RM2	λ, t_a	t_a, N_P	N_P	Use the effects that were estimated for TLR7-Vac study, including interaction	778.3
RM2b	λ, t_a	t_a, N_P	t_a, N_P	Use the effects that were estimated for TLR7-Vac study, including interactions on both t_a and N_P	776.6

Table 6.7: Details of TLR7-agnosist dosing variations received by animals in Study 2[148]

ID	ID #	ART duration 76 wks? (vs 115 wks)	Description	GS-9620? (vs GS-986)	# doses	Cumulative dose (mg/kg)	Average dose (mg/kg)	Max dose (mg/kg)
1	156-08	1	GS-986 0.1 mg/kg	0	7	1.8	0.26	0.3
2	166-08	1	wk 70, 0.2 mg/kg	0	7	1.8	0.26	0.3
3	280-09	1	wk 72, 0.3 mg/kg	0	7	1.8	0.26	0.3
4	310-09	1	wks 74:2:82	0	7	1.8	0.26	0.3
5	205-08	1	Control	0	0	0	0	0
6	267-08	1		0	0	0	0	0
7	105-09	1		0	0	0	0	0
8	234-09	1		0	0	0	0	0
9	322-09	1		0	0	0	0	0
10	374-09	1		0	0	0	0	0
11	162-09	0	Control	0	19	0	0	0
12	305-10	0		0	19	0	0	0
13	280-10	0	GS-986 0.1 mg/kg	0	19	1.9	0.1	0.1
14	288-10	0	wks 72:2:90,	0	19	1.9	0.1	0.1
15	344-10	0	108:2:124	0	19	1.9	0.1	0.1
16	293-09	0	GS-9620 0.05	1	19	0.95	0.05	0.05
17	295-10	0	mg/kg wks 72:2:90,	1	19	0.95	0.05	0.05
18	304-10	0	108:2:124	1	19	0.95	0.05	0.05
19	177-10	0	GS-9620 0.15mg/kg wks 72:2:90	1	10	1.5	0.15	0.15
20	341-10	0		1	10	1.5	0.15	0.15
21	412-10	0		1	10	1.5	0.15	0.15

Table 6.8: Comparing population parameter values in the absense of treatment for SIV vs SHIV vs HIV.

	SIV		SHIV		HIV	
Parameter	Mean	Random effect (fold-change)	Mean	Random effect (fold-change)	Mean	Random effect (fold-change)
λ	71	1.2	39		330	1.2
β	4.4×10^{-7}	1.1	1.0×10^{-6}	1.3	5.30×10^{-7}	1.1
t_a	0.30	2.7	3.2	2.3	0.014	20
m	32	5.4	1.9	23	3.8	12
p	0.85	1.8	3.6		3.3	2.7
N_P	1.6×10^{6}	3.5	7.7×10^{4}	1.9	9.3×10^{5}	4.4

common, the inferred a value is linearly related to the frequency of reactivation, whereas when reactivation is rare, it is $\log(a)$ which is proportional to reactivation rate. Another problem is that the variance between individuals in the inferred a value is expected to increase dramatically when reactivation is rare, since the combination of inter-individual variation in reservoir size and the stochastic waiting time until the first reactivation will contribute to the observed time of rebound. One solution could be to fit our data to one of the fully or fully or partially-stochastic models for viral rebound that have been developed previously, however, these models are not amenable to the statistically-rigorous group level fitting approaches we wish to employ here. Therefore, as detailed below, we develop a parameter transformation approach that allows us to capture the expected rebound kinetics for any rate of reservoir reactivation. The main idea of this approach is to replace the continuous rate a with a variable t_a which describes the average time between reactivation events for latently infected cells.

Throughout this derivation, we will consider a model of only a single varible - actively infected cells. This is an approximation for the regime where target cells are not yet limited, an effective immune response has not yet kicked in, and free virus is proportional to infected cells. At the end, we will incorporate the results with the full model.

6.8.2 Rebound kinetics in the limit of rare reactivation

When latent cells reactivate rarely, the reactivation process can be well described consisting of first a waiting time t_a until the first latent cell reactivates and produces an instantaneous jump in infected cell count to level I_1 (concentration equivalent of 1 infected cell), followed by growth. The differential equation for this process is

$$\dot{I}^{rare} = \begin{cases} 0 & t < t_a \\ bI - d_I I & t \geq t_a \end{cases}$$

The solution to this equation is

$$I(t)^{rare} = \begin{cases} 0 & t < t_a \\ I_1 e^{rt} & t \geq t_a \end{cases}$$

were we define $R_0 = b/d_I$ (the basic reproductive ratio) and $r = d_I(R_0 - 1)$ (the asymptotic growth rate in the absence of reservoir reactivation, target cell limitation, or immune responses).

The time until rebound, defined as $I(t) = I_r$, can be solved as

$$t_r^{rare} = t_a + \frac{1}{r} \ln\left(\frac{I_r}{I_1}\right) \tag{6.6}$$

6.8.3 Rebound kinetics in the limit of frequent reactivation

When latent cells reactivate frequently, the reactivation process is well described as a continuous rate, α, at which cells exit the latent reservoir. If each cell contributes a concentration equivalent of I_1, then the dynamics follow a single differential equation

$$\begin{aligned}\dot{I} &= \alpha I_1 + bI - d_I I \\ &= a + rI\end{aligned}$$

were we define $a = \alpha I_1$ (the concentration of latent cells exiting the reservoir per day), and as before, $R_0 = b/d_I$ and $r = d_I(R_0 - 1)$. This equation has the solution, for all $t > 0$

$$I(t)^{freq} = \frac{a}{r}\left(e^{rt} - 1\right) + I_0 e^{rt} \tag{6.7}$$

The initial condition, I_0, is the equilibrium value of Eq. 6.8.3 with $R_0 = 0$, which gives the value of I during administration of ART and therefore at the time of ART stop

$$I_0 = \frac{a}{d_I}$$

which results in the solution

$$I(t)^{freq} = \frac{a}{r}\left(e^{rt} - 1\right) + \frac{a}{d_I} e^{rt}$$

The time until rebound, defined as $I(t) = I_r$, can be solved as

$$t_r^{freq} = \frac{1}{r} \ln\left(\frac{I_r r/a + 1}{r/d_I + 1}\right) \tag{6.8}$$

6.8.4 Discontinuity between the two regimes

First, we will show that there is a discontinuity between these regimes in terms of the time to rebound as a function of the amount of reactivation. To compare them, first note that if cells exit the latent reservoir at rate a, and these events are independent and time homogeneous, then the average

time between reactivation events is $\tau_a = 1/\alpha = I_1/a$, and the distribution of individual waiting times t_a follows $p(t_a) = (1/\tau_a)e^{-t_a/\tau_a}$.

If we look at the limit of rare reactivation dynamics, then the average time to rebound, for a given average waiting time τ_a between reactivation events, is

$$\begin{aligned} E[t_r]^{rare} &= \int_0^\infty p(t_a)\, t_r(t_a)\, dt_a \\ &= \int_0^\infty (1/\tau_a) e^{-t_a/\tau_a} \left(t_a + \frac{1}{r} \ln\left(\frac{I_r}{I_1}\right)\right) dt_a \\ &= \tau_a + \frac{1}{r}\ln\left(\frac{I_r}{I_1}\right) \end{aligned} \tag{6.9}$$

Next, we look at the formula for frequent reactivation dynamics, and see what happens if reactivation becomes rarer. Does it approach Eq. 6.9 ? We replace a using $\tau_a = I_1/a$ in Eq. 6.8 and take the limit of τ_a approaching zero

$$\begin{aligned} \lim_{\tau_a \to \infty} t_r^{freq}(\tau_a) &= \lim_{\tau_a \to \infty} \frac{1}{r} \ln\left(\frac{\tau_a r I_r/I_1 + 1}{r/d_I + 1}\right) \\ &= \frac{1}{r}\ln\left(\frac{\tau_a r I_r/I_1}{r/d_I+1}\right) \\ &= \frac{1}{r}\ln\left(\frac{I_r}{I_1}\right) + \frac{1}{r}\ln(\tau_a) + \frac{1}{r}\ln\left(\frac{r}{r/d_I+1}\right) \\ &= \frac{1}{r}\ln\left(\frac{I_r}{I_1}\right) + \frac{1}{r}\ln(\tau_a) + \frac{1}{r}\ln(d_I) + \frac{1}{r}\ln\left(1 - \frac{1}{R_0}\right) \end{aligned} \tag{6.10}$$

We can see that Equation 6.9 and Equation 6.10 do not match. Rebound time should grow linearly with τ_a in the rare reactivation regime (Eq. 6.9) but in Eq. 6.10 it only grows logarithmically. Generally, Eq. 6.10 underestimates the rebound time, since $I_r \gg I_1$ and $R_0 > 1$.

We can understand qualitatively why the two models don't match. The rare reactivation model assumes that even in the case of instantaneous reactivation ($a \to \infty$ or $\tau_a \to 0$), the infection only

starts growing from a level I_1. Only a single reactivating cell contributes to rebound. However, in the frequent reactivation model, if reactivation is high then the initial condition is much larger than I_1, since there would have been many reactivated cells around prior to ART stop which can begin growing immediately upon ART stop. And, cells can continue reactivating during rebound, which further increases the rate at which infection grows beyond just at rate r.

*Make a figure to show this

6.8.5 Bridging the two regimes

We can bridge the two regimes by thinking of a model that would apply in the regime of intermediate a, where reactivation occurs relatively quickly and a few reactivation events contribute to rebound. Assume that cells reactivate exactly every Δ time steps. Each cell that reactivates starts at level I_1 and then grows, according to Eq. 6.8.2, exponentially at rate r. This gives a formula for the size of the total infection

$$\begin{aligned} I(t) &= I_1 e^{r(t-t_0)} + I_1 e^{r(t-t_0-\Delta)} + I_1 e^{r(t-t_0-2\Delta)} + ... + I_1 e^{r(t-t_0-k\Delta)} \\ &= \sum_{n=0}^{k} I_1 e^{r(t-t_0)} e^{-rn\Delta} \\ &= I_1 e^{r(t-t_0)} \left(\frac{1 - e^{-r(k+1)\Delta}}{1 - e^{-r\Delta}} \right) \\ &= I_1 e^{r(t-t_0)} \left(\frac{e^{r\Delta} - e^{-rk\Delta}}{e^{r\Delta} - 1} \right) \end{aligned} \tag{6.11}$$

where t_0 is the time of the first reactivation, and $k+1$ is the number of reactivations that happen before time t. It is the highest integer such that $t - t_0 - k\Delta \geq 0$, which implies that $(t-t_0)/\Delta - 1 < k \leq (t-t_0)/\Delta$. We set $t_0 = \Delta = t_a$, because we want t_a to have the interpretation of being the time of the first reactivation and will assume that it is also representative of the average waiting time.

We choose the highest possible k, so $k = (t - t_0)/\Delta = t/t_a - 1$.

$$\begin{aligned} I(t) &= I_1 e^{r(t-t_a)} \left(\frac{e^{rt_a} - e^{-r(t/t_a - 1)t_a}}{e^{rt_a} - 1} \right) \\ &= I_1 \left(\frac{e^{rt} - 1}{e^{rt_a} - 1} \right) \end{aligned} \tag{6.12}$$

Compare this equation to Eq. 6.7. They are equivalent when

$$\begin{aligned} a_{eff} &= \frac{I_1 r}{e^{rt_a} - 1} \\ I_0 &= 0 \end{aligned} \tag{6.13}$$

Thus, conceptually, this method for including multiple reactivations has made our formula for rare reactivations (Eq. 6.8.2) closer to the one for frequent reactivations (Eq. 6.7) by accounting for the contributions of multiple reactivating cells. However, it would still underestimate rebound time when reactivation is really common because it still assumes the initial cell level is zero.

We want to choose a value of the initial condition I_0 which is a function of t_a and can cover all the regimes. When t_a is large, we want $I_0 = 0$. When t_a is small, we want $I_0 = a/d_I = I_1/(d_I t_a)$. One option is

$$\begin{aligned} I_0 &= w \frac{I_1}{t_a d_I} \\ w &= 2^{-(t_a d_I)^n} \end{aligned} \tag{6.14}$$

This function describes a sigmoidal curve that goes from one to zero, switching at $t_a = 1/d_I$. The constant n controls the sharpness of the interpolation (higher n, sharper transition). $n = 2$ gives

reasonable behavior for biological parameters.

The rational behind this function is that when $t_a d_I \ll 1$, there are many cells reactivated even at the moment ART is stopped (at average level $\frac{I_1}{t_a d_I}$), while when $t_a d_I \gg 1$, there are usually no latent cells activated at the time ART is stopped. In reality, the number of cells present at ART stop is random variable, but this is a reasonable approximation to the average behavior and reproduces the average rebound time when t_a is the mean time between reactivations.

Note: With this new I_0, it will not be true that $I(t_a) = I_0$. As soon as there is a non-zero initial condition, there will be some growth that happens before the first post-ART reactivation. This growth is due to cells that reactivated before ART stopped. Therefore, $I(t_a)$ is equal to I_0 plus this older growth.

We can analytically calculate the rebound time for this model

$$\begin{aligned} t_r &= \frac{1}{r} \ln \left(\frac{I_r r + a_{eff}}{I_0 r + a_{eff}} \right) \\ &= \frac{1}{r} \ln \left(\frac{(I_r/I_1)(e^{rt_a} - 1) + 1}{w/(t_a d_I)(e^{rt_a} - 1) + 1} \right) \\ &= \frac{1}{r} \ln \left(\frac{(I_r/I_1)(e^{rt_a} - 1) + 1}{2^{-(t_a d_I)^n}/(t_a d_I)(e^{rt_a} - 1) + 1} \right) \end{aligned} \tag{6.15}$$

We can check that in the limit of large t_a, the rebound time will approach the value for the rare reactivation model (t_r^{rare}, Eq 6.6):

$$\begin{aligned} \lim_{t_a \to \infty} t_r &= \\ &= \begin{cases} t_a + \frac{1}{r} \ln \left(\frac{I_r}{I_1} \right) & n > 1 \\ \frac{1}{r} \ln \left(\frac{I_r}{I_1} \right) + \frac{1}{r} \ln(t_a d_I) + \frac{1}{r} \ln(2)(t_a d_I)^n & n < 1 \end{cases} \end{aligned} \tag{6.16}$$

The first expression, with $n > 1$, gives the correct limit. However, if n is too small (less than one), then the weighing function w doesn't decay fast enough with t_a, and so the initial condition I_0 doesn't decay to zero fast enough, and the time to rebound is underestimated (rebound happens too fast). With our value of $n = 2$, rebound times will be correct for large t_a.

We can put all of this together to define a model that works in both regimes:

Two-regime model vi

Composite parameters:

$$\begin{aligned} r &= d_I(R_0 - 1) \\ a_{eff} &= \frac{I_1 r}{e^{r t_a} - 1} \\ w &= 2^{-(t_a d_I)^n} \end{aligned} \tag{6.17}$$

Initial conditions:

$$I_0 = w\frac{I_1}{t_a d_I} \tag{6.18}$$

Equations:

$$\dot{I} = a_{eff} + rI \tag{6.19}$$

Note that these equations work even when $R_0 < 1$ ($r < 0$), and $a_{eff} = I_1$ when $R_0 = 1$ ($r = 0$) (although numerically it may be undefined).

6.8.6 Conditioning on survival reactivating cells in the stochastic regime

We previously analyzed the dynamics of rebound assuming that cells reactivate from latency every t_a days (i.e. at rate $1/t_a$) and that infection then grows exponentially towards rebound. This model is a simplification, since in reality infection dynamics are a fully stochastic process. Developing a model to track every stochastic reaction between a cell and virus is beyond the scope of this work, and such a model would not be identifiable from typical *in vivo* measurements of viral kinetics. However, recognizing the underlying stochastic nature of these dynamics leads to an important correction to our work.

While some reactivating latent cells will produce a chain of infection that eventually leads to rebound, others will - simply by chance - end up going extinct. Without specifying any details of the underlying stochastic processes, we can define the probability of long-term survival of the infection started from a single reactivating cell (often called the "establishment" or "survival" probability) as $p_{surv} \in [0, 1]$. Then the average time between *surviving* reactivations is t_a/p_{surv}.

While overall, the expected dynamics averaged over *all* reactivating cells is described by the deterministic equations,

$$\mathbb{E}[I(t)] = I_1 e^{rt}.$$

If we condition on survival of the reactivated cell, then the expected dynamics are larger by a factor of $1/p_{surv}$:

$$\mathbb{E}[I(t)|I(t) > 0] = \frac{I_1}{p_{surv}} e^{rt}.$$

Together, this means that Eqs. 6.17-6.19 can be updated with an effectively longer interval between reactivating latent cells ($t_a \rightarrow t_a/p_{surv}$) and an effectively higher initial concentration of

actively infected cells ($I_1 \rightarrow I_1/p_{surv}$).

