Dissertation
submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the Ruperto Carola University Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Presented by
Verena Körber, M.Sc.
born in Schramberg, Germany
Oral examination: .....

Somatic mutations and tissue turnover in glioblastoma and hematopoiesis

Referees:
Prof. Dr. Thomas Höfer
PD Dr. Frank Westermann

Zusammenfassung

Mutationen entstehen in erster Linie während der Zellteilung, weshalb in somatischen Geweben eine genetische Vielfalt besteht. Das Ausmaß dieser Heterogenität wird von der zellulären Dynamik während Wachstum und Homöostase bestimmt, was sich im Mutationsprofil einer Gewebeprobe widerspiegelt. In der vorliegenden Arbeit werden mathematische Methoden entwickelt, die eine Abschätzung des Wachstumsverhaltens maligner Tumore und physiologischer Gewebe anhand von Ganzgenomsequenzierdaten erlauben.

Im ersten Teil der Arbeit wird die klonale Evolution im primären Glioblastom, einem hochaggressiven Hirntumor, untersucht. Dazu wird zunächst ein Multinomialmodell entwickelt, das anhand von tiefen Ganzgenomsequenzierdaten die phylogenetische Entwicklung eines Gewebes nachvollzieht. Mithilfe dieses Modells wird anschließend die Entstehungsgeschichte adulter Glioblastome aus 21 gepaarten Primär- und Rezidivtumorproben rekonstruiert. Trotz ausgeprägter Heterogenität im Mutationsprofil individueller Tumore werden in allen Tumoren, bis auf einen, frühe Kopienzahlveränderungen auf Chromosom 7 (Gewinn), Chromosom 9p (Verlust) oder Chromosom 10 (Verlust) nachgewiesen. Dementgegen befinden sich Mutationen in der Promoterregion des *TERT*-Gens, das für eine Telomerase-Untereinheit kodiert, häufig auf subklonaler Ebene, was auf einen späteren Entstehungszeitpunkt in der Geschichte des Tumors hindeutet. Unsere Daten legen nahe, dass Tumorrezidive typischerweise aus mehreren Unterpopulationen des Primärtumors auswachsen und kein charakteristisches Mutationsprofil aufweisen, sodass keine Evidenz für ein selektives Auswachsen resistenter Unterpopulationen nach Standardtherapie besteht. Eine populationsdynamische Modellierung des Tumorwachstums erlaubt die Abschätzung des Tumoralters auf mehrere Jahre und erklärt die lange Entwicklungszeit mit einer hohen Zelltodrate, die erst durch Mutationen im *TERT*-Promoter reduziert wird. Die hier gewonnenen Einblicke in die Entwicklungsgeschichte von Glioblastomen können zur Entwicklung neuer Methoden zur Früherkennung eingesetzt werden.

Im zweiten Teil der vorliegenden Arbeit wird die Anhäufung von Mutationen in

normalen Geweben betrachtet. Dazu wird zunächst die bereits bestehende Theorie zur Mutationslast in wachsenden Geweben auf einen zweistufigen Prozess, der sich aus einer embryonalen Wachstumsphase und anschließender Homöostase zusammensetzt, erweitert. Mittels stochastischer Simulationen wird die Aussagekraft des Modells für selbsterneuernde Zellpopulationen, wie besipielsweise Stammzellen, bestätigt. Das aufgrund der Modellvorhersage erwartete Mutationsprofil sich neutral entwickelnder Gewebe wird anschließend anhand von Ganzgenomsequenzierdaten muriner Granulozyten und humaner Leukozyten geprüft und für die Mehrheit der Proben bestätigt. Allerdings wird die unter neutraler Dynamik erwartete Mutationslast in einigen Blutproben deutlich übertroffen. Da diese Fälle zusätzlich eine oder mehrere mit Leukämien assoziierte Mutationen aufweisen, liegt hier die Vermutung einer hämatopoietischen Perturbation oder einer prä-leukämischen Expansion nahe. Die vorgestellte Analyse der Mutationslast unter normaler und gestörter Blutbildung kann zum besseren Verständis der frühen Tumorentstehung beitragen.

Abstract

Somatic mutations accumulate in tissues primarily through cell divisions. This observation opens the opportunity to use somatic mutations as clonal markers for inferring the past dynamics of cell turnover during tissue growth and homeostasis. In this thesis, I develop mathematical approaches to the inference problems that are stimulated by deep genome sequencing data of malignant growth and physiological tissue turnover.

In the first part of this thesis I reconstruct the evolutionary history of adult glioblastoma, a highly aggressive brain cancer, prior to and after standard therapy. To this end, I develop a likelihood-based multinomial model that jointly infers genetic subclones and their phylogenetic relationships from whole genome sequencing data. Applied to 21 sample pairs from primary and recurrent glioblastomas, the model infers a common path of early tumorigenesis characterized by three pervasive copy number changes on chromosome 7 (gain), chromosome 9p (loss) and chromosome 10 (loss). *TERT* promoter mutations are subclonal in one third of the tumor pairs and are thus placed at a later stage of tumorigenesis. Our data indicate that recurrent tumors typically re-grow from multiple subclones of the primary tumor with no evidence for a 'resistance genotype' induced by therapy. Combining the results from phylogenetic inference with population dynamics models of tumor growth, I estimate that glioblastomas originate several years prior to initial diagnosis but reach detectable sizes only after *TERT* promoter mutations stabilized cellular survival. This project provides new insights into the evolutionary history of glioblastoma that may ultimately aid early diagnosis.

In the second part I analyze the mutation frequency distribution in normal tissues. To this end, I extend existing theory on mutation accumulation in exponentially growing tissues to a two-stage situation of initial embryonic expansion and subsequent homeostasis during adulthood. Based on stochastic simulations I show that the theoretical framework recovers the average mutation frequency spectrum in stem cell populations. Whole genome sequencing data from murine granulocytes and human leukocytes from subjects of different ages without

diagnosed leukemia confirm the model prediction in the majority of cases but reveal an unexpectedly high mutational burden in a smaller subgroup. These cases were associated with one or several leukemic driver mutations, suggesting that perturbed hematopoiesis or pre-leukemic expansions caused the deviation of the mutation frequency spectrum from neutrality. The comprehensive analysis of the mutation frequency spectra in normal and perturbed hematopoiesis may aid the understanding of tumor initiation *in vivo*.

Contents

CHAPTER 1

Introduction

Mutations are changes of the genetic code that occur during DNA replication or are introduced by mutagenic processes independent of replication. The former are mostly single base pair substitutions while the latter involve more complex alterations such as base oxidation by reactive oxygen species, DNA crosslinks by UV irradiation or tobacco-induced DNA-adducts (Chatterjee and Walker, 2017; Fujii et al., 1999; Hecht, 2003; Kunkel, 2004). As different mutagenic processes cause distinct patterns of single nucleotide variants in the genome, the spectrum of somatic mutations mirrors a cell's history of divisions and mutagenic exposure (Alexandrov et al., 2013a).

Most mutations are functionally neutral (Kimura, 1991) but in rare cases a gene's expression or the structure of its encoded protein are changed in a functionally relevant way. Pathological mutations in the germline can cause hereditary diseases, often associated with developmental dysfunction, while mutations in somatic cells are the leading cause of cancer (Alberts et al., 2002). Nevertheless, accumulation of mutations per se is not a hallmark of disease but may generate phenotypic variation on which natural selection can act (Fisher, 1999).

Traditionally, the spread of a variant has been predominantly studied on the organismal level and population genetics models have been developed for sexually and asexually reproducing populations (reviewed, e.g. in Nei, 1975; Nowak, 2006). With the advances of next generation sequencing, genome-wide profiling of mutations in tissues or even single cells has become possible, and research now increasingly focuses on evolutionary processes within individual organisms, associated with somatic mutations (Huang et al., 2018; Martincorena and Campbell, 2015; Meyerson et al., 2010). A major interest is to understand the evolutionary dynamics within tumors, thus associating individual mutations with a selective advantage during tumorigenesis or under therapy (Bozic et al., 2016; Martincorena et al., 2018). At the same time, more and more

studies attempt to trace the spread of mutations in normal tissues (Blokzijl et al., 2016; Lee-Six et al., 2018; Martincorena et al., 2015; O'Huallachain et al., 2012).

Intra-organismal 'somatic' evolution is experimentally and theoretically easier to tackle than organismal evolution. In contrast to mating populations, intra-organismal evolution is solely driven by cell division, death and mutation, which simplifies population genetics theory. Moreover, intra-organismal evolution is rebooted at every organismal generation so that multiple outcomes of the same evolutionary 'experiment' can be measured (Sidow and Spies, 2015). In the context of cancer, this means that analyzing the mutational profile of many cancers might reveal pervasive evolutionary pathways during tumorigenesis. Likewise, analysis of many normal tissues is expected to provide a null model of mutation accumulation in the absence of disease. Notably, as mutations can be interpreted as cellular labels, studying the mutational profile of a tissue goes beyond classical attempts to link genotype and phenotype but might reveal the dynamics of tissue growth and homeostasis.

1.1 Outline of this thesis

The two projects presented in this thesis both center around somatic mutagenesis. The first project analyzes evolutionary dynamics in adult glioblastoma, a highly aggressive brain tumor, while the second project focuses on somatic mutagenesis in hematopoiesis.

Chapter 1 provides general background on the mechanisms of mutation and DNA repair and introduces cancer as a mutation-driven disease. More detailed background underlying the individual projects is outlined in Chapter 2 and 3 prior to the respective results sections.

Chapter 2 focuses on clonal evolution in adult glioblastoma. An introductory section summarizes clinical and biological aspects of the disease and reviews computational methods to infer tumor heterogeneity from whole genome sequencing data. Next, a likelihood-based multinomial model for phylogenetic reconstruction in cancer is developed and applied to deep whole genome sequencing data from 21 pairs of primary and recurrent glioblastomas. The inferred evolutionary trajectories are then interpreted with population dynamics models, revealing characteristics of tumor growth and the selective advantage of mutations in the promoter region of the telomerase gene *TERT*. The results are discussed in the context of the current state of research in the concluding section of this chapter.

The dynamics of somatic mutations in normal tissues are analyzed in Chapter 3. Here, the hematopoietic system is chosen as a model system to test the theoretical predictions. The introductory part of this chapter provides biological background on the hematopoietic system and mutagenesis in dynamic tissues. Subsequently, the existing theory on the spread of mutations in growing tissues is introduced and extended to a coupled system of tissue growth and homeostasis. The model predictions are then compared to whole genome sequencing data of murine granulocytes and of blood samples obtained from neuro- and glioblastoma patients as

controls for the genome sequencing of tumor tissue. Potentials and limitations of the presented approach to delineate normal from pre-cancerous situations are discussed in the end of this chapter.

In the final Chapter 4, the results of Chapters 2 and 3 are placed in a larger context and the potential of whole genome sequencing data to probe evolutionary dynamics in somatic tissues is discussed more generally.

Overall, this thesis covers functional and dynamical aspects of mutation accumulation in health and disease. Deep whole genome sequencing data are analyzed with the aid of mathematical models to delineate the growth dynamics of healthy and cancerous tissues. Beyond specific insights into the somatic mutagenesis in glioblastomas and hematopoiesis, this thesis highlights the potential of mutational analysis to address diverse biological questions, ranging from cancer evolution, to tissue heterogeneity, and to early detection of pre-cancerous lesions.

1.2 Mechanisms of mutation & DNA repair

Genetic mutations are changes of the DNA sequence that are introduced by endogenous or exogenous mutagenic processes. Endogenous mutagenic processes comprise replication errors, spontaneous deamination and DNA damage due to reaction with naturally occurring reactants such as reactive oxygen species (ROS). Exogenous mutagenic processes are caused by chemical or physical agents that are not naturally occurring in the cell, but taken in through food consumption, respiration or exposition to irradiation. Mutations vary in complexity, ranging from single nucleotide exchanges to small insertions/deletions to chromosomal rearrangements, aneuploidy and polyploidy (Alberts et al., 2002; Chatterjee and Walker, 2017). In the following, I will introduce the most frequent mutagenic processes along with the DNA repair mechanisms that have evolved in response (c.f. Fig. 1.1).

1.2.1 Replication errors

DNA replication is prerequisite for each somatic cell division. Using one parental strand as a template, DNA polymerases catalyze the copying of the parental DNA by complementary base pairing between the purine bases adenine and guanine, and their pyrimidine counterparts thymine and cytosine. Since the two copies consist of one newly synthesized and one parental DNA strand, replication is semi-conservative (Alberts et al., 2002).

Despite an energetic optimum for complementary base pairing, DNA replication is prone to error, because the difference in free energy between matches and mismatches is low (Loeb and Kunkel, 1982). Moreover, misalignments between the template and the newly synthesized strand introduce small insertions and deletions, especially at replicative sites. However, intrinsic proofreading mechanisms of the DNA polymerase together with a mismatch repair system increase fidelity of DNA replication tremendously, yielding an error rate of approximately one mismatch in 10^8 base pairs only (Kunkel, 2009).

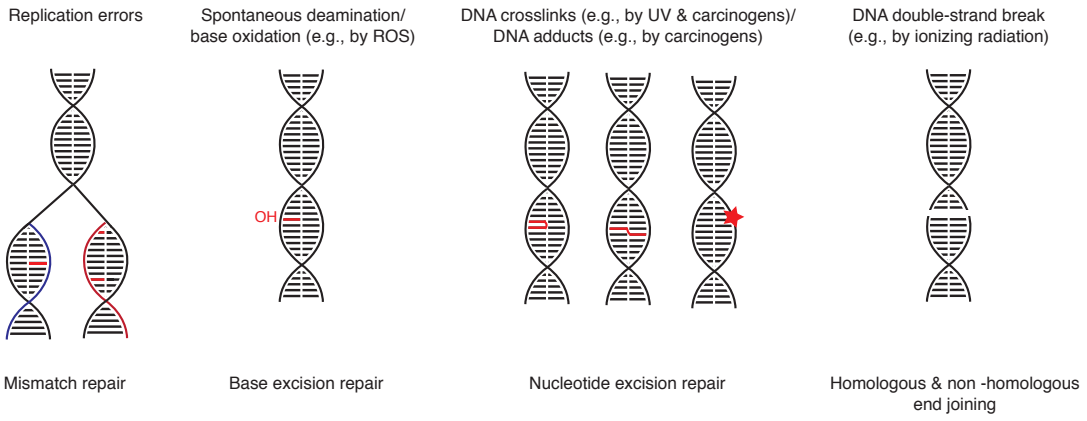


Figure 1.1: Major mechanisms of mutation. Changes of the DNA sequence are highlighted in red; associated major repair pathways are indicated below. Illustration based on Fig. 1 in Weeden and Asselin-Labat, 2018.

1.2.2 Spontaneous deamination

A major source of replication-independent mutation is spontaneous deamination of adenine, guanine, cytosine and 5-methylcytosine to hypoxanthine, xanthine, uracil and thymine, respectively (Kow, 2002). While most deamination products are rapidly corrected by base excision repair, thymine-guanine mismatches are less efficiently repaired, reflected in an overrepresentation of C>T substitutions at methylated cytosine sites (Waters and Swann, 2000).

1.2.3 Mutagen-induced mutations

In addition to replication errors and spontaneous deamination, the human DNA is continuously exposed to mutagenic chemicals and irradition. Among these are naturally occuring substances such as reactive oxygen species (ROS) or extrinsic hazards like UV irradiation or tobacco smoke. Exposition to mutagens causes a wide range of DNA damages. For example, reaction of the DNA with ROS radicals has been linked with base oxidation, DNA intra-strand-crosslinks and DNA-protein-crosslinks (Cadet and Wagner, 2013). Ionizing irradiation is mutagenic on two accounts - indirectly, by promoting the generation of ROS and directly by introducing single- and double-strand DNA breaks (Desouky et al., 2015). By contrast, the principal DNA-damage that is attributed to UV-irradiation, is dimerization of adjacent pyrimidines (McGregor, 1999). Finally, DNA can be damaged by exogenous chemicals, which introduce DNA cross-links or DNA adducts. These chemicals are primarily alkylating agents, present for example in tobacco smoke, fermented food products and chemotherapeutics (Fu et al., 2012).

1.2.4 DNA repair

In response to the many sources of mutation a broad variety of cellular repair systems has evolved. Depending on the type of lesion, the damage is either directly corrected, or a DNA damage response is initiated by one of the cell cycle checkpoints. If DNA repair is not possible, the cell either dies by apoptosis or incorporates the mutation (Jackson and Bartek, 2009).

Modified DNA bases, generated for example by spontaneous deamination, are directly corrected by the 'base excision repair' (BER) pathway, which operates during the entire cell cycle (Mjelle et al., 2015). In BER, glycosylases remove the faulty base and endonucleases cut the DNA backbone causing an intermediate single strand break. Finally, DNA polymerases resynthesize the strand and DNA ligases seal the break (Krokan and Bjørås, 2013).

Base mismatches arising from replication errors or spontaneous deamination of methylcytosine to thymine are subjected to the mismatch repair pathway (MMR), which is most active during S-phase (Schroering et al., 2007). In MMR, the erroneous strand is identified and up to several hundreds of nucleotides around the mismatch are excised by exonucleases, before the lesion is resynthesized by DNA polymerases (Kunkel and Erie, 2015).

Bulky DNA lesions, introduced by UV irradiation or alkylating agents, are removed by the nucleotide excision repair pathway (NER). NER is active during the entire cell cycle, but the final steps depend on enzymes active in S-phase only (Mjelle et al., 2015). NER works on a global level but in addition can be activated during transcription by a stalled RNA polymerase. Upon damage recognition, the vicinity of the damage is opened by DNA helicases and the damaged site is removed by endonucleases from both sides. DNA polymerases ultimately resynthesize the strand and ligases connect the 3' end to the remaining DNA (Schärer, 2013).

Finally, double-strand breaks are repaired by homologous or non-homologous end joining (abbreviated as HEJ and NHEJ, respectively). While NHEJ is active during interphase, HEJ operates in G2- and S-phase (Hustedt and Durocher, 2017). NHEJ recruits a DNA repair complex to the double-strand break, which prepares the breakage for ligation (Davis and Chen, 2013). By contrast, HEJ corrects the breakage by copying the missing sequence from the sister chromosome, which may result in loss of heterozygosity due to crossing-over (Li and Heyer, 2008).

Despite the many ways of DNA repair and DNA damage response, some mutations will remain undetected or unrepaired and are stably incorporated in the genome. In the majority of cases this will not affect cellular survival but in rare cases, the cell will be malignantly transformed and cancer will be initiated.

1.3 Cancer as a mutation-driven disease

The many mutagenic processes that confront the DNA on a daily basis leave their mark on the 10^{13} cells in the human body (Savage, 1977). As cancer develops from mutations in proto-oncogenes or tumor suppressor genes, mutagenesis is a constant threat to every multicellular

organism. Nevertheless, most mutations are phenotypically neutral. Since only 1% of the human genome is coding for proteins (Lander et al., 2001), 99% of the mutations *a priori* target non-coding regions and thus do not change protein structure. Moreover, the genetic code is 'redundant', as only 20 amino acids and the stop codons are encoded by the 64 possible trinucleotides in the DNA. Approximately one third of all single base pair substitutions are therefore synonymous and do not change the amino-acid sequence (Alberts et al., 2002). In case of a non-synonymous mutation, the functional impact depends on the change in the protein structure, the expression level of the gene and on the environmental context. Naturally, if the gene is not expressed, a protein changing mutation will not have a phenotypic effect. Similarly, the functional impact of the very same mutation might differ depending on the differentiation state of the cell, the cellular micro-environment, or on extrinisic stressors. Thus non-neutrality of a non-synonymous mutation is not a universal feature.

1.3.1 Somatic mutations in normal tissues and cancer

A prime example for a disease caused by non-neutral genetic mutations is cancer. Cancer is a tissue hyperplasia that grows via uncontrolled cell divisions, invades adjacent tissues and has the ability to metastasize to other parts of the body (Alberts et al., 2002). In addition, self-sufficiency with respect to growth factors, the capacity to divide unlimitedly, induction of neoangiogenesis, immune evasion and the potential to resist apoptosis have been attributed to cancer cells (Hanahan and Weinberg, 2000, 2011). Likely, the many phenotypic properties of cancer cells are caused by more than a single somatic mutation and, indeed, the mutational burden in cancer is often very high. To give an example, the median mutational burden in glioblastoma was estimated to 2.2 coding mutations per megabase, i.e. approximately 7,300 mutations per genome (Brennan et al., 2013). Although this is remarkable, most mutations in cancer are probably functionally neutral and only few are expected to be responsible for tumorigenesis, as discussed above. These mutations are likely targeting proto-oncogenes or tumor suppressor genes and are commonly called driver mutations. The remaining mutations hitchhike on the functional advantage provided by the driver mutations and are accordingly called passenger mutations (Stratton et al., 2009).

To distinguish driver from passenger mutations, one typically compares the frequency of a mutation among cancer samples to the theoretical expectation due to random hits (Lawrence et al., 2013). This approach identifies genes that are targeted by a mutation more often than expected by chance, suggesting that the mutant protein provides a selective advantage. Alternatively, the ratio between non-synonymous and synonymous substitutions in a gene of interest can be compared (corrected for the number of target loci). If the ratio is significantly greater than one, the gene is likely a driver gene whose mutant form is under positive selection (e.g., Kosakovsky Pond and Frost, 2005). Surprisingly, sequencing of 274 tissue biopsies of sun-exposed eyelids identified that cancer-associated driver mutations are frequently found in normal skin (Martincorena et al., 2015). Notably, re-analysis of the same dataset using

population-dynamics models suggested that driver mutation expansions in most biopsies are actually consistent with neutral dynamics (Simons, 2016). Thus distinction between neutral drift, expansion of a mutant due to stochastic fluctuations, and natural selection, expansion of a mutant due to a selective advantage, is not trivial.

1.3.2 Cancer initiation

The apparent contradiction between statistical and dynamical arguments used to assess positive selection of driver mutations (Martincorena et al., 2015; Simons, 2016) may partly be explained by sequential acquisition of multiple driver mutations prior to tumor initiation (Fig. 1.2A). That is, single driver mutations will confer only a transient selective advantage (e.g., because homeostatic mechanisms at the tissue level kick in; Lander et al., 2009) and further mutations are required before the clone transforms malignantly. The hypothesis of a multistep process at tumor initiation was originally postulated by two pioneering studies. In 1953, Carl Nordling noticed that cancer mortality increases with the sixth power of age, an observation which he explained with a multistep model of tumor initiation (Nordling, 1953). 18 years later, Alfred Knudson arrived at a similar conclusion upon analysing the age distribution of patients with hereditary and spontaneous retinoblastomas. Patients with a hereditary mutation were typically younger than those with spontaneous tumors, which Knudson explained with a two-hit model of tumorigenesis (Knudson, 1971). Subsequently, the multistep model of tumor initiation was supported by data on colorectal cancer, which often develops from benign adenomas. Here, the mutational burden in driver genes increases proportionally with the progression from adenoma to carcinoma (Fearon and Vogelstein, 1990).

1.3.3 The clonal evolution model of cancer

It is generally thought that most cancers grow from a monoclonal origin, that is, all cancer cells are progeny of a single somatic cell (Cheng et al., 2011). Comparison of cancer incidence across tissues with age suggests that the cell of origin is a tissue stem cell that accumulated driver mutations during normal replications (Tomasetti and Vogelstein, 2015). Since the acquisition of driver mutations likely does not stop upon tumor initiation, Peter Nowell argued in 1976 that tumor progression resembles an evolutionary process during which malignancy increases over iterative cycles of mutation and selection (Fig. 1.2B). As mutation acquisition is continuous, each tumor cell has a unique mutational profile. In the presence of a selective pressure such as lack of oxygen, spatial constrains or therapeutic intervention, some cells may adapt better than others due to mutations acquired by chance. These cells will be positively selected, rendering the tumor increasingly aggressive during the course of tumorigenesis (Nowell, 1976). Aspects of the clonal evolution theory have been proven in a plethora of experimental studies (Greaves and Maley, 2012), with the clearest evidence being the presence of resistant subpopulations that are selected for by targeted therapy in some tumors (Francis et al., 2014; Turke et al.,

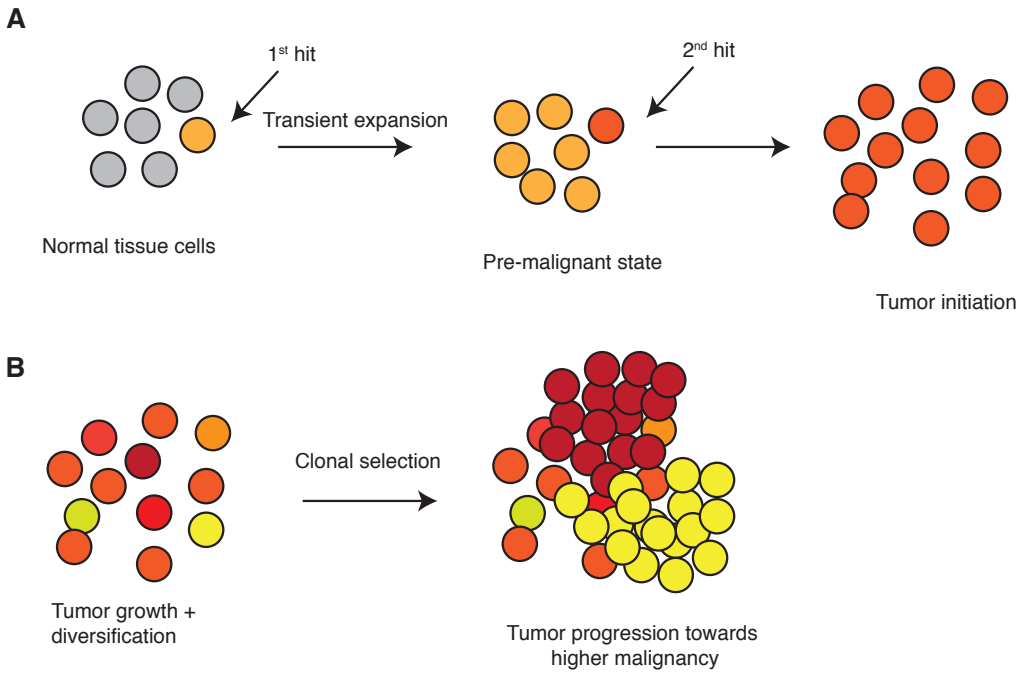


Figure 1.2: Multistep model of tumorigenesis. **A** According to the multistep hypothesis of tumor initiation, multiple driver mutations are required prior to tumor initiation. **B** The clonal evolution model of tumorigenesis states that tumors diversify into subpopulations through continuous mutation. Of these, fitter subclones (yellow, darkred) are selected for, by which tumor malignancy increases.