The value of the survival probability and its relationship to the other model parameters depends on the details of the underlying stochastic process. In all cases, p_{surv} will be higher when R_0 is higher. As an example, we consider the relatively generic stochastic process used to describe reactivation from latent infection in Hill et al[108]:

$$\begin{aligned} &\emptyset \rightarrow Y\text{... rate constant: } \alpha \\ &Y \rightarrow NY\text{... rate constant: } b \times p_\lambda(N) \\ &Y \rightarrow \emptyset\text{... rate constant: } d \end{aligned} \tag{6.20}$$

This model tracks only actively infected cells. In this notation, Y represents an individual cells and $\emptyset$ represents no cells, and the arrows represent events that change the number of cells. Mathematically this process is a type of burst-death-immigration branching process. A reactivation event from latency produces an actively infected cell at rate α, where $t_a = 1\alpha$. This cell can either die (at rate d) or produce a collection of virions (at rate b) that results in the infection of N other cells, where N is a Poisson-distributed random variable with parameter λ, $p_\lambda(N) = \left(e^{-\lambda}\lambda^N\right)/(N!)$. After an infection event, the original cell dies.

The overall survival probability for a single reactivated cell is the weighted sum of the probability of producing N offspring and the probability that at least one of these offspring survives.

$$
\begin{aligned}
p_{surv} &= \frac{d}{d+b} * 0 + \frac{b}{d+b} \sum_{N=0}^{\infty} p_\lambda(N) * (1 - (1 - p_{surv})^N) \\
&= \frac{b}{d+b} \left(1 - e^{-\lambda} \sum_{N=0}^{\infty} \frac{\lambda^N}{N!} * (1 - p_{surv})^N \right) \\
&= \frac{b}{b+d} \left(1 - e^{-\lambda p_{surv}} \right)
\end{aligned}
$$

Since it is generally impossible to measure the underlying stochastic parameters (e.g. b, d, and λ), we want to replace them in this equation with composite parameters that are more accessible. We use the basic reproductive ratio, $R_0 = \frac{b\lambda}{b+d}$, which describes the average number of secondary infections arising from a single cell, and ρ, the so-called Fano factor, which is the ratio of the variance to the average for the same quantity ($\rho = 1 + d\lambda/(b+d)$). Previous studies have suggested that $\rho \sim 10$ (reviewed in[108]). Then, we find that the survival probability satisfies the implicit condition

$$
p_{surv}(R_0 + \rho - 1) = R_0(1 - e^{-p_{surv}(R_0+\rho-1)}) \tag{6.21}
$$

Note that the survival probability is only non-zero when $R_0 > 1$. Otherwise extinction is guaranteed.

The solution to this implicit definition can be expressed in terms of the Lambert W function:

$$
p_{surv} = \frac{1}{R_0 + \rho - 1} \left(W(-R_0 e^{-R_0}) + R_0 \right) \tag{6.22}
$$

Conditioning on survival for reactivating lineages in the stochastic regime leads to the following combined model. Note that for the $R_0 < 1$ case $p_{surv} = 0$ and we have taken the limit of expressions as $p_{surv} \to 0$.

Two-regime model v2

Composite parameters:

$$
\begin{aligned}
r &= d_I(R_0 - 1) \\
p_{surv} &= \begin{cases} 0 & R_0 \leq 1 \\ \dfrac{1}{R_0 + \rho - 1}\left(W(-R_0 e^{-R_0}) + R_0\right) & R_0 > 1 \end{cases} \\
a_{eff} &= \begin{cases} \dfrac{I_1}{t_a} & R_0 \leq 1 \\ \dfrac{I_1 r}{p_{surv}(e^{r t_a / p_{surv}} - 1)}) & R_0 > 1 \end{cases} \\
w &= \begin{cases} 0 & R_0 \leq 1 \\ 2^{-(t_a d_I / p_{surv})^n} & R_0 > 1 \end{cases}
\end{aligned}
\tag{6.23}
$$

Initial conditions:

$$
I_0 = w \frac{I_1}{t_a d_I} \tag{6.24}
$$

Equations:

$$
\dot{I} = a_{eff} + rI \tag{6.25}
$$

6.8.7 Accounting for numerical errors

For all differential equation solvers we have tested, numerical errors often arise when evaluating Eqs. 6.19 or 6.25 when t_a is large. This occurs due to the extremely small values of a_{eff} and w (and

therefore I_0) and the extremely small values of $I(t)$ inferred for $t < t_a$ with this continuous approximation. We can get around this by choosing not to start integration for $I(t)$ right at $t = 0$ if both a_{eff} and I_0 are very small, but instead, choosing a time t_{go} such that at least one of them is larger. We also want to make sure this time is not too large, because then, infection levels could have grown out of the simple regime we are considering now (no target cell limitation, no immune control).

An easy solution is to choose a t_{go} such that $I(t_{go}) = I_1$. Since I_1 is the true biological minimum of infection, we know that for $I(t) < I_1$ any approximations about exponential viral growth are valid.

$$\begin{aligned} I(t_{go}) &= I_1 \\ \frac{a_{eff}}{r}\left(e^{rt_{go}} - 1\right) + I_0 e^{rt_{go}} &= I_1 \\ t_{go} &= \frac{1}{r}\ln\left(\frac{I_1 + a_{eff}/r}{I_0 + a_{eff}/r}\right) \\ &= t_a - \frac{1}{r}\ln\left(\frac{w(e^{rt_a} - 1)}{t_a d_I} + 1\right) \end{aligned} \tag{6.26}$$

Then we integrate the equation

$$\dot{I} = \begin{cases} 0 & t < t_{go} \\ a_{eff} + rI & t \geq t_{go} \end{cases} \tag{6.27}$$

with initial condition

$$I(t_{go}) = I_1 \tag{6.28}$$

We will only use t_{go} if $I(0) < I_1$, which guarantees that $t_{go} > 0$.

This leads to the updated model

Two-regime model v3

Composite parameters:

$$
\begin{aligned}
r &= d_I(R_0 - 1) \\
p_{surv} &= \begin{cases} 0 & R_0 \leq 1 \\ \dfrac{1}{R_0 + \rho - 1}\left(W(-R_0 e^{-R_0}) + R_0\right) & R_0 > 1 \end{cases} \\
a_{eff} &= \begin{cases} \dfrac{I_1}{t_a} & R_0 \leq 1 \\ \dfrac{I_1 r}{p_{surv}(e^{r t_a / p_{surv}} - 1)}) & R_0 > 1 \end{cases} \\
w &= \begin{cases} 0 & R_0 \leq 1 \\ 2^{-(t_a d_I / p_{surv})^n} & R_0 > 1 \end{cases}
\end{aligned}
\tag{6.29}
$$

Functions to avoid small number errors :

$$
t_{go} = \begin{cases} 0 & R_0 \leq 1 \\ \max\left(0, \dfrac{t_a}{p_{surv}} - \dfrac{1}{r}\ln\left(w\dfrac{p_{surv}}{t_a d_I}\left(e^{r t_a / p_{surv}} - 1\right) + 1\right)\right) & R_0 > 1 \end{cases}
\tag{6.30}
$$

Initial conditions:

$$
I_0 = w\frac{I_1}{t_a d_I}
\tag{6.31}
$$

EQUATIONS: For all $t > t_{go}$

$$\dot{I} = a_{eff} + rI \tag{6.32}$$

6.8.8 CALCULATING THE LONG-TERM GROWTH RATE WHEN FREE VIRUS IS INCLUDED

The model we have shown tracks only infected cells, as well as ignoring the immune response and target cell limitation. The final model can be augmented to includes these effects by simply adding the extra variables and terms to the system of equation for $I(t)$. However, in doing so, we must make a slight alteration in our expression for r, the early growth rate of infection (before target cell limitation or immune response has set in). Previously, we had used the expression $r = d_I(R_0 - 1)$, but this is not valid when we track free virus as well as infected cells.

Consider a set of viral dynamics equations tracking infected cells along with free virus:

$$\dot{I} = bV - d_I I$$

$$\dot{V} = kI - cV$$

Which can be expressed in matrix form, with $\vec{x} = \begin{bmatrix} I \\ V \end{bmatrix}$, as

$$\dot{\vec{x}} = \begin{bmatrix} -d_I & b \\ k & -c \end{bmatrix} \vec{x}$$

The eigenvalues λ of this matrix satisfy the polynomial

$$(-d_I - \lambda)(-c - \lambda) - bk = 0$$

Which has the solutions

$$\begin{aligned}\lambda &= \frac{-(c+d_I) \pm \sqrt{(c+d_I)^2 - 4(cd_I - bk)}}{2} \\ &= \frac{-(c+d_I) \pm \sqrt{(c-d_I)^2 + 4d_I c R_0)}}{2}\end{aligned}$$

The positive root of this equation dominates the solution in the long term, so we set the parameter r in our model to be

$$r = \frac{-(c+d_I) + \sqrt{(c-d_I)^2 + 4bk)}}{2}$$

In this model, the basic reproductive ratio is $R_0 = bk/(cd_I)$, and so r can be expressed in terms of R_0 as

$$r = \frac{-(c+d_I) + \sqrt{(c-d_I)^2 + 4d_I c R_0)}}{2} \tag{6.33}$$

If $R_0 > 1$, then $r > 0$. This equation can be re-arranged for R_0, as

$$R_0 = \left(1 + \frac{r}{c}\right)\left(1 + \frac{r}{d_I}\right) \tag{6.34}$$

which reduces to the simplified formula, $R_0 = r/d_I + 1$ or $r = d_I(R_0 - 1)$ only if $r \ll c$. While estimates of R_0 and r using the simplified formulas will only be off by around 5% for values of $c = 23$ /day and $r \sim 1$/day observed in this study, this will significantly alter our estimates of a_{eff} (Eq. 6.17) and lead to biased estimates of other parameter values.

6.8.9 Two regime model including all variables

Variables: (Observed) V, (Unobserved) T, I, P, E

Basic parameters: (Fit) λ, β, t_a, N_P, p, m, (Fixed) d_T, d_I, d_P, d_E, c, k, N_E, f, and washout time (t_w), viral load equivalent of one cell ($I_1 = 4 * 10^{-5}$), mean-to-variance ratio for virus production

($\rho = 10$), and interpolation constant ($n = 3$).

Time-averaged values of variables during ART ($\beta = 0$) :

$$
\begin{aligned}
T_{ART} &= \frac{\lambda}{d_T} \\
I_{ART} &= \frac{I_1}{t_a d_I} \\
V_{ART} &= \frac{I_1}{t_a d_I}\frac{k}{c} \\
P_{ART} &= \begin{cases} \dfrac{m}{d_P - p(1-f)\,V_{ART}/(\,V_{ART} + N_P)} & d_P > p(1-f)\,V_{ART}/(\,V_{ART} + N_P) \\ 1000 N_E & d_P \leq p(1-f)\,V_{ART}/(\,V_{ART} + N_P) \end{cases} \\
E_{ART} &= \frac{pfP_{ART}V_{ART}}{d_E(\,V_{ART} + N_p)}
\end{aligned}
\tag{6.35}
$$

Functions to connect stochastic and deterministic regimes:

Basic reproductive ratio:

$$R_0 = \frac{\lambda \beta k}{(1 + E_{ART}/N_E)\,d_T d_I c} \tag{6.36}$$

Early exponential growth rate:

$$r = \frac{-(c + d_I) + \sqrt{(c - d_I)^2 + 4 c d_I R_0}}{2} \tag{6.37}$$

Survival probability starting from single actively-infected cell:

$$p_{surv} = \begin{cases} 0 & R_0 \leq 1 \\ \dfrac{1}{R_0 + \rho - 1}\left(LambertW(-R_0 e^{-R_0}) + R_0\right) & R_0 > 1 \end{cases} \tag{6.38}$$

Effective reactivation rate of latently-infected cells:

$$a = \begin{cases} \dfrac{I_1}{t_a} & R_0 \leq 1 \\ \dfrac{I_1 r}{p_{surv}\left(e^{rt_a/p_{surv}} - 1\right)} & R_0 > 1 \end{cases} \tag{6.39}$$

Interpolation function:

$$w = \begin{cases} 0 & R_0 \leq 1 \\ 2^{-(t_a d_I/p_{surv})^n} & R_0 > 1 \end{cases} \tag{6.40}$$

Functions to avoid small number errors :

$$t_{go} = \begin{cases} 0 & R_0 \leq 1 \\ \max\left(0, \dfrac{t_a}{p_{surv}} - \dfrac{1}{r}\ln\left(w\dfrac{p_{surv}}{t_a d_I}\left(e^{rt_a/p_{surv}} - 1\right) + 1\right)\right) & R_0 > 1 \end{cases} \tag{6.41}$$

Initial conditions:

$$\begin{aligned} T_0 &= T_{ART} \\ I_0 &= \begin{cases} I_{ART} & R_0 \leq 1 \\ \max(wI_{ART}, I_1/p_{surv}) & R_0 > 1 \end{cases} \\ V_0 &= \frac{k}{c} I_0 \\ P_0 &= P_{ART} \\ E_0 &= E_{ART} \end{aligned} \tag{6.42}$$

EQUATIONS: For all $t > t_w + t_{go}$

$$\begin{aligned}
\dot{T} &= \lambda - \beta TV - d_T T \\
\dot{I} &= a + \frac{\beta TV}{1 + (E/N_E)} - d_I I \\
\dot{V} &= kI - cV \\
\dot{P} &= m + p(1-f)\frac{V}{V + N_P} P - d_P P \\
\dot{E} &= pf\frac{V}{V + N_P} P - d_E E
\end{aligned} \tag{6.43}$$

References

[1] Adib-Conquy, M., Scott-Algara, D., Cavaillon, J.-M., & Souza-Fonseca-Guimaraes, F. (2014). TLR-mediated activation of NK cells and their role in bacterial/viral immune responses in mammals. *Immunology & Cell Biology*, 92(3), 256–262.

[2] Akondy, R. S., Fitch, M., Edupuganti, S., Yang, S., Kissick, H. T., Li, K. W., Youngblood, B. A., Abdelsamed, H. A., McGuire, D. J., Cohen, K. W., Alexe, G., Nagar, S., McCausland, M. M., Gupta, S., Tata, P., Haining, W. N., McElrath, M. J., Zhang, D., Hu, B., Greenleaf, W. J., Goronzy, J. J., Mulligan, M. J., Hellerstein, M., & Ahmed, R. (2017). Origin and differentiation of human memory CD8 T cells after vaccination. *Nature*, 552(7685), 362.

[3] Alexandrov, L., Nik-Zainal, S., Wedge, D., Campbell, P., & Stratton, M. (2013a). Deciphering signatures of mutational processes operative in human cancer. *Cell Reports*, 3(1), 246–259.

[4] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, V. A., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Imielinsk, M., Jäger, N., Jones, D. T. W., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, V. J., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R., van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Initiative, A. P. C. G., Consortium, I. B. C., Consortium, I. M.-S., PedBrain, I., Zucman-Rossi, J., Futreal, P. A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J., & Stratton, M. R. (2013b). Signatures of mutational processes in human cancer. *Nature*, 500(7463), 415–21.

[5] Altrock, P. M., Liu, L. L., & Michor, F. (2015). The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer*, 15(12), 730–745.

[6] Amikura, K., Kobari, M., & Matsuno, S. (1995). The time of occurrence of liver metastasis in carcinoma of the pancreas. *International journal of pancreatology : official journal of the International Association of Pancreatology*, 17(2), 139–46.

[7] Anagnostou, V., Smith, K. N., Forde, P. M., Niknafs, N., Bhattacharya, R., White, J., Zhang, T., Adleff, V., Phallen, J., Wali, N., Hruban, C., Guthrie, V. B., Rodgers, K., Naidoo, J., Kang, H., Sharfman, W., Georgiades, C., Verde, F., Illei, P., Li, Q. K., Gabrielson, E., Brock, M. V., Zahnow, C. A., Baylin, S. B., Scharpf, R. B., Brahmer, J. R., Karchin, R., Pardoll, D. M., & Velculescu, V. E. (2017). Evolution of Neoantigen Landscape during Immune Checkpoint Blockade in Nonâ□□Small Cell Lung Cancer. *Cancer Discovery*, 7(3), 264–276.

[8] Apetrei, C., Pandrea, I., & Mellors, J. W. (2012). Nonhuman Primate Models for HIV Cure Research. *PLOS Pathog*, 8(8), e1002892.

[9] Archin, N. M., Vaidya, N. K., Kuruc, J. D., Liberty, A. L., Wiegand, A., Kearney, M. F., Cohen, M. S., Coffin, J. M., Bosch, R. J., Gay, C. L., Eron, J. J., Margolis, D. M., & Perelson, A. S. (2012). Immediate antiviral therapy appears to restrict resting CD4+ cell HIV-1 infection without accelerating the decay of latent infection. *Proceedings of the National Academy of Sciences of the United States of America*, 109(24), 9523–9528.

[10] Athreya, K. B. & Ney, P. E. (1972). *Branching Processes*. Springer-Verlag.

[11] Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K.-S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., Rubio-Perez, C., Nagarajan, N., Cortés-Ciriano, I., Zhou, D. C., Liang, W.-W., Hess, J. M., Yellapantula, V. D., Tamborero, D., Gonzalez-Perez, A., Suphavilai, C., Ko, J. Y., Khurana, E., Park, P. J., Allen, V. E. M., Liang, H., Lawrence, M. S., Godzik, A., Lopez-Bigas, N., Stuart, J., Wheeler, D., Getz, G., Chen, K., Lazar, A. J., Mills, G. B., Karchin, R., & Ding, L. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2), 371 – 385.e18.

[12] Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A.-M., Gingras, M.-C., Miller, D. K., Christ, A. N., Bruxner, T. J. C., Quinn, M. C., Nourse, C., Murtaugh, L. C., Harliwong, I., Idrisoglu, S., Manning, S., Nourbakhsh, E., Wani, S., Fink, L., Holmes, O., Chin, V., Anderson, M. J., Kazakoff, S., Leonard, C., Newell, F., Waddell, N., Wood, S., Xu, Q., Wilson, P. J., Cloonan, N., Kassahn, K. S., Taylor, D., Quek, K., Robertson, A., Pantano, L., Mincarelli, L., Sanchez, L. N., Evers, L., Wu, J., Pinese, M., Cowley, M. J., Jones, M. D., Colvin, E. K., Nagrial, A. M., Humphrey, E. S., Chantrill, L. A., Mawson, A., Humphris, J., Chou, A., Pajic, M., Scarlett, C. J., Pinho, V. A., Giry-Laterriere, M., Rooman, I., Samra, J. S., Kench, J. G., Lovell, J. A., Merrett, N. D., Toon, C. W., Epari, K., Nguyen, N. Q., Barbour, A., Zeps, N., Moran-Jones, K., Jamieson, N. B., Graham, J. S., Duthie, F., Oien, K., Hair, J., Grützmann, R., Maitra, A., Iacobuzio-Donahue, C. A., Wolfgang, C. L., Morgan, R. A., Lawlor, R. T., Corbo, V., Bassi, C., Rusev, B., Capelli, P., Salvia, R., Tortora, G., Mukhopadhyay, D., Petersen, G. M., Munzy, D. M., Fisher, W. E., Karim, S. A., Eshleman, J. R., Hruban, R. H., Pilarsky, C., Morton, J. P., Sansom, O. J., Scarpa, A., Musgrove, E. A., Bailey, U.-M. H., Hofmann, O., Sutherland, R. L., Wheeler, D. A., Gill, A. J., Gibbs, R. A., Pearson, V. J., Waddell,

N., Biankin, V. A., & Grimmond, S. M. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, 531(7592), 47–52.

[13] Barouch, D. H., Liu, J., Li, H., Maxfield, L. F., Abbink, P., Lynch, D. M., Iampietro, M. J., SanMiguel, A., Seaman, M. S., Ferrari, G., Forthal, D. N., Ourmanov, I., Hirsch, V. M., Carville, A., Mansfield, K. G., Stablein, D., Pau, M. G., Schuitemaker, H., Sadoff, J. C., Billings, E. A., Rao, M., Robb, M. L., Kim, J. H., Marovich, M. A., Goudsmit, J., & Michael, N. L. (2012). Vaccine protection against acquisition of neutralization-resistant SIV challenges in rhesus monkeys. *Nature*, 482(7383), 89–93.

[14] Barouch, D. H., Stephenson, K. E., Borducchi, E. N., Smith, K., Stanley, K., McNally, A. G., Liu, J., Abbink, P., Maxfield, L. F., Seaman, M. S., Dugast, A.-S., Alter, G., Ferguson, M., Li, W., Earl, P. L., Moss, B., Giorgi, E. E., Szinger, J. J., Eller, L. A., Billings, E. A., Rao, M., Tovanabutra, S., Sanders-Buell, E., Weijtens, M., Pau, M. G., Schuitemaker, H., Robb, M. L., Kim, J. H., Korber, B. T., & Michael, N. L. (2013). Protective efficacy of a global HIV-1 mosaic vaccine against heterologous SHIV challenges in rhesus monkeys. *Cell*, 155(3), 531–539.

[15] Bashashati, A., Ha, G., Tone, A., Ding, J., Prentice, L. M., Roth, A., Rosner, J., Shumansky, K., Kalloger, S., Senz, J., Yan, W., McConechy, M., Melnyk, N., Anglesio, M., et al. (2013). Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *The Journal of Pathology*, 231(1), 21–34.

[16] Bassani-Sternberg, M., Bräunlein, E., Klar, R., Engleitner, T., Sinitcyn, P., Audehm, S., Straub, M., Weber, J., Slotta-Huspenina, J., Specht, K., Martignoni, M. E., Werner, A., Hein, R., H. Busch, D., Peschel, C., Rad, R., Cox, J., Mann, M., & Krackhardt, A. M. (2016). Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nature Communications*, 7, 13404.

[17] Basturk, O., Hong, S.-M., Wood, L. D., Adsay, N. V., Albores-Saavedra, J., Biankin, V. A., Brosens, L. A., Fukushima, N., Goggins, M., Hruban, R. H., Kato, Y., Klimstra, D. S., Klöppel, G., Krasinskas, A., Longnecker, D. S., Matthaei, H., Offerhaus, G. J. A., Shimizu, M., Takaori, K., Terris, B., Yachida, S., Esposito, I., & Furukawa, T. (2015). A revised classification system and recommendations from the baltimore consensus meeting for neoplastic precursor lesions in the pancreas. *The American Journal of Surgical Pathology*, 39(12), 1730–1741. 022720172.pdf.

[18] Beerenwinkel, N., Antal, T., Dingli, D., Traulsen, A., Kinzler, K. W., Velculescu, V. E., Vogelstein, B., & Nowak, M. A. (2007). Genetic Progression and the Waiting Time to Cancer. *PLOS Computational Biology*, 3(11), e225.

[19] Beerenwinkel, N., Schwarz, R. F., Gerstung, M., & Markowetz, F. (2015). Cancer evolution: mathematical models and computational inference. *Systematic Biology*, 64(1), e1—e25.

[20] Bellu, G., Saccomani, M. P., Audoly, S., & D'Angiò, L. (2007). Daisy: A new software tool to test global identifiability of biological and physiological systems. *Computer methods and programs in biomedicine*, 88(1), 52–61.

[21] Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289–300.

[22] Bernard, S., Bélair, J., & Mackey, M. C. (2003). Oscillations in cyclical neutropenia: new evidence based on mathematical modeling. *Journal of Theoretical Biology*, 223(3), 283–298.

[23] Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M., Mc Henry, K. T., Pinchback, R. M., Ligon, A. H., Cho, Y.-J., Haery, L., Greulich, H., Reich, M., Winckler, W., Lawrence, M. S., Weir, B. A., Tanaka, K. E., Chiang, D. Y., Bass, A. J., Loo, A., Hoffman, C., Prensner, J., Liefeld, T., Gao, Q., Yecies, D., Signoretti, S., Maher, E., Kaye, F. J., Sasaki, H., Tepper, J. E., Fletcher, J. A., Tabernero, J., Baselga, J., Tsao, M.-S., Demichelis, F., Rubin, M. A., Janne, P. A., Daly, M. J., Nucera, C., Levine, R. L., Ebert, B. L., Gabriel, S., Rustgi, A. K., Antonescu, C. R., Ladanyi, M., Letai, A., Garraway, L. A., Loda, M., Beer, D. G., True, L. D., Okamoto, A., Pomeroy, S. L., Singer, S., Golub, T. R., Lander, E. S., Getz, G., Sellers, W. R., & Meyerson, M. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283), 899–905.

[24] Biankin, V. A., Waddell, N., Kassahn, K. S., Gingras, M.-C., Muthuswamy, L. B., Johns, A. L., Miller, D. K., Wilson, P. J., Patch, A.-M., Wu, J., Chang, D. K., Cowley, M. J., Gardiner, B. B., Song, S., Harliwong, I., Idrisoglu, S., Nourse, C., Nourbakhsh, E., Manning, S., Wani, S., Gongora, M., Pajic, M., Scarlett, C. J., Gill, A. J., Pinho, V. A., Rooman, I., Anderson, M., Holmes, O., Leonard, C., Taylor, D., Wood, S., Xu, Q., Nones, K., Fink, J. L., Christ, A., Bruxner, T., Cloonan, N., Kolle, G., Newell, F., Pinese, M., Mead, R. S., Humphris, J. L., Kaplan, W., Jones, M. D., Colvin, E. K., Nagrial, A. M., Humphrey, E. S., Chou, A., Chin, V. T., Chantrill, L. A., Mawson, A., Samra, J. S., Kench, J. G., Lovell, J. A., Daly, R. J., Merrett, N. D., Toon, C., Epari, K., Nguyen, N. Q., Barbour, A., Zeps, N., Kakkar, N., Zhao, F., Wu, Y. Q., Wang, M., Muzny, D. M., Fisher, W. E., Brunicardi, F. C., Hodges, S. E., Reid, J. G., Drummond, J., Chang, K., Han, Y., Lewis, L. R., Dinh, H., Buhay, C. J., Beck, T., Timms, L., Sam, M., Begley, K., Brown, A., Pai, D., Panchal, A., Buchner, N., De Borja, R., Denroche, R. E., Yung, C. K., Serra, S., Onetto, N., Mukhopadhyay, D., Tsao, M.-S., Shaw, P. A., Petersen, G. M., Gallinger, S., Hruban, R. H., Maitra, A., Iacobuzio-Donahue, C. A., Schulick, R. D., Wolfgang, C. L., Morgan, R. A., Lawlor, R. T., Capelli, P., Corbo, V., Scardoni, M., Tortora, G., Tempero, M. A., Mann, K. M., Jenkins, N. A., Perez-Mancera, P. A., Adams, D. J., Largaespada, D. A., Wessels, L. F. A., Rust, A. G., Stein, L. D., Tuveson, D. A., Copeland, N. G., Musgrove, E. A., Scarpa, A., Eshleman, J. R., Hudson, T. J., Sutherland, R. L., Wheeler, D. A., Pearson, V. J., McPherson, J. D., Gibbs, R. A., &

Grimmond, S. M. (2012). Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*, 491(7424), 399–405.

[25] Blankson, J. (2011). The study of elite controllers: a pure academic exercise or a potential pathway to an HIV-1 vaccine? *Current Opinion in HIV and AIDS*, 6(3), 147–150.

[26] Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M., Boymans, S., Kuijk, E., Prins, P., Nijman, I. J., Martincorena, I., Mokry, M., Wiegerinck, C. L., Middendorp, S., Sato, T., Schwank, G., Nieuwenhuis, E. E. S., Verstegen, M. M. A., van der Laan, L. J. W., de Jonge, J., IJzermans, J. N. M., Vries, R. G., van de Wetering, M., Stratton, M. R., Clevers, H., Cuppen, E., & van Boxtel, R. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, 538(7624), 260–264.

[27] Boer, R. J. D., Homann, D., & Perelson, A. S. (2003). Different Dynamics of CD4+ and CD8+ T Cell Responses During and After Acute Lymphocytic Choriomeningitis Virus Infection. *The Journal of Immunology*, 171(8), 3928–3935.

[28] Bonet, M., Steel, M., Warnow, T., & Yooseph, S. (1998). Better methods for solving parsimony and compatibility. *Journal of Computational Biology*, 5(3), 391–407.

[29] Borducchi, E. N., Cabral, C., Stephenson, K. E., Liu, J., Abbink, P., Ng'ang'a, D., Nkolola, J. P., Brinkman, A. L., Peter, L., Lee, B. C., Jimenez, J., Jetton, D., Mondesir, J., Mojta, S., Chandrashekar, A., Molloy, K., Alter, G., Gerold, J. M., Hill, A. L., Lewis, M. G., Pau, M. G., Schuitemaker, H., Hesselgesser, J., Geleziunas, R., Kim, J. H., Robb, M. L., Michael, N. L., & Barouch, D. H. (2016). Ad26/MVA therapeutic vaccination with TLR7 stimulation in SIV-infected rhesus monkeys. *Nature*, 540(7632), 284–287.

[30] Borducchi, E. N., Liu, J., Nkolola, J. P., Cadena, A. M., Yu, W.-H., Fischinger, S., Broge, T., Abbink, P., Mercado, N. B., Chandrashekar, A., Jetton, D., Peter, L., McMahan, K., Moseley, E. T., Bekerman, E., Hesselgesser, J., Li, W., Lewis, M. G., Alter, G., Geleziunas, R., & Barouch, D. H. (2018). Antibody and TLR7 agonist delay viral rebound in SHIV-infected monkeys. *Nature*, 563(7731), 360.

[31] Bozic, I., Allen, B., & Nowak, M. A. (2012). Dynamics of targeted cancer therapy. *Trends in Molecular Medicine*, 18(6), 311–316.

[32] Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K. W., Vogelstein, B., & Nowak, M. A. (2010). Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107(43), 18545.

[33] Bozic, I., Gerold, J. M., & Nowak, M. A. (2016a). Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. *PLOS Computational Biology*, 12(2), e1004731.

[34] Bozic, I., Gerold, J. M., & Nowak, M. A. (2016b). Quantifying clonal and subclonal passenger mutations in cancer evolution. *PLoS Computational Biology*, 12(2).

[35] Bozic, I. & Nowak, M. A. (2013). Unwanted Evolution. *Science*, 342(6161), 938–939.

[36] Bozic, I. & Nowak, M. A. (2014). Timing and heterogeneity of mutations associated with drug resistance in metastatic cancers. *Proceedings of the National Academy of Sciences*, 111(45), 15964–15968.

[37] Bozic, I., Reiter, J. G., Allen, B., Antal, T., Chatterjee, K., Vogelstein, B., & Nowak, M. A. (2013). Evolutionary dynamics of cancer in response to targeted combination therapy. *eLife*, 2, e00747.

[38] Brastianos, P. K., Carter, S. L., Santagata, S., Cahill, D. P., Taylor-Weiner, A., Jones, R. T., Van Allen, E. M., Lawrence, M. S., Horowitz, P. M., Cibulskis, K., et al. (2015). Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer discovery*.

[39] Broder, S. (2010). The development of antiretroviral therapy and its impact on the HIV-1/AIDS pandemic. *Antiviral research*, 85(1), 1.

[40] Brodin, J., Zanini, F., Thebo, L., Lanz, C., Bratt, G., Neher, R. A., & Albert, J. (2016). Establishment and stability of the latent HIV-1 DNA reservoir. *eLife*, 5, e18889.

[41] Brown, D., Smeets, D., Székely, B., Larsimont, D., Szász, A. M., Adnet, P.-Y., Rothé, F., Rouas, G., Nagy, Z. I., Faragó, Z., Tőkés, A.-M., Dank, M., Szentmártoni, G., Udvarhelyi, N., Zoppoli, G., Pusztai, L., Piccart, M., Kulka, J., Lambrechts, D., Sotiriou, C., & Desmedt, C. (2017). Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations. *Nature Communications*, 8, 14944.

[42] Bui, J. K., Sobolewski, M. D., Keele, B. F., Spindler, J., Musick, A., Wiegand, A., Luke, B. T., Shao, W., Hughes, S. H., Coffin, J. M., Kearney, M. F., & Mellors, J. W. (2017). Proviruses with identical sequences comprise a large fraction of the replication-competent HIV reservoir. *PLOS Pathogens*, 13(3), e1006283.

[43] Burg, D., Rong, L., Neumann, A. U., & Dahari, H. (2009). Mathematical modeling of viral kinetics under immune control during primary HIV-1 infection. *Journal of Theoretical Biology*, 259(4), 751–759.

[44] Campbell, P. J., Pleasance, E. D., Stephens, P. J., Dicks, E., Rance, R., Goodhead, I., Follows, G. A., Green, A. R., Futreal, P. A., & Stratton, M. R. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci USA*, 105(35), 13081–13086.

[45] Campbell, P. J., Yachida, S., Mudie, L. J., Stephens, P. J., Pleasance, E. D., Stebbings, L. A., Morsberger, L. A., Latimer, C., McLaren, S., Lin, M.-L., McBride, D. J., Varela, I., Nik-Zainal, S. A., Leroy, C., Jia, M., Menzies, A., Butler, A. P., Teague, J. W., Griffin, C. A., Burton, J., Swerdlow, H., Quail, M. A., Stratton, M. R., Iacobuzio-Donahue, C., & Futreal,

P. A. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 467(7319), 1109–1113.

[46] Cardozo, E. F., Andrade, A., Mellors, J. W., Kuritzkes, D. R., Perelson, A. S., & Ribeiro, R. M. (2017). Treatment with integrase inhibitor suggests a new interpretation of HIV RNA decay curves that reveals a subset of cells with slow integration. *PLOS Pathogens*, 13(7), e1006478.