2010). By contrast, intratumoral heterogeneity, often viewed as evidence for tumor evolution, does not necessarily arise from non-neutral evolution towards more aggressive phenotypes. Rather, subclonal diversification through neutral mutations is expected in any cellular expansion. Whether additional driver mutations are acquired during tumorigenesis depends on the time scales of tumor progression and the rate of driver mutation acquisition. Indeed, a comparative study over 14 cancer types suggests that neutral tumor evolution is frequent across cancer entities (Williams et al., 2016). In an extreme variant of this model, all malignant events are acquired prior to tumor initiation (Sottoriva et al., 2015). To which extent neutral evolution guides tumor progression is currently subject to active debate (Davis et al., 2017; Heide et al., 2018; Tarabichi et al., 2018).

1.4 Mutational profiling with next generation sequencing

The possibility to profile whole genomes at reasonable costs has fueled research on cancer genomics tremendously within the last ten years. High-throughput next-generation sequencing is achieved by parallelization of the sequencing process through DNA fragmentation. In

most technologies, the actual sequencing step is performed by synthesizing or hybridizing complementary strands to the DNA fragments and identifying individual nucleotides by fluorescent labeling or ionic concentration (Goodwin et al., 2016). Confidence in signal detection is gained by PCR amplification and measurement of thousands of identical copies in parallel (Goodwin et al., 2016). Moreover, by sequencing each DNA locus several times, the tissue fraction harboring a specific mutation can be estimated from the fraction of reads carrying the mutation (Strom, 2016). Thus, beyond the sole identification of mutations, next-generation sequencing provides a probe of tissue heterogeneity. In Chapter 2 and 3 we will exploit this property to analyze the evolutionary dynamics in glioblastoma and hematopoiesis.

CHAPTER 2

Clonal evolution in IDH-wildtype glioblastoma

In this chapter, the clonal evolution of isocitrate dehydrogenase wildtype (IDH^{WT}) glioblastomas is analyzed using whole genome sequencing (WGS) data of paired primary and relapsed tumor samples. The first part provides an overview on clinical, cellular and molecular aspects of the disease with a special focus on intratumoral heterogeneity. This is complemented with a review on existing computational methods to infer tumor evolution from deep sequencing data. In the second part, we develop a likelihood-based multinomial model that reconstructs the phylogenetic structure of tumor subclones from deeply sequenced sample pairs. This approach is then applied to WGS data from 21 primary/relapsed glioblastoma sample pairs. The inferred phylogenies are coupled with population-dynamic models of mutation acquisition and tumor growth to learn dynamic aspects of tumor evolution in glioblastoma. Finally, the results are discussed in the context of the current state of research.

All data shown in this chapter were collected within the scope of the 'SysGlio' consortium, headed by Peter Lichter (German Cancer Research Center, Heidelberg). Bioinformatic preprocessing (sequence alignment, variant calling and mutational signature analysis) was done by Jing Yang from the group of Matthias Schlesner (German Cancer Research Center, Heidelberg), while I performed all downstream analyses. Additional contributions are acknowledged in the respective subsections. Parts of this chapter were published in Körber et al., 2019, and were written by myself.

2.1 Background

2.1.1 Glioblastoma

Tumors of the central nervous system are a heterogenous group of neoplasias that differ substantially in their clinical prognosis. Approximately one third of all brain tumors are malignant and of predominantly glial origin. These are further subdivided into slowly progressing gliomas (~40%) and highly aggressive glioblastomas (~60%; Ostrom et al., 2017). Based on histological and morphological features one distinguishes oligodendrogliomas, astrocytomas and, less frequently, ependymomas or oligoastrocytomas among the gliomas (Louis et al., 2016). By contrast, glioblastomas typically have an astrocytic cellular morphology and oligodendroglial features are rare (Alexander and Cloughesy, 2017). Approximately 10% of all glioblastomas develop from low-grade gliomas and are called secondary glioblastomas, while the majority of glioblastomas (~90%) arises spontaneously (primary glioblastomas; Ohgaki and Kleihues, 2013). Primary and secondary glioblastomas are more accurately distinguished by mutations in the *IDH* gene which are found in most gliomas and secondary glioblastomas but absent in primary glioblastomas (Louis et al., 2016).

Clinical epidemiology

Glioblastoma, albeit being an infrequent diesease (incidence rate of 3.2 per 100,000), has a strikingly poor prognosis with a 5-years survival rate of 5.5% (Ostrom et al., 2017). IDH^{WT} glioblastomas are primarily diagnosed in elderly people (mean age at diagnosis: 61 years) with no anatomical predominance; IDH-mutant glioblastomas occur in younger patients (mean age at diagnosis: 48 years) and pre-eminently develop in the frontal lobe (Lai et al., 2011; Ohgaki and Kleihues, 2013). Characteristics of glioblastomas are pseudopalisading nuclei around a necrotic core (Wippold et al., 2006) and a diffuse growth behavior with frequent infiltration throughout the brain (Lefranc et al., 2005). Metastasis outside of the central nervous system is, however, rare (Bernstein and Woodard, 1995; Schweitzer et al., 2001).

Cellular pathogenesis

Over the last decade there has been increasing evidence of a functional heterogeneity within glioblastomas, reminiscent of normal stem cell hierarchy. Initially coined in leukemia, the term 'cancer stem cell' (CSC) has by now been widely accepted to describe a cancer cell subpopulation capable of self-renewal and tumor initiation (Batlle and Clevers, 2017). Early evidence for CSCs in glioblastoma came from two studies showing that the potential of unrestricted self-renewal and of tumor initiation upon xenotransplantation into immunocompromised mice is restricted to cells expressing the neural stem cell surface marker CD133 (Singh et al., 2003, 2004). Follow-up studies identified alternative markers for CSCs and suggested that CD133 expression is associated with better prognosis and less aggressive tumor growth, rather than being a pre-requisite for tumor

initiation and self-renewal (Bhat et al., 2013; Joo et al., 2008). This refined view distinguishes $CD133^+$ and $CD133^-$ CSCs, which due to their resemblance with fetal and adult neural stem cells, respectively, are classified as 'proneural' and 'mesenchymal' CSCs (Lottaz et al., 2010).

The resemblance between tumor initiating cells and normal stem cells prompts the hypothesis that glioblastomas originate from neural precursor cells. However, whether de-differentiation of astrocytes or malignant transformation of glial stem cells initiates tumorigenesis remains subject to debate. Most attempts at identifying the cell of origin have relied on the introduction of oncogenic driver mutations to glial precursors and astrocytes in mice. In 1998, Holland et al. showed that retroviral transfer of a mutant epidermal growth factor receptor (*EGFR*) gene to glial lineage cells of mice with a disrupted cyclin-dependent kinase inhibitor 2A (*Cdkn2A*) locus induces lesions that resemble human glioblastomas (Holland et al., 1998). Based on the glial stem cell marker nestin and the astrocytic marker glial fibrillary acidic protein (GFAP) the authors noticed that tumors develop more frequently from glial precursors than from differentiated cells, indicating that glial stem cells are the likely cells of origin in glioblastoma. These findings were corroborated by a later study which showed that astrocytomas can be induced in neural progenitor cells, but not in astrocytes, upon inactivation of the tumor suppressor genes *p53*, *Nf1* and *Pten* (Llaguno et al., 2009). Moreover, a recent study identified tumor-specific driver mutations at low levels in neural stem cells from the subventricular zone of the same patient (Lee et al., 2018b). There is, however, also evidence that glioblastomas may arise from de-differentiation of mature astrocytes or neurons. Lentiviral disruption of *p53* and *Nf1* in cortical neurons has been shown sufficient to trigger glioblastoma induction in mice (Friedmann-Morvinski et al., 2012). Similarly, transplantation of mature astrocytes with a depleted *Cdkn2a* locus and a mutant *EGFR* gene induced high-grade gliomas in immunodeficient mice (Bachoo et al., 2002). Both studies notice de-differentiation during tumor formation, suggesting that specific molecular dysregulation can induce glioblastomas in different stages of cellular differentiation. Whether this holds also true for glioblastomas that arise spontaneously *in vivo*, remains to be elucidated.

Molecular pathogenesis

Karyotype analyzes and DNA sequencing studies have uncovered characteristic patterns of chromosomal and genetic alterations in glioblastoma. Pioneering studies in the 1980s noticed that gains of the entire chromosome 7, focal amplifications of the *EGFR* locus via double minutes, losses of chromosome 10 and deletions or structural rearrangements of chromosome 9p are frequent events in glioblastoma (Bigner et al., 1986, 1984, 1987). These chromosomal alterations target prominent oncogenes — *EGFR, CDK6* and *MET* on chromosome 7, *CDKN2A/B* on chromsome 9p and *PTEN* on chromosome 10, to name the most striking examples (Brennan et al., 2013) — and were confirmed to be among the most frequent events in glioblastoma overall (Brennan et al., 2013; Dahlback et al., 2009; Vranová et al., 2007). More detailed investigation distinguishing glioblastoma subtypes revealed frequent loss of chromosome 19q but less frequent

loss of chromosome 10 in secondary glioblastomas, suggesting that different genetic pathways drive primary and secondary glioblastomas (Fujisawa et al., 2000; Nakamura et al., 2000). This hypothesis is corroborated by studies on the single-gene level, revealing that mutations in the tumor suppressor gene *PTEN* are characteristic for primary glioblastomas, whereas mutations in the metabolic enzyme *IDH* and the tumor suppressor gene *TP53* are more frequently found in secondary glioblastomas (Balss et al., 2008; Ohgaki et al., 2004; Tohma et al., 1998).

The most comprehensive analysis of the glioblastoma genome has been performed by The Cancer Genome Atlas Research Network (TCGA), who analyzed the exomes of 291 glioblastoma samples and the transcriptomes of 164 RNA samples with next generation sequencing. In addition to identifying several new mutations (among these somatic mutations in *LZTR1, SPTA1* and *ATRX1*), the network could link therapy-induced hypermutator phenotypes to mutations in mismatch repair genes and define molecular core pathways in glioblastoma. Strikingly, alterations in the RB, TP53 and RTK pathways, all involved in cell cycle progression and cell survival, were found in approximately 80% of the patients. Moreover, the authors noticed that activating mutations in the promoter region of the telomerase reverse transcriptase gene (*TERT*) and inactivating mutations in *ATRX*, associated with alternative telomere lengthening, were mutually exclusive. The authors concluded that alternative ways of telomere maintenance exist in glioblastoma (Brennan et al., 2013; The Cancer Genome Atlas Research Network et al., 2008).

Treatment

Standard of care The standard of care for glioblastoma patients to date is surgical resection of the tumor followed by radiation therapy and chemotherapy with the alkylating agent temozolomide. Nevertheless, the disease progresses within 12 months after initial diagnosis in approximately 75% of the patients (Stupp et al., 2005). To improve progression-free survival the factors predicting treatment response and patient prognosis need to be identified. Insufficient surgical resection due to infiltrative growth is clearly a major contributor to disease recurrence and more gross resection has been linked to prolonged progression-free survival (Sanai et al., 2011). Moreover, global hypermethylation on CpG sites, referred to as 'glioma-CpG island methylator phenotype' (G-CIMP), is associated with better prognosis overall (Noushmehr et al., 2010).

Temozolomide acts by linking methyl groups to the purine bases adenine and guanine in the DNA (Drabløs et al., 2004). The primary toxic product, O6-methylguanine, mismatches with thymine, which is directly repaired by the O-6-methylguanine-DNA methyltransferase (MGMT). In the absence of MGMT expression, the mismatch is recognized by the DNA mismatch repair, which fails to correct lesions on the template strand and induces double strand breaks of which the cell eventually dies (Zhang et al., 2012). In line with its mode of action, the response to temozolomide treatment could be positively correlated with functional MMR and silencing of the *MGMT* gene by promoter methylation (Cahill et al., 2007; Hegi et al., 2005).

Personalized treatment based on molecular subtyping In an attempt to capture inter-individual heterogeneity in treatment response more accurately, Verhaak et al., 2010, developed a molecular classification system of gene expression in glioblastoma, based on which four major glioblastoma subtypes were identified. These subtypes were re-analyzed by Brennan et al., 2013, collectively revealing the following characteristics:

- A 'classical' subtype is linked to joined high-level *EGFR* amplification, focal chromosome 9p deletion and chromosome 10 loss, while *TP53* mutations are typically absent; in addition, neural stem cell markers are highly expressed in tumors classified as of the classical subtype (Verhaak et al., 2010). Promoter methylation of the *MGMT* gene in glioblastomas of the classical, but not other subtypes, is associated with better prognosis (Brennan et al., 2013).
- A 'mesenchymal' subtype is primarily linked to focal *NF1* deletions and the expression of mesenchymal and astrocytic markers, among them *CD44* (Verhaak et al., 2010). The MAP kinase signaling pathway has been described as "moderately upregulated" in this subtype (Brennan et al., 2013).
- A 'proneural' subtype is mainly characterized by mutations in *PDGFRA*, *IDH1* and *TP53*, absence of chromosome 7 amplification and chromosome 10 loss, and expression of proneural development genes (Verhaak et al., 2010). Proneural glioblastomas show increased activation of the PI3K pathway and are frequently of the G-CIMP phenotype (Brennan et al., 2013). Overall, they have been associated with better survival and comprise most secondary glioblastomas (Brennan et al., 2013; Phillips et al., 2006; Verhaak et al., 2010).
- A 'neural' subtype is linked to the expression of neural markers and little infiltration with non-tumor cells (Verhaak et al., 2010).

Capper and co-workers recently proposed an alternative classification approach, which focuses on the DNA methylation status rather than gene expression and distinguishes eight subclasses in glioblastoma. The receptor tyrosine kinase I and II and the mesenchymal subtypes are most frequent among these and resemble the proneural, classical and mesenchymal expression subtypes (Capper et al., 2018).

Comparison of molecular subtypes in pairs of primary and relapsed glioblastomas revealed frequent subtype switching between primary and recurrent tumors, mostly towards the mesenchymal or the proneural subtype (Wang et al., 2016, 2017). Interestingly, classification of glioblastomas as mesenchymal could be linked to immune cell infiltration by primarily macrophages and is overall associated with poorer survival (Engler et al., 2012; Wang et al., 2017). Nevertheless, advances in robust molecular subtyping have not led to an improvement of patient survival to date (Lee et al., 2018a).

Personalized treatment based on mutational profiles In theory, a cancer genome should hold key to the pathogenic mutations and precise targeting of these should exterminate the disease. In practice, molecular targeting has proven widely disappointing in glioblastoma. The most prominent example is precision medicine targeting mutant EGF-receptors. Tyrosine kinase inhibitors or monoclonal antibodies against EGFR have been successful across a wide range of cancers but have consistently failed in glioblastoma, despite the high frequency of *EGFR* mutations (Westphal et al., 2017). Alternative approaches have primarily targeted neovascularization and cell cycle progression with small molecule inhibitors against the endothelial growth factor receptors KDR and FLT1 and the kinases mTOR, PKCβ and PDGFR, but turned out to be majorily unsuccessful also (De Witt Hamer, 2010).

The reasons why precision medicine has been failing in glioblastoma are numerous and range from difficulties in drug delivery, due to the blood-brain barrier, to redundancy of pathway inactivating mutations, to intratumoral heterogeneity (Prados et al., 2015). This is especially true for *EGFR*, which is frequently amplified at cell-specific levels or targeted by distinct, often mutually exclusive mutations in different subpopulations of the tumor (Francis et al., 2014; Snuderl et al., 2011). Understanding intratumoral heterogeneity and genetic evolution in glioblastoma has thus become a prime interest in therapeutic development (Qazi et al., 2017).

2.1.2 Intratumoral heterogeneity and clonal evolution in glioblastoma

Intratumoral heterogeneity is considered a major cause of treatment failure in glioblastoma (Qazi et al., 2017). As discussed in Section 2.1.1, glioblastoma cells are hierarchically organized with a few cells having the capacity to self-renew, while most of their progeny divides transiently and eventually dies off. Such functional heterogeneity is most likely due to transcriptional or epigenetic differences, and, indeed, single-cell transcriptomic analysis revealed highly variable expression of gene sets related to proliferation, immune response or hypoxia among glioblastoma cells (Patel et al., 2014). Intratumoral heterogeneity can, however, also be studied on the genomic level, which has been increasingly done in recent years. Two studies, which analyzed single-cell genomes with fluorescence in situ hybridization (FISH) and single-cell sequencing, found distinct *EGFR* mutations in different cells of the same tumor and, moreover, various amplification levels of *EGFR, MET* or *PDGFRA* in a mutually exclusive way (Francis et al., 2014; Snuderl et al., 2011). In a comprehensive study of the genetic and transcriptional heterogeneity in eleven primary glioblastomas, Sottoriva et al., 2013 compared multiple regions per tumor and found that copy number changes of chromosomes 7 and 9 (involving the *EGFR* and *CDKN2A* locus) were usually present in the entire tumor, while aberrations on chromosomes 4 and 10 (involving the *PDGFRA* and *PTEN* locus) were frequently shared by a subset of the tumor only. Moreover, the authors noticed that several expression subtypes can co-exist within a single tumor, an observation that was corroborated by two studies of single-cell transcriptomes in glioblastoma (Patel et al., 2014; Wang et al., 2017).

In an attempt to link intratumoral heterogeneity with clonal evolution under standard

therapy, recent studies have started to compare the genetic, epigenetic and transcriptional profiles of matched pairs of primary and recurrent glioblastomas. Interestingly, whole exome sequencing of sample pairs from initially low-grade gliomas that progressed to secondary glioblastomas at recurrence indicated clear branching between primary and recurrent tumor with frequent replacement of driver mutations (Johnson et al., 2014). Moreover, approximately 50% of the patients treated with temozolomide acquired a hypermutation genotype at recurrence, indicating ongoing evolution under therapy (Johnson et al., 2014). A similar approach, including primary and secondary glioblastomas, suggested highly branched evolution of both subtypes, by which the authors argued that clonal replacements of key driver mutations between primary and recurrent tumors are frequent (Wang et al., 2016). By contrast, comparative whole exome sequencing of pairs with local or distant tumor recurrence suggested that clear branching between primary and recurrent tumor occurs in the latter, but not in the former (Kim et al., 2015). Notwithstanding the insights into the gentic architecture of glioblastomas gained by these studies, none of them combined intra- with inter-sample heterogeneity to obtain a comprehensive model of the phylogenetic relationships within the tumors. Moreover, intratumoral heterogeneity has been barely used to infer the dynamics of tumor growth. Thus further analysis is necessary to understand the evolutionary dynamics underlying glioblastoma growth and recurrence.

2.1.3 Existing algorithms for phylogenetic inference in cancer

Phylogenetic inference of genetic subclones in cancer can be performed at various levels of resolution. A crude analysis compares two or more tissue samples of the same tumor, treats each as a single clone and attempts to find the phylogenetic tree that explains the mutational profile with least error, while being maximally parsimonious. More elaborate methods take the variant allele frequencies (VAFs) of measured mutations into account, that is the fraction of sequencing reads reporting the mutation. Knowing the copy number state at a given locus, VAFs can be transformed to cancer cell fractions, from which genetic heterogeneity within samples can be inferred. Ultimately, the highest resolution can be achieved by sequencing single cells to learn the phylogenetic relationships between these.

Irrespective of the level of resolution, methods for phylogenetic inference frequently rely on a set of basic assumptions (Schwartz and Schäffer, 2017):

- **Maximum parsimony** The maximum parsimony criterion assumes that the simplest phylogenetic tree explaining the data is the most likely solution. Phylogenetic inference with maximum parsimony thus minimizes the number of evolutionary steps needed to generate the data (Schwartz and Schäffer, 2017).
- **Infinite sites model** According to the infinite sites model the probability of a specific mutation happening twice is negligibly small (Kimura, 1969). Conversely, the probability for backmutation can be neglected as well.

- **Pigeonhole principle** If the sum of two subclones exceeds 100%, one subclone must be contained in the other (Beerenwinkel et al., 2014).

Phylogenetic inference from bulk sequencing data

Learning the genetic subclones and their phylogenetic relationships in a tumor sample from bulk sequencing data is a two-dimensional inference problem. As DNA is sheared into small pieces in next generation sequencing, information on the groups of mutations belonging to the same subclones in the tumor is lost. Reconstructing the genetic profiles of tumor subclones is therefore the first inference problem. On the other hand, individual subclones need to be mapped on a phylogenetic tree, while fulfilling the pigeonhole criterion, which states the second inference problem. Aneupoloidy, gains or losses of (parts) of chromosomes in the entire tumor or a subset of it, further aggravates phylogenetic reconstruction. In order to address these problems, a multitude of computational methods have been developed in the last decade. Early attempts mostly relied on clustering of single nucleotide variants (SNVs), while assuming that copy number variations (CNVs) are monoclonal (e.g., Roth et al., 2014). These methods require a transformation of the VAFs to cancer cell fractions, on which the cluster analysis is performed. Alternative approaches define subclones on the level of CNVs, while neglecting SNVs (e.g., Oesper et al., 2013). Here, read counts at individual loci in the genome are clustered using a multinomial model. Both approaches lack the combination of subclonal SNVs with subclonal CNVs and many methods refrain from inferring a phylogenetic structure that relates identified subclones to each other. In a more sophisticated approach, both SNVs and CNVs are hence combined to jointly infer the subclonal composition and its underlying phylogenetic tree (Deshwar et al., 2015). However, copy numbers and the subclonal fraction harboring a CNA must be estimated beforehand, which introduces a potential bias in the analysis.

Phylogenetic inference from single cell sequencing data

In contrast to bulk sequencing, phylogenetic inference from single cell sequencing data has the advantage that the mutational profile of each cell is measured. Thus genetic subclones are in principle known and it only remains to find the most likely tree explaining the phylogenetic relationships between them. In practice, however, sequencing single cells is more prone to sequencing artefacts so that error models incorporating false positive and false negative rates for mutation calling are necessary. Jahn and co-workers developed a computational framework that addresses this challenge by scanning the space of all possible trees with a Markov chain Monte Carlo sampling scheme that simultaneously estimates the false-postive rate in mutation calling (Jahn et al., 2016). Although single cell sequencing provides a higher resolution than bulk sequencing, it is more prone to undersampling the tumor due to higher costs and technical limitations. Thus single cell sequencing might become a powerful alternative to bulk sequencing in the future but is not yet suitable for high throughput analyses to date.

2.2 A likelihood-based multinomial model for phylogenetic reconstruction in cancer

2.2.1 General considerations

The aim of the phylogenetic inference approach introduced below is to infer the major phylogenetic structure within pairs of primary and recurrent tumor samples from deep whole genome sequencing data. The methodology relies on the assumption that a tumor grows from a monoclonal origin and that all SNVs and small insertions/deletions are unique and irreversible events (formally known as the infinite sites model; Kimura, 1969). Copy number changes often span larger regions so that the probability of two independent alterations in the same region is not negligible. We address this by allowing up to two independent copy number changes at a single locus, as described below in more detail.