[47] Chan, P. L., Jacqmin, P., Lavielle, M., McFadyen, L., & Weatherley, B. (2011). The use of the saem algorithm in monolix software for estimation of population pharmacokinetic-pharmacodynamic-viral dynamics parameters of maraviroc in asymptomatic hiv subjects. *Journal of pharmacokinetics and pharmacodynamics*, 38(1), 41–61.

[48] Chang, M. T., Asthana, S., Gao, S. P., Lee, B. H., Chapman, J. S., Kandoth, C., Gao, J., Socci, N. D., Solit, D. B., Olshen, A. B., Schultz, N., & Taylor, B. S. (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature biotechnology*, 34(2), 155–63. 030220173.pdf.

[49] Chatterjee, A., Guedj, J., & Perelson, A. S. (2012). Mathematical modelling of HCV infection: what can it teach us in the era of direct-acting antiviral agents? *Antiviral Therapy*, 17(6 Pt B), 1171–1182.

[50] Chauvin, J., Ayral, G., & Traynard, P. (2018). Cossac (conditional sampling use for stepwise approach based on correlation tests) method for covariate search. *Poster PAGE, montreux Switzerland*, III-44.

[51] Chen, H. Y., Mascio, M. D., Perelson, A. S., Ho, D. D., & Zhang, L. (2007). Determination of virus burst size in vivo using a single-cycle SIV in rhesus macaques. *Proceedings of the National Academy of Sciences*, 104(48), 19079–19084.

[52] Choo, D. K., Murali-Krishna, K., Anita, R., & Ahmed, R. (2010). Homeostatic Turnover of Virus-Specific Memory CD8 T Cells Occurs Stochastically and Is Independent of CD4 T Cell Help. *The Journal of Immunology*, 185(6), 3436–3444.

[53] Chun, T.-W., Moir, S., & Fauci, A. S. (2015). HIV reservoirs as obstacles and opportunities for an HIV cure. *Nature Immunology*, 16(6), 584–589.

[54] Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., & Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3), 213–219.

[55] Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., & Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, 45(10), 1127–1133.

[56] Clapham, H. E., Tricou, V., Van Vinh Chau, N., Simmons, C. P., & Ferguson, N. M. (2014). Within-host viral dynamics of dengue serotype 1 infection. *Journal of the Royal Society Interface*, 11(96).

[57] Coldman, A. J. & Goldie, J. H. (1983). A model for the resistance of tumor cells to cancer chemotherapeutic agents. *Mathematical Biosciences*, 65(2), 291–307.

[58] Collins, F. (2007). Cancer: A Disease of the Genome. *Cancer Research*, 67(9 Supplement), PL01–01–PL01–01.

[59] Conway, J. M. & Coombs, D. (2011). A Stochastic Model of Latently Infected Cell Reactivation and Viral Blip Generation in Treated HIV Patients. *PLOS Computational Biology*, 7(4), e1002033.

[60] Conway, J. M. & Perelson, A. S. (2015). Post-treatment control of HIV infection. *Proceedings of the National Academy of Sciences*, 112(17), 5467–5472.

[61] Cooper, C. S., Eeles, R., Wedge, D. C., Van Loo, P., Gundem, G., Alexandrov, L. B., Kremeyer, B., Butler, A., Lynch, A. G., Camacho, N., Massie, C. E., Kay, J., Luxton, H. J., Edwards, S., Kote-Jarai, Z., Dennis, N., Merson, S., Leongamornlert, D., Zamora, J., Corbishley, C., Thomas, S., Nik-Zainal, S., O'Meara, S., Matthews, L., Clark, J., Hurst, R., Mithen, R., Bristow, R. G., Boutros, P. C., Fraser, M., Cooke, S., Raine, K., Jones, D., Menzies, A., Stebbings, L., Hinton, J., Teague, J., McLaren, S., Mudie, L., Hardy, C., Anderson, E., Joseph, O., Goody, V., Robinson, B., Maddison, M., Gamble, S., Greenman, C., Berney, D., Hazell, S., Livni, N., the ICGC Prostate Group, Fisher, C., Ogden, C., Kumar, P., Thompson, A., Woodhouse, C., Nicol, D., Mayer, E., Dudderidge, T., Shah, N. C., Gnanapragasam, V., Voet, T., Campbell, P., Futreal, A., Easton, D., Warren, A. Y., Foster, C. S., Stratton, M. R., Whitaker, H. C., McDermott, U., Brewer, D. S., & Neal, D. E. (2015). Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat Genet*, 47(4), 367–372.

[62] Cui, W. & Kaech, S. M. (2010). Generation of effector CD8+ T cells and their conversion to memory T cells. *Immunological reviews*, 236, 151–166.

[63] Curtius, K., Hazelton, W. D., Jeon, J., & Luebeck, E. G. (2015). A Multiscale Model Evaluates Screening for Neoplasia in Barrett's Esophagus. *PLOS Computational Biology*, 11(5), e1004272.

[64] Dahari, H., Shudo, E., Ribeiro, R. M., & Perelson, A. S. (2009). Modeling complex decay profiles of hepatitis B virus during antiviral therapy. *Hepatology*, 49(1), 32–38.

[65] Davoli, T., Xu, A. W., Mengwasser, K. E., Sack, L. M., Yoon, J. C., Park, P. J., & Elledge, S. J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, 155(4), 948–62.

[66] Day, W. H. E. & Sankoff, D. (1986). Computational complexity of inferring phylogenies by compatibility. *Systematic Biology*, 35(2), 224–229.

[67] De Boer, R. J. & Perelson, A. S. (2013). Quantifying T lymphocyte turnover. *Journal of Theoretical Biology*, 327, 45–87.

[68] de Bruin, E. C., McGranahan, N., Mitter, R., Salm, M., Wedge, D. C., Yates, L., Jamal-Hanjani, M., Shafi, S., Murugaesu, N., Rowan, A. J., Gronroos, E., Muhammad, M. A., Horswell, S., Gerliner, M., Varela, I., Jones, D., Marshall, J., Voet, T., Van Loo, P., Rassl, D. M., Rintoul, R. C., Janes, S. M., Lee, S. M., Forster, M., Ahmad, T., Lawrence, D., Falzon, M., Capitanio, A., Harkins, T. T., Lee, C. C., Tom, W., Teefe, E., Chen, S. C., Begum, S., Rabinowitz, A., Phillimore, B., Spencer-Dene, B., Stamp, G., Szallasi, Z., Matthews, N., Stewart, A., Campbell, P., & Swanton, C. (2014). Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*, 346(6206), 251–256.

[69] Deeks, S. G., Autran, B., Berkhout, B., Benkirane, M., Cairns, S., Chomont, N., Chun, T.-W., Churchill, M., Mascio, M. D., Katlama, C., Lafeuillade, A., Landay, A., Lederman, M., Lewin, S. R., Maldarelli, F., Margolis, D., Markowitz, M., Martinez-Picado, J., Mullins, J. I., Mellors, J., Moreno, S., O'Doherty, U., Palmer, S., Penicaud, M.-C., Peterlin, M., Poli, G., Routy, J.-P., Rouzioux, C., Silvestri, G., Stevenson, M., Telenti, A., Lint, C. V., Verdin, E., Woolfrey, A., Zaia, J., Barré-Sinoussi, F., Deeks, S. G., Autran, B., Berkhout, B., Benkirane, M., Cairns, S., Chomont, N., Chun, T.-W., Churchill, M., Mascio, M. D., Katlama, C., Lafeuillade, A., Landay, A., Lederman, M., Lewin, S. R., Maldarelli, F., Margolis, D., Markowitz, M., Martinez-Picado, J., Mullins, J. I., Mellors, J., Moreno, S., O'Doherty, U., Palmer, S., Penicaud, M.-C., Peterlin, M., Poli, G., Routy, J.-P., Rouzioux, C., Silvestri, G., Stevenson, M., Telenti, A., Lint, C. V., Verdin, E., Woolfrey, A., Zaia, J., & Barré-Sinoussi, F. (2012). Towards an HIV cure: a global scientific strategy. *Nature Reviews Immunology*, 12(8), 607–614. PMCID 3595991.

[70] Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a Stochastic Approximation Version of the EM Algorithm. *The Annals of Statistics*, 27(1), 94–128.

[71] DePristo, M. A., Banks, E., Poplin, R., Garimella, V. K., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature Genetics*, 43(5), 491–498.

[72] Derrien, T., Estellé, J., Marco Sola, S., Knowles, D. G., Raineri, E., Guigó, R., & Ribeca, P. (2012). Fast computation and applications of genome mappability. *PloS one*, 7(1), e30377—e30377.

[73] Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L., & Morris, Q. (2015). Phylowgs: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1), 35.

[74] Dewanji, A., Luebeck, E. G., & Moolgavkar, S. H. (2005). A generalized Luria-Delbrück model. *Mathematical Biosciences*, 197(2), 140–152.

[75] Diaz Jr, L. A., Williams, R. T., Wu, J., Kinde, I., Hecht, J. R., Berlin, J., Allen, B., Bozic, I., Reiter, J. G., Nowak, M. A., Kinzler, K. W., Oliner, K. S., Vogelstein, B., Diaz, L. A., Williams, R. T., Wu, J., Kinde, I., Hecht, J. R., Berlin, J., Allen, B., Bozic, I., Reiter, J. G., Nowak, M. A., Kinzler, K. W., Oliner, K. S., & Vogelstein, B. (2012). The molecular evolution of acquired resistance to targeted egfr blockade in colorectal cancers. *Nature*, 486(7404), 537–40.

[76] Durrett, R. (2013). Population genetics of neutral mutations in exponentially growing cancer cell populations. *The annals of applied probability : an official journal of the Institute of Mathematical Statistics*, 23(1), 230–250.

[77] Durrett, R. (2015). Branching process models of cancer. In *Branching Process Models of Cancer* (pp. 1–63). Cham: Springer International Publishing.

[78] Durrett, R., Foo, J., Leder, K., Mayberry, J., & Michor, F. (2011). Intratumor Heterogeneity in Evolutionary Models of Tumor Progression. *Genetics*, 188(2), 461–477.

[79] Durrett, R. & Moseley, S. (2010). Evolution of resistance and progression to disease during clonal expansion of cancer. *Theoretical Population Biology*, 77(1), 42–48.

[80] El-Kebir, M., Oesper, L., Acheson-Field, H., & Raphael, B. J. (2015). Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12), i62—i70.

[81] El-Kebir, M., Satas, G., & Raphael, B. J. (2018). Inferring parsimonious migration histories for metastatic cancers. *Nature Genetics*, 50(5), 718.

[82] Eyal, N. & Kuritzkes, D. R. (2013). Challenges in clinical trial design for HIV-1 cure research. *The Lancet*, 382(9903), 1464–1465.

[83] Fearon, E. R. & Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, 61(5), 759–67.

[84] Felsenstein, J. (2004). *Inferring phylogenies*, volume 2. Sinauer Associates, Sunderland, MA.

[85] Fennessey, C. M., Pinkevych, M., Immonen, T. T., Reynaldi, A., Venturi, V., Nadella, P., Reid, C., Newman, L., Lipkey, L., Oswald, K., Bosche, W. J., Trivett, M. T., Ohlen, C., Ott, D. E., Estes, J. D., Del Prete, G. Q., Lifson, J. D., Davenport, M. P., & Keele, B. F. (2017).

Genetically-barcoded SIV facilitates enumeration of rebound variants and estimation of reactivation rates in nonhuman primates following interruption of suppressive antiretroviral therapy. *PLoS pathogens*, 13(5), e1006359.

[86] Frank, S. A. & Nowak, M. A. (2003). Developmental predisposition to cancer. *Nature*, 422(6931), 494.

[87] Funk, E., Kottilil, S., Gilliam, B., & Talwani, R. (2014). Tickling the TLR7 to cure viral hepatitis. *Journal of Translational Medicine*, 12, 129.

[88] Furukawa, H., Iwata, R., & Moriyama, N. (2001). Growth rate of pancreatic adenocarcinoma: initial clinical experience. *Pancreas*, 22(4), 366–369.

[89] Gadhamsetty, S., Coorens, T., & Boer, R. J. d. (2016). Notwithstanding circumstantial alibis, cytotoxic T cells can be major killers of HIV-1 infected cells. *Journal of Virology*, (pp. JVI.00306–16).

[90] Ganusov, V. V. & Boer, R. J. D. (2013). A mechanistic model for bromodeoxyuridine dilution naturally explains labelling data of self-renewing T cell populations. *Journal of The Royal Society Interface*, 10(78), 20120617.

[91] Gerlinger, M., Horswell, S., Larkin, J., Rowan, A. J., Salm, M. P., Varela, I., Fisher, R., McGranahan, N., Matthews, N., Santos, C. R., et al. (2014). Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics*, 46(3), 225–233.

[92] Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, 366(10), 883–892.

[93] Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., & Beerenwinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature Communications*, 3, 811.

[94] Gibson, W. J., Hoivik, E. A., Halle, M. K., Taylor-weiner, A., Cherniack, A. D., Berg, A., Holst, F., Zack, T. I., Werner, H. M. J., Staby, K. M., Rosenberg, M., Stefansson, I. M., Kusonmano, K., Chevalier, A., Mauland, K. K., Trovik, J., Krakstad, C., Giannakis, M., Hodis, E., Woie, K., Bjorge, L., Vintermyr, O. K., Wala, J. A., Lawrence, M. S., Getz, G., & Carter, S. L. (2016). The genomic landscape and evolution of endometrial carcinoma progression and abdominopelvic metastasis. *Nat Genet*, 48, 848–855.

[95] Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25), 2340–2361.

[96] Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E., & Desai, M. M. (2017). The dynamics of molecular evolution over 60,000 generations. *Nature*, 551, 45–50.

[97] Greaves, M. & Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, 481(7381), 306–313.

[98] Griffiths, R. C. & Pakes, A. G. (1988). An Infinite-Alleles Version of the Simple Branching Process. *Advances in Applied Probability*, 20(3), 489–524.

[99] Gros, A., Parkhurst, M. R., Tran, E., Pasetto, A., Robbins, P. F., Ilyas, S., Prickett, T. D., Gartner, J. J., Crystal, J. S., Roberts, I. M., Trebska-McGowan, K., Wunderlich, J. R., Yang, J. C., & Rosenberg, S. A. (2016). Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. *Nature Medicine*, advance online publication.

[100] Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L. B., Tubio, J. M. C., Papaemmanuil, E., Brewer, D. S., Kallio, H. M. L., Högnäs, G., Annala, M., Kivinummi, K., Goody, V., Latimer, C., O'Meara, S., Dawson, K. J., Isaacs, W., Emmert-Buck, M. R., Nykter, M., Foster, C., Kote-Jarai, Z., Easton, D., Whitaker, H., Group, I. P. U. K., Neal, D. E., Cooper, C. S., Eeles, R. A., Visakorpi, T., Campbell, P. J., McDermott, U., Wedge, D. C., & Bova, S. (2015). The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547), 353–357.

[101] Gusfield, D. (1991). Efficient algorithms for inferring evolutionary trees. *Networks*, 21(1), 19–28.

[102] Haeno, H., Gonen, M., Davis, M. B., Herman, J. M., Iacobuzio-Donahue, C. A., & Michor, F. (2012). Computational modeling of pancreatic cancer reveals kinetics of metastasis suggesting optimum treatment strategies. *Cell*, 148(1), 362–375.

[103] Hajirasouliha, I. & Raphael, B. J. (2014). Reconstructing mutational history in multiply sampled tumors using perfect phylogeny mixtures. *International Workshop on Algorithms in Bioinformatics*, (pp. 354–367).

[104] Hammarlund, E., Lewis, M. W., Hansen, S. G., Strelow, L. I., Nelson, J. A., Sexton, G. J., Hanifin, J. M., & Slifka, M. K. (2003). Duration of antiviral immunity after smallpox vaccination. *Nature Medicine*, 9(9), 1131–1137.

[105] Harwood, J., Tachibana, A., Davis, R., Bhattacharyya, N. P., & Meuth, M. (1993). High rate of multilocus deletion in a human tumor cell line. *Human Molecular Genetics*, 2(2), 165–171.

[106] Henrich, T., Hanhauser, E., Marty, F., Sirignano, M., Keating, S., Lee, T., Robles, Y., Davis, B., Li, J., Heisey, A., et al. (2014). Antiretroviral-free hiv-1 remission and viral rebound after allogeneic stem cell transplantationreport of 2 caseshiv-1 remission and viral rebound after allogeneic hsct. *Annals of internal medicine*, 161(5), 319–327.

[107] Henrich, T. J., Hatano, H., Bacon, O., Hogan, L. E., Rutishauser, R., Hill, A., Kearney, M. F., Anderson, E. M., Buchbinder, S. P., Cohen, S. E., Abdel-Mohsen, M., Pohlmeyer, C. W., Fromentin, R., Hoh, R., Liu, A. Y., McCune, J. M., Spindler, J., Metcalf-Pate, K., Hobbs, K. S., Thanh, C., Gibson, E. A., Kuritzkes, D. R., Siliciano, R. F., Price, R. W., Richman, D. D., Chomont, N., Siliciano, J. D., Mellors, J. W., Yukl, S. A., Blankson, J. N., Liegler, T., & Deeks, S. G. (2017). HIV-1 persistence following extremely early initiation of antiretroviral therapy (ART) during acute HIV-1 infection: An observational study. *PLOS Medicine*, 14(11), e1002417.

[108] Hill, A. L., Rosenbloom, D. I. S., Fu, F., Nowak, M. A., & Siliciano, R. F. (2014). Predicting the outcomes of treatment to eradicate the latent reservoir for HIV-1. *Proceedings of the National Academy of Sciences*, 111(37), 13475–13480.

[109] Hill, A. L., Rosenbloom, D. I. S., Nowak, M. A., & Siliciano, R. F. (2018). Insight into treatment of HIV infection from viral dynamics models. *Immunological Reviews*, 285(1), 9–25.

[110] Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M., Markowitz, M., et al. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, 373(6510), 123–126.

[111] Hong, M. K. H., Macintyre, G., Wedge, D. C., Van Loo, P., Patel, K., Lunke, S., Alexandrov, L. B., Sloggett, C., Cmero, M., Marass, F., Tsui, D., Mangiola, S., Lonieand, A., Naeem, H., Sapre, N., Phal, P. M., Kurganovs, N., Chin, X., Kerger, M., Warren, A. Y., Neal, D., Gnanapragasam, V., Rosenfeld, N., Pedersen, J. S., Ryan, A., Haviv, I., Costello, A. J., Corcoran, N. M., & Hovens, C. M. (2015a). Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nat Commun*, 6, 6605.

[112] Hong, W. S., Shpak, M., & Townsend, J. P. (2015b). Inferring the origin of metastases from cancer phylogenies. *Cancer Research*, 75(19), 4021–4025.

[113] Hosoda, W., Chianchiano, P., Griffin, J. F., Pittman, M. E., Brosens, L. A., Noë, M., Yu, J., Shindo, K., Suenaga, M., Rezaee, N., Yonescu, R., Ning, Y., Albores-Saavedra, J., Yoshizawa, N., Harada, K., Yoshizawa, A., Hanada, K., Yonehara, S., Shimizu, M., Uehara, T., Samra, J. S., Gill, A. J., Wolfgang, C. L., Goggins, M. G., Hruban, R. H., & Wood, L. D. (2017). Genetic analyses of isolated high-grade pancreatic intraepithelial neoplasia (hg-panin) reveal paucity of alterations in *TP53* and *SMAD4*. *The Journal of Pathology*, 242(1), 16–23.

[114] Hruban, R. H., Goggins, M., Parsons, J., & Kern, S. E. (2000). Progression model for pancreatic cancer. *Clinical cancer research*, 6(8), 2969–72.

[115] Hruban, R. H., Takaori, K., Klimstra, D. S., Adsay, N. V., Albores-Saavedra, J., Biankin, V. A., Biankin, S. A., Compton, C., Fukushima, N., Furukawa, T., Goggins, M., Kato, Y., Klöppel, G., Longnecker, D. S., Lüttges, J., Maitra, A., Offerhaus, G. J. A., Shimizu, M., &

Yonezawa, S. (2004). An illustrated consensus on the classification of pancreatic intraepithelial neoplasia and intraductal papillary mucinous neoplasms. *The American journal of surgical pathology*, 28(8), 977–87.

[116] Irvine, D. (2016). Materializing the future of vaccines and immunotherapy. *Nature Reviews Materials*, 1, 15008.

[117] Iwasa, Y., Nowak, M. A., & Michor, F. (2006). Evolution of Resistance During Clonal Expansion. *Genetics*, 172(4), 2557–2566.

[118] Jamal-Hanjani, M., Wilson, G. A., McGranahan, N., Birkbak, N. J., Watkins, T. B. K., Veeriah, S., Shafi, S., Johnson, D. H., Mitter, R., Rosenthal, R., Salm, M., Horswell, S., Escudero, M., Matthews, N., Rowan, A., Chambers, T., Moore, D. A., Turajlic, S., Xu, H., Lee, S.-M., Forster, M. D., Ahmad, T., Hiley, C. T., Abbosh, C., Falzon, M., Borg, E., Marafioti, T., Lawrence, D., Hayward, M., Kolvekar, S., Panagiotopoulos, N., Janes, S. M., Thakrar, R., Ahmed, A., Blackhall, F., Summers, Y., Shah, R., Joseph, L., Quinn, A. M., Crosbie, P. A., Naidu, B., Middleton, G., Langman, G., Trotter, S., Nicolson, M., Remmen, H., Kerr, K., Chetty, M., Gomersall, L., Fennell, D. A., Nakas, A., Rathinam, S., Anand, G., Khan, S., Russell, P., Ezhil, V., Ismail, B., Irvin-Sellers, M., Prakash, V., Lester, J. F., Kornaszewska, M., Attanoos, R., Adams, H., Davies, H., Dentro, S., Taniere, P., O'Sullivan, B., Lowe, H. L., Hartley, J. A., Iles, N., Bell, H., Ngai, Y., Shaw, J. A., Herrero, J., Szallasi, Z., Schwarz, R. F., Stewart, A., Quezada, S. A., Le Quesne, J., Van Loo, P., Dive, C., Hackshaw, A., & Swanton, C. (2017). Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine*, 376(22), 2109–2121.

[119] Jones, B. R., Kinloch, N. N., Horacsek, J., Ganase, B., Harris, M., Harrigan, P. R., Jones, R. B., Brockman, M. A., Joy, J. B., Poon, A. F. Y., & Brumme, Z. L. (2018). Phylogenetic approach to recover integration dates of latent HIV sequences within-host. *Proceedings of the National Academy of Sciences*, 115(38), E8958–E8967.

[120] Jones, S., Zhang, X., Parsons, D. W., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., Hong, S.-M., Fu, B., Lin, M.-T., Calhoun, E. S., Kamiyama, M., Walter, K., Nikolskaya, T., Nikolsky, Y., Hartigan, J., Smith, D. R., Hidalgo, M., Leach, S. D., Klein, A. P., Jaffee, E. M., Goggins, M., Maitra, A., Iacobuzio-Donahue, C., Eshleman, J. R., Kern, S. E., Hruban, R. H., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E., & Kinzler, K. W. (2008a). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science (New York, N.Y.)*, 321(5897), 1801–6.

[121] Jones, S. S., Chen, W.-d., Parmigiani, G., Diehl, F., Beerenwinkel, N., Antal, T., Traulsen, A., Nowak, M. A., Siegel, C., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., Willis, J., & Markowitz, S. D. (2008b). Comparative lesion sequencing provides insights into tumor evolution. *Proceedings of the National Academy of Sciences*, 105(11), 4283–4288.

[122] Julien, J.-P., Sok, D., Khayat, R., Lee, J. H., Doores, K. J., Walker, L. M., Ramos, A., Diwanji, D. C., Pejchal, R., Cupo, A., Katpally, U., Depetris, R. S., Stanfield, R. L., McBride, R., Marozsan, A. J., Paulson, J. C., Sanders, R. W., Moore, J. P., Burton, D. R., Poignard, P., Ward, A. B., & Wilson, I. A. (2013). Broadly Neutralizing Antibody PGT121 Allosterically Modulates CD4 Binding via Recognition of the HIV-1 gp120 V3 Base and Multiple Surrounding Glycans. *PLOS Pathogens*, 9(5), e1003342.

[123] Kanda, M., Matthaei, H., Wu, J., Hong, S.-M., Yu, J., Borges, M., Hruban, R. H., Maitra, A., Kinzler, K., Vogelstein, B., & Goggins, M. (2012). Presence of somatic mutations in most early-stage pancreatic intraepithelial neoplasia. *Gastroenterology*, 142(4), 730–733.e9.

[124] Karp, R. M. (1972). Reducibility among combinatorial problems. In *Complexity of Computer Computations*, The IBM Research Symposia Series (pp. 85–103). Springer US.

[125] Kasturi, S. P., Kozlowski, P. A., Nakaya, H. I., Burger, M. C., Russo, P., Pham, M., Kovalenkov, Y., Silveira, E. L. V., Havenar-Daughton, C., Burton, S. L., Kilgore, K. M., Johnson, M. J., Nabi, R., Legere, T., Sher, Z. J., Chen, X., Amara, R. R., Hunter, E., Bosinger, S. E., Spearman, P., Crotty, S., Villinger, F., Derdeyn, C. A., Wrammert, J., & Pulendran, B. (2017). Adjuvanting a Simian Immunodeficiency Virus Vaccine with Toll-Like Receptor Ligands Encapsulated in Nanoparticles Induces Persistent Antibody Responses and Enhanced Protection in TRIM5α Restrictive Macaques. *Journal of Virology*, 91(4).

[126] Keller, P. & Antal, T. (2015). Mutant number distribution in an exponentially growing population. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(1), P01011.

[127] Kerr, J. F. R. & Searle, J. (1972). A suggested explanation for the paradoxically slow growth rate of basal-cell carcinomas that contain numerous mitotic figures. *The Journal of Pathology*, 107(1), 41–44.

[128] Kessler, D. A. & Levine, H. (2013). Large population solution of the stochastic Luria-Delbrück evolution model. *Proceedings of the National Academy of Sciences*, 110(29), 11682–11687.

[129] Kessler, D. A. & Levine, H. (2015). Scaling solution in the large population limit of the general asymmetric stochastic Luria-Delbrück evolution process. *Journal of statistical physics*, 158(4), 783–805.

[130] Kim, R., Emi, M., & Tanabe, K. (2007). Cancer immunoediting from immune surveillance to immune escape. *Immunology*, 121(1), 1–14.

[131] Kim, T.-M., Jung, S.-H., An, C. H., Lee, S. H., Baek, I.-P., Kim, M. S., Park, S.-W., Rhee, J.-K., Lee, S.-H., & Chung, Y.-J. (2015). Subclonal genomic architectures of primary and metastatic colorectal cancer based on intratumoral genetic heterogeneity. *Clinical Cancer Research*, 21(19), 4461–4472.

[132] Kimmel, M. & Axelrod, D. E. (2002). *Branching Processes in Biology*. Interdisciplinary Applied Mathematics. New York: Springer-Verlag.

[133] Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications*, 13(3), 235–248.

[134] Kleeff, J., Korc, M., Apte, M., La Vecchia, C., Johnson, C. D., Biankin, V. A., Neale, R. E., Tempero, M., Tuveson, D. A., Hruban, R. H., & Neoptolemos, J. P. (2016). Pancreatic cancer. *Nature reviews. Disease primers*, 2, 16022. 071420171-s.pdf.

[135] Klein, W. M., Hruban, R. H., Klein-Szanto, A. J., & Wilentz, R. E. (2002). Direct correlation between proliferative activity and dysplasia in pancreatic intraepithelial neoplasia (panin): Additional evidence for a recently proposed model of progression. *Modern Pathology*, 15(4), 441–447.

[136] Komarova, N. L., Lengauer, C., Vogelstein, B., & Nowak, M. A. (2002). Dynamics of genetic instability in sporadic and familial colorectal cancer.

[137] Komarova, N. L., Wu, L., & Baldi, P. (2007). The fixed-size LuriaDelbrück model with a nonzero death rate. *Mathematical Biosciences*, 210(1), 253–290.

[138] Kuhn, E. & Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49(4), 1020–1038.

[139] Kumar, A., Coleman, I., Morrissey, C., Zhang, X., True, L. D., Gulati, R., Etzioni, R., Bolouri, H., Montgomery, B., White, T., Lucas, J. M., Brown, L. G., Dumpit, R. F., DeSarkar, N., Higano, C., Yu, E. Y., Coleman, R., Schultz, N., Fang, M., Lange, P. H., Shendure, J., Vessella, R. L., & Nelson, P. S. (2016). Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nature Medicine*, 22(4), 369–378.

[140] Landau, D., Carter, S., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., Wan, Y., Zhang, W., Shukla, S., Vartanov, A., Fernandes, S., Saksena, G., Cibulskis, K., Tesar, B., Gabriel, S., Hacohen, N., Meyerson, M., Lander, E., Neuberg, D., Brown, J., Getz, G., & Wu, C. (2013). Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell*, 152(4), 714–726.

[141] Landau, D. A., Tausch, E., Taylor-Weiner, A. N., Stewart, C., Reiter, J. G., Bahlo, J., Kluth, S., Bozic, I., Lawrence, M., Bottcher, S., Cibulskis, K., Mertens, D., Sougnez, C., Rosenberg, M., Hess, J. M., Carter, S. L., Edelmann, J., Kless, S., Kneba, M., Ritgen, M., Fink, A., Fischer, K., Gabriel, S., Lander, E. S., Nowak, M. A., Dohner, H., Hallek, M., Neuberg, D., Getz, G., Stilgenbauer, S., & Wu, C. J. (2015). Mutations driving cll and their evolution in progression and relapse. *Nature*.

[142] Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, V. G., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Lee, R. S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S., & Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457), 214–8.

[143] Le, D. T., Uram, J. N., Wang, H., Bartlett, B. R., Kemberling, H., Eyring, A. D., Skora, A. D., Luber, B. S., Azad, N. S., Laheru, D., Biedrzycki, B., Donehower, R. C., Zaheer, A., Fisher, G. A., Crocenzi, T. S., Lee, J. J., Duffy, S. M., Goldberg, R. M., de la Chapelle, A., Koshiji, M., Bhaijee, F., Huebner, T., Hruban, R. H., Wood, L. D., Cuka, N., Pardoll, D. M., Papadopoulos, N., Kinzler, K. W., Zhou, S., Cornish, T. C., Taube, J. M., Anders, R. A., Eshleman, J. R., Vogelstein, B., & Diaz, L. A. J. (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine*, 372(26), 2509–2520.

[144] Leder, K., Foo, J., Skaggs, B., Gorre, M., Sawyers, C. L., & Michor, F. (2011). Fitness Conferred by BCR-ABL Kinase Domain Mutations Determines the Risk of Pre-Existing Resistance in Chronic Myeloid Leukemia. *PLOS ONE*, 6(11), e27682.

[145] Lengauer, C., Kinzler, K. W., & Vogelstein, B. (1998). Genetic instabilities in human cancers. *Nature*, 396(6712), 643–9.

[146] Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14), 1754–1760.

[147] Li, J. Z., Etemad, B., Ahmed, H., Aga, E., Bosch, R. J., Mellors, J. W., Kuritzkes, D. R., Lederman, M. M., Para, M., & Gandhi, R. T. (2016). The size of the expressed HIV reservoir predicts timing of viral rebound after treatment interruption. *AIDS*, 30(3), 343–353.

[148] Lim, S.-Y., Osuna, C. E., Hraber, P. T., Hesselgesser, J., Gerold, J. M., Barnes, T. L., Sanisetty, S., Seaman, M. S., Lewis, M. G., Geleziunas, R., Miller, M. D., Cihlar, T., Lee, W. A., Hill, A. L., & Whitney, J. B. (2018). TLR7 agonists induce transient viremia and reduce the viral reservoir in SIV-infected rhesus macaques on antiretroviral therapy. *Science Translational Medicine*, 10(439), eaao4521.

[149] Lopatin, U., Wolfgang, G., Tumas, D., Frey, C. R., Ohmstede, C., Hesselgesser, J., Kearney, B., Moorehead, L., Subramanian, G. M., & McHutchison, J. G. (2013). Safety, pharmacoki-

netics and pharmacodynamics of GS-9620, an oral Toll-like receptor 7 agonist. *Antiviral Therapy*, 18(3), 409–418.

[150] Lu, C.-L., Murakowski, D. K., Bournazos, S., Schoofs, T., Sarkar, D., Halper-Stromberg, A., Horwitz, J. A., Nogueira, L., Golijanin, J., Gazumyan, A., Ravetch, J. V., Caskey, M., Chakraborty, A. K., & Nussenzweig, M. C. (2016). Enhanced clearance of HIV-1-infected cells by anti-HIV-1 broadly neutralizing antibodies in vivo. *Science (New York, N.Y.)*, 352(6288), 1001–1004.