In principle, a genetic subclone can be defined at arbitrary levels ranging from the entire tumor down to single cells (Fig. 2.1). In practice, resolution is limited by the operating principle of WGS, in which each genomic locus is sampled and sequenced multiple times. If a mutation is uniformly present in the tissue, it is likely detected, whereas mutations present in a few cells only might be overlooked. Although the detection limit increases with the sequencing coverage (i.e., the average number of reads spanning each locus), the capacity to distinguish two subclones from each other decreases inversely with the subclonal size. To see this, assume a very high resolution, allowing the reliable detection of mutations present in as little as 1% of the tumor cells. In this *gedankenexperiment* up to 100 subclones, each of approximately 1% relative size, co-exist. Even if all subclones are of distinct sizes (thus slightly varying around 1%), the measured variant allele frequencies will all cluster around 0.5% in a diploid genome. Therefore, the mutational profiles of very small subclones are indistinguishable even at very high sequencing depths. Keeping these practical limitations in mind we define a subclone as a tumor subpopulation sharing a set of passenger and driver mutations down to a resolution of approximately 10% of the sample size. As the measured subclone sizes will vary around their true value, accuracy increases with the number of datapoints, meaning that both driver and passenger mutations are informative for our analysis.

2.2.2 Mathematical description of the model

Measuring a subclonal mutation in whole genome sequencing depends on two sampling steps (Fig. 2.2). First, a tissue sample is taken from which DNA is extracted. Second, sequencing reads are sampled from the fragmented DNA. As solid tissues are not well mixed, the frequencies of subclonal mutations in the tissue sample may differ from the true frequencies in the tissue. Moreover, mutations which are present in a certain region of the tissue only, may be entirely missed by the first sampling step. The impact of tissue sampling on the measured subclonal

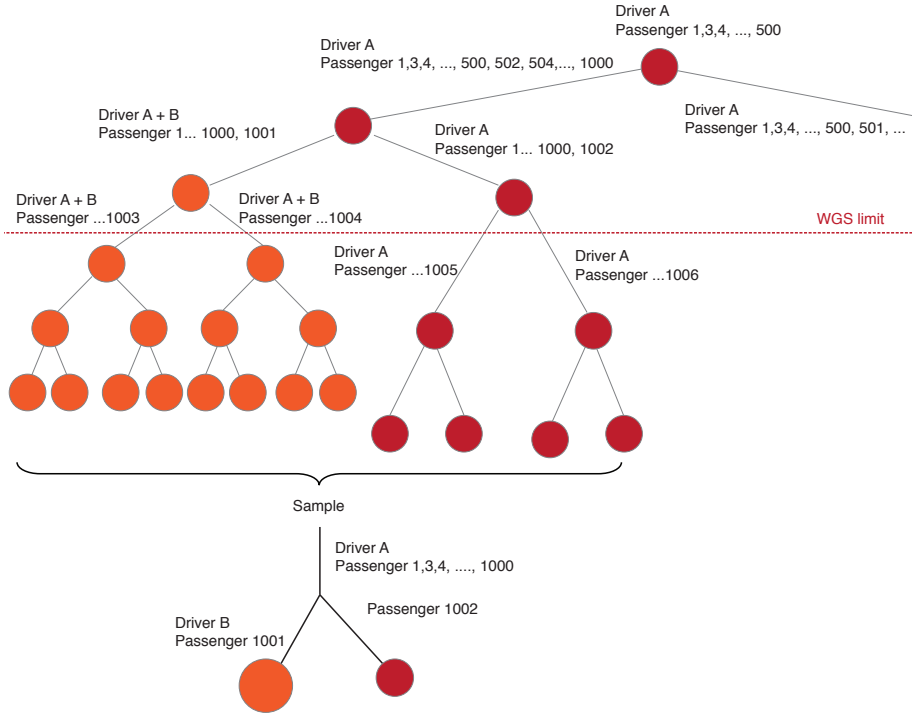


Figure 2.1: Clonal evolution in tumorigenesis. Clones can be defined as the entire tumor, as cells sharing driver mutations (marked here by different colors) or as cells sharing passenger mutations (corresponding to single cells). The resolution of whole genome sequencing is limited to major subclonal branchings, which can be visualized by a schematic summary tree (bottom; circle sizes scale with the relative subclonal sizes).

distribution is exclusively controlled by the experiment itself and can only be assessed if multiple tissue samples are taken. By contrast, the second sampling step depends entirely on the sequencing strategy, since VAFs are measured from the sampled sequencing reads. Due to the finite sample size, a measured VAF does not exactly correspond to the fraction of mutated DNA in the tissue sample, but is binomially distributed around it. Thus the more sequencing reads cover the mutated locus, i.e., the higher the sequencing coverage, the smaller the variance around the true value. Likewise, the more mutations are characteristic for a subclone, the more reliably can the subclonal size be inferred from the measured distribution of mutated sequencing reads.

We will now describe the second sampling step mathematically. As multiple subclones can co-exist within a tumor sample, we model read counts as sampling from a multinomially distributed pool, whose different categories represent the genetic subclones in the tumor. At each genomic locus, the probability to sample reads from a distinct subclone scales with its relative size and its copy number state. Let $SC_i, i = 1, ..., K$ denote the i-th subclone in a heterogeneous tumor with K subclones, and μ_i the proportion of cells in the sample originating from the $i-$th

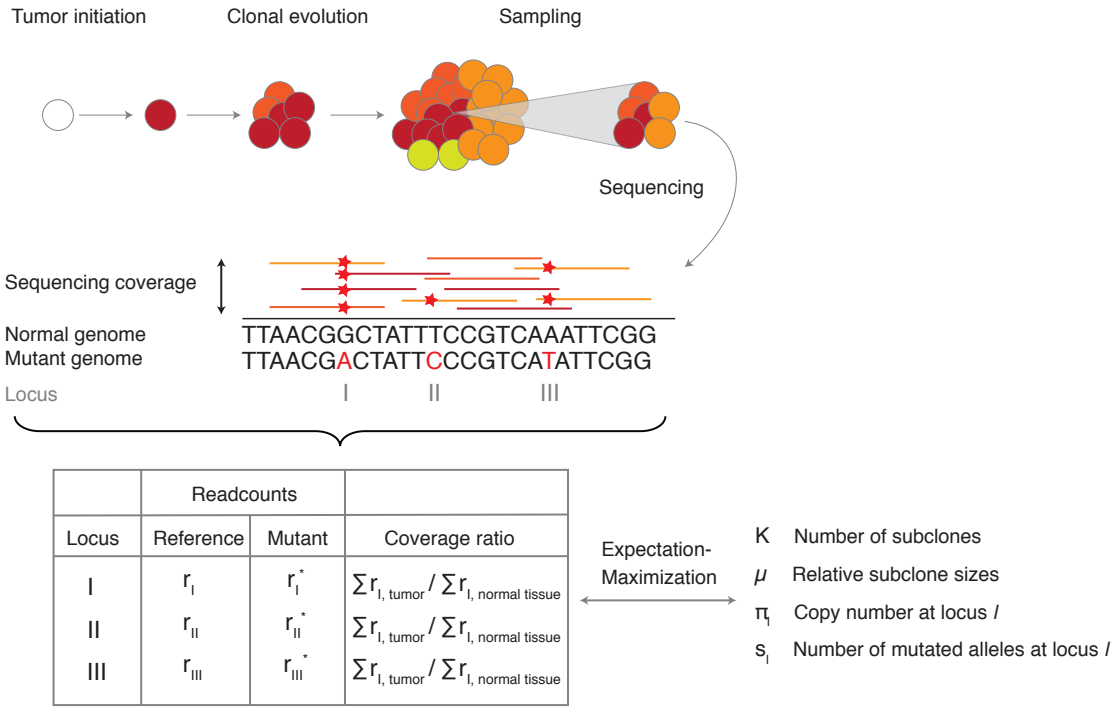


Figure 2.2: Subclonal inference from whole genome sequencing data. Two sampling steps determine the measured read counts at mutated loci. First, a set of tumor cells is sampled from which DNA is extracted. Second, sequencing reads are sampled at the target coverage. While the first sampling step is not accessible from the measured data, the second sampling step can be modeled with a multinomial distribution. Subclonal sizes and phylogenetic structure can be inferred by fitting the model parameters to the measured read count distribution using an expectation-maximization algorithm.

subclone. Then, at each locus l, the probability to sample a read from SC_i in a Bernoulli trial can be written as

$$p_{i,l} = \mu_i \frac{\pi_{i,l}}{\sum_{i'=1}^{K} \mu_{i'} \pi_{i',l}}, \tag{2.1}$$

where $\pi_{i,l}$ denotes the integer-valued copy number of the $i-$th subclone at locus l. Now, let $s_{i,l} \in \{0, \ldots, \pi_{i,l}\}$ be integers that denote the number of mutated alleles at locus l in each subclone. Accordingly, the probabilities to sample a read supporting the reference genome or a mutation, respectively, are given by

$$p_{i,l}^{\text{ref}} = p_{i,l} \left(1 - \frac{s_{i,l}}{\pi_{i,l}}\right) \qquad \text{and} \qquad p_{i,l}^{\text{mut}} = p_{i,l} \frac{s_{i,l}}{\pi_{i,l}}. \tag{2.2}$$

In bulk sequencing, the genomes from all subclones are intermingled. Consequently, r_l^{ref} reference reads and r_l^{mut} mutated reads at locus l originate from the K subclones in the sample, such that

$$r_l^{\text{ref}} = \sum_{i=1}^{K} r_{i,l}^{\text{ref}} \qquad \text{and} \qquad r_l^{\text{mut}} = \sum_{i=1}^{K} r_{i,l}^{\text{mut}}. \tag{2.3}$$

With this, the likelihood function for the measured numbers of reference and mutated reads at locus l is given by

$$\mathcal{L}_l(p_l | r_l^{\text{ref}}, r_l^{\text{mut}}) = P_l(r_l^{\text{ref}}, r_l^{\text{mut}} | p_l) = \frac{(r_l^{\text{ref}} + r_l^{\text{mut}})\,!}{\prod_{i=1}^{K} r_{i,l}^{\text{ref}}\,!\, r_{i,l}^{\text{mut}}\,!} \prod_{i=1}^{K} \left(p_{i,l}^{\text{ref}}\right)^{r_{i,l}^{\text{ref}}} \left(p_{i,l}^{\text{mut}}\right)^{r_{i,l}^{\text{mut}}}, \tag{2.4}$$

with the corresponding log-likelihood function l_l

$$l_l = \log \mathcal{L}_l = C + \sum_{i=1}^{K} \left[r_{i,l}^{\text{ref}} \log\left(p_{i,l}^{\text{ref}}\right) + r_{i,l}^{\text{mut}} \log\left(p_{i,l}^{\text{mut}}\right) \right], \tag{2.5}$$

where C is a constant that solely depends on the read counts and is thus irrelevant for finding the parameter values maximizing l_l.

Solution space and phylogenetic tree design

We assume that tumors evolve from a monoclonal origin and can be visualized by a phylogenetic tree. This restricts the solution space for subclonal inference as we will see in the following. Invoking the infinite sites hypothesis (Kimura, 1969), we require that the combinations of subclones carrying a mutation can be explained by a single event in the tree and are present on one of the two parental alleles only. We thus require

$$0 \leq s_{i,l} \leq \max(B_{i,l}, \pi_{i,l} - B_{i,l}), \tag{2.6}$$

where $B_{i,l}$ is the number of B-alleles in subclone i at locus l.

If an SNV collocates with a CNV, $s_{i,l}$ becomes further restricted by the following criteria:

- if the copy number change precedes the mutation in the phylogenetic tree, the mutation can only be present on one allele,
- if the mutation precedes the copy number change, the mutation must either be present on all A-alleles or on all B-alleles,
- if the order of the mutation and the copy number change is unclear, the mutation can be present on any number of A- or B-alleles.

Parameter estimation and model selection

To obtain the best tree explaining our data, we need to find the most likely tree structure along with the relative subclone sizes. To address this two-layered problem, we successively estimate the best parameter set for each tree template contained in a set of pre-defined candidate trees (Fig. 2.3A). Note that by treating normal tissue as an additional subclone, we automatically account for sample purity. Moreover, we design the candidate trees in such a way that, in general, all subclones present in the primary sample are different from the ones in the relapse sample. However, this also comprises solutions in which the same subclone is present in both samples if the branches separating the two subclones are collapsed (Fig. 2.3B, two leftmost panels). Similarly, these trees can also be collapsed into topologies of linear evolution (Fig. 2.3B, two rightmost panels). The most likely tree among the fitted candidate trees is selected by requiring a good fit of the tumor stem and further employing a modified Bayesian Information Criterion.

Parameter estimation with a nested expectation-maximization algorithm

The expectation-maximization algorithm (EM-algorithm) is a local optimization method for parameter estimation with known and missing data, x and y, respectively (Held, 2008). In our case, the model parameters Θ are the clone sizes of genetic subclones in the primary and recurrent tumor (μ), the available data are the number of reference and mutant sequencing reads per locus, r_l^{ref} and r_l^{mut}, and the missing data are the number of sequencing reads stemming from each subclone, $r_{i,l}^{\text{ref}}$ and $r_{i,l}^{\text{mut}}$. The probability density function for the known and missing data, $f(x, y)$, is

$$f(x, y) = f(y|x)f(x), \tag{2.7}$$

and, accordingly, the log-likelihood functions read as

$$\log L(\Theta; x, y) = \log L(\Theta; y|x) + \log L(\Theta; x). \tag{2.8}$$

We now guess an initial value of Θ, which we call Θ^0, and compute the expectation over all possible values of the missing data y:

$$\begin{aligned}\sum_y f(y|x;\Theta^0)\log L(\Theta;x,y) &= \sum_y f(y|x;\Theta^0)\log L(\Theta;y|x) \\ &\quad + \underbrace{\sum_y f(y|x;\Theta^0)\log L(\Theta;x)}_{=\log L(\Theta)}\end{aligned} \tag{2.9}$$

$$\underbrace{\mathrm{E}\left[\log L(\Theta;x,y)|\Theta^0\right]}_{=Q(\Theta;\Theta^0)} = \underbrace{\mathrm{E}\left[\log L(\Theta;y|x)|\Theta^0\right]}_{=C(\Theta;\Theta^0)} + \underbrace{\log L(\Theta;x)}_{=\log L(\Theta)}. \tag{2.10}$$

Eqn. 2.10 holds for any value of Θ and hence we can compute

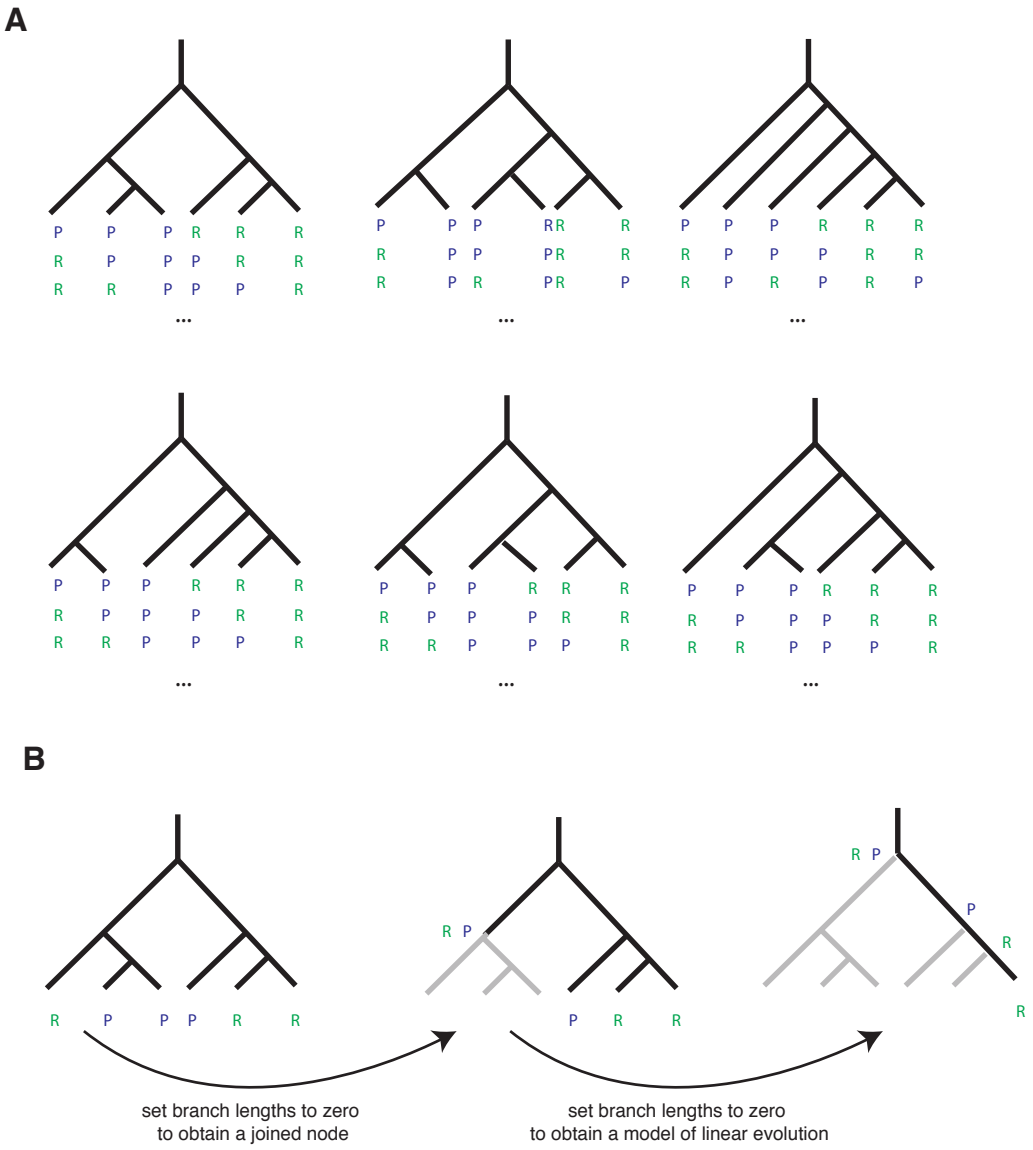


Figure 2.3: Tree templates for phylogenetic inference. **A** Candidate trees consisting of three primary (P) and three recurrent subclones (R). All combinations of assigning up to three primary and recurrent subclones each to specific tips are tested in the inference algorithm (three unique combinations are displayed each). **B** By setting individual branch lengths to zero, this setting inherently accounts for scenarios in which the same subclone is present in both tumors (two leftmost panels) and of linear evolution (two rightmost panels). Figure modified from Körber et al., 2019.

$$Q(\Theta;\Theta^0) - Q(\Theta^0;\Theta^0) = C(\Theta;\Theta^0) - C(\Theta^0;\Theta^0) + \log L(\Theta) - \log L(\Theta^0). \tag{2.11}$$

We know from Gibb's inequality that $C(\Theta^0;\Theta^0) \geq C(\Theta;\Theta^0)$. Thus if $Q(\Theta;\Theta^0) \geq Q(\Theta^0;\Theta^0)$, it follows that $L(\Theta) \geq L(\Theta^0)$. This leads to the EM-algorithm with start parameter Θ^0:

1. **Expecation**: Evaluate $Q(\Theta;\Theta^0)$.
2. **Maximization**: Maximize $Q(\Theta;\Theta^0)$ w.r.t. Θ, to obtain Θ^1
3. Abrogate if $|\Theta^0 - \Theta^1| < \epsilon$ (ϵ needs to be defined); else set $\Theta^0 = \Theta^1$ and repeat step 1 and 2.

As Θ is updated by maximization of $Q(\Theta;\Theta^0)$, it follows that $Q(\Theta^1;\Theta^0) \geq Q(\Theta^0;\Theta^0)$ and thus the algorithm increases the log-likelihood with every iteration. However, expectation-maximization optimizes the likelihood locally, so that the algorithm needs to be run from multiple start conditions in order to find the global maximum.

In our tree inference problem, we jointly estimate the parameters of matched primary and relapse samples at a given evolutionary tree with an expectation-maximization approach. As we will see in the following, we slightly modify the algorithm by nesting the estimation of the copy number state within the expectation step. We then maximize the likelihood and generate the new input to the expectation step. Both steps are iteratively repeated until convergence (required as $\sum(\mu^i - \mu^{i-1})^2 < 5 \cdot 10^{-4}$, where i is the index of the iteration). In order to identify the global maximum, optimization is repeated 100 times at random starting conditions for each candidate tree. Expectation and maximization steps are described in detail in the following.

Expectation

The algorithm is initiated with random values of μ_i, which are then updated recursively. In each expectation step, the expected counts of mutated and reference reads per subclone are iteratively calculated for every mutated locus after inferring the copy number state as follows:

Copy numbers and B-allele frequencies. We assume that there is at most one dominating copy number change, $CN_{\text{aberr},l}$, per locus and sample. This change does not have to be present in all subclones, allowing for tumor heterogeneity. While we allow different copy number changes to dominate the primary and the relapse sample at a specific locus, we assume that there is at most one copy number change per locus within a sample.

$CN_{\text{aberr},l}$ is determined from the normalized coverage ratios between tumor and blood along with the measured B-allele frequencies, BAF_l, in the tumor[1]. To this end, we apply the following criteria:

[1]If no information on the coverage ratio is available, we assume normal ploidy (2 on autosomes and female sex chromosomes, 1 on male sex chromosomes.). If no information on the B-allele frequency is available, we assume a B-allele frequency of 0.5 on autosomes and female sex chromosomes and of 0 on male sex chromosomes

- We assume that loci with coverage ratios in the interval [0.9, 1.1] and a B-allele frequency in the interval [0.45, 0.55] (or [0, 0.05] in case of male sex chromosomes) reflect normal copy number states, CN_{norm}, such that $CN_{\text{aberr},l} = CN_{\text{norm}} = 2$ on autosomes and female sex chromosomes and $CN_{\text{aberr},l} = CN_{\text{norm}} = 1$ on male sex chromosomes. The cutoffs are chosen based on the expected standard deviation of 8% in Poisson distributed read counts at a coverage of 150x.

- At all other loci, we infer the copy numbers and B-allele numbers by minimizing the squared errors between the expected and observed coverage ratios and B-allele frequencies. To this end, we start with a single copy number which we iteratively increase. At each copy number we then test different subclonal distributions $\boldsymbol{f} \in F$ of the copy number change, where $\boldsymbol{f}$ is a vector $\boldsymbol{f} = (f_1, f_2, \ldots, f_K)$, whose elements are binary indicators of a copy number change in the respective subclone, i.e. $f_i \in \{0, 1\}$. The different combinations are restricted by the candidate tree and comprise solutions in which the same or two different copy number changes dominate the primary and the relapse sample, respectively. We compute the expected B-allele frequency for B-allele counts, $B_{\text{aberr},l}$, in the interval [0, $CN_{\text{aberr},l}$] with

$$\mathrm{E}[BAF_l] = \frac{\sum_i \left[f_i \mu_i B_{\text{aberr},l} + (1 - f_i)\mu_i B_{\text{norm}}\right]}{\sum_i \left[f_i \mu_i CN_{\text{aberr},l} + (1 - f_i)\mu_i CN_{\text{norm}}\right]}, \tag{2.12}$$

  and choose the B-allele count that minimizes the squared error between expected and observed B-allele frequencies. Likewise, we compute the expected coverage ratio, cr_l, with

$$\mathrm{E}[cr_l] = \frac{\sum_i \left[f_i \mu_i CN_{\text{aberr},l} + (1 - f_i)\mu_i CN_{\text{norm}}\right]}{CN_{\text{norm}}}. \tag{2.13}$$

  The algorithm is aborted once $(\mathrm{E}[BAF_l] - BAF_{l,\text{obs}})^2 + (\mathrm{E}[cr_l] - cr_{l,\text{obs}})^2 < 0.01$ (the threshold of 0.01 corresponds to the expected Poisson noise at sequencing depths of 150x)[2]. $B_{i,l}$ and $\pi_{i,l}$ are then determined for each $i \in \{1, 2, \ldots, K\}$ to

$$B_{i,l} = B_{\text{aberr},l} f_i + B_{\text{norm}}(1 - f_i) \tag{2.14}$$

$$\pi_{i,l} = CN_{\text{aberr},l} f_i + CN_{\text{norm}}(1 - f_i) \tag{2.15}$$

[2] While only the most likely intratumoral distribution of a CNVs is selected here, alternative solutions are accounted for in data presentation and analysis as discussed below.