[151] Luebeck, E. G., Curtius, K., Jeon, J., & Hazelton, W. D. (2013). Impact of Tumor Progression on Cancer Incidence Curves. *Cancer Research*, 73(3), 1086–1096.

[152] Luo, R., Piovoso, M. J., Martinez-Picado, J., & Zurakowski, R. (2012). HIV Model Parameter Estimates from Interruption Trial Data including Drug Efficacy and Reservoir Dynamics. *PLoS ONE*, 7(7), e40198.

[153] Luria, S. E. & DelbrÃŒck, M. (1943). Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics*, 28(6), 491–511.

[154] Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences*, 107(3), 961–968.

[155] Ma, J., Ratan, A., Raney, B. J., Suh, B. B., Miller, W., & Haussler, D. (2008). The infinite sites model of genome evolution. *Proc Natl Acad Sci USA*, 105(38), 14254–14261.

[156] Macallan, D. C., Wallace, D. L., Irvine, A. J., Asquith, B., Worth, A., Ghattas, H., Zhang, Y., Griffin, G. E., Tough, D. F., & Beverley, P. C. (2003). Rapid turnover of T cells in acute infectious mononucleosis. *European Journal of Immunology*, 33(10), 2655–2665.

[157] Maddipati, R. & Stanger, B. Z. (2015). Pancreatic cancer metastases harbor evidence of polyclonality. *Cancer Discovery*, 5(10), 1086–1097.

[158] Makohon-Moore, A. & Iacobuzio-Donahue, C. A. (2016). Pancreatic cancer biology and genetics from an evolutionary perspective. *Nature reviews. Cancer*, 16(9), 553–65.

[159] Makohon-Moore, A. P., Matsukuma, K., Zhang, M., Reiter, J. G., Gerold, J. M., Jiao, Y., Sikkema, L., Attiyeh, M. A., Yachida, S., Sandone, C., Hruban, R. H., Klimstra, D. S., Papadopoulos, N., Nowak, M. A., Kinzler, K. W., Vogelstein, B., & Iacobuzio-Donahue, C. A. (2018). Precancerous neoplastic cells can move through the pancreatic ductal system. *Nature*, 561(7722), 201.

[160] Makohon-Moore, A. P., Zhang, M., Reiter, J. G., Bozic, I., Allen, B., Kundu, D., Chatterjee, K., Wong, F., Jiao, Y., Kohutek, Z. A., Hong, J., Attiyeh, M., Javier, B., Wood, L. D., Hruban, R. H., Nowak, M. A., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., & Iacobuzio-Donahue, C. A. (2017). Limited heterogeneity of known driver gene mutations

among the metastases of individual patients with pancreatic cancer. *Nature Genetics*, 49(3), 358–366.

[161] Makohon-Moore, A. P., Zhang, M., Reiter, J. G., Bozic, I., Allen, B., Kundu, D., Chatterjee, K., Wong, F., Jiao, Y., Kohutek, Z. A., Hong, J., McMahon, B., Wood, L. D., Hruban, R. H., Nowak, M. A., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., & Iacobuzio-Donahue, C. A. (2016). There is little heterogeneity of known driver genes among the metastases of individual pancreatic cancer patients. *Nature Genetics*, (in press).

[162] Maley, C. C., Galipeau, P. C., Li, X., Sanchez, C. A., Paulson, T. G., & Reid, B. J. (2004). Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett's Esophagus. *Cancer Research*, 64(10), 3414–3427.

[163] Malikic, S., McPherson, A. W., Donmez, N., & Sahinalp, C. S. (2015). Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9), 1349–1356.

[164] Markowitz, M., Louie, M., Hurley, A., Sun, E., Di Mascio, M., Perelson, A. S., & Ho, D. D. (2003). A Novel Antiviral Intervention Results in More Accurate Assessment of Human Immunodeficiency Virus Type 1 Replication Dynamics and T-Cell Decay In Vivo. *Journal of Virology*, 77(8), 5037–5038.

[165] Marty, R., Kaabinejadian, S., Rossell, D., Slifker, M. J., van de Haar, J., Engin, H. B., de Prisco, N., Ideker, T., Hildebrand, W. H., Font-Burgada, J., & Carter, H. (2017). MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell*, 171(6), 1272–1283.e15.

[166] Marty Pyke, R., Thompson, W. K., Salem, R. M., Font-Burgada, J., Zanetti, M., & Carter, H. (2018). Evolutionary Pressure against MHC Class II Binding Cancer Mutations. *Cell*, 175(2), 416–428.e13.

[167] Massagué, J. & Obenauf, A. C. (2016). Metastatic colonization by circulating tumour cells. *Nature*, 529(7586), 298–306.

[168] Matsuda, Y., Furukawa, T., Yachida, S., Nishimura, M., Seki, A., Nonaka, K., Aida, J., Takubo, K., Ishiwata, T., Kimura, W., Arai, T., & Mino-Kenudson, M. (2017). The prevalence and clinicopathological characteristics of high-grade pancreatic intraepithelial neoplasia. *Pancreas*, 46(5), 658–664. 022720171.pdf.

[169] McDermott, D., Lebbé, C., Hodi, F. S., Maio, M., Weber, J. S., Wolchok, J. D., Thompson, J. A., & Balch, C. M. (2014). Durable benefit and the potential for long-term survival with immunotherapy in advanced melanoma. *Cancer Treatment Reviews*, 40(9), 1056–1064.

[170] McGranahan, N., Furness, A. J. S., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., Jamal-Hanjani, M., Wilson, G. A., Birkbak, N. J., Hiley, C. T., Watkins, T. B. K., Shafi, S., Murugaesu, N., Mitter, R., Akarca, A. U., Linares, J., Marafioti, T., Henry, J. Y., Allen, E.

M. V., Miao, D., Schilling, B., Schadendorf, D., Garraway, L. A., Makarov, V., Rizvi, N. A., Snyder, A., Hellmann, M. D., Merghoub, T., Wolchok, J. D., Shukla, S. A., Wu, C. J., Peggs, K. S., Chan, T. A., Hadrup, S. R., Quezada, S. A., & Swanton, C. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*, 351(6280), 1463–1469.

[171] McGranahan, N. & Swanton, C. (2015). Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell*, 27(1), 15–26.

[172] McPherson, A., Roth, A., Laks, E., Masud, T., Bashashati, A., Zhang, A. W., Ha, G., Biele, J., Yap, D., Wan, A., Prentice, L. M., Khattra, J., Smith, M. A., Nielsen, C. B., Mullaly, S. C., Kalloger, S., Karnezis, A., Shumansky, K., Siu, C., Rosner, J., Chan, H. L., Ho, J., Melnyk, N., Senz, J., Yang, W., Moore, R., Mungall, A. J., Marra, M. A., Bouchard-Cote, A., Gilks, C. B., Huntsman, D. G., McAlpine, J. N., Aparicio, S., & Shah, S. P. (2016). Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics*.

[173] Merlo, L. M. F., Pepper, J. W., Reid, B. J., & Maley, C. C. (2006). Cancer as an evolutionary and ecological process. *Nature Reviews Cancer*, 6(12), 924–935.

[174] Miller, C. A., White, B. S., Dees, N. D., Griffith, M., Welch, J. S., Griffith, O. L., Vij, R., Tomasson, M. H., Graubert, T. A., Walter, M. J., et al. (2014). Sciclone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Computational Biology*, 10(8), e1003665.

[175] Mose, L. E., Wilkerson, M. D., Hayes, D. N., Perou, C. M., & Parker, J. S. (2014). Abra: improved coding indel detection via assembly-based realignment. *Bioinformatics (Oxford, England)*, 30(19), 2813–5.

[176] Murillo, L. N., Murillo, M. S., & Perelson, A. S. (2013). Towards multiscale modeling of influenza infection. *Journal of Theoretical Biology*, 332, 267–290.

[177] Murphy, S. J., Hart, S. N., Lima, J. F., Kipp, B. R., Klebig, M., Winters, J. L., Szabo, C., Zhang, L., Eckloff, B. W., Petersen, G. M., Scherer, S. E., Gibbs, R. A., McWilliams, R. R., Vasmatzis, G., & Couch, F. J. (2013). Genetic alterations associated with progression from pancreatic intraepithelial neoplasia to invasive pancreatic tumor. *Gastroenterology*, 145(5), 1098–1109.e1. 2017000001.pdf 2017000001notes.pdf.

[178] Murray, A. J., Kwon, K. J., Farber, D. L., & Siliciano, R. F. (2016). The Latent Reservoir for HIV-1: How Immunologic Memory and Clonal Expansion Contribute to HIV-1 Persistence. *The Journal of Immunology*, 197(2), 407–417.

[179] Nakhleh, L., Ringe, D., & Warnow, T. (2005). Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, 81(2), 382–420.

[180] Naxerova, K., Brachtel, E., Salk, J. J., Seese, A. M., Power, K., Abbasi, B., Snuderl, M., Chiang, S., Kasif, S., & Jain, R. K. (2014). Hypermutable dna chronicles the evolution of human colon cancer. *Proc Natl Acad Sci USA*, 111(18), E1889—E1898.

[181] Naxerova, K. & Jain, R. K. (2015). Using tumour phylogenetics to identify the roots of metastasis in humans. *Nature Reviews Clinical Oncology*, 12(238), 258–272.

[182] Nemhauser, G. L. & Wolsey, L. A. (1988). *Integer and combinatorial optimization*, volume 18. Wiley New York.

[183] Neumann, A. U., Lam, N. P., Dahari, H., Gretch, D. R., Wiley, T. E., Layden, T. J., & Perelson, A. S. (1998). Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-α therapy. *Science*, 282(5386), 103–107.

[184] Nicholson, M. D. & Antal, T. (2019). Competing evolutionary paths in growing populations with applications to multidrug resistance. *PLOS Computational Biology*, 15(4), e1006866.

[185] Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., Shlien, A., Cooke, S. L., Hinton, J., Menzies, A., Stebbings, L. A., Leroy, C., Jia, M., Rance, R., Mudie, L. J., Gamble, S. J., Stephens, P. J., McLaren, S., Tarpey, P. T. S., Papaemmanuil, E., HR, D., Varela, I., McBride, D. J., Bignell, G. R., Leung, K., & Butler, A. P. (2012). The life history of 21 breast cancers. *Cell*, 149(5), 994–1007.

[186] Niknafs, N., Beleva-Guthrie, V., Naiman, D. Q., & Karchin, R. (2015). Subclonal hierarchy inference from somatic mutations: Automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing. *PLoS Comput Biol*, 11(10), e1004416.

[187] Nishimura, Y., Gautam, R., Chun, T.-W., Sadjadpour, R., Foulds, K. E., Shingai, M., Klein, F., Gazumyan, A., Golijanin, J., Donaldson, M., Donau, O. K., Plishka, R. J., Buckler-White, A., Seaman, M. S., Lifson, J. D., Koup, R. A., Fauci, A. S., Nussenzweig, M. C., & Martin, M. A. (2017). Early antibody therapy can induce long-lasting immunity to SHIV. *Nature*, 543(7646), 559–563.

[188] Nishimura, Y., Igarashi, T., Donau, O. K., Buckler-White, A., Buckler, C., Lafont, B. A. P., Goeken, R. M., Goldstein, S., Hirsch, V. M., & Martin, M. A. (2004). Highly pathogenic SHIVs and SIVs target different CD4+ T cell subsets in rhesus monkeys, explaining their divergent clinical courses. *Proceedings of the National Academy of Sciences*, 101(33), 12324–12329.

[189] Notta, F., Chan-Seng-Yue, M., Lemire, M., Li, Y., Wilson, G. W., Connor, A. A., Denroche, R. E., Liang, S.-B., Brown, A. M. K., Kim, J. C., Wang, T., Simpson, J. T., Beck, T., Borgida, A., Buchner, N., Chadwick, D., Hafezi-Bakhtiari, S., Dick, J. E., Heisler, L., Hollingsworth, M. A., Ibrahimov, E., Jang, G. H., Johns, J., Jorgensen, L. G. T., Law, C., Ludkovski, O.,

Lungu, I., Ng, K., Pasternack, D., Petersen, G. M., Shlush, L. I., Timms, L., Tsao, M.-S., Wilson, J. M., Yung, C. K., Zogopoulos, G., Bartlett, J. M. S., Alexandrov, L. B., Real, F. X., Cleary, S. P., Roehrl, M. H., McPherson, J. D., Stein, L. D., Hudson, T. J., Campbell, P. J., & Gallinger, S. (2016). A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature*, 538(7625), 378–382. 2017000002.pdf 011617Journal.docx.

[190] Nowak, M. & Schuster, P. (1989). Error thresholds of replication in finite populations mutation frequencies and the onset of muller's ratchet. *Journal of Theoretical Biology*, 137(4), 375–395.

[191] Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Cambridge, Mass: Belknap Press, first edition edition edition.

[192] Nowak, M. A., Bonhoeffer, S., Hill, A. M., Boehme, R., Thomas, H. C., & McDade, H. (1996). Viral dynamics in hepatitis B virus infection. *Proceedings of the National Academy of Sciences*, 93, 4398–4402.

[193] Nowak, M. A., Komarova, N. L., Sengupta, A., Jallepalli, P. V., Shih, I.-M., Vogelstein, B., & Lengauer, C. (2002). The role of chromosomal instability in tumor initiation. *Proceedings of the National Academy of Sciences*, 99(25), 16226–16231.

[194] Nowak, M. A. & May, R. M. C. (2000). *Virus dynamics: mathematical principles of immunology and virology*. Oxford University Press, USA.

[195] Nowak, M. A., Michor, F., & Iwasa, Y. (2003). The linear process of somatic evolution. *Proceedings of the National Academy of Sciences*, 100(25), 14966–14969.

[196] Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, 194(4260), 23–28.

[197] Nuraini, N., Tasman, H., Soewono, E., & Sidarto, K. A. (2009). A with-in host Dengue infection model with immune response. *Mathematical and Computer Modelling*, 49(5–6), 1148–1155.

[of Health Statistics & Informatics] of Health Statistics, D. & Informatics, W. Global health risks: mortality and burden of disease attributable to selected major risks.

[199] Patel, M. C., Shirey, K. A., Pletneva, L. M., Boukhvalova, M. S., Garzino-Demo, A., Vogel, S. N., & Blanco, J. C. (2014). Novel drugs targeting Toll-like receptors for antiviral therapy. *Future Virology*, 9(9), 811–829.

[200] Patel, S. J., Sanjana, N. E., Kishton, R. J., Eidizadeh, A., Vodnala, S. K., Cam, M., Gartner, J. J., Jia, L., Steinberg, S. M., Yamamoto, T. N., Merchant, A. S., Mehta, G. U., Chichura, A., Shalem, O., Tran, E., Eil, R., Sukumar, M., Guijarro, E. P., Day, C.-P., Robbins, P., Feldman, S., Merlino, G., Zhang, F., & Restifo, N. P. (2017). Identification of essential genes for cancer immunotherapy. *Nature*, advance online publication.

[201] Pea, A., Yu, J., Rezaee, N., Luchini, C., He, J., Dal Molin, M., Griffin, J. F., Fedor, H., Fesharakizadeh, S., Salvia, R., Weiss, M. J., Bassi, C., Cameron, J. L., Zheng, L., Scarpa, A., Hruban, R. H., Lennon, A. M., Goggins, M., Wolfgang, C. L., & Wood, L. D. (2017). Targeted dna sequencing reveals patterns of local progression in the pancreatic remnant following resection of intraductal papillary mucinous neoplasm (ipmn) of the pancreas. *Annals of surgery*, 266(1), 133–141.

[202] Pectasides, E., Stachler, M. D., Derks, S., Liu, Y., Maron, S., Islam, M., Alpert, L., Kwak, H., Kindler, H., Polite, B., Sharma, M. R., Allen, K., O'Day, E., Lomnicki, S., Maranto, M., Kanteti, R., Fitzpatrick, C., Weber, C., Setia, N., Xiao, S.-Y., Hart, J., Nagy, R. J., Kim, K.-M., Choi, M.-G., Min, B.-H., Nason, K. S., O\textquoterightKeefe, L., Watanabe, M., Baba, H., Lanman, R., Agoston, A. T., Oh, D. J., Dunford, A., Thorner, A. R., Ducar, M. D., Wollison, B. M., Coleman, H. A., Ji, Y., Posner, M. C., Roggin, K., Turaga, K., Chang, P., Hogarth, K., Siddiqui, U., Gelrud, A., Ha, G., Freeman, S. S., Rhoades, J., Reed, S., Gydush, G., Rotem, D., Davison, J., Imamura, Y., Adalsteinsson, V., Lee, J., Bass, A. J., & Catenacci, V. D. (2018). Genomic heterogeneity as a barrier to precision medicine in gastroesophageal adenocarcinoma. *Cancer Discovery*, 8(1), 37–48.