SNVs and small indels. The expected read counts for each possible combination of $s_{i,l}$ and $\pi_{i,l}$[3] are computed with

$$E[r_{i,l}^{\text{ref}}] = p_{i,l}'^{\text{ref}} r_l^{ref}, \qquad p_{i,l}'^{\text{ref}} = \frac{p_{i,l}^{\text{ref}}}{\sum_{i'=1}^{K} p_{i',l}^{\text{ref}}}, \tag{2.16}$$

$$E[r_{i,l}^{\text{mut}}] = p_{i,l}'^{\text{mut}} r_l^{\text{mut}}, \qquad p_{i,l}'^{ref} = \frac{p_{i,l}^{\text{mut}}}{\sum_{i'=1}^{K} p_{i',l}^{\text{mut}}}, \tag{2.17}$$

where $p_{i,l}'^{\text{ref}}$ and $p_{i,l}'^{\text{mut}}$ are the conditional probabilities of a sampled reference or mutated read originating from SC_i, provided that μ, s and π are known. We then compute the corresponding likelihood of μ, s and π:

$$\mathcal{L}_l(p_l | r_l^{\text{ref}}, r_l^{\text{mut}}) = \sum_{(r_{i,l}^{\text{ref}}, r_{i,l}^{\text{mut}})} \mathcal{L}_l(p_l | r_{i,l}^{\text{ref}}, r_{i,l}^{\text{mut}}), \tag{2.18}$$

and select the solution with the highest likelihood[4]. Since DNA is fragmented before amplification and mapping, the read count distributions at different loci are independent of each other, so that the expectation step can be independently evaluated at each mutated locus. Of note, independence of measured coverage ratios is not guaranteed, since copy number variations can span multiple loci. This is already accounted for during segmentation and therefore does not affect the inference procedure.

Maximization

In the maximization step, the log-likelihood function (Eqn. 2.5) at the expected readcount distribution (Eqn. 2.16, Eqn. 2.17) is maximized w.r.t. μ. This is approached by summing up the log-likelihoods (Eqn. 2.5) at each locus and by introducing the constraint $\sum_{I=1}^{K} \mu_i = 1$ with a Lagrange multiplier before maximization:

$$\widetilde{l} = \sum_l l_l + \lambda \left(1 - \sum_{i=1}^{K} \mu_i \right). \tag{2.19}$$

After inserting Eqn. 2.2 into Eqn. 2.19, deviation with respect to μ_i and λ yields:

$$\frac{\partial \widetilde{l}}{\partial \mu_i} = \sum_l \left(\frac{1}{\mu_i} \left(r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}} \right) - \pi_{i,l} \frac{\sum_{i'=1}^{K} r_{i',l}^{\text{ref}} + r_{i',l}^{\text{mut}}}{\sum_{i'=1}^{K} \mu_{i'} \pi_{i',l}} \right) - \lambda \tag{2.20}$$

$$\frac{\partial \widetilde{l}}{\partial \lambda} = 1 - \sum_{i=1}^{K} \mu_i. \tag{2.21}$$

[3] Possible combinations are predefined by the candidate tree. This is explained in more detail above.

[4] For data representation and analysis, alternative solutions are considered also if the best solution accounts for less than 90% of the total likelihood.

We find the maximum of the log-likelihood by setting Eqn. 2.20 and Eqn. 2.21 equal to zero and solving for λ and μ:

$$\lambda\mu_i = \sum_l \left(r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}} - \pi_{i,l}\mu_i \frac{\sum_{i'=1}^{K} r_{i',l}^{\text{ref}} + r_{i',l}^{\text{mut}}}{\sum_{i'=1}^{K} \mu_{i'}\, \pi_{i',l}} \right) \tag{2.22}$$

$$\sum_{i=1}^{K} \mu_i = 1. \tag{2.23}$$

Summing up Eqn. 2.22 over all subclones yields

$$\lambda \sum_{i=1}^{K} \mu_i = \sum_{i=1}^{K} \sum_l \left(r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}} - \pi_{i,l}\mu_i \frac{\sum_{i=1}^{K} r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}}}{\sum_{i=1}^{K} \mu_i\, \pi_{i,l}} \right). \tag{2.24}$$

With Eqn. 2.23 this reduces to

$$\lambda = \sum_{i=1}^{K} \sum_l \left(r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}} - \pi_{i,l}\mu_i \frac{\sum_{i=1}^{K} r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}}}{\sum_{i=1}^{K} \mu_i\, \pi_{i,l}} \right), \tag{2.25}$$

and consequently, by inserting Eqn. 2.25 into Eqn. 2.22, μ_i can be determined as

$$\mu_i = \frac{\sum_l r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}}}{\sum_l \left(\pi_{i,l} \frac{\sum_{i'=1}^{K} r_{i',l}^{\text{ref}} + r_{i',l}^{\text{mut}}}{\sum_{i'=1}^{K} \mu_{i'}\, \pi_{i',l}} \right) + \lambda}, \tag{2.26}$$

which reduces to

$$\mu_i = \frac{\sum_l \frac{r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}}}{\pi_{i,l}}}{\sum_{i'=1}^{K} \sum_l \frac{r_{i',l}^{\text{ref}} + r_{i',l}^{\text{mut}}}{\pi_{i',l}}}. \tag{2.27}$$

Model selection

To select the most likely tree among the candidate topologies, we first require a good fit of the tumor trunk. This is achieved by discarding trees in which more than 50% of the clonal (truncal) mutations are ambiguously mapped and, further, by discarding trees whose average squared error of clonal VAFs lies outside the 10% quantile of all candidate trees. We then assess the likelihoods of the remaining trees with a modified Bayesian Information Criterion (*BIC*) (Chen and Chen, 2008). Briefly, the modified *BIC* incorporates increasing model complexity with increasing numbers of parameters,

$$BIC_\gamma = -2 \log \mathcal{L}_n + \upsilon \log n + 2\gamma \log \tau, \tag{2.28}$$

where n is the number of data points, v the number of parameters and τ a parameter accounting for increasing model complexity weighted by γ. In our case, the number of data points is the number of readcounts ($r_l^{\text{ref}} + r_l^{\text{mut}}$), the number of parameters is the number of subclones, K, and τ is obtained by summing up all possible values of s. We choose $\gamma = 0.9$ to stringently incorporate the increasing model complexity when increasing the number of parameters.

Ambiguous solutions

To avoid bias in data interpretation due to potentially ambiguous solutions, we account for three types of ambiguity:

- **CNVs**. If model inference suggests two independent copy number changes in primary and relapse tumor, respectively, but the squared error of a joined solution is less than twice the least squared error, we account for the joined solution in data analysis and interpretation. Likewise, if the squared error for a clonal copy number change is less than twice the error of a subclonal copy number change, we account for the clonal solution.
- **SNVs**. If the location of a mutation to the phylogenetic tree is non-unique, i.e., if the best solution carries less than 90% of the total likelihood at this locus, we sort solutions by decreasing likelihood and account for all solutions that jointly yield at least 90% of the total likelihood in data analysis and interpretation.
- **Tree structure**. We account for all solutions with $BIC_\gamma \leq \min(BIC_\gamma) + 10$.

2.2.3 Method validation

To validate our phylogenetic inference algorithm we test its performance computationally and experimentally against a known 'ground thruth'.

Simulated data

First, we evaluate our method computationally on 100 simulated test sets of up to three primary and recurrent subclones, respectively (Algorithm 1). Three exemplary simulations and the matching inference results are shown in Fig. 2.4. In the example in Fig. 2.4A, the correct tree structure was inferred and the algorithm had only minor problems in the mapping of individual mutations to their position in the phylogenetic tree. This is expected due to the sampling noise in the data, which renders accurate assignment of mutations to subclones ambiguous if several subclones have similar sizes. In the two examples in Fig. 2.4B and C, the inferred tree structure is slightly different from the true structure, but major branching events are recovered in the inferred trees. Summarized over all simulated trees, the algorithm provides reliable estimates of the tumor cell content (Fig. 2.5A) and the subclone sizes (Fig. 2.5B). The correct number of subclones is estimated in the majority of cases, though overestimation and, more rarely, underestimation

of the subclonal number occurs in some cases (Fig. 2.5C-E). In the vast majority of the tested trees, clonal mutations are inferred at small false negative rates, though at the cost of higher false positive rates (Fig. 2.5F, G). Thus, the inference algorithm is conservative towards clonal mutations. Key characteristics of the tree structure, such as the mutational distance between the most recent common ancestors of primary and recurrent tumors, or the mean number of mutations per subclone, correlate well with their true counterparts, though one should be aware of some uncertainty in these estimates (Fig. 2.5H, I).

In summary, key characteristics of tumor phylogenies can be reliably inferred from simulated sequencing data. However, as the phylogenetic inference problem cannot be unambiguously solved in every detail, inferred phylogenies should not be read as the true tree structure, but rather as simplified representations of major branchings in the phylogenetic tree.

Performance on artificially mixed cell lines

In order to assess the performance of the algorithm on real data, we analyze the variant allele frequencies in two neuroblastoma cell lines, SMS-KCN and SMS-KCNR, and in a 1:4 and a 1:10 mixture of the two cell lines (SMS-KCN : SMS-KCNR). Both cell lines were generated from the same patient, so that the mixed samples model a genetically heterogenous tumor with a monoclonal origin.

Sample preparation and mutation calling. Sample preparation, DNA extraction and sequencing library preparation were performed by Selina Jansky from the group of Frank Westermann (German Cancer Research Center, Heidelberg), and whole genome sequencing at a target coverage of 80x was conducted by the DKFZ Genomics and Proteomics Core Facility. Sequence alignment, variant calling and copy number estimation were done by myself. Briefly, sequencing reads were aligned against the human reference genome hs37d5 using burrows-wheeler alignment (Li and Durbin, 2009), and duplicates were marked using GATK Markduplicates v4.0.8.1 at default parameters (McKenna et al., 2010). Mutations were called against a normal blood sample from a neuroblastoma patient (kindly provided by Frank Westermann) using Strelka v2.8.4 at default parameters (Saunders et al., 2012) and annotated using annovar (Wang et al., 2010). Low-quality variants and mutations in repeat regions (extracted from UCSC table browser, https://genome.ucsc.edu/cgi-bin/hgTables, selecting 'RepeatMasker' and 'Simple Repeats') were filtered upon mutation calling. Finally, coverage ratios and copy number estimates were obtained with Control-FREEC (Boeva et al., 2011; with breakPointThreshold=0.6, maximal SubclonePresence=0.2, coefficientOfVariation=0.5 and the normal blood sample as control).

Algorithm 1 Simulate sequencing data from heterogenous tumors

1: Sample between one and three primary and recurrent clones (K_{prim} and K_{rec}).
2: Sample the corresponding subclone sizes, μ_{prim} and μ_{rec}, at a minimal tumor cell content of 0.5.
3: Sample a random tree from Figure 2.3.
4: Sample K_{prim} and K_{rec} nodes from this tree .
5: **for** each subclone **do**
6: Sample uniformly between 1 and 200 mutations for this subclone.
7: **end for**
8: **for** each sampled mutation **do**
9: Sample a copy number, π_l according to
10: P(1 allele) = 0.1
11: P(2 alleles) = 0.75
12: P(3 alleles) = 0.1
13: P(4 to 8 alleles) = 0.05 each
14: P(0 alleles | > 8 alleles) = 0
15: **end for**
16: **for** each copy number other than 2 **do**
17: Sample the subclones from the tree structure that carry the copy number change and store these in the indicator vector $f_{i,l}$.
18: Sample whether the mutation was affected by the copy number change; from this determine s_l .
19: Simulate the coverage ratios according to $cr_l = \frac{\sum_i \pi_l f_{i,l} \mu_i + 2(1-f_{i,l})\mu_i}{2} + \mathcal{N}(\mu = 0.05, \sigma = 10^{-4})$, thus, adding Gaussian noise to the simulated coverage ratios.
20: **end for**
21: **for** each sampled mutation **do**
22: Sample the read depth per locus from a Poisson distribution with $\lambda = 150$.
23: Sample the number of mutated reads from a binomial distribution with sampling probabilities according to Eqn. 2.2.
24: **end for**

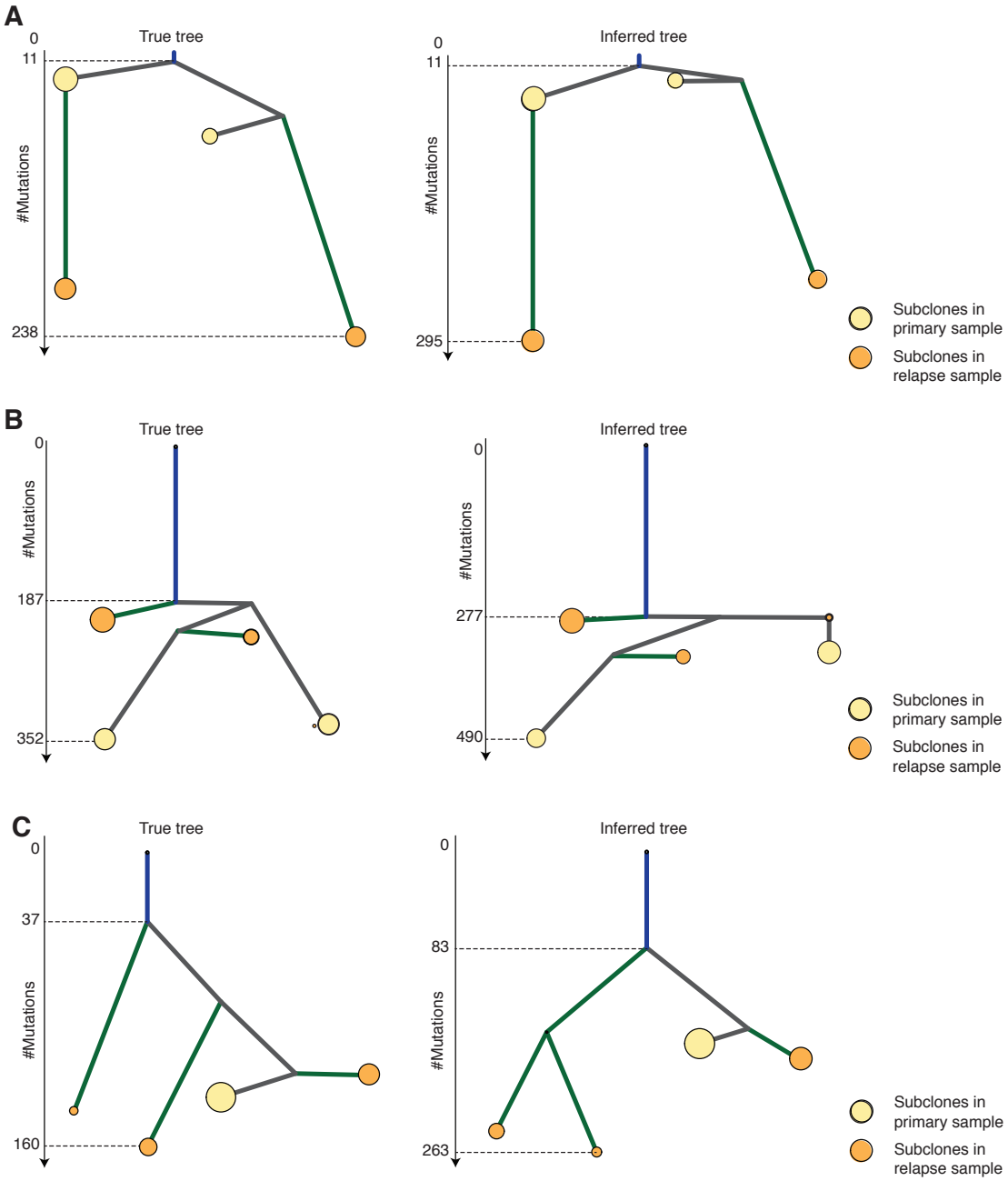


Figure 2.4: Three examples of true (left panels) and inferred (right panels) phylogenetic trees based on simulated data. Subclones present/inferred in the primary and recurrent sample are colored in yellow and orange, respectively. Circle areas scale with the relative subclone size and vertical branch lengths with the number of mutations.

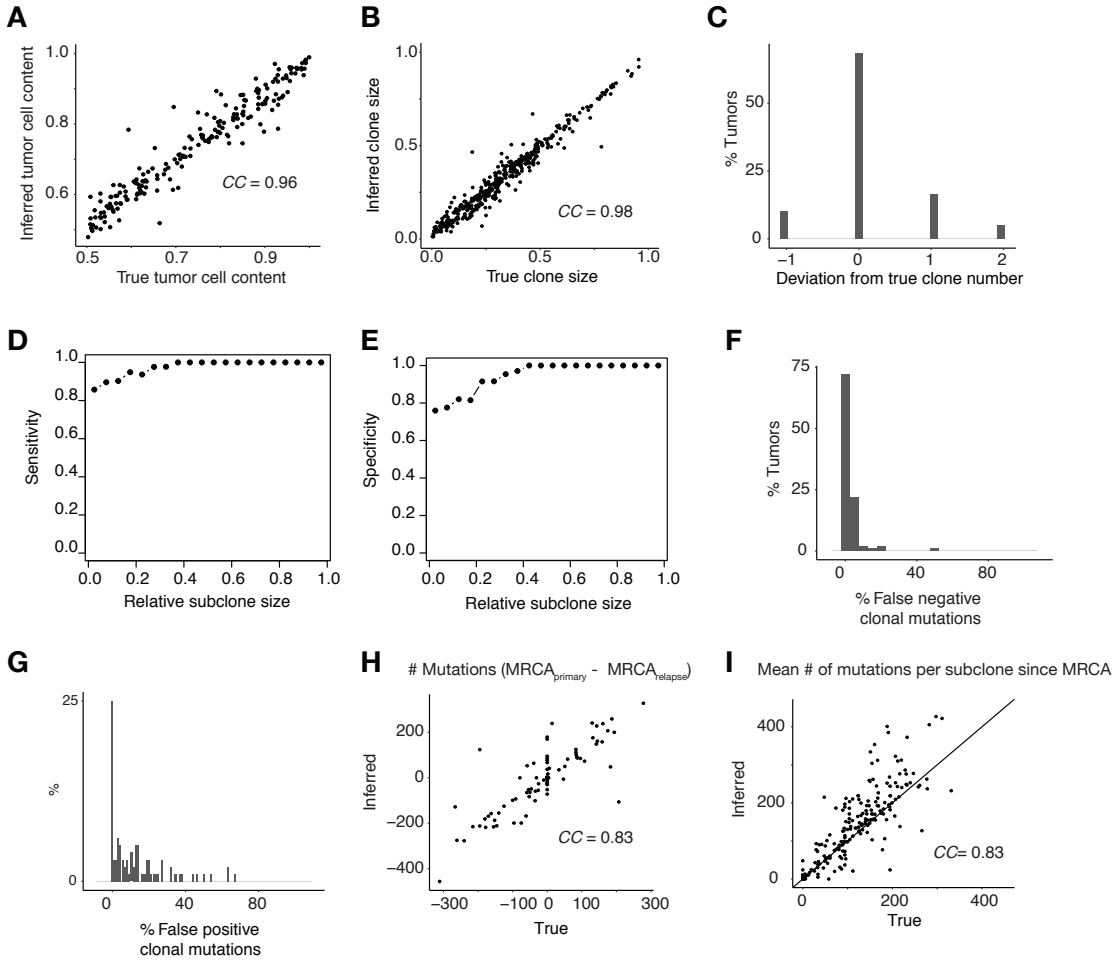


Figure 2.5: Phylogenetic inference on simulated data. **A** True and inferred tumor cell content with Pearson's correlation coefficient. **B** True and inferred clone size with Pearson's correlation coefficient. Shown are only cases, in which the correct number of subclones had been called. **C** Deviation from the true number of subclones. **D** Sensitivity of detecting a subclone in dependence of the relative subclone size. **E** Specificity of detecting a subclone in dependence of the relative subclone size. **F, G** False negative and false positive clonal mutations if neglecting subclones $\leq$ 10%. **H** True and inferred difference between the most recent common ancestors of primary and relapse sample, measured in mutation counts. Pearson's correlation coefficient is indicated in the plot. **I** True and inferred mean number of mutations per subclone since the most recent common ancestor population of the tumor. Shown are the results from 100 simulations of up to three subclones per primary and relapse sample, respectively. Figure adjusted from Körber et al., 2019.

Variant allele frequencies of clonal and subclonal mutations. We begin by testing whether VAFs measured with whole genome sequencing are binomially distributed around their true value, a basic assumption of our phylogenetic inference algorithm. Mutations present in both SMS-KCN and SMS-KCNR cells are most likely clonal mutations, comprising germline mutations and somatic mutations acquired in the common ancestor cell of the two cell lines. As shown in Fig. 2.6A, the measured VAF distribution of these mutations indeed conforms to the theoretical prediction according to a binomial distribution, and thus a binomial model can be used to identify clonal mutations. To test whether this holds also true for subclonal mutations, we next analyze the VAF distribution of mutations that are private to either of the two cell lines. Among these, mutations with a VAF ≥ 0.5 are likely clonal in the respective cell line, but, since they are absent in the other cell line, were acquired after the two cell lines split from their common ancestor cell. Analyzing the VAF distribution of these mutations in different dilutions hence reveals whether subclonal mutations can be modeled with binomial distributions also. The selection of high-confidence clonal mutations that are private in either of the two cell lines is shown in Fig. 2.6B. As before, the distribution of these 'privately clonal' mutations agrees well with the theoretical expectation according to a binomial distribution (Fig. 2.6B, lower panels). Moreover, the VAFs of 'privately clonal' mutations remain binomially distributed upon dilution in both mixed populations, supporting a multinomial model to analyze genetic heterogeneity in tumors (Fig. 2.6C). Of note, average VAFs of 'privately clonal' mutations suggest higher fractions of SMS-KCN cells among the cell line mixtures than experimentally mixed (Table 2.1). Nevertheless, the estimated ratios add up to 102% in the 1:10 and to 99% in the 1:4 mixture, indicating that the fractions of SMS-KCN and SMS-KCNR cells are accurately determined from the VAF distribution.

Next, we assess how well we can distinguish the two cell lines without prior knowledge on the genetic heterogeneity within the samples. To this end, we subject the two mixed samples to our phylogenetic inference algorithm, treating the 1:10 mixture as the 'primary tumor sample' and the 1:4 mixture as the 'relapse tumor sample'. Tree inference is run on a subset of 595 mutations, consisting of 100 likely clonal mutations (identified from co-occurrence in pure SMS-KCN and SMS-KCNR cells), 400 mutations randomly sampled from non-clonal mutations in the 1:10 mixture (consisting of 305 mutations that were identified in both mixtures and 95 mutations that were identified in the 1:10 mixture only) and 95 non-clonal mutations identified in the 1:4 mixture only. Thus 500 mutations from each 'tumor sample' are used for tree learning.

We know from the experimental design that the 'true tree' consists of a common tumor origin from which two major branches split off, yielding the cell lines SMS-KCN and SMS-KCNR. Moreover, the two branches are present in different proportions in the two 'tumor samples' and add up to 100%, respectively, as there is no contamination with normal tissue (Fig. 2.6C, left panel). We now ask how well these key characteristics are recovered by phylogenetic inference. The best fit (c.f. Section 2.2.2) suggests an asymmetric tree that consists of three subclones per sample, which add up to 99% and 98%, respectively (Fig. 2.6C, right panel). Clonal mutations are identified with 100% specificity and 91% sensitivity, and almost all mutations that are 'privately

Table 2.1: Experimentally mixed vs inferred cell line fractions of mixed SMS-KCN and SMS-KCNR cells. Experimental data: Selina Jansky.

	Experimentally mixed fraction	Inferred fraction
SMS-KCN	0.1	0.23
SMS-KCNR	0.9	0.79
SMS-KCN	0.25	0.44
SMS-KCNR	0.75	0.55

Inferred fractions were determined by multiplying the average VAFs of 'privately clonal' mutations by two, thus transforming VAFs at heterozygous loci to cell fractions.

clonal' in SMS-KCNR cells are among the false positive clonal mutations in the joined tree. Thus the major branching between SMS-KCN and SMS-KCNR cells is not inferred, because 'privately clonal' mutations in SMS-KCNR cells are wrongly assigned to the common trunk of the phylogenetic tree. This is most likely due to the paucitiy of these mutations (2% of the mutations used for tree inference), which reduces the statistical power to identify the branching event. By contrast, the more frequent 'privately clonal' mutations in SMS-KCN cells (19% of the mutations used for tree inference) are reliably classified as subclonal, and 88% of them are sorted to the major subclonal branch present in both tumor samples. The relative size of this branch is estimated to 19% in the 1:10 and to 35% in the 1:4 mixture (Fig. 2.6C, right panel); compared to the previous estimates based on the VAF distribution only (23% SMS-KCN cells in the 1:10 mixture and 44% SMS-KCN cells in the 1:4 mixture; c.f. Table 2.1), phylogenetic inference reliably estimates the fraction of SMS-KCN cells in the 'primary tumor', while slightly underestimating it in the 'relapse sample'.

In summary, phylogenetic inference on artificially mixed cell lines reveals that, although subclonal details of very small populations are blurred in bulk sequencing, clonal mutations and major subclonal branchings can be inferred if supported by sufficiently many mutations.

2.3 Evolutionary trajectories of IDH-wildtype glioblastomas

2.3.1 Tumor samples

The data underlying this work were collected from the German Glioma Network (GGN, www.gliomnetzwerk.de) and the database of the Central Nervous System (CNS) tumor tissue bank at the Department of Neuropathology, Heinrich Heine University Düsseldorf, Germany. All patients included in the study provided their written informed consent. The study was approved by the institutional review board of the Medical Faculty, Heinrich Heine University, Düsseldorf, Germany (study number 4940). Tumor samples were contributed by Katrin Lamszus (University

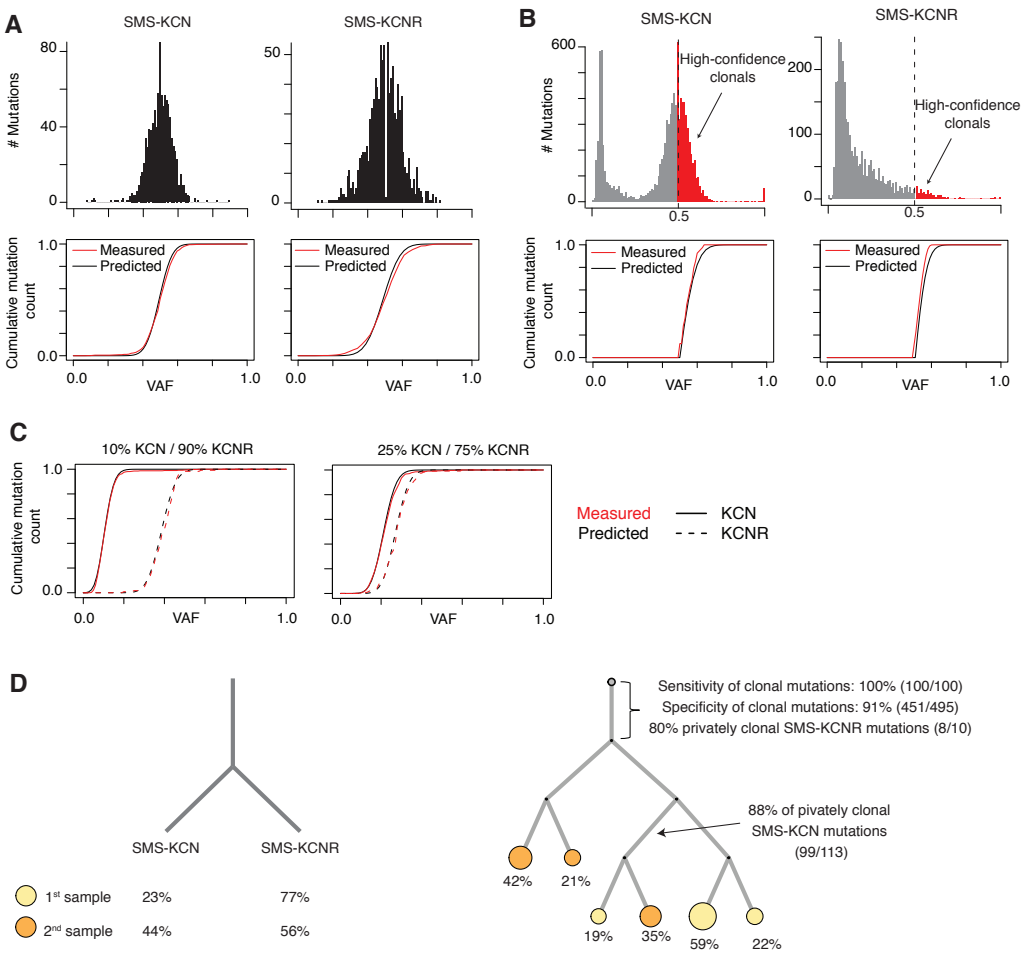


Figure 2.6: Whole genome sequencing of mixed SMS-KCN and SMS-KCNR cells. **A** Variant allele frequencies of putatively clonal mutations that were acquired prior to the split between SMS-KCN and SMS-KCNR cells. Shown are the VAF histograms (top panels) and the cumulative distributions (bottom panels) for 1,000 heterozygous variants, randomly sampled from loci with normal copy numbers (left, SMS-KCN; right, SMS-KCNR). Black lines show cumulative binomial distributions with success probabilities 0.5 and sample sizes corresponding to the median sequencing depth at the displayed variants (72 and 42 for SMS-KCN and SMS-KCNR, respectively). **B** Variant allele frequencies of private mutations in SMS-KCN and SMS-KCNR cells. Among these, mutations with a VAF≥ 0.5 are clonal in the respective cell line with high probability ('high-confidence clonals', colored in red). Lower panels show the cumulative distribution of heterozygous 'high-confidence clonal' mutations (red lines). Black lines mark the theoretical VAF distributions for mutations with a VAF≥ 0.5 according to binomial sampling with success probability 0.5 and sample size corresponding to the median sequencing depth at the displayed variants (71 and 40 for SMS-KCN and SMS-KCNR, respectively).

Caption continues on next page.

Figure 2.6 *(continued)*: C Cumulative VAF distributions of 'privately clonal' mutations (selected from the 'high-confidence clonal' mutations in (B)) in the 1:10 (left) and the 1:4 mixture (right). Black lines, measured distributions; red lines, binomial distributions with sampling size according to the median sequencing depth measured at the displayed variants and success probability $0.5 \times \rho$, where ρ is the inferred cell line fraction according to Table 2.1. **D** Phylogenetic inference if treating the 1:10 mixture as a 'primary tumor sample' and the 1:4 mixture as a 'relapse tumor sample'. Left, ground truth with subclonal sizes estimated from the distributions in (C) (c.f. Table 2.1); right, inferred tree.

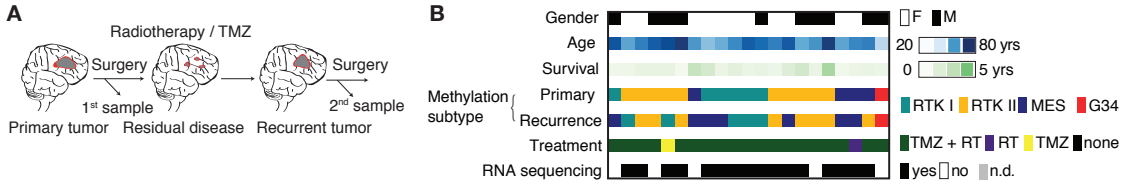


Figure 2.7: Study design. **A** Tumor samples were obtained after primary (treatment-naïve) and secondary resection (exposed to treatment). **B** Overview of clinical parameters of the patient cohort. Figure adjusted from Körber et al., 2019; patient data: Kerstin Kaulich, Guido Reifenberger; methylation subtyping: Damian Stichel.

Medical Center, Hamburg-Eppendorf), Jörg-Christian Tonn (Ludwig Maximilians University, Munich), Christel Herold-Mende (University Hospital Heidelberg), Gabriele Schackert (Technical University Dresden), Michael Sabel (Heinrich Heine University, Düsseldorf), Bettina Hentschel (University Leipzig) and Michael Weller (University Hospital Zürich)

The dataset comprises 21 sample triplets, consisting of matched primary and recurrent tumor samples along with normal blood controls from patients with IDH^{WT} glioblastomas. Primary samples were retrieved from treatment-naïve tumors after initial surgery. With few exceptions, recurrent tumors had been exposed to both radiation therapy and concomitant chemotherapy with temozolomide (TMZ, Fig. 2.7). Tumor samples were histologically classified as grade IV glioblastomas by Jörg Felsberg and Guido Reifenberger (Heinrich Heine University, Düsseldorf) and further stratified into four DNA-methylation-based subtypes ('Receptor tyrosine kinase I', 'Receptor tyrosine kinase II', 'Mesenchymal' and 'H3F3A-G34-mutant') by Damian Stichel from the group of Andreas von Deimling (German Cancer Research Center, Heidelberg, following the classification scheme by Capper et al., 2018). Clinical metadata was collected by Kerstin Kaulich from the group of Guido Reifenberger (Heinrich Heine University, Düsseldorf) and is summarized in Table 2.2.

2.3.2 Sequencing strategy

Whole genomes of the tumor samples and the matched blood controls were sequenced at an average coverage of 149x and 78x, respectively, and the methylation status of all tumor samples

Table 2.2: Clinical characteristics of the glioblastoma patient cohort. Modified from Körber et al., 2019. Data: Kerstin Kaulich, Gudio Reifenberger.

Age at diagnosis (years)		
Median (range)	60 (33-73)	
Gender		
Male	11 (52%)	
Female	10 (48%)	
Tumor location		
Frontal	4 (19%)	
Temporal	6 (29%)	
Parietal	3 (14%)	
Occipital	1 (5%)	
More than 1 cerebral lobe	7 (33%)	
Local relapse	20 (95%)	
Extent of initial surgery		
Gross total resection	11 (52%)	
Subtotal resection	10 (48%)	
MGMT promoter methylation	Primary tumor	Relapse tumor
Methylated	11 (52%)	12 (57%)
Unmethylated	10 (48%)	9 (43%)
DNA methylation subgroup	Primary tumor	Relapse tumor
RTK I	6 (29%)	4 (19%)
RTK II	10 (48%)	9 (43%)
Mesenchymal	4 (19%)	7 (33%)
H3-G34	1 (5%)	1 (5%)
First-line therapy		
Radiotherapy alone	1 (5%)	
Temozolomide alone	1 (5%)	
Radiotherapy plus temozolomide	19 (90%)	
Survival data (days), median (range)		
Interval between first and second surgery	280 (46-994)	
Overall survival	580 (261-1783)	
Patients alive at last follow-up	10 (48%)	

Abbreviations used: RTK I, receptor tyrosine kinase I group; RTK II, receptor tyrosine kinase II group; H3-G34, H3F3A-G34-mutant group.

was assessed with 450k or EPIC methylation bead chip arrays (Illumina, Hayward, CA). In addition, we sequenced the transcriptomes of 16 tumor pairs, of which sufficient RNA could be extracted (see also Fig. 2.7B). DNA extraction and sequencing were performed by David Jones, Bernhard Radlwimmer, Yonghe Wu and Andreas von Deimling (all German Cancer Research Center, Heidelberg, with support from the DKFZ Genomics and Proteomics Core Facility). Sequence alignment, tumor-specific variant calling and mutational signature analysis, as well as detection of copy number alterations and structural variants were performed by Jing Yang, Pankaj Barah, Daniel Hübschmann and Matthias Schlesner (all German Cancer Research Center, Heidelberg, with support from the DKFZ Omics IT and Data Management Core Facility). All downstream analyses, including the analysis of the mutational profiles, the selection of driver genes, phylogenetic inference and modeling of clonal dynamics were performed by myself.

2.3.3 Mutational burden in primary and recurrent tumors

To begin with, we analyze the mutational burden of single nucleotide variants (SNVs) and small insertions/deletions (indels) per tumor. Upon removal of putative germline mutations, detected from comparison with the blood controls, we find tumor-specific SNVs and small indels in the order of 10^4 in most tumors (median 12,800, Fig. 2.8A). Four tumors stand out with many more mutations ($> 10^5$), which we classify as 'hypermutated'. Notably, three of these hypermutated tumors acquired the hypermutation genotype upon exposure to treatment, while one was already hypermutated at primary resection. Except for the hypermutated cases, the mutational burden is comparable between primary and matched recurrent samples, and approximately as many mutations are shared between a sample pair as are private to one of the sample (Fig. 2.8A-C).

While the overall mutational burden per tumor provides a rough estimate of its exposure to mutagenic processes, more detailed information can be obtained from mutational signature analysis. The idea behind this approach is to infer different mutagenic processes such as replication-errors, defective DNA repair or exposure to mutagens, from the pattern of SNVs and their sequence context by non-negative matrix factorization (Alexandrov et al., 2013b). Estimating the contribution of previously described mutational signatures to the observed SNV profile in non-hypermutated tumors, we find a clear dominance of mutational signature 1 (previously associated with the age at cancer diagnosis; Alexandrov et al., 2015, 2013a; Forbes et al., 2016) among SNVs that are shared between primary and recurrent samples (Fig. 2.8D; analysis performed by Jing Yang and Daniel Hübschmann). Among SNVs that are private to either of the tumor samples, the contribution of signature 1 decreases at the expense of signatures associated with DNA double-strand break-repair (signature 3), defective DNA mismatch repair (signatures 15, 26) and one of unknown etiology (signature 5; Fig. 2.8E). Notably, the mutational signatures between primary- and relapse-specific SNVs are majorily comparable among non-hypermutated tumors, while there is a clear dominance of signature 11 in relapse, but not in primary samples of hypermutated tumors. This signature has been associated with exposure to alkylating agents (Alexandrov et al., 2013a), suggesting that the hypermutation genotype of

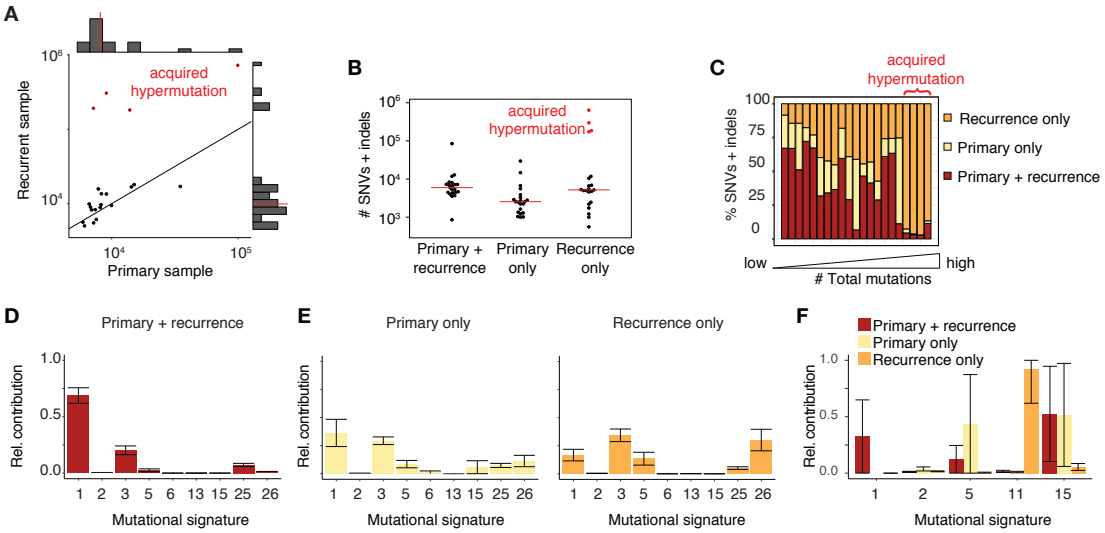


Figure 2.8: Mutational burden in IDH^{WT} glioblastomas. **A** Single nucleotide variants and small insertions/deletions per tumor after correcting for germline mutations (red lines, median). **B** Like (A), but distinguishing mutations shared between the samples of a tumor pair from private ones (red lines, median). **C** Like (B), but relative proportions and sorted per tumor. **D, E** Mutational signatures among mutations shared between the samples of a tumor pair (D) or among mutations present in the primary (E, left panel) or the recurrent tumor (E, right panel) only. Shown are non-hypermutated tumors only. Data are presented as mean ± SEM. **F** Like (D, E), but for hypermutated tumors. Figure and legend modified from Körber et al., 2019. Whole genome sequencing: Yonghe Wu. Mutation calling and mutational signature analysis: Jing Yang & Daniel Hübschmann.

these tumors developed under therapy.

In summary, primary and recurrent glioblastomas harbor comparable numbers and patterns of somatic mutations, suggesting that there is little genetic evolution under therapy. Sole exception are three tumors, which acquired a hypermutation genotype under temozolomide and radiation therapy.

Patterns of driver mutations

The vast majority of mutations in cancer are likely passengers of no functional importance, while only a minority drives malignancy. This raises the question, whether the pattern of driver mutations changes between primary and recurrent tumors, despite their high similarity with respect to the mutational burden and signatures overall. To address this question, we need to distinguish passengers and drivers among the identified mutations. Clearly, direct evidence for the functional impact of a mutation requires experimental proof. However, given the large amount of mutations in cancer, we need to predict a mutation's impact in a faster and cheaper way. Therefore, we search for indirect evidence of driver mutations with statistical methods, relying on the assumption that mutations are randomly acquired. If this holds true, neutral

mutations should be evenly distributed along the genome, whereas clustering of mutations in distinct genes is likely a footprint of selection. This approach works best with large cohorts as the statistical power improves with the number of samples available. In the present study, we have a limitied sample size of 21 tumor pairs and thus, in order to overcome the limitations of our relatively small cohort, we build our set of candidate driver genes from three sources:

1. We use a list of genes that have been previously identified as likely drivers in glioblastoma. This list has been curated by the online platform 'intogen' and encompasses 74 genes (downloaded on January 24th, 2018 from www.intogen.org; Gonzalez-Perez et al., 2013). In addition, we include mutations in the *TERT* promoter region as likely drivers (Barthel et al., 2018).

2. We complement our candidate list with our own dataset, including genes in which mutations are significantly overrepresented according to the computational framework OncodriveFML (Mularoni et al., 2016). OncodriveFML divides the region of interest (in our case the coding region; https://bitbucket.org/bbglab/oncodrivefml/downloads/) into short intervals (in our case individual genes). For each interval, a functional impact score is computed from the observed number of mutations, weighted by the specific base pair substitution. In this way, OncodriveFML corrects for the overall frequency of specific substitutions in the dataset. The algorithm then simulates the same number of mutations per interval by randomly drawing individual substitutions and computes an empirical p-value for the functional impact score per gene by comparing the observed and simulated scores. We accept genes as putative drivers if $p_{\text{adj}} \leq 0.1$, obtained by correcting the p-values for multiple sampling with the Benjamini-Hochberg procedure. This adds another five genes - *ZNF835, CIC, FUBP1, NOTCH1* and *ATM* - to our driver gene list.

3. As the functional impact of mutations in non-coding regions is less well understood, we conservatively accept non-coding RNAs as putative drivers if mutated in more than five patients.

The most pervasively mutated driver genes in our dataset are *PTEN*, *EGFR* and *CDKN2A* (Fig. 2.9A, B). These are primarily targeted by gain of chromosome 7 (including the *EGFR* locus), loss of chromosome 9p (or focal deletion of the *CDKN2A* locus, predominantly homozygous) and loss of chromosome 10 (predominantly hemizygous). In addition, *EGFR* is frequently targeted by SNVs, high-level amplifications or structural variants (Fig. 2.9A). In seven primary and four relapse tumors the active variant EGFRvIII is generated from deletion of exons 2-7 (Brennan et al., 2013 and Fig. 2.9C). Similarly, hemizygous deletions of chromosome 10 frequently co-occur with a small mutation on the second allele of *PTEN*. These observations agree with previous reports of mutations in *PTEN*, *EGFR* and *CDKN2A* as likely drivers in glioblastoma (e.g., Brennan et al., 2013). All three alterations may increase cell proliferation via activation of the AKT-signaling pathway (by gain of function of *EGFR* and/or loss of function of *PTEN*) or via cell-cycle-checkpoint-inhibition (by loss of function of the *CDKN2A* transcripts p14/p16; The Cancer

Genome Atlas Research Network et al., 2008; Wee and Wang, 2017; Fig. 2.10).

Mutations in non-coding RNAs (ncRNAs) are overall more case-specific and no ncRNA stands out as being pervasively mutated among our cohort (Fig. 2.9A). By contrast, all but one tumor harbor mutations in the *TERT* promoter region. These are predominantly one of two 'canonical' substitutions that have frequently been observed in glioblastoma and in cancer overall (Barthel et al., 2018; Vinagre et al., 2013). Both mutations are G>A substitutions upstream of the transcription start site, which generate an ETS transcription factor binding motif (Vinagre et al., 2013). *TERT* promoter mutations have been associated with enhanced transcriptional activity and improved cellular survival (Maciejowski and de Lange, 2017). Interestingly, the two tumors without canonical *TERT* promoter mutations, harbor frame-shift mutations in *ATRX* (Fig. 2.9A), which is associated with alternative telomere lengthening (Brosnan-Cashman et al., 2018). Our data thus indicate that stabilization of cellular survival is a critical step in glioblastoma evolution.

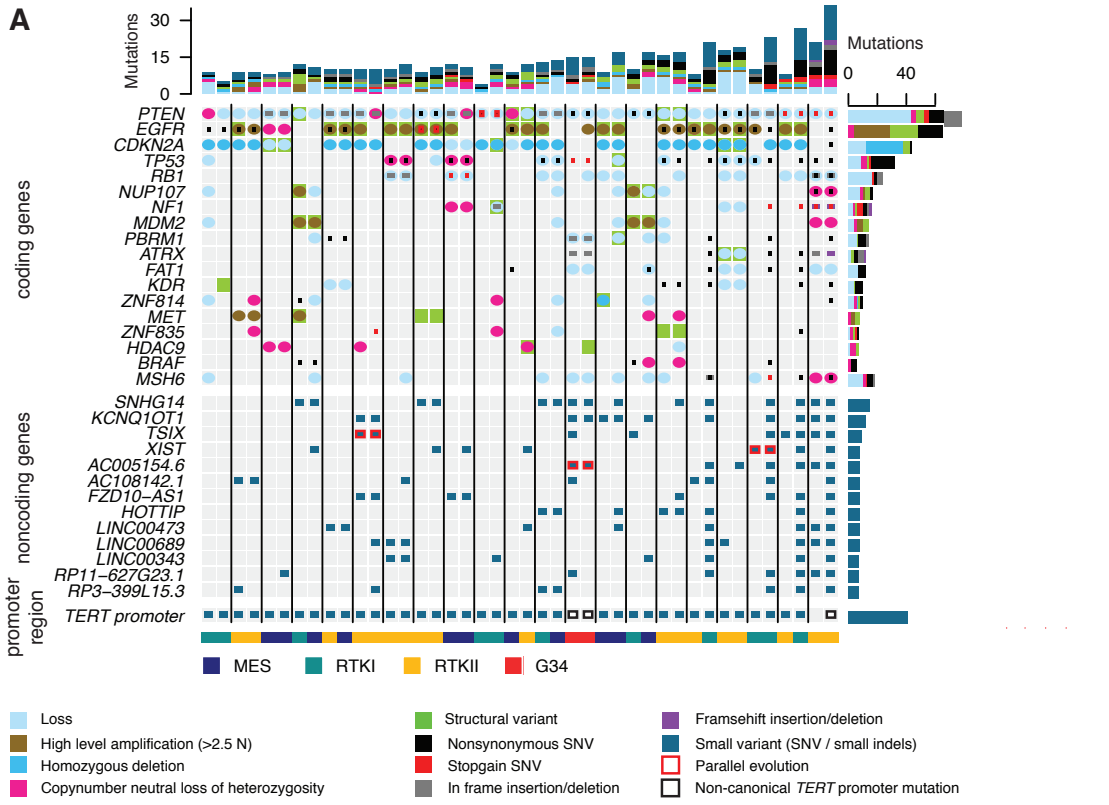
In contrast to the prevalence of pervasive driver mutations that are shared between both samples of a tumor pair, few driver mutations are recurrently acquired in the relapse tumor (Fig. 2.9A). Moreover, primary and recurrent samples of a tumor pair harbor comparable numbers of driver mutations if neglecting hypermutated cases. In the latter, the number of driver mutations increases upon tumor recurrence; however, this does not necessarily reflect clonal selection, as the mutation rate in these tumors is higher overall. This is corroborated by the presence of mutations in the mismatch repair gene *MSH6* in all hypermutated tumors, indicating that these tumors have a defective DNA repair and are thus more likely to acquire driver mutations by chance.

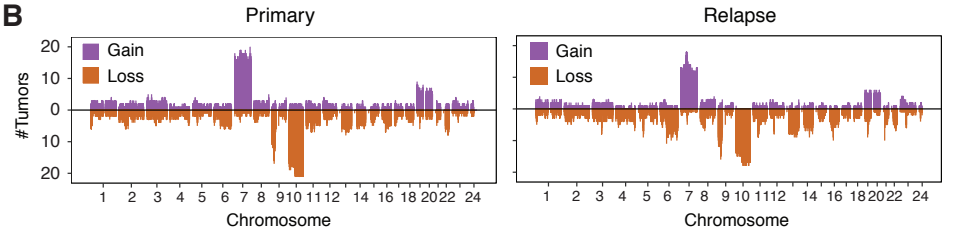
In summary, the spectrum of driver mutations in IDH^{WT}-glioblastomas reveals a similar pattern between primary and recurrent tumors. Mutations in *PTEN*, *EGFR*, *CDKN2A* and the *TERT* promoter are the most frequent events and are typically present in both samples of a tumor pair. By contrast, no stereotypical pattern of relapse-specific mutations becomes apparent.

2.3.4 Evolutionary history

To understand how clonal dynamics shape the mutational patterns in glioblastoma, we complement our study with a phylogenetic analysis of the tumor pairs. Assuming that glioblastomas grow from a single cell of origin, we attempt to infer which mutations lay at tumor initiation and which were acquired later during tumor progression. As early and late mutations will be found at clonal and subclonal levels, respectively, information on the temporal order of mutations can be obtained from the tissue fraction in which a mutation is present. Although we cannot directly measure tissue fractions with bulk sequencing, we can estimate them from the combined information of the measured VAFs and the coverage ratios at each SNV and small insertion/deletion (indel). Both metrics are readily available from the sequencing read count distribution at mutated loci.