[203] Perelson, A. S. & Ribeiro, R. M. (2004). Hepatitis B Virus Kinetics and Mathematical Modeling. *Seminars in Liver Disease*, 24(S 1), 11–16.

[204] Persaud, D., Gay, H., Ziemniak, C., Chen, Y. H., Piatak, M., Chun, T.-W., Strain, M., Richman, D., & Luzuriaga, K. (2013). Absence of Detectable HIV-1 Viremia after Treatment Cessation in an Infant. *New England Journal of Medicine*, 369(19), 1828–1835.

[205] Phillips, A. N. (1996). Reduction of HIV concentration during acute infection: independence from a specific immune response. *Science*, 271(5248), 497.

[206] Pinkevych, M., Cromer, D., Tolstrup, M., Grimm, A. J., Cooper, D. A., Lewin, S. R., Søgaard, O. S., Rasmussen, T. A., Kent, S. J., Kelleher, A. D., & Davenport, M. P. (2015). HIV Reactivation from Latency after Treatment Interruption Occurs on Average Every 5-8 Days—Implications for HIV Remission. *PLoS Pathog*, 11(7), e1005000.

[207] Popic, V., Salari, R., Hajirasouliha, I., Kashef-Haghighi, D., West, R. B., & Batzoglou, S. (2015). Fast and scalable inference of multi-sample cancer lineages. *Genome biology*, 16(1), 91.

[208] Potten, C. S., Booth, C., & Hargreaves, D. (2003). The small intestine as a model for evaluating adult tissue stem cell drug targets1. *Cell Proliferation*, 36(3), 115–129.

[209] Prague, M., Commenges, D., Drylewicz, J., & Thiébaut, R. (2012). Treatment Monitoring of HIV-Infected Patients based on Mechanistic Models. *Biometrics*, 68(3), 902–911.

[210] Proost, J. H. (2017). Combined proportional and additive residual error models in population pharmacokinetic modelling. *European Journal of Pharmaceutical Sciences*, 109, S78–S82.

[211] Rajasagi, M., Shukla, S. A., Fritsch, E. F., Keskin, D. B., DeLuca, D., Carmona, E., Zhang, W., Sougnez, C., Cibulskis, K., Sidney, J., Stevenson, K., Ritz, J., Neuberg, D., Brusic, V., Gabriel, S., Lander, E. S., Getz, G., Hacohen, N., & Wu, C. J. (2014). Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood*, 124(3), 453–462.

[212] Ramratnam, B., Bonhoeffer, S., Binley, J., Hurley, A., Zhang, L., Mittler, J. E., Markowitz, M., Moore, J. P., Perelson, A. S., & Ho, D. D. (1999). Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. *The Lancet*, 354(9192), 1782–1785.

[213] Raphael, B. J., Hruban, R. H., Aguirre, A. J., Moffitt, R. A., Yeh, J. J., Stewart, C., Robertson, A. G., Cherniack, A. D., Gupta, M., Getz, G., Gabriel, S. B., Meyerson, M., Cibulskis, C., Fei, S. S., Hinoue, T., Shen, H., Laird, P. W., Ling, S., Lu, Y., Mills, G. B., Akbani, R., Loher, P., Londin, E. R., Rigoutsos, I., Telonis, A. G., Gibb, E. A., Goldenberg, A., Mezlini, A. M., Hoadley, K. A., Collisson, E., Lander, E., Murray, B. A., Hess, J., Rosenberg, M., Bergelson, L., Zhang, H., Cho, J., Tiao, G., Kim, J., Livitz, D., Leshchiner, I., Reardon, B., Van Allen, E., Kamburov, A., Beroukhim, R., Saksena, G., Schumacher, S. E., Noble, M. S., Heiman, D. I., Gehlenborg, N., Kim, J., Lawrence, M. S., Adsay, V., Petersen, G., Klimstra, D., Bardeesy, N., Leiserson, M. D., Bowlby, R., Kasaian, K., Birol, I., Mungall, K. L., Sadeghi, S., Weinstein, J. N., Spellman, P. T., Liu, Y., Amundadottir, L. T., Tepper, J., Singhi, A. D., Dhir, R., Paul, D., Smyrk, T., Zhang, L., Kim, P., Bowen, J., Frick, J., Gastier-Foster, J. M., Gerken, M., Lau, K., Leraas, K. M., Lichtenberg, T. M., Ramirez, N. C., Renkel, J., Sherman, M., Wise, L., Yena, P., Zmuda, E., Shih, J., Ally, A., Balasundaram, M., Carlsen, R., Chu, A., Chuah, E., Clarke, A., Dhalla, N., Holt, R. A., Jones, S. J., Lee, D., Ma, Y., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Tse, K., Wong, T., Brooks, D., Auman, J. T., Balu, S., Bodenheimer, T., Hayes, D. N., Hoyle, A. P., Jefferys, S. R., Jones, C. D., Meng, S., Mieczkowski, P. A., Mose, L. E., Perou, C. M., Perou, A. H., Roach, J., Shi, Y., Simons, V. J., Skelly, T., Soloway, M. G., Tan, D., Veluvolu, U., Parker, J. S., Wilkerson, M. D., Korkut, A., Senbabaoglu, Y., Burch, P., McWilliams, R., Chaffee, K., Oberg, A., Zhang, W., Gingras, M.-C., Wheeler, D. A., Xi, L., Albert, M., Bartlett, J., Sekhon, H., Stephen, Y., Howard, Z., Judy, M., Breggia, A., Shroff, R. T., Chudamani, S., Liu, J., Lolla, L., Naresh, R., Pihl, T., Sun, Q., Wan, Y., Wu, Y., Jennifer, S., Roggin, K., Becker, K.-F., Behera, M., Bennett, J., Boice, L., Burks, E., Carlotti Junior, C. G., Chabot, J., Pretti da Cunha Tirapelli, D., Sebastião dos Santos, J., Dubina, M., Eschbacher, J., Huang, M., Huelsenbeck-Dill, L., Jenkins, R., Karpov, A., Kemp, R., Lyadov, V., Maithel, S., Manikhas, G., Montgomery, E., Noushmehr, H., Osunkoya, A., Owonikoko, T., Paklina, O., Potapova, O., Ramalingam, S., Rathmell, W. K., Rieger-Christ, K., Saller, C., Setdikova, G., Shabunin, A., Sica, G., Su, T., Sullivan, T., Swanson, P., Tarvin, K., Tavobilov, M., Thorne, L. B., Urbanski, S., Voronina, O., Wang, T., Crain, D., Curley, E., Gardner, J., Mallery, D., Morris, S., Paulauskis, J., Penny, R., Shelton, C., Shelton, T., Janssen, K.-P., Bathe, O., Bahary, N., Slotta-Huspenina, J., Johns, A., Hibshoosh,

H., Hwang, R. F., Sepulveda, A., Radenbaugh, A., Baylin, S. B., Berrios, M., Bootwalla, M. S., Holbrook, A., Lai, P. H., Maglinte, D. T., Mahurkar, S., Triche, T. J., Van Den Berg, D. J., Weisenberger, D. J., Chin, L., Kucherlapati, R., Kucherlapati, M., Pantazi, A., Park, P., Saksena, G., Voet, D., Lin, P., Frazer, S., Defreitas, T., Meier, S., Chin, L., Kwon, S. Y., Kim, Y. H., Park, S.-J., Han, S.-S., Kim, S. H., Kim, H., Furth, E., Tempero, M., Sander, C., Biankin, A., Chang, D., Bailey, P., Gill, A., Kench, J., Grimmond, S., Johns, A., Cancer Genome Initiative (APGI, A. P., Postier, R., Zuna, R., Sicotte, H., Demchok, J. A., Ferguson, M. L., Hutter, C. M., Mills Shaw, K. R., Sheth, M., Sofia, H. J., Tarnuzzer, R., Wang, Z., Yang, L., Zhang, J. J., Felau, I., & Zenklusen, J. C. (2017). Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell*, 32(2), 185–203.e13. 091220173-s.pdf.

[214] Rausch, T., Zichner, T., Schlattl, A., Stutz, A. M., Benes, V., & Korbel, J. O. (2012). Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18), i333–i339.

[215] Reiter, J. G., Bozic, I., Allen, B., Chatterjee, K., & Nowak, M. A. (2013a). The effect of one additional driver mutation on tumor progression. *Evolutionary Applications*, 6(1), 34–45.

[216] Reiter, J. G., Bozic, I., Chatterjee, K., & Nowak, M. A. (2013b). Ttp: Tool for tumor progression. volume 8044 (pp. 101–106).: Springer Berlin Heidelberg.

[217] Reiter, J. G. & Iacobuzio-Donahue, C. A. (2016). Pancreatic cancer: Pancreatic carcinogenesis — several small steps or one giant leap? *Nature Reviews Gastroenterology and Hepatology*, 14(1), 7–8.

[218] Reiter, J. G., Makohon-Moore, A. P., Gerold, J. M., Bozic, I., Chatterjee, K., Iacobuzio-Donahue, C. A., Vogelstein, B., & Nowak, M. A. (2017). Reconstructing metastatic seeding patterns of human cancers. *Nature Communications*, 8, 14114.

[219] Reiter, J. G., Makohon-Moore, A. P., Gerold, J. M., Heyde, A., Attiyeh, M. A., Kohutek, Z. A., Tokheim, C. J., Brown, A., DeBlasio, R. M., Niyazov, J., Zucker, A., Karchin, R., Kinzler, K. W., Iacobuzio-Donahue, C. A., Vogelstein, B., & Nowak, M. A. (2018). Minimal functional driver gene heterogeneity among untreated metastases. *Science*, 361(6406), 1033–1037.

[220] Ribas A, Hamid O, Daud A, & et al (2016). Association of pembrolizumab with tumor response and survival among patients with advanced melanoma. *JAMA*, 315(15), 1600–1609.

[221] Rizvi, N. A., Hellmann, M. D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J. J., Lee, W., Yuan, J., Wong, P., Ho, T. S., Miller, M. L., Rekhtman, N., Moreira, A. L., Ibrahim, F., Bruggeman, C., Gasmi, B., Zappasodi, R., Maeda, Y., Sander, C., Garon, E. B., Merghoub, T., Wolchok, J. D., Schumacher, T. N., & Chan, T. A. (2015). Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science*, 348(6230), 124–128.

[222] Robert, C. & Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

[223] Roberts, N. J., Norris, A. L., Petersen, G. M., Bondy, M. L., Brand, R., Gallinger, S., Kurtz, R. C., Olson, S. H., Rustgi, A. K., Schwartz, A. G., Stoffel, E., Syngal, S., Zogopoulos, G., Ali, S. Z., Axilbund, J., Chaffee, K. G., Chen, Y.-C., Cote, M. L., Childs, E. J., Douville, C., Goes, F. S., Herman, J. M., Iacobuzio-Donahue, C., Kramer, M., Makohon-Moore, A., McCombie, R. W., McMahon, K. W., Niknafs, N., Parla, J., Pirooznia, M., Potash, J. B., Rhim, A. D., Smith, A. L., Wang, Y., Wolfgang, C. L., Wood, L. D., Zandi, P. P., Goggins, M., Karchin, R., Eshleman, J. R., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., Hruban, R. H., & Klein, A. P. (2016). Whole genome sequencing defines the genetic heterogeneity of familial pancreatic cancer. *Cancer Discovery*, 6(2), 166–175.

[224] Rodriguez-Brenes, I. A., Komarova, N. L., & Wodarz, D. (2013). Tumor growth dynamics: insights into evolutionary processes. *Trends in Ecology & Evolution*, 28(10), 597–604.

[225] Rodriguez-Brenes, I. A., Wodarz, D., & Komarova, N. L. (2015). Characterizing inhibited tumor growth in stem-cell-driven non-spatial cancers. *Mathematical Biosciences*, 270, 135–141.

[226] Roerink, S. F., Sasaki, N., Lee-Six, H., Young, M. D., Alexandrov, L. B., Behjati, S., Mitchell, T. J., Grossmann, S., Lightfoot, H., Egan, D. A., Pronk, A., Smakman, N., van Gorp, J., Anderson, E., Gamble, S. J., Alder, C., van de Wetering, M., Campbell, P. J., Stratton, M. R., & Clevers, H. (2018). Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature*, 556(7702), 457–462.

[227] Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., & Hacohen, N. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell*, 160(1-2), 48–61.

[228] Rosenbloom, D. I. S., Hill, A. L., Rabi, S. A., Siliciano, R. F., & Nowak, M. A. (2012). Antiretroviral dynamics determines HIV evolution and predicts therapy outcome. *Nature Medicine*, 18(9), 1378–1385.

[229] Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., & Shah, S. P. (2014). Pyclone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11, 396–401.

[230] Rouzioux, C., Hocqueloux, L., & Sáez-Cirión, A. (2015). Posttreatment controllers: what do they tell us? *Current Opinion in HIV and AIDS*, 10(1), 29–34.

[231] Salari, R., Saleh, S. S., Kashef-Haghighi, D., Khavari, D., Newburger, D. E., West, R. B., Sidow, A., & Batzoglou, S. (2013). Inference of tumor phylogenies with improved somatic mutation discovery. *Journal of Computational Biology*, 20(11), 933–944.

[232] Samson, A., Lavielle, M., & Mentré, F. (2006). Extension of the saem algorithm to left-censored data in nonlinear mixed-effects model: Application to hiv dynamics model. *Computational Statistics & Data Analysis*, 51(3), 1562–1574.

[233] Samstein, R. M., Lee, C.-H., Shoushtari, A. N., Hellmann, M. D., Shen, R., Janjigian, Y. Y., Barron, D. A., Zehir, A., Jordan, E. J., Omuro, A., Kaley, T. J., Kendall, S. M., Motzer, R. J., Hakimi, A. A., Voss, M. H., Russo, P., Rosenberg, J., Iyer, G., Bochner, B. H., Bajorin, D. F., Al-Ahmadie, H. A., Chaft, J. E., Rudin, C. M., Riely, G. J., Baxi, S., Ho, A. L., Wong, R. J., Pfister, D. G., Wolchok, J. D., Barker, C. A., Gutin, P. H., Brennan, C. W., Tabar, V., Mellinghoff, I. K., DeAngelis, L. M., Ariyan, C. E., Lee, N., Tap, W. D., Gounder, M. M., D'Angelo, S. P., Saltz, L., Stadler, Z. K., Scher, H. I., Baselga, J., Razavi, P., Klebanoff, C. A., Yaeger, R., Segal, N. H., Ku, G. Y., DeMatteo, R. P., Ladanyi, M., Rizvi, N. A., Berger, M. F., Riaz, N., Solit, D. B., Chan, T. A., & Morris, L. G. T. (2019). Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature Genetics*, (pp.1).

[234] Sanborn, J. Z., Chung, J., Purdom, E., Wang, N. J., Kakavand, H., Wilmott, J. S., Butler, T., Thompson, J. F., Mann, G. J., Haydu, L. E., Saw, R. P. M., Busam, K. J., Lo, R. S., Collisson, E. A., Hur, J. S., Spellman, P. T., Cleaver, J. E., Gray, J. W., Huh, N., Murali, R., Scolyer, R. A., Bastian, B. C., & Cho, R. J. (2015). Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. *Proc Natl Acad Sci USA*, 112(35), 10995–11000.

[235] Sanli, O., Dobruch, J., Knowles, M. A., Burger, M., Alemozaffar, M., Nielsen, M. E., & Lotan, Y. (2017). Bladder cancer. *Nature Reviews Disease Primers*, 3, 17022.

[236] SAS, L. (2018). Monolix.

[237] Sato-Kaneko, F., Yao, S., Ahmadi, A., Zhang, S. S., Hosoya, T., Kaneda, M. M., Varner, J. A., Pu, M., Messer, K. S., Guiducci, C., Coffman, R. L., Kitaura, K., Matsutani, T., Suzuki, R., Carson, D. A., Hayashi, T., & Cohen, E. E. W. (2017). Combination immunotherapy with TLR agonists and checkpoint inhibitors suppresses head and neck cancer. *JCI Insight*, 2(18).

[238] Schiffer, J. T., Swan, D. A., Magaret, A., Corey, L., Wald, A., Ossig, J., Ruebsamen-Schaeff, H., Stoelben, S., Timmler, B., Zimmermann, H., Melhem, M. R., Van Wart, S. A., Rubino, C. M., & Birkmann, A. (2016). Mathematical modeling of herpes simplex virus-2 suppression with pritelivir predicts trial outcomes. *Science Translational Medicine*, 8(324), 324ra15.

[239] Schliep, K. P. (2011). phangorn: Phylogenetic analysis in R. *Bioinformatics*, 27(4), 592–593.