To get an idea about the extent of subclonality in our dataset, we will start with a simple analysis, in which we compare the VAFs in primary and relapse samples at loci with normal


A
Mutations
30
15
0
Mutations
0
40
coding genes
noncoding genes
promoter region
PTEN
EGFR
CDKN2A
TP53
RB1
NUP107
NF1
MDM2
PBRM1
ATRX
FAT1
KDR
ZNF814
MET
ZNF835
HDAC9
BRAF
MSH6
SNHG14
KCNQ1OT1
TSIX
XIST
AC005154.6
AC108142.1
FZD10-AS1
HOTTIP
LINC00473
LINC00689
LINC00343
RP11-627G23.1
RP3-399L15.3
TERT promoter
MES
RTKI
RTKII
G34
Loss
High level amplification (>2.5 N)
Homozygous deletion
Copynumber neutral loss of heterozygosity
Structural variant
Nonsynonymous SNV
Stopgain SNV
In frame insertion/deletion
Framesehift insertion/deletion
Small variant (SNV / small indels)
Parallel evolution
Non-canonical TERT promoter mutation


B
Primary
Relapse
Gain
Loss
#Tumors
20
10
0
10
20
1 2 3 4 5 6 7 8 9 10 11 12 14 16 18 20 22 24
Chromosome

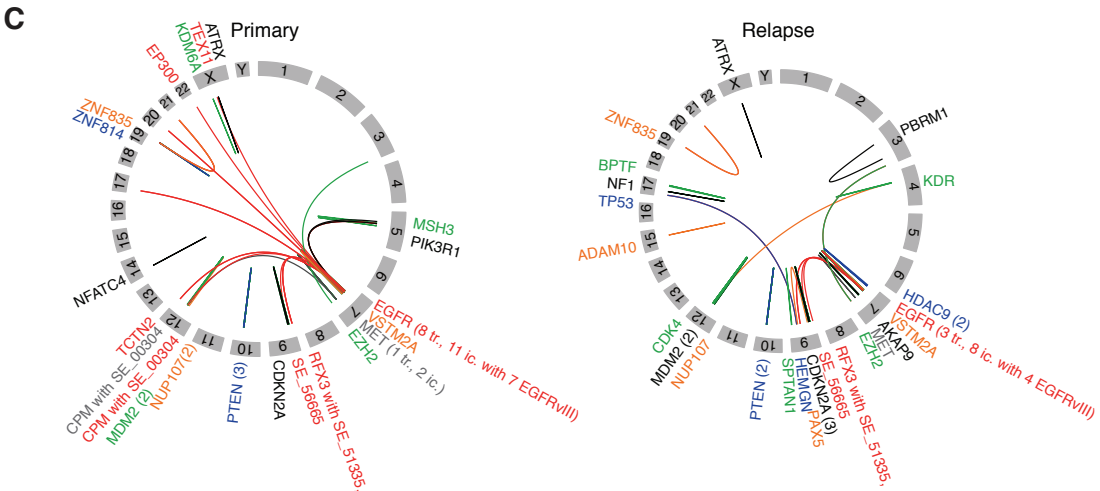

C
Primary
Relapse
ATRX
TEX11
KDM6A
EP300
ZNF835
ZNF814
NFATC4
TCTN2
CPM with SE_00304
CPM with SE_00304
MDM2 (2)
NUP107(2)
PTEN (3)
CDKN2A
SE_56665
RFX3 with SE_51335,
MSH3
PIK3R1
EGFR (8 tr., 11 ic. with 7 EGFRvIII)
VSTM2A
MET (1 tr., 2 ic.)
EZH2
PBRM1
KDR
HDAC9 (2)
EGFR (3 tr., 8 ic. with 4 EGFRvIII)
VSTM2A
AKAP9
MET
EZH2
RFX3 with SE_51335,
SE_56665
CDKN2A (3)
HEMGN
SPTAN1
PTEN (2)
CDK4
MDM2 (2)
NUP107
ADAM10
TP53
NF1
BPTF
ZNF835

Figure 2.9 ***(previous page)*****:** Driver mutations in IDH^{WT} glioblastomas. **A** Mutational spectrum in candidate driver genes. Coding genes are shown if targeted by a structural variant, SNV, small indel, homozygous deletion or a high-level amplification in at least two patients, excluding hypermutated samples. Mutations in *MSH6* are shown due to their correlation with a hypermutation genotype. Non-coding genes are shown if targeted by an SNV or a small indel in more than five patients. *TERT* promoter mutations are shown at the bottom. Tumor samples are displayed in pairs (left, primary tumor; right, recurrent tumor). Different types of mutations and methylation subtypes (bottom legend) are marked by different colors. **B** Chromosomal gains and losses in primary (left) and recurrent glioblastomas (right). **C** Structural variants in primary and relapse tumors. Intra-chromosomal variants are shown if targeting a driver gene. Translocation partners of inter-chromosomal variants are highlighted if targeting a gene or the vicinity of a super-enhancer (based on dbSUPER; Khan and Zhang, 2015). Numbers in brackets indicate the number of recurrences of a structural variant (tr., inter-chromosomal translocation; ic., intra-chromosomal variant). Figure and legend modified from Körber et al., 2019; whole genome sequencing: Yonghe Wu; mutation calling and oncoprint: Jing Yang.

copy numbers. To this end, we first need to correct the measured VAFs for tumor cell content. There are basically two ways to estimate the tumor cell content from whole genome sequencing data. The first one looks at the clonal peak of the measured VAF histogram. In a pure sample of a diploid tumor, the VAFs at heterozygous mutations are expected to peak at approximately 0.5. If the sample is contaminated with normal tissue, the peak is shifted towards smaller VAFs. Thus the extent of contamination with normal tissue can be estimated from the position of the clonal VAF peak if disregarding mutations in regions with aberrant copy number (Fig. 2.11A). In an alternative approach, the tumor cell content is estimated from the ratio between the measured coverage in distinct genomic intervals, and the overall genomic coverage in the tumor. This method relies on the assumption that most copy number changes are clonal and jointly adjusts the tumor ploidy and the tumor cell content until the measured coverage ratios can be explained by integral copy number changes (implemented e.g. in ACEseq, 'allele-specific copy number estimation from whole genome sequencing'; Kleinheinz et al., 2017). The phylogenetic inference algorithm introduced in Section 2.2 combines elements from both methods, but loosens the assumption that most copy number changes are clonal by accounting for subclonal copy number changes also. In the following, we will use the tumor cell content estimates obtained with this approach, but we will see later on that these estimates correlate well with the estimates based on ACEseq.

With an estimate of the tumor cell content, ρ, at hand, we now correct the number of sequencing reads originating from the tumor to $r^{\text{total,corrected}} = r^{\text{total}} \times \rho$. Obviously, the mutated read counts, r^{mut} are not affected by this correction as we are considering tumor-specific mutations only (assessed by comparison with a blood control). The corrected VAF then reads $VAF_{\text{corrected}} = \frac{r^{\text{mut}}}{r^{\text{total,corrected}}}$ and can be used to assess whether a mutation was subclonal. To do this, we rely on the simple assumption that the measured readcounts can be modeled by binomial sampling (c.f. Fig. 2.6A-C). If a mutation is clonal, its true VAF is 0.5 (after correction for tumor

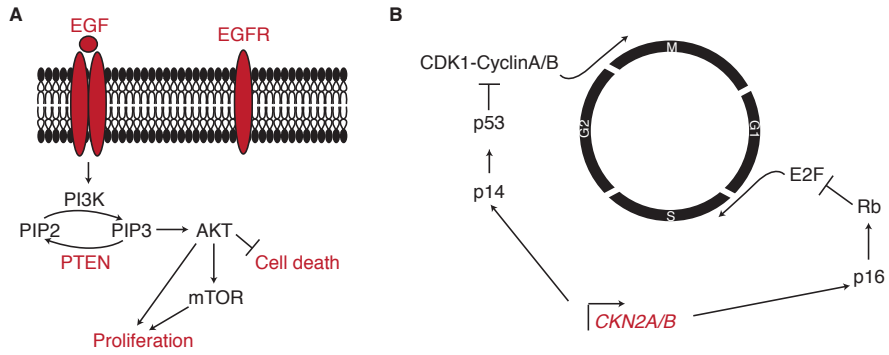


Figure 2.10: Key driver mutations in IDH^{WT} glioblastoma. **A** Signaling through the epidermal growth factor receptor (EGFR) signaling pathway is frequently aberrant in glioblastoma. Ligand binding to the receptor tyrosine kinase EGFR initiates signaling through receptor dimerization and autophosphorylation. EGFR signaling promotes cell proliferation and inhibits cell death via cascade-activation of PI3K, AKT and mTOR. EGFR signaling is frequently upregulated in glioblastoma either due to receptor overexpression or due to activating mutations in *EGFR*, often in combination with loss-of-function mutations in its antagonising phosphatase PTEN. Illustration based on Fig. 1 in Mellinghoff et al., 2007. **B** Cell cycle progression by *CDKN2A* deletion. *CDKN2A*, frequently deleted in glioblastoma, encodes two tumor suppressor proteins, p14 and p16. p16 inhibits G1/S-transition via activation of the E2F-inhibitor Rb, while p14 inhibits G2/M-transition via activation of the CDK1-CyclinA/B-complex-inhibitor p53. Illustration based on Fig. 1 in Al-Kaabi et al., 2014.

cell content) and the measured VAFs should vary around it. To detect clonal mutations at a 5% false negative rate, we thus require

$$P(X \leq r^{\text{mut}}) = \sum_{k=0}^{r^{\text{mut}}} \binom{r^{\text{total, corrected}}}{r^{\text{mut}}} (0.5 \cdot \rho) r^{\text{mut}} (1 - 0.5 \cdot \rho)^{r^{\text{total, corrected}} - r^{\text{mut}}} < 0.05. \tag{2.29}$$

If excluding hypermutated samples, 73% of all mutations are classified as subclonal in at least one sample of a tumor pair, indicating that the majority of mutations in a tumor are subclonal (Fig. 2.12A). Notably, the assessment of subclonality improves with a second tumor sample, as 17% of the mutations would have been erroneously classified as clonal if measuring the primary sample only (Fig. 2.12, orange dots). Since, however, a transition from clonality to subclonality violates a monoclonal tumor origin, these mutations must have been subclonal already in the primary tumor, but were classified as clonal due to undersampling of the tumor (Fig. 2.12B). Thus, we estimate the false positive rate of clonal mutations due to incomplete tumor sampling to

$$P(\text{subclonal in tumor} | \text{clonal in tumor sample}) = \frac{n_{\text{cPsR}}}{n_{\text{cPsR}} + n_{\text{cPcR}}} = 39\%, \tag{2.30}$$

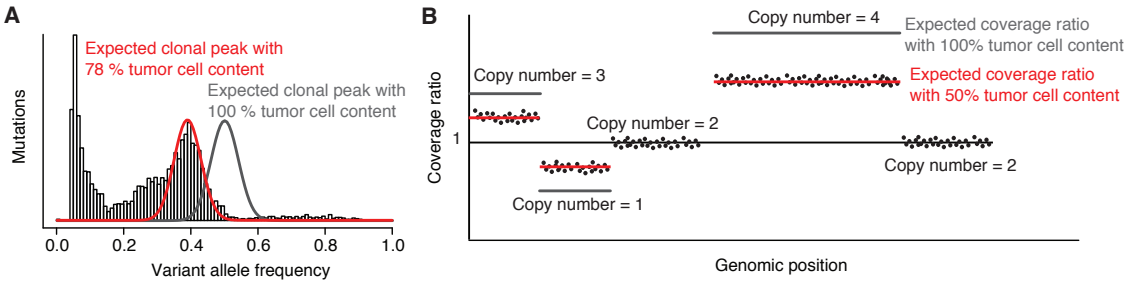


Figure 2.11: Tumor cell content estimation with whole genome sequencing. **A** The tumor cell content is estimated by adjusting the theoretical distribution of the clonal VAF peak to the measured data (grey line, binomial distribution if sampling from n=150 with success probability 0.5; red line, binomial distribution if sampling from n=150 with success probability 0.39). The measured VAF distribution shown exemplarily here is taken from the primary tumor sample of Fig. 2.13K. **B** The tumor cell content is estimated by adjusting the expected coverage ratios at integral copy number changes to the measured coverage ratio (grey lines, expected coverage ratios at the indicated copy number changes in samples with 100% tumor cell content; red lines, expected coverage ratios at the indicated copy number changes in samples with 50% tumor cell content; points, measured ratios).

where n_{cPsR} is the number of mutations classified as clonal in the primary and subclonal in the recurrent sample, and n_{cPcR} is the number of mutations classified as clonal in both samples. The high fraction of false positive clonal mutations emphasizes the advantage of multiple tumor samples to reliably identify subclonal mutations.

In order to analyze subclonality in more detail, we next employ the phylogenetic inference algorithm introduced in Section 2.2 and estimate the subclonal composition along with the underlying phylogenetic structure in the primary/relapsed tumor pairs of our dataset[5]. To briefly recapitulate the basic principle of the inference algorithm, Fig. 2.13A illustrates how SNVs and small indels are sorted to specific positions on the phylogenetic tree based on the measured sequencing read count distribution. In the example shown, most mutations are heterozygous and map to the trunk of the phylogenetic tree. Two separate sets of mutations are present in either of the two subclones only and map to the branches of the tree. In few cases, the copy number deviates from two and therefore shifts the fraction of mutated alleles away from 0.5.

Applied to our dataset, the algorithm typically recovers two to three subclones per tumor sample, of which one dominates the sample in size (Fig. 2.13B, C and Fig. A.1,A.2). Inherent estimates of the tumor cell content from phylogenetic inference correlate well with estimates by ACEseq (Pearson's r = 0.95, Fig. 2.13D), and indicate a high sample purity for most tumors (median tumor cell content of 0.8 in primary and 0.78 in relapse samples, respectively, Fig. 2.13E).

[5]The fits were performed on 500 mutated loci (including coding mutations and filled up by randomly chosen non-coding mutations). Subsequently, all mutations were mapped on the inferred phylogenetic structure. In order to identify large gains and losses and to adjust the solution in the case of inference problems (due to high level amplifications or homozygous deletions) the fitting results were manually inspected.

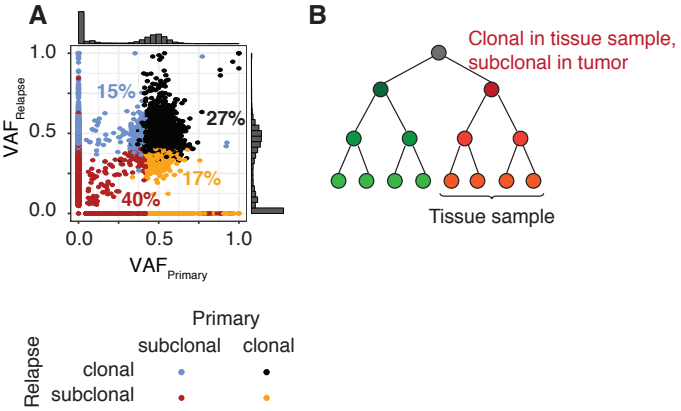


Figure 2.12: Variant allele frequencies measured in non-hypermutated samples. **A** Variant allele frequencies of all non-hypermutated tumors, classified as clonal or subclonal according to Eqn. 2.29 (shown VAFs are corrected for tumor cell content; mutations in regions with a copy number change were excluded. 10,000 sampled data points are shown for better visualization, whereas the indicated fractions were computed from the entire dataset). Figure modified from Körber et al., 2019; variant calling and copy number analysis: Jing Yang. **B** Incomplete tumor sampling may result in erroneous classification of mutations as clonal.

Three exemplary trees are visualized in Fig. 2.13G-I, according to the display format outlined in Fig. 2.13F (the inferred phylogenies of the remaining tumors are shown in Fig. 2.13K-B1). The tumor of Fig. 2.13G lost one allele of chromosome 10 and, on the second allele, captured a mutation in *PTEN* at tumor initiation. Additional driver mutations — gain of chromosome 7 and a canonical *TERT* promoter mutation — were acquired at subclonal levels. The relapse tumor re-grew from both branches of the primary tumor and acquired additional putative driver mutations in *FAT1* and the noncoding RNAs *KCNQ1OT1* and *LINC00343*. Notably, *LINC00343* was independently mutated in both branches of the relapse tumor, indicating convergent evolution.

Fig. 2.13H shows an example where no additional driver mutations were acquired upon primary resection. This tumor, like the previous example, lost one allele of chromosome 10 in the common trunk. However, the cell of origin accumulated many more driver mutations prior to initial branching. These comprise additional copy number changes (gain of chromosome 7 and loss of chromosome 9p), a small mutation in *EGFR* and, notably, a canonical *TERT* promoter mutation. Thus *TERT* promoter mutations can be found at clonal or subclonal levels in glioblastoma.

In contrast to an oligoclonal relapse origin as discussed so far, we also find one case in which the relapse tumor re-grew from a single branch of the primary tumor (Fig. 2.13I). In this example, the mutational distance between the two ancestor populations of primary and relapse tumor is rather long and several new driver mutations are acquired in the recurrent tumor. Like in the other two examples, we find copy number alterations at tumor initiation (gain of chromosome 7

and loss of chromosome 9) along with small mutations in *PTEN*, the ncRNA *KCNQ1OT1* and the *TERT* promoter.

Note that all three examples harbor specific copy number alterations (gain of chromosome 7 and loss of chromosome 9p/10) at tumor initiation, and a *TERT* promoter mutation at an early or an intermediate stage of tumorigenesis. Moreover, the three examples indicate that tumors can re-grow from a single, or from multiple subclones upon primary resection. To judge whether these differences and commonalities hold true for the entire dataset, we next compare key metrics between all tumor phylogenies.

First, in order to analyze the structural origin of relapse tumors, we measure the mutational distance between the most recent common ancestor (MRCA) of the primary and the recurrent tumor, respectively. In case of an oligoclonal relapse-origin (as observed in Fig. 2.13G and H), the MRCAs of both tumors fall together. By contrast, the two MRCAs are distinct if the relapse tumor re-grows from a monoclonal origin (compare Fig. 2.13I). Summarized over the entire dataset, we find that most relapse tumors re-grew from multiple branches of the primary tumor, reflected in a short mutational distance between the two MRCAs (Fig. 2.14A). Our data therefore indicate that selective re-growth of a specific subclone upon standard treatment is rare in glioblastoma. Rather, relapsing glioblastomas typically re-grow from the intratumoral heterogeneity established prior to primary resection.

Second, in order to identify potential commonalities at tumor initiation, we compare clonal driver mutations between individual tumors. In all but one tumor, we find at least one of three distinct chromosomal changes clonally (gain of chromosome 7, loss of chromosome 9p, including the *CDKN2A* locus and loss of chromosome 10), indicating that large chromosomal gains and losses on these chromosomes might be tumor initiating events (Fig. 2.14B). The single tumor without any clonal chromosomal alteration captured a clonal mutation in *PTEN* before acquiring a gain of chromosome 7 and a loss of chromosome 9p subclonally (Fig. 2.13T), which may either reflect an alternative path to tumorigenesis or be an artefactual observation due to an inference error. In contrast to pervasive copy number changes at tumor initiation, single nucleotide variants and small indels are often found at subclonal levels, suggesting that small mutations in driver genes may be later events in glioblastoma (Fig. 2.14C). Interestingly, *TERT* promoter mutations, though being found in all but one of the tumor samples, are placed at a subclonal level in at least one third of our dataset and hence might be dispensable at tumor initiation, but selected for at a later stage in tumorigenesis.

In summary, our dataset reveals a common path of early tumorigenesis in glioblastoma that is characterized by pervasive copy number changes on chromosome 7 (gain), 9 and 10 (losses). By contrast, *TERT* promoter mutations are found at an early to intermediate stage, reflected in frequent subclonality of these mutations. Finally, we find that relapse tumors typically re-grow from an oligoclonal origin with little addition of new driver mutations. Relapsing glioblastomas hence primarily mirror the heterogeneity already established in the primary tumor, and thus there is little evidence for selective pressure exerted by standard therapy (Fig. 2.14D).

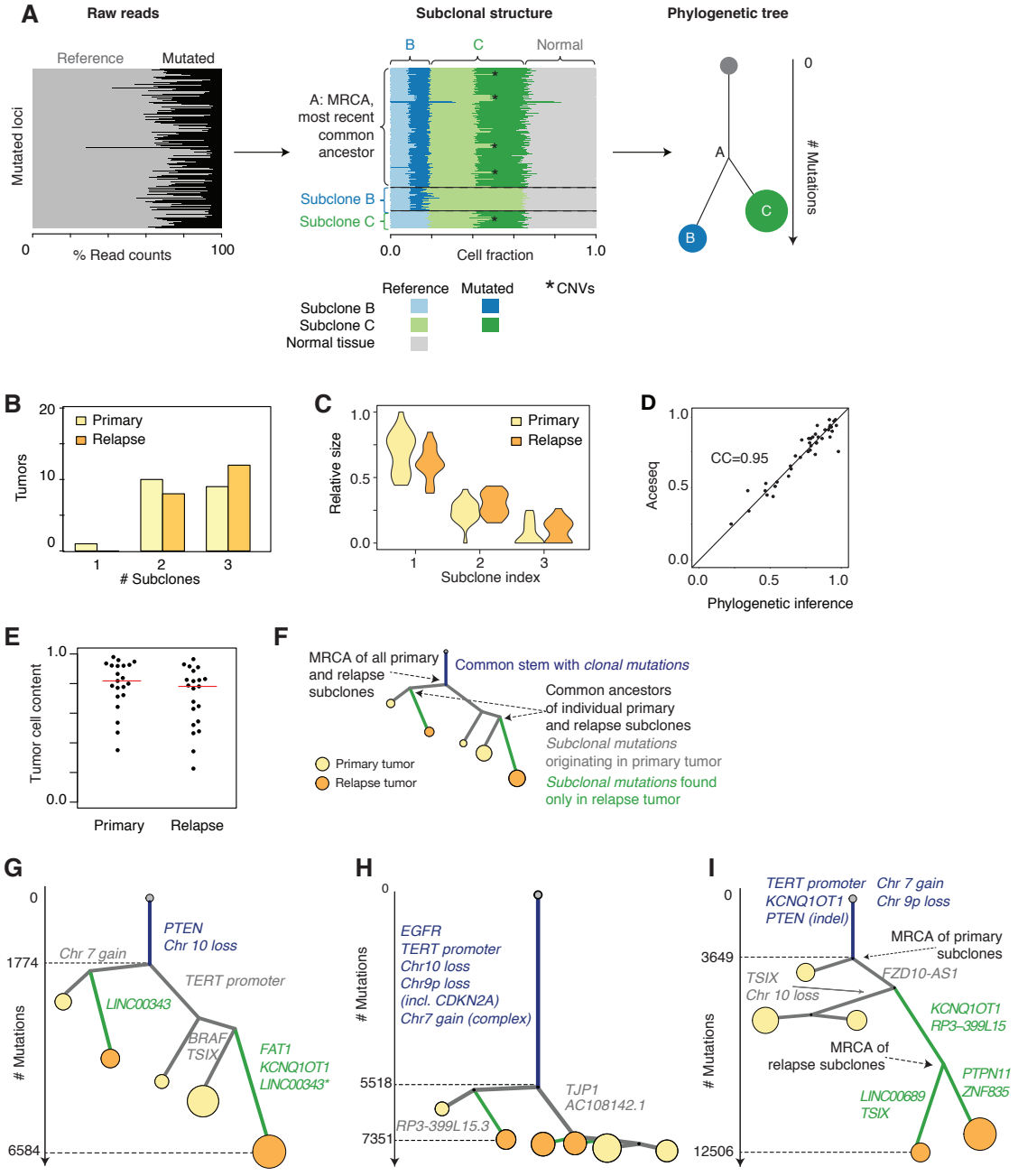
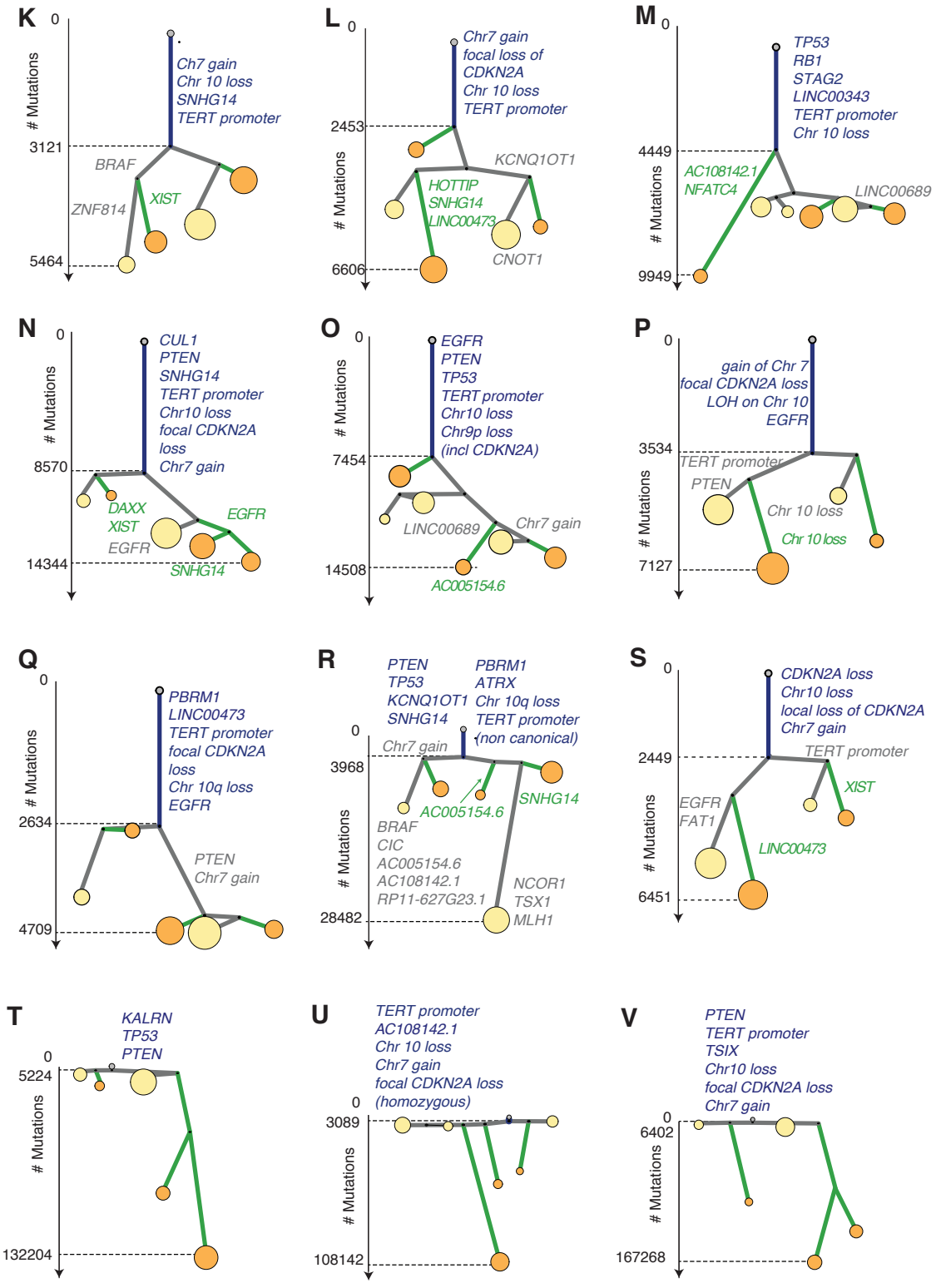


Figure 2.13: Phylogenetic inference. **A** Overview of the inference algorithm. Read count fractions at mutated loci before (left panel; rows correspond to individual loci, mutated read fraction is colored in black) and after sorting mutations to tumor subclones (middle panel; different colors indicate individual subclones, darker shades correspond to mutated read fraction; grey fraction corresponds to normal tissue infiltration). The corresponding phylogenetic tree is illustrated in the rightmost panel and consists of a common trunk that, at the most recent common ancestor (MRCA), 'A', branches into two subclones (blue and green).

Caption continues on page 51.


K
Mutations
0
3121
5464
Ch7 gain
Chr 10 loss
SNHG14
TERT promoter
BRAF
XIST
ZNF814
L
Mutations
0
2453
6606
Chr7 gain
focal loss of
CDKN2A
Chr 10 loss
TERT promoter
KCNQ1OT1
HOTTIP
SNHG14
LINC00473
CNOT1
M
Mutations
0
4449
9949
TP53
RB1
STAG2
LINC00343
TERT promoter
Chr 10 loss
AC108142.1
NFATC4
LINC00689
N
Mutations
0
8570
14344
CUL1
PTEN
SNHG14
TERT promoter
Chr10 loss
focal CDKN2A
loss
Chr7 gain
DAXX
XIST
EGFR
EGFR
SNHG14
O
Mutations
0
7454
14508
EGFR
PTEN
TP53
TERT promoter
Chr10 loss
Chr9p loss
(incl CDKN2A)
LINC00689
Chr7 gain
AC005154.6
P
Mutations
0
3534
7127
gain of Chr 7
focal CDKN2A loss
LOH on Chr 10
EGFR
TERT promoter
PTEN
Chr 10 loss
Chr 10 loss
Q
Mutations
0
2634
4709
PBRM1
LINC00473
TERT promoter
focal CDKN2A
loss
Chr 10q loss
EGFR
PTEN
Chr7 gain
R
Mutations
0
3968
28482
PTEN
TP53
KCNQ1OT1
SNHG14
PBRM1
ATRX
Chr 10q loss
TERT promoter
(non canonical)
Chr7 gain
AC005154.6
SNHG14
BRAF
CIC
AC005154.6
AC108142.1
RP11-627G23.1
NCOR1
TSX1
MLH1
S
Mutations
0
2449
6451
CDKN2A loss
Chr10 loss
local loss of CDKN2A
Chr7 gain
TERT promoter
XIST
EGFR
FAT1
LINC00473
T
Mutations
0
5224
132204
KALRN
TP53
PTEN
U
Mutations
0
3089
108142
TERT promoter
AC108142.1
Chr 10 loss
Chr7 gain
focal CDKN2A loss
(homozygous)
V
Mutations
0
6402
167268
PTEN
TERT promoter
TSIX
Chr10 loss
focal CDKN2A loss
Chr7 gain

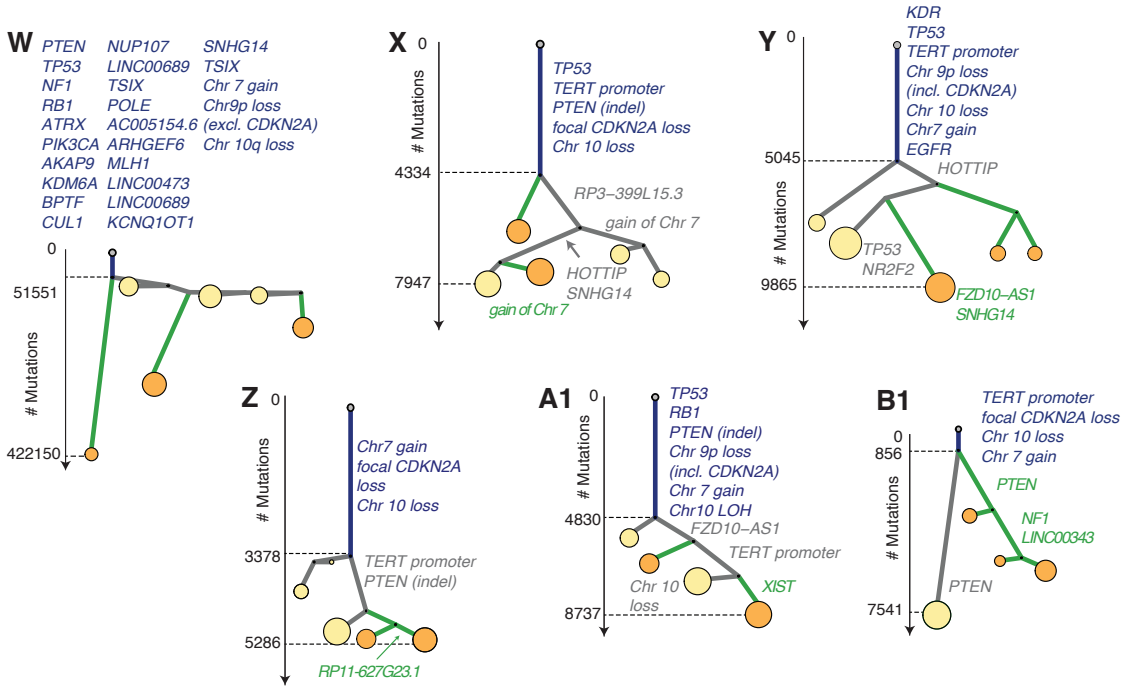


Figure 2.13 ***(continued)*****: B** Number of inferred subclones and **C**, corresponding relative sizes of the 21 tumor pairs analyzed. **D** Comparison of the estimated tumor cell content to the estimate based on ACEseq (performed by Jing Yang; *CC*, Pearson's correlation coefficient). **E** Estimated tumor cell content in primary and recurrent samples (red lines denote medians). **F** Display format for the phylogenetic trees. Primary subclones are colored in yellow, recurrent subclones in orange (circle areas scale with their relative sizes). Vertical branch lengths (green, relapse-specific branches; blue, common trunk) scale with the number of mutations. **G-I** Three examples of phylogenetic trees with an oligoclonal relapse origin (G, H) and a monoclonal origin (I; trees are designed as outlined in (F)). Putative driver mutations are indicated at their respective positions in the phylogenetic tree (asterisk marks convergent evolution). If mutations could not be unambiguously placed on the phylogenetic tree as clonal or subclonal, they were conservatively counted as clonal if the probability to place them in the tumor stem was >10%. **K-B2** Phylogenetic trees of the remaining tumors of the dataset (T-W, hypermutated cases; here only the clonal driver mutations are shown). Figure and legend modified from Körber et al., 2019; whole genome sequencing: Yonghe Wu; bioinformatic pre-processing: Jing Yang.

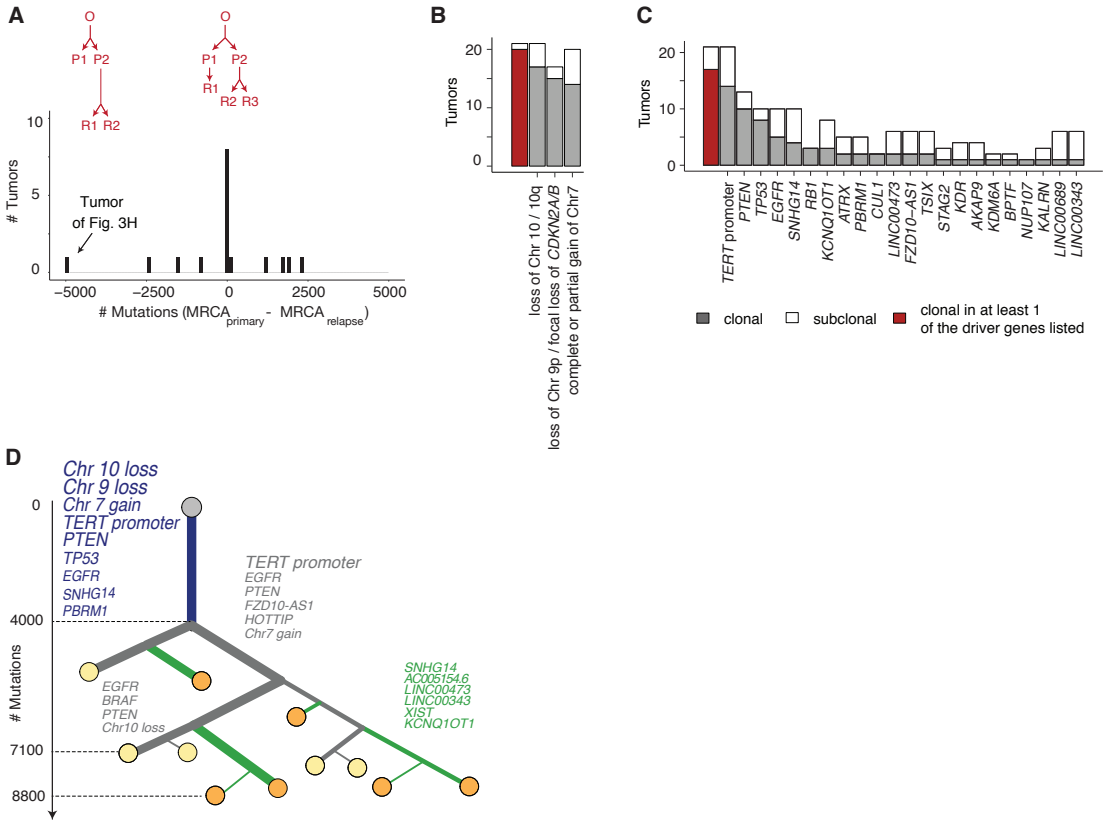


Figure 2.14: Phylogenetic metrics of the 21 glioblastoma pairs. **A** Origin of the recurrent tumor relative to the origin of the primary tumor. The distance between the most recent common ancestors (MRCA) of primary and recurrent samples is measured in mutation counts and provided as a measure for tumor origin. **B** Number of tumors with a clonal and subclonal copy number change on chromosome 7, 9 and 10. Number of tumors harboring at least one of the indicated copy number changes are displayed in darkred (if at least one change is clonal) or white (if none of the changes is clonal) on the leftmost side. **C** Number of tumors with a clonal and subclonal mutation (SNV or indel) in driver genes that were clonal in at least one tumor. Number of tumors harboring at least one of indicated mutations are displayed in darkred (if at least one change is clonal) or white (if none of the changes is clonal) on the leftmost side. If mutations could not be unambiguously placed on the phylogenetic tree as clonal or subclonal, they were conservatively counted as clonal if the probability to place them in the tumor stem was >10%. **D** Consensus tree over all sample pairs. Branch widths and font sizes scale with the number of cases supporting a connection and mutation, respectively. The median number of clonal mutations and the median of the maximal number of mutations per subclone are indicated for primary and recurrent tumors. Driver mutations are indicated at particular tree branches [clonal, subclonal, present in primary or recurrence only] if they are found there in at least two tumors and are frequent overall [present at any position in at least three (coding genes) or five (non-coding genes) among non-hypermutated tumors.] Figure and legend modified from Körber et al., 2019; whole genome sequencing: Yonghe Wu; bioinformatic pre-processing: Jing Yang.

2.3.5 Dynamics of tumor growth

Phylogenetic inference can be coupled with population dynamics models to analyze tumor growth behaviour beyond the ordering of mutations to different stages of tumorigenesis. In a simplistic view, the evolutionary dynamics during tumor growth are determined by two major processes, proliferation and mutation (Fig. 2.15A). As most mutations are acquired during cell divisions, and as mutations in driver genes affect proliferation, the two processes are mutually dependent. Nevertheless, if we assume (i) that the mutation rate is constant over time and (ii) that cells divide at a constant rate during tumorigenesis, single cells accumulate mutations according to a Poisson process and the number of mutations per cell thus serves as a mitotic clock. Denoting the division rate with λ and the mutation rate per cell division with μ, the average number of mutations per single cell, $m(t)$, reads

$$m(t) = \mu\lambda t. \tag{2.31}$$

Accordingly, the time it takes to accumulate $\hat{m}$ mutations in a single cell is

$$T(\hat{m}) = \frac{\hat{m}}{\mu\lambda}. \tag{2.32}$$

On the other hand, the number of effective cell divisions, i.e., the number of divisions leading to two surviving daughter cells, can be estimated from the number of tumor cells at resection. Assuming that tumors grow exponentially from a single founder cell, the number of tumor cells at time t, $N(t)$, is given by

$$N(t) = e^{\lambda\left(1-\frac{\delta}{\lambda}\right)t}, \qquad N(0) = 1, \qquad \lambda > \delta, \tag{2.33}$$

where λ and δ denote the cellular division and death rate, respectively. We solve again for the time it takes to grow the tumor to a known number of cells, $\hat{N}$:

$$T(\hat{N}) = \frac{\log(\hat{N})}{\lambda\left(1-\frac{\delta}{\lambda}\right)}. \tag{2.34}$$

At given estimates of m, N and μ, we now compute the ratio between cellular death and division rates, $\tilde{\delta} = \frac{\delta}{\lambda}$:

$$T(\hat{m}) \stackrel{!}{=} T(\hat{N}) \tag{2.35}$$

$$\tilde{\delta} = 1 - \frac{\mu}{\hat{m}} \log \hat{N}. \tag{2.36}$$

From the median mutation count between the MRCA of the tumor and the founder cells of primary subclones in the phylogenetic trees, we estimate that a single tumor cell acquired at least $\hat{m} = 2300$ mutations during tumorigenesis (Fig. 2.15B, C, excluding hypermutated cases). Note that this is a conservative estimate, as (i) tumor initiation might have started prior to the MRCA and (ii) the resolution of WGS is limited, so that most likely more mutations will be present in a single cell. Moreover, we estimate the number of tumor cells at resection to 10^9 cells, as previously suggested for the number of tumor cells per cubic centimeter (Del Monte, 2009; Devita Jr et al., 1975). This number agrees with a brain tumor size in the order of $20 - 80$ cm^3 (Goldberg-Zimring et al., 2005) and 10^{12} glia cells in a brain of 1500 cm^3 (Drachman, 2005; Herculano-Houzel et al., 2006; Milo et al., 2010; Pakkenberg and Gundersen, 1997). Finally, we take the somatic mutation rate as $(2.6 - 10.6) \times 10^{-9}$ mutations per base pair and cell division and account for 3.3×10^9 basepairs in the human genome. The lower bound of the mutation rate was estimated by Milholland et al., 2017, who measured mutation accumulation in single cell expansions; the upper bound accounts for a putative four-fold increase in the mutation rate during tumorigenesis, since the contribution of the clock-like mutational signature 1 drops from 75% among clonal mutations to 15% among tip-specific mutation (Fig. 2.15D). At these estimates, tumor size and mutational burden are only reconciled if the progeny generated by 69-92% of the cell divisions eventually died (Fig. 2.15E, F). Our data thus indicate that glioblastoma growth is accompanied by extensive cell death.

Survival advantage by *TERT* promoter mutations

The notion that most cell divisions during glioblastomas growth are unsuccessful emphasizes the need for survival-stabilizing mutations during tumorigenesis. In our dataset we observed that the vast majority of all tumors (19/21) acquired a 'canonical' *TERT* promoter mutation, which is associated with improved cellular survival via telomere elongation (Chiba et al., 2017). The remaining tumors had frameshift mutations in *ATRX*, associated with alternative lengthening of telomeres (Amorim et al., 2016). Together, this suggests a large selective advantage of survival-stabilizing mutations during glioblastoma growth. Moreover, as typically at least one pro-proliferative mutation (gain of *EGFR*, loss of *PTEN* and loss of *CDKN2A*) was clonal, whereas *TERT* promoter mutations were subclonal in a sizeable fraction of the dataset ($\geq 1/3$ of all tumors), the former presumably precedes the latter. To test this idea, we first confirm whether the inferred subclonality of *TERT* promoter mutations is a robust result. To do this, we compare the measured VAFs of the *TERT* promoter mutations to germline variants at loci with comparable sequencing coverage (median 123x and interquartile range [106, 140]) as a positive control for clonality. As discussed in Section 2.3.4, the measured VAFs must be corrected for normal tissue infiltration prior to analysis and hence, errors in the estimation of the tumor cell content are a major source for erroneous classificiation of a mutation as subclonal. Thus, in order to gain confidence in our estimate, we bootstrap the tumor cell content from the two independent estimates obtained with ACEseq and with phylogenetic inference, and compute the corrected

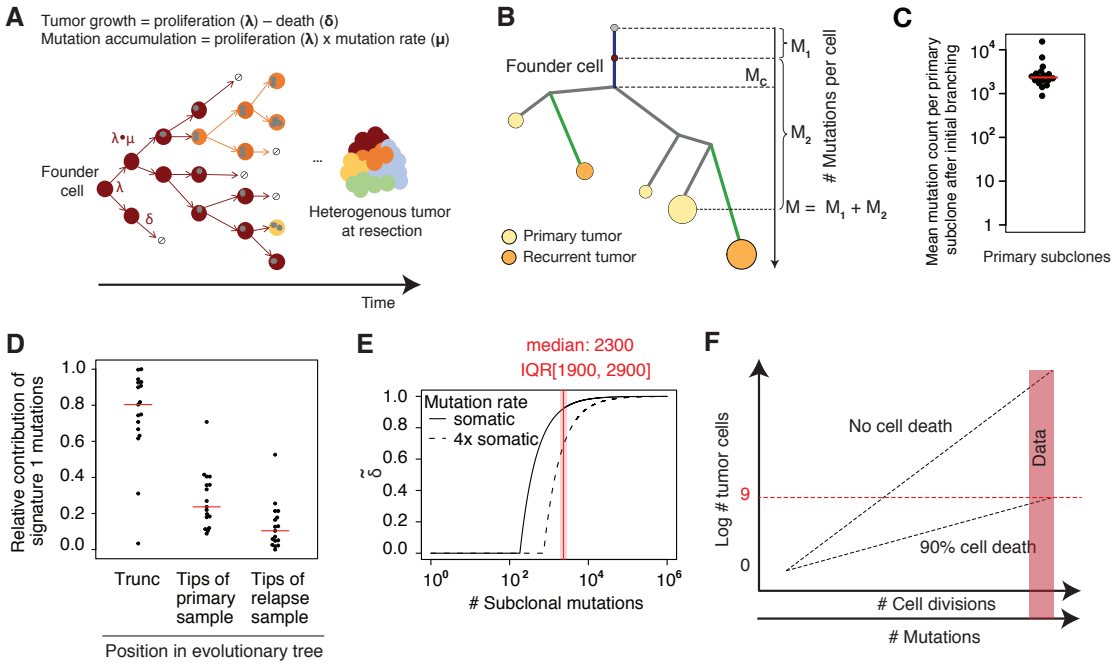


Figure 2.15: Dynamics of tumor growth and mutation accumulation. **A** Model illustration (stars, mutations; different colors, genetic subclones). **B** Estimation of the mutational burden per single cell from phylogenetic inference. A single tumor cell accumulates M_2 mutations during tumor growth. The number of mutations already present in the founder cell of the tumor, M_1, is not accessible, but can be conservatively estimated with the number of clonal mutations, M_c. With this, M_2 can be estimated to $M - M_c$. **C** Mean mutation count per primary subclone after initial branching, corresponding to $M - M_c$ in (B) (red line, median); hypermutated tumors are excluded. **D** Contribution of the clock-like mutational signature 1 to clonal mutations and to mutations present in the tips of the phylogenetic trees (excluding hypermutated tumors). **E** Extent of cell death (cellular death rate relative to cell division rate) required to reconcile subclonal mutation counts with a tumor size of 10^9 cells (red line, median number of subclonal mutations inferred from the data; shaded area, inter-quartile range). **F** Schematic illustration of the interplay between cell divisions, death and mutation accumulation. The number of tumor cells scales logarithmically with the number of cell divisions, while the number of cell divisions scales linearly with the number of mutations per cell. The slope of the growth curve decreases with the extent of cell death during growth and can be estimated from the combined information on the mutational burden and the tumor size. Figure and legend modified from Körber et al., 2019; whole genome sequencing: Yonghe Wu; bioinformatic pre-processing & mutational signature analysis: Jing Yang.

VAFs with the respective estimate. As shown in Fig. 2.16A the subclonal shoulder of the *TERT* promoter-mutant VAF distribution is robustly inferred independent of the respective tumor cell content estimate. This allows us to extend our model of glioblastoma growth to a two-step model of tumorigenesis (Fig. 2.16B). In a first step, cells divide at rates that are only marginally larger

than the death rate, meaning that cellular turnover is high. The growth law for this phase reads in accordance with Eqn. 2.3.5 (and re-scaling λ to 1)

$$N_{TERT^{\mathrm{WT}}}(t) = e^{(1-\tilde{\delta}_0)t}. \tag{2.37}$$

TERT promoter mutations can be acquired at each cell division at fixed probabilities according to the single-basepair substitution rate, μ, and improve cellular survival, hence reaching clonal dominance over time. The growth law for the number of cells in the tumor carrying the *TERT* promoter mutation is consequently given by

$$N_{TERT^*}(t) = \begin{cases} 0 & t < T_{0,TERT^*} \\ e^{(1-\tilde{\delta}_{TERT^*})(t-T_{0,TERT^*})} & t \geq T_{0,TERT^*} \end{cases}, \tag{2.38}$$

where $T_{0,TERT^*}$ denotes the time point at which the mutation is acquired in a single founder cell. The probability for at least one 'canonical' *TERT* promoter mutation, $P(TERT^*)$, increases with the number of division in the system, n_{div}, and thus reads

$$P(TERT^*) = 1 - (1-2\mu)^{n_{\mathrm{div}}}, \tag{2.39}$$

where the scaling factor of two accounts for the two canonical *TERT* promoter mutations recurrently observed in glioblastoma (chr5, 1,295,228 C>T and 1,295,250 C>T). As the number of cell divisions per time unit scales with the population size, we can express the total number of divisions in the system, n_{div} ,as a function of time:

$$\dot{n}_{\mathrm{div}} = \lambda N(t) \tag{2.40}$$

$$n_{\mathrm{div}} = \frac{1}{1-\tilde{\delta}_0}\left(e^{(1-\tilde{\delta}_0)t} - 1\right). \tag{2.41}$$

Inserting Eqn. 2.41 in Eqn. 2.39 then yields the probability distribution of at least one *TERT* promoter mutation at time point t.