[240] Schuh, A., Becq, J., Humphray, S., Alexa, A., Burns, A., Clifford, R., Feller, S. M., Grocock, R., Henderson, S., Khrebtukova, I., Kingsbury, Z., Luo, S., McBride, D., Murray, L., Menju, T., Timbs, A., Ross, M., Taylor, J., & Bentley, D. (2012). Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*, 120(20), 4191–4196.

[241] Shen, R. & Seshan, V. E. (2016). Facets: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput dna sequencing. *Nucleic Acids Research*, 44(16), e131–e131.

[242] Shihab, H. A., Gough, J., Cooper, D. N., Day, I. N. M., & Gaunt, T. R. (2013). Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics*, 29(12), 1504–1510.

[243] Siliciano, J. D. & Siliciano, R. F. (2006). The latent reservoir for HIV-1 in resting CD4+ T cells: a barrier to cure. *Current Opinion in HIV and AIDS*, 1(2), 121–128.

[244] Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., Marjoram, P., Siegmund, K., Press, M. F., Shibata, D., & Curtis, C. (2015a). A big bang model of human colorectal tumor growth. *Nature Genetics*, 47(3), 209–216.

[245] Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., Marjoram, P., Siegmund, K., Press, M. F., Shibata, D., & Curtis, C. (2015b). A Big Bang model of human colorectal tumor growth. *Nature Genetics*, 47(3), 209–216.

[246] Stadler, T., Vaughan, T. G., Gavryushkin, A., Guindon, S., KÃŒhnert, D., Leventhal, G. E., & Drummond, A. J. (2015). How well can the exponential-growth coalescent approximate constant-rate birth-death population dynamics? *Proceedings of the Royal Society B: Biological Sciences*, 282(1806).

[247] Stafford, M. A., Corey, L., Cao, Y., Daar, E. S., Ho, D. D., & Perelson, A. S. (2000). Modeling Plasma Virus Concentration during Primary HIV Infection. *Journal of Theoretical Biology*, 203(3), 285–301.

[248] Stevceva, L. (2011). Toll-like receptor agonists as adjuvants for HIV vaccines. *Current Medicinal Chemistry*, 18(33), 5079–5082.

[249] Stockenstrom, S. v., Odevall, L., Lee, E., Sinclair, E., Bacchetti, P., Killian, M., Epling, L., Shao, W., Hoh, R., Ho, T., Faria, N. R., Lemey, P., Albert, J., Hunt, P., Loeb, L., Pilcher, C., Poole, L., Hatano, H., Somsouk, M., Douek, D., Boritz, E., Deeks, S. G., Hecht, F. M., & Palmer, S. (2015). Longitudinal Genetic Characterization Reveals That Cell Proliferation Maintains a Persistent HIV Type 1 DNA Pool During Effective HIV Therapy. *Journal of Infectious Diseases*, 212(4), 596–607.

[250] Strino, F., Parisi, F., Micsinai, M., & Kluger, Y. (2013). Trap: a tree approach for fingerprinting subclonal tumor composition. *Nucleic acids research*, 41(17), e165.

[251] Talmadge, J. E. & Fidler, I. J. (2010). The biology of cancer metastasis: Historical perspective. *Cancer Research*, 70(14), 5649–5669.

[252] The Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407), 330–337.

[253] Thiébaut, R., Drylewicz, J., Prague, M., Lacabaratz, C., Beq, S., Jarne, A., Croughs, T., Sekaly, R.-P., Lederman, M. M., Sereti, I., Commenges, D., & Lévy, Y. (2014). Quantifying and Predicting the Effect of Exogenous Interleukin-7 on CD4+T Cells in HIV-1 Infection. *PLoS computational biology*, 10(5), e1003630.

[254] Tokheim, C. & Karchin, R. (2018). Enhanced context reveals the scope of somatic missense mutations driving human cancers. *bioRxiv*, (pp. 313296).

[255] Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., & Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences of the United States of America*, 113(50), 14330–14335.

[256] Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G., & Vogelstein, B. (2015). Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences*, 112(1), 118–123.

[257] Tomasetti, C., Vogelstein, B., & Parmigiani, G. (2013). Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proceedings of the National Academy of Sciences of the United States of America*, 110(6), 1999–2004.

[258] Turajlic, S., McGranahan, N., & Swanton, C. (2015). Inferring mutational timing and reconstructing tumour evolutionary histories. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1855(2), 264–275.

[259] Turajlic, S. & Swanton, C. (2016). Metastasis as an evolutionary process. *Science*, 352(6282), 169–175.

[260] UNAIDS (2017). *Fact sheet - Latest statistics on the status of the AIDS epidemic*. Technical report.

[261] Urosevic, J., Garcia-Albéniz, X., Planet, E., Real, S., Céspedes, M. V., Guiu, M., Fernandez, E., Bellmunt, A., Gawrzak, S., Pavlovic, M., Mangues, R., Dolado, I., Barriga, F. M., Nadal, C., Kemeny, N., Batlle, E., Nebreda, A. R., & Gomis, R. R. (2014). Colon cancer cells colonize the lung from established liver metastases through p38 mapk signalling and pthlh. *Nature Cell Biology*, 16(7), 685–694.

[262] van Heek, N. T., Meeker, A. K., Kern, S. E., Yeo, C. J., Lillemoe, K. D., Cameron, J. L., Offerhaus, G. J. A., Hicks, J. L., Wilentz, R. E., Goggins, M. G., De Marzo, A. M., Hruban, R. H., & Maitra, A. (2002). Telomere shortening is nearly universal in pancreatic intraepithelial neoplasia. *The American journal of pathology*, 161(5), 1541–7.

[263] Vogelstein, B., Fearon, E. R., Hamilton, S. R., Kern, S. E., Preisinger, A. C., Leppert, M., Smits, A. M. M., & Bos, J. L. (1988). Genetic alterations during colorectal-tumor development. *New England Journal of Medicine*, 319(9), 525–532.

[264] Vogelstein, B. & Kinzler, K. W. (2015). The path to cancer –three strikes and you're out. *The New England journal of medicine*, 373(20), 1895–8.

[265] Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013a). Cancer genome landscapes. *Science (New York, N.Y.)*, 339(6127), 1546–58. 012720173.pdf 012720173notes.pdf.

[266] Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013b). Cancer Genome Landscapes. *Science*, 339(6127), 1546–1558.

[267] Waclaw, B., Bozic, I., Pittman, M. E., Hruban, R. H., Vogelstein, B., & Nowak, M. A. (2015). A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature*, 525(7568), 261–264.

[268] Waddell, N., Pajic, M., Patch, A.-M., Chang, D. K., Kassahn, K. S., Bailey, P., Johns, A. L., Miller, D., Nones, K., Quek, K., Quinn, M. C. J., Robertson, A. J., Fadlullah, M. Z. H., Bruxner, T. J. C., Christ, A. N., Harliwong, I., Idrisoglu, S., Manning, S., Nourse, C., Nourbakhsh, E., Wani, S., Wilson, P. J., Markham, E., Cloonan, N., Anderson, M. J., Fink, J. L., Holmes, O., Kazakoff, S. H., Leonard, C., Newell, F., Poudel, B., Song, S., Taylor, D., Waddell, N., Wood, S., Xu, Q., Wu, J., Pinese, M., Cowley, M. J., Lee, H. C., Jones, M. D., Nagrial, A. M., Humphris, J., Chantrill, L. A., Chin, V., Steinmann, A. M., Mawson, A., Humphrey, E. S., Colvin, E. K., Chou, A., Scarlett, C. J., Pinho, V. A., Giry-Laterriere, M., Rooman, I., Samra, J. S., Kench, J. G., Pettitt, J. A., Merrett, N. D., Toon, C., Epari, K., Nguyen, N. Q., Barbour, A., Zeps, N., Jamieson, N. B., Graham, J. S., Niclou, S. P., Bjerkvig, R., Grützmann, R., Aust, D., Hruban, R. H., Maitra, A., Iacobuzio-Donahue, C. A., Wolfgang, C. L., Morgan, R. A., Lawlor, R. T., Corbo, V., Bassi, C., Falconi, M., Zamboni, G., Tortora, G., Tempero, M. A., Biankin, V. A., Brancato, M.-A. L., Rowe, S. J., Simpson, S. H., Martyn-Smith, M., Thomas, M. T., Chin, V. T., Humphris, J. L., Scott Mead, R., Pettit, J., Tao, J., DiPietro, R., Watson, C., Steinmann, A., Ching Lee, H., Wong, R., Daly, R. J., Musgrove, E. A., Sutherland, R. L., Grimmond, S. M., Miller, D. K., Gongora, M., Anderson, M., Xu, C., Lynn Fink, J., Christ, A., Bruxner, T., Pearson, V. J., Quinn, M., Nagaraj, S., Kazakoff, S., Krisnan, K., Wood, D., Gill, A. J., Pavlakis, N., Guminski, A., Asghari, R., Pavey, D., Das, A., Cosman, P. H., Ismail, K., O'Connnor, C., Lam Duncan McLeod, V. W., Pleass, H. C., Richardson, A., James, V., Cooper, C. L., Joseph, D., Sandroussi, C., Crawford, M., Gallagher, J., Texler, M., Forest, C., Laycock, A., Epari, K. P., Ballal, M., Fletcher, D. R., Mukhedkar, S., Spry, N. A., DeBoer, B., Chai, M., Beilin, M., Feeney, K., Nguyen, N. Q., Ruszkiewicz, A. R., Worthley, C., Tan, C. P., Debrencini, T., Chen, J., Brooke-Smith, M. E., Papangelis, V., Tang, H., Barbour, A. P., Clouston, A. D., Martin, P., O'Rourke, T. J., Chiang, A., Fawcett, J. W., Slater, K., Yeung, S., Hatzifotis, M., Hodgkinson, P., Christophi, C., Nikfarjam, M., Mountain, A., Eshleman, J. R., Schulick, R. D., Morgan, R. A., Hodgin, M., Scarpa, A., Beghelli, S., Scardoni, M., Oien, K., Hair, J., & Pilarsky, C. (2015). Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*, 518(7540), 495–501.

[269] Walker, L. M., Huber, M., Doores, K. J., Falkowska, E., Pejchal, R., Julien, J.-P., Wang, S.-K., Ramos, A., Chan-Hui, P.-Y., Moyle, M., Mitcham, J. L., Hammond, P. W., Olsen, O. A., Phung, P., Fling, S., Wong, C.-H., Phogat, S., Wrin, T., Simek, M. D., Principal Investigators, P. G., Koff, W. C., Wilson, I. A., Burton, D. R., & Poignard, P. (2011). Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature*, 477(7365), 466–470.

[270] Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., Lifson, J. D., Bonhoeffer, S., Nowak, M. A., & Hahn, B. H. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, 373(6510), 117–122.

[271] Whitney, J. B., Hill, A. L., Sanisetty, S., Penaloza-MacMaster, P., Liu, J., Shetty, M., Parenteau, L., Cabral, C., Shields, J., Blackmore, S., Smith, J. Y., Brinkman, A. L., Peter, L. E., Mathew, S. I., Smith, K. M., Borducchi, E. N., Rosenbloom, D. I. S., Lewis, M. G., Hattersley, J., Li, B., Hesselgesser, J., Geleziunas, R., Robb, M. L., Kim, J. H., Michael, N. L., & Barouch, D. H. (2014). Rapid seeding of the viral reservoir prior to SIV viraemia in rhesus monkeys. *Nature*, 512(7512), 74–77.

[272] Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A., & Sottoriva, A. (2016). Identification of neutral tumor evolution across cancer types. *Nature Genetics*.

[273] Williams, M. J., Werner, B., Heide, T., Curtis, C., Barnes, C. P., Sottoriva, A., & Graham, T. A. (2018). Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics*, 50, 895–903.

[274] Witkiewicz, A. K., McMillan, E. A., Balaji, U., Baek, G., Lin, W.-C., Mansour, J., Mollaee, M., Wagner, K.-U., Koduru, P., Yopp, A., Choti, M. A., Yeo, C. J., McCue, P., White, M. A., & Knudsen, E. S. (2015). Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nature communications*, 6, 6744.

[275] Wodarz, D. & Komarova, N. L. (2005). *Computational Biology of Cancer: Lecture Notes and Mathematical Modeling*. World Scientific Pub Co Inc.

[276] Wodarz, D. & Komarova, N. L. (2014). *Dynamics of Cancer: Mathematical Foundations of Oncology*. Hackensack New Jersey: WSPC, 1 edition edition.

[277] Wu, H., Huang, Y., Acosta, E. P., Rosenkranz, S. L., Kuritzkes, D. R., Eron, J. J., Perelson, A. S., & Gerber, J. G. (2005). Modeling long-term HIV dynamics and antiretroviral response: effects of drug potency, pharmacokinetics, adherence, and drug resistance. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 39(3), 272–283.

[278] Wu, X. & Kimmel, M. (2013). Modeling Neutral Evolution Using an Infinite-Allele Markov Branching Process.

[279] Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R. H., Eshleman, J. R., Nowak, M. A., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., & Iacobuzio-Donahue, C. A. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, 467(7319), 1114–1117.

[280] Yamasaki, S., Suda, K., Nobukawa, B., & Sonoue, H. (2002). Intraductal spread of pancreatic cancer. clinicopathologic study of 54 pancreatectomized patients. *Pancreatology : official journal of the International Association of Pancreatology (IAP) ... [et al.]*, 2(4), 407–12. 12132017i-s.pdf.

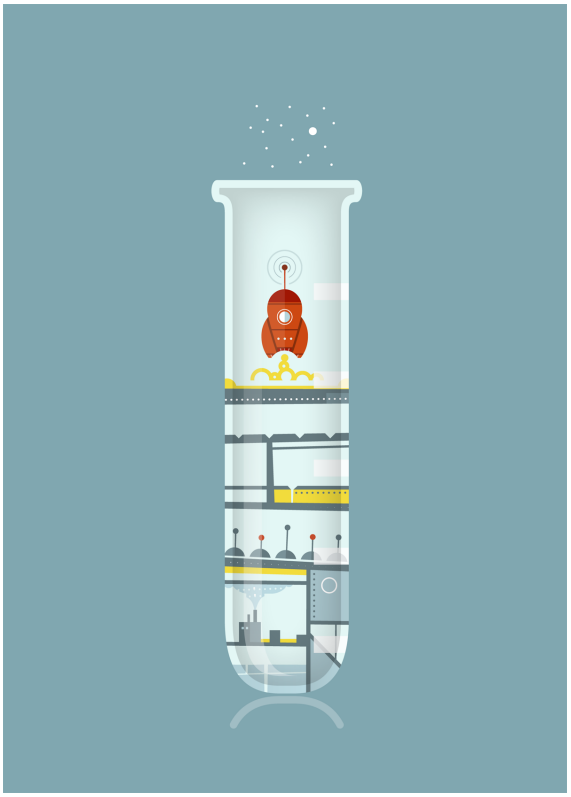
[281] Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L. B., Larsimont, D., Davies, H., Li, Y., Ju, Y. S., Ramakrishna, M., Haugland, H. K., Lilleng, P. K., Nik-Zainal, S., McLaren, S., Butler, A., Martin, S., Glodzik, D., Menzies, A., Raine, K., Hinton, J., Jones, D., Mudie, L. J., Jiang, B., Vincent, D., Greene-Colozzi, A., Adnet, P.-Y., Fatima, A., Maetens, M., Ignatiadis, M., Stratton, M. R., Sotiriou, C., Richardson, A. L., Lonning, P. E., Wedge, D. C., & Campbell, P. J. (2015). Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med*, 21(7), 751–759.

[282] Yuan, K., Sakoparnig, T., Markowetz, F., & Beerenwinkel, N. (2015). Bitphylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome biology*, 16(1), 36.

[283] Yukl, S., Boritz, E., Busch, M., Bentsen, C., Chun, T., Douek, D., Eisele, E., Haase, A., Ho, Y., Hütter, G., et al. (2013). Challenges in detecting hiv persistence during potentially curative interventions: a study of the Berlin patient. *PLoS Pathogens*, 9(5), e1003347.

[284] Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., Lawrence, M. S., Zhang, C.-Z., Wala, J., Mermel, C. H., Sougnez, C., Gabriel, S. B., Hernandez, B., Shen, H., Laird, P. W., Getz, G., Meyerson, M., & Beroukhim, R. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, 45(10), 1134–1140.

[285] Zhang, J., Fujimoto, J., Zhang, J., Wedge, D. C., Song, X., Zhang, J., Seth, S., Chow, C.-W., Cao, Y., Gumbs, C., Gold, K. A., Kalhor, N., Little, L., Mahadeshwar, H., Moran, C., Protopopov, A., Sun, H., Tang, J., Wu, X., Ye, Y., William, W. N., Lee, J. J., Heymach, V. J., Hong, W. K., Swisher, S., Wistuba, I. I., & Futreal, P. A. (2014). Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science*, 346(6206), 256–259.



This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, "Science Experiment 02", was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.