In order to assess the selective advantage provided by *TERT* promoter mutations, we now need to compare the predicted distribution of the *TERT* promoter-mutant tumor fraction with the observed fraction in our dataset. To do this, we employ semi-stochastic simulations of the two-step model of glioblastoma evolution, scanning the parameter space over $\tilde{\delta}_0$ and $\tilde{\delta}_{TERT^*}$ and requiring $\tilde{\delta}_0 \geq \tilde{\delta}_{TERT^*}$ (thus assuring that the *TERT* promoter mutation decreases cellular death). At each value of $\tilde{\delta}_0$, we sample 1000 instances of the time point at which the *TERT* promoter mutation is acquired according to Eqn. 2.39. For each instance, we simulate the number of tumor cells with and without *TERT* promoter mutation according to Eqns. 2.37 and 2.38, and abort the simulation once the total tumor cell count equals 10^9 cells. We then compare the mean

and variance of the *TERT* promoter-mutant tumor fraction to their observed counterparts by computing the sum of the weighted squared residuals:

$$RSS = \frac{\bar{f} - \bar{f}^{\text{obs}}}{\sigma^2_{\bar{f}^{\text{obs}}}} + \frac{\text{Var}(f) - \text{Var}(f^{\text{obs}})}{\sigma^2_{\text{Var}(f^{\text{obs}})}}, \tag{2.42}$$

$$f = \frac{N_{TERT^*}(t_{\text{res}})}{N_{TERT^*}(t_{\text{res}}) + N_{TERT^{\text{WT}}}(t_{\text{res}})}, \tag{2.43}$$

where f and f^{obs} denote the simulated and observed *TERT* promoter-mutant tumor fraction at the time point of resection, t_{res}, respectively. The standard deviation of the mean and variance of f^{obs} ($\sigma_{\bar{f}^{\text{obs}}}$ and $\sigma_{\text{Var}(f^{\text{obs}})}$, respectively) is assessed by bootstrapping the measured data 10,000 times. The best fit corresponds to the parameter combination that minimizes Eqn. 2.42. Fits with an $RSS \leq RSS_{\text{min}} + 5.99$ lie within the 95% confidence interval of the best fit, according to a Chi-squared distribution with two degrees of freedom.

As illustrated in Fig. 2.16C, D, the exact parameter combination is not identifiable from our data, but rather we obtain good fits for $\delta_0 \geq 0.87$ at both normal and four-fold elevated mutation rates. However, since we estimated that on average 69% - 92% of the cell divisions fail to produce two surviving daughter cells (four-fold elevated and normal mutation rate, respectively; see Section 2.3.5), and since two thirds of all tumors harbor the *TERT* promoter mutation clonally, we require $\tilde{\delta}_{TERT^*} \geq 0.92$ at normal mutation rates and $\tilde{\delta}_{TERT^*} \geq 0.69$ at elevated mutation rates (corresponding to the black lines in Fig. 2.16C, D). With this we infer that the *TERT* promoter mutation reduces cell death by 6 to 26% (Fig. 2.16E), which translates to an increase in the effective division rate by a factor of four to five,

$$\text{normal mutation rate:} \quad \frac{1 - \tilde{\delta}_{TERT*}}{1 - \tilde{\delta}_0} = \frac{1 - 0.92}{1 - 0.98} \approx 4,$$

$$\text{four-fold elevated mutation rate:} \quad \frac{1 - \tilde{\delta}_{TERT*}}{1 - \tilde{\delta}_0} = \frac{1 - 0.69}{1 - 0.942} \approx 5,$$

suggesting that *TERT* promoter mutations speed up the dynamics of tumor growth tremendously. In this view, they will eventually reach clonality in all tumors but may still be subclonal at clinical detection.

Finally, we define the selective advantage associated with tumor cells before and after the *TERT* promoter mutation, s_0 and s_{TERT^*}, as the factor by which the decision between division and death is biased towards division (Bozic et al., 2016):

$$\frac{\tilde{\delta}_0}{\lambda + \tilde{\delta}_0} = \frac{1}{2}(1 - s_0) \tag{2.44}$$

$$\frac{\tilde{\delta}_{TERT^*}}{\lambda + \tilde{\delta}_{TERT^*}} = \frac{1}{2}(1 - s_0)(1 - s_{TERT^*}). \tag{2.45}$$

Solving for s_0 and s_{TERT^*} and setting $\lambda = 1$, as before, yields

$$s_0 = 1 - 2\frac{\tilde{\delta}_0}{1 + \tilde{\delta}_0}, \tag{2.46}$$

$$s_{TERT^*} = 1 - \frac{\tilde{\delta}_{TERT^*}}{1 + \tilde{\delta}_{TERT^*}} \frac{1 + \tilde{\delta}_0}{\tilde{\delta}_0}. \tag{2.47}$$

According to this definition, our estimated rates of $\tilde{\delta}_0$ and $\tilde{\delta}_{TERT^*}$ suggest that *TERT* promoter mutations provide a selective advantage of 0.03 - 0.16. This is very large in comparison to the average selective advantage of driver mutations of 0.004 (Bozic et al., 2016).

Tumor age

Our estimates of the cellular death rates relative to cell division rates during glioblastoma growth are informative on the total number of cell divisions during tumorigenesis. Finally, we attempt to link these estimates to real time, using the time span between primary and secondary resection to roughly estimate the cell division rate. To this end, we take the difference between the maximal number of mutations accumulated in primary and relapse subclones, respectively, as a proxy for the number of mutations which a single cell accumulated between the two surgeries. Excluding hypermutated tumors and six cases in which the difference was negative, we estimate a median of 1600 mutations and an inter-quartile range of [1000, 5400] mutations that a single cell accumulated between the two surgeries. Dividing these numbers by the mutation rate and the time span between the surgeries yields an estimate of the division rate in real time. We obtain a median of 0.17 d^{-1} and an inter-quartile range of [0.08, 0.23] d^{-1} if, as before, a four-fold increased mutation rate of $4 \times 0.27 \times 10^{-9}$ is assumed at tumor resection. Thus, at the lower bound, tumor cells divide approximately once in ten days. Assuming a relative cell death rate of $0.69 \leq \frac{\lambda}{\delta} \leq 0.92$ and a tumor size of 10^9 cells, as before, we estimate an upper bound of two to seven years of glioblastoma growth by solving

$$e^{(\lambda-\delta)t} \stackrel{!}{=} 10^9. \tag{2.48}$$

Thus our data suggest that IDH^{WT} glioblastomas evolve for several years before being clinically detected.

2.4 Discussion

The prognosis for patients with IDH^{WT} glioblastoma is exceptionally poor and has been unimproved for years. As most patients succumb to the rapidly relapsing tumor within a few months to years after diagnosis, an improved understanding of the clonal dynamics underlying

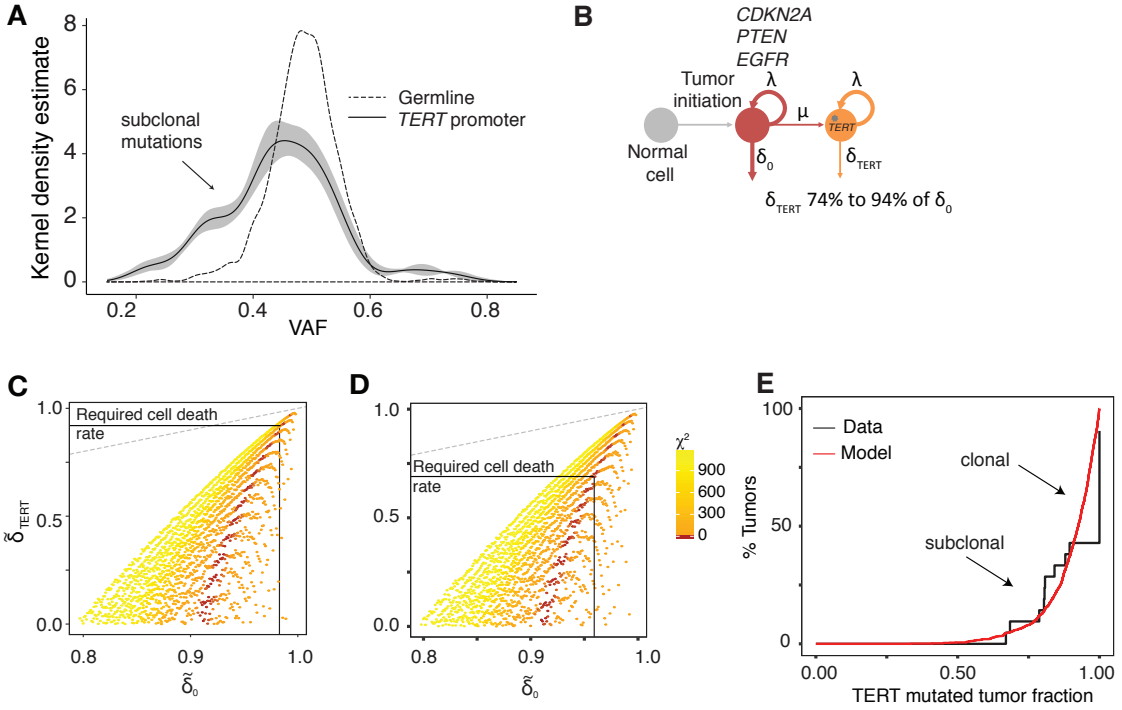


Figure 2.16: Selective advantage by 'canonical' *TERT* promoter mutations. **A** Variant allele frequencies (VAFs) of *TERT* promoter mutations after correction for tumor cell content (mean and 95% confidence interval of the kernel density estimate after sampling the tumor cell contents 10,000 times from the estimates from phylogenetic inference and ACEseq; tumors with non-neutral copy number at the *TERT* promoter were excluded) and of germline mutations from loci with read coverage within the inter-quartile-range of the coverage at the *TERT* promoter (IQR=[106,140], down-sampled to 1,000 loci for better visualization). **B** Two-step model of tumorigenesis with an initial phase of rapid turnover and reduced cellular death after random acquisition of the *TERT* promoter mutation. **C, D** Residual sum of squares when comparing the mean and variance of measured and simulated *TERT* promoter-mutant tumor fractions at different values of $\tilde{\delta}_0$ and $\tilde{\delta}_{TERT^*}$ (grey dashed lines, bisetrices). For each point the simulated mean and variance of 1000 simulations were compared to the measurement. Shown are the results for the somatic mutation rate (C) and the four-fold somatic mutation rate (D). Estimates lying within the 95% confidence interval of the best fit are shaded in dark red. Horizontal black lines mark the minimal cell death rate (relative to cell divisions), required to reconcile tumor sizes of 10^9 cells with mutation accumulation during glioblastoma growth (c.f. Fig. 2.15). **E** Measured (black) and simulated (red) *TERT* promoter mutant tumor fractions at the best-fit parameters. Figure and legend modified from Körber et al., 2019; whole genome sequencing: Yonghe Wu; bioinformatic pre-processing: Jing Yang.

glioblastoma growth and recurrence is needed to guide the development of novel treatment approaches. The objective of this work was therefore to characterize the genomic evolution of IDH^{WT} glioblastomas before and after standard therapy. To this end, a likelihood-based multinomial model for phylogenetic reconstruction from deep whole genome sequencing data

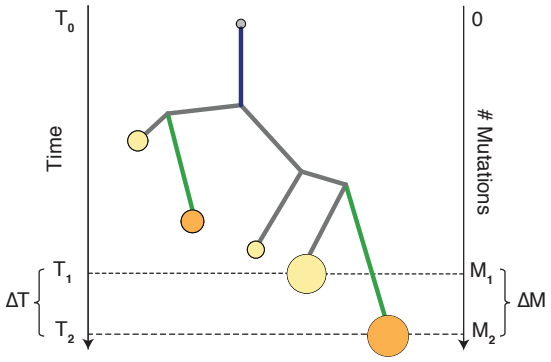


Figure 2.17: Estimation of the cellular division rate from the time span between primary and secondary resection, ΔT, and the number of mutations acquired during this time span, ΔM. Note that ΔM measures the number of mutations acquired in a single cell.

was developed and applied to primary and relapse sample pairs from 21 glioblastoma patients. This revealed a common path of early tumorigenesis and case specific tumor recurrence from an oligoclonal origin. Moreover, coupling mutation counts and tumor size with population dynamics models allowed an estimation of the tumor age and a characterization of the clonal dynamics during glioblastoma evolution.

In Section 2.2, we developed a likelihood-based multinomial model for phylogenetic inference from whole genome sequencing data. The model attempts to infer major subclonal branchings from the combined information of variant allele frequencies and coverage ratios at SNVs and small indels. In contrast to most existing frameworks for phylogenetic inference (e.g., Deshwar et al., 2015; Oesper et al., 2013; Roth et al., 2014), the here introduced model accounts for both subclonality of SNVs and CNVs without prior information other than the readcounts and the coverage ratios. This provides a high level of control over the assumptions flowing into phylogenetic inference, as no additional computational tools are required to run the anaylsis. Trained and evaluated on simulated data and a cell line mixture experiment, the strenghts and weaknesses of this approach become apparent. The power of the methodology lies in the inference of major subclonal branching, reliable estimation of the tumor cell content and solid identification of clonal mutations, without prior assumptions on the growth behaviour of the tumor. The performance improves with the subclonal size and the number of mutations supporting a subclone, whereas subclones that are characterized by few mutations are likely missed, as our model selection criterion favours parsimonious solutions. Likewise, mutations with small VAFs cannot be unambiguously assigned to subclones. This is due to the very nature of bulk whole genome sequencing, as mutual information between two variants is lost when shearing the DNA into short pieces. Therefore, accurate resolution of very small subclones cannot be achieved by bulk sequencing and requires a single cell sequencing approach. Nevertheless, most fine-grained subclonal diversification is due to neutral mutations that are

of no interest for the current project, in which we primarily ask how major drivers shape the evolutionary trajectory of a tumor and how this trajectory is affected by standard therapy. Thus phylogenetic inference from bulk whole genome sequencing data with the here presented likelihood-based model provides the resolution required for our purpose.

In Section 2.3, we applied the inference algorithm to deep whole genome sequencing data from 21 pairs of primary and relapsed glioblastomas and typically found that (locally) re-growing tumors recover the genetic heterogeneity of the primary tumor. This was reflected in an oligoclonal origin of the recurrent tumor and little acquisition of driver mutations upon primary resection, contrasting with two previous studies that suggested highly branched evolution between primary and relapsed glioblastomas (Kim et al., 2015; Wang et al., 2016). The different conclusions drawn by these studies might be partially due to the characteristics of the patient cohort, as Kim et al. linked branched evolution to a distant tumor relapse, whereas our cohort was biased towards locally re-growing tumors (95%). Moreover, apparent subclonal branching between primary and recurrent tumors may in fact represent two subsamples of the same subclone that differ solely by neutral mutations, consistent with the idea of predominantly "neutral tumor evolution across cancer types" (Williams et al., 2016). Nevertheless, proof or disproof for selection under standard therapy is difficult as, in theory, each tumor may find its own way of adaptation. If this is the case, common trajectories of mutations associated with resistance to treatment will become apparent in very large datasets only. Overall, however, such a scenario is unlikely as most mutations are of little functional relevance (Bozic et al., 2016). Moreover, our observation that tumors typically re-grow from all branches of the primary tumor argues against directed selection under therapy.

In contrast to a paucity of driver mutations that were recurrently acquired after primary resection, we found highly recurrent driver mutations at tumor initiation; all but one tumor pair harbored gains of chromosome 7 or losses of chromosome 9p or 10 clonally, suggesting that these alterations might be tumor initiating events. This finding is supported by the high prevalence of these copy number changes among glioblastomas overall (Brennan et al., 2013; Gerstung et al., 2017) and by two studies that sequenced multiple samples within a single tumor and from matched primary/relapsed pairs (Brastianos et al., 2017; Sottoriva et al., 2013). As opposed to this, small mutations were overall more diverse and frequently found at subclonal levels. This included prominent driver mutations such as mutations in *PTEN*, *EGFR* and *TP53*, in good agreement with previous reports (Francis et al., 2014; Ozawa et al., 2014; Spiteri et al., 2018; Wang et al., 2016). *TERT* promoter mutations, despite being found in all tumors, were also subclonal in one third of the tumor pairs, suggesting that these mutations may be acquired later during tumorigenesis. This is a surprising finding, since *TERT* promoter mutations are discussed as tumor initiating events in glioblastoma (Barthel et al., 2018). However, subclonal levels of *TERT* promoter mutations have also been reported in thyroid cancer and meningioma (Juratli et al., 2017; Landa et al., 2016), and few studies have assessed the clonality of *TERT* promoter mutations based on variant allele frequencies, which is necessary to account for intratumoral

heterogeneity. Indeed, a recent study that analyzed the VAF distribution of driver mutations in glioblastoma suggested that *TERT* promoter mutations can also be subclonal in glioblastoma (Brastianos et al., 2017), supporting the temporal order of early copy number changes and subsequent *TERT* promoter mutations as suggested by our dataset.

In Section 2.3.5, we employed population dynamics models to combine the inferred evolutionary trajectories with tumor growth. We showed that mutation accumulation serves as a mitotic clock that can be used to date back the tumor origin if calibrated with the time span between primary and secondary resection. Unexpectedly, the combined information on glioblastoma size, mutational burden and division rate suggested that glioblastomas have been growing for years before they were clinically detected. This contrasts with reports of glioblastoma or high-grade glioma patients who had negative MRI scans when presenting with seizures several months prior to initial diagnosis (e.g., Landy et al., 2000; Nishi et al., 2009). However, our population dynamics model suggests that glioblastomas grow below the clinical detection limit for years, since only a minority of cell divisions contributes to tumor growth. In this view, glioblastoma is initiated by mutations in cell cycle genes that accelerate cell divisions, but fail to stabilize the survival of the daughter cells. The three pervasive copy number changes identified at tumor initiation in our dataset (gain of chromosome 7, loss of chromsome 9p or 10) are attractive candidates for these mutations as all of them target pivotal cell cycle genes (*EGFR* on chromosome 7, *CDKN2A* on chromosome 9p and *PTEN* on chromosome 10). Despite extensive cell death during this initial phase, the rapid turnover of the tumor cell population increases the probability to capture a survival stabilizing mutation in the *TERT* promoter by chance. As the *TERT* promoter mutant tumor fraction grows faster than the founder population, it will eventually reach clonality, but, depending on the time it had to expand, may still be subclonal at resection.

Our population dynamics model indicates that the *TERT* promoter mutant tumor fraction should by tendency increase between primary and secondary resection. However, we found no such trend in our dataset and, in fact, only two of the seven cases with a subclonal VAF of the *TERT* promoter mutation progressed to clonality under treatment. Two recent studies have associated *TERT* promoter mutations with better prognosis in *MGMT* promoter methylated glioblastomas (Arita et al., 2016; Nguyen et al., 2016), raising the question of how the selective advantage associated with *TERT* promoter mutations changes under therapy. As *TERT* promoter mutations have been shown to support the survival of differentiated cells, but not of stem cells (Chiba et al., 2015), it will be interesting to stratify the fate of *TERT* promoter mutant cells based on transcriptional and epigenetic heterogeneity. Indeed, the large extent of cell death during tumor growth predicted by our population dynamics model supports the idea of a cellular hierarchy in glioblastoma. In this view, tumor growth is driven by self-renewing divisions in a stem cell-like compartment, while transiently amplifying cells in more downstream compartments survive for a limited time span. A recent study on the clonal dynamics of glioblastoma xenografts corroborates this interpretation by estimating approximately 10-15% symmetric divisions among

stem cell-like glioblastoma cells (Lan et al., 2017). Together, these findings call for a combined analysis of genetic evolution and cellular hierarchies to unravel treatment failure in glioblastoma.

In summary, the presented project provides a comprehensive analysis on the genomic evolution of IDH^{WT} glioblastomas. Linking genomic data with population dynamics models, a common path of early tumorigenesis was reconstructed that happens years ahead of diagnosis and is characterized by large chromosomal changes on chromosome 7 (gains), 9p or 10 (losses). This common early path is accompanied by rapid cellular turnover, followed by the acquisition of *TERT* promoter mutations that stabilize cellular survival and eventually reach clonal dominance. Upon primary resection and standard therapy, glioblastomas re-grow oligoclonally from residual cancer cells, with no evidence for a directed evolution. These findings are of clinical relevance as they suggest that treatment resistance of relapsing glioblastomas might be due to transcriptional or epigenetic heterogeneity rather than genomic evolution. Moreover, the long phase of glioblastoma evolution prior to clinical presentation may open up new opportunities for early detection and treatment design.

CHAPTER 3

Mutation accumulation in growing and homeostatic tissues

While mouse models have played a pivotal role in cancer research they have been of very limited utility in understanding how cancer arises initially, without the artificial introduction of oncogenes by the experiments. It is therefore of particular interest to understand how tumors arise *de novo* and how this knowledge can be used to guide early diagnosis. Mutational profiling with whole genome sequencing provides a minimally invasive way to study clonal dynamics in human tissues and may aid this understanding. However, accurate classification of clonal expansions as pre-cancerous requires reliable distinction from expansions expected under neutral drift, and thus a thorough understanding of the clonal dynamics in normal tissues.

In this chapter, we will develop a theoretical expectation of the mutation frequency distribution in growing and homeostatic tissues, exemplified by adult hematopoiesis. In the first part, biological and theoretical background on hematopoiesis and mutation acquisition is provided. Next, existing theory on mutational dynamics in exponentially growing tissues is reviewed. We extend this model to a two-stage situation of initial tissue expansion, followed by a phase of homeostatic turnover. The theoretical expectation is then compared to deep whole genome sequencing data of granulocytes from healthy mice, and of leukocytes from neuroblastoma and glioblastoma patients. Finally, the potentials and limitations of this approach to identify pre-leukemic states in peripheral blood are discussed.

The data used in this chapter were provided by different collaborators. All murine data were generated by Ruzhica Bogeska and bioinformatically pre-processed by Megan Druce, both from the group of Michael Milsom (German Cancer Research Center, Heidelberg). Human data were obtained from neuroblastoma and glioblastoma patients, whose blood samples were originally sequenced as normal controls for tumor sequencing. The glioblastoma data were collected and sequenced by the SysGlio Consortium, headed by Peter Lichter (German Cancer Research Center,

Heidelberg). The neuroblastoma data were collected and sequenced by Frank Westermann, Moritz Gartlgruber and Sabine Hartlieb (German Cancer Research Center, Heidelberg). All human sequencing data had already been mapped to the human genome, while I performed all downstream bioinformatic analyses.

3.1 Background

3.1.1 Brief introduction to the hematopoietic system

The formation of blood and immune cells is a hierarchically organized process termed hematopoiesis. Hematopoietic stem cells (HSCs) reside at the apex of the hematopoietic system and are capable of self-renewal and multi-lineage output of all cell types of the blood (Fig. 3.1; Dick et al., 1985; Ford et al., 1956; Keller et al., 1985; Pei et al., 2017). Lineage differentiation runs through multiple intermediate states of increasing proliferative activity and decreasing capacity to self-renew (Busch et al., 2015). Myeloid and lymphoid lineages split from a multipotent progenitor state (MPP; Akashi et al., 2000; Kondo et al., 1997).

During embryogenesis, hematopoiesis develops in two waves. The first hematopoietic progenitors are produced in the yolk sac at day E7.5 of murine embryogenesis and give rise to primitive erythrocytes and myeloid cells (Golub and Cumano, 2013), while the second wave of embryonic hematopoiesis is initiated in the aortic-gonado-mesonephros at day E9.5. From there, hematopoietic progenitors migrate to the fetal liver, where the definitive hematopoiesis takes place (Medvinsky et al., 1993; Samokhvalov et al., 2007; Zovein et al., 2008). Starting from day E17.5, hematopoietic stem cells colonize the bone marrow and fully establish adult hematopoiesis by week three after birth (Bowie et al., 2007; Christensen et al., 2004).

3.1.2 Clonal hematopoiesis

A prominent example of a pre-cancerous lesion that may progress to malignancy is 'clonal hematopoiesis' in the elderly. Clonal hematopoiesis refers to large expansions of single hematopoietic stem cells (typically $\geq$ 20%; Jaiswal et al., 2014) and has been reported for >50% of subjects older than 85 years (Zink et al., 2017). Clonal hematopoiesis was first detected from skewed X-inactivation in the blood of elderly females (Busque et al., 1996) and later confirmed with next generation sequencing of larger cohorts (e.g., Genovese et al., 2014; Jaiswal et al., 2014). These studies identified known leukemic driver mutations, primarily in the DNA (de-)methylation genes *DNMT3A* and *TET2* and the chromatin remodeling gene *ASXL1*, at high frequencies in the blood of elderly people. Follow up studies revealed an elevated risk for initially asymptomatic persons with clonal hematopoiesis to develop hematological cancers and also coronary heart disease and stroke (Genovese et al., 2014; Jaiswal et al., 2014). Surprisingly, a recent study showed that a substantial fraction of individuals with clonal hematopoiesis does not harbor any known driver muation, but nevertheless has a higher overall mortality risk (Zink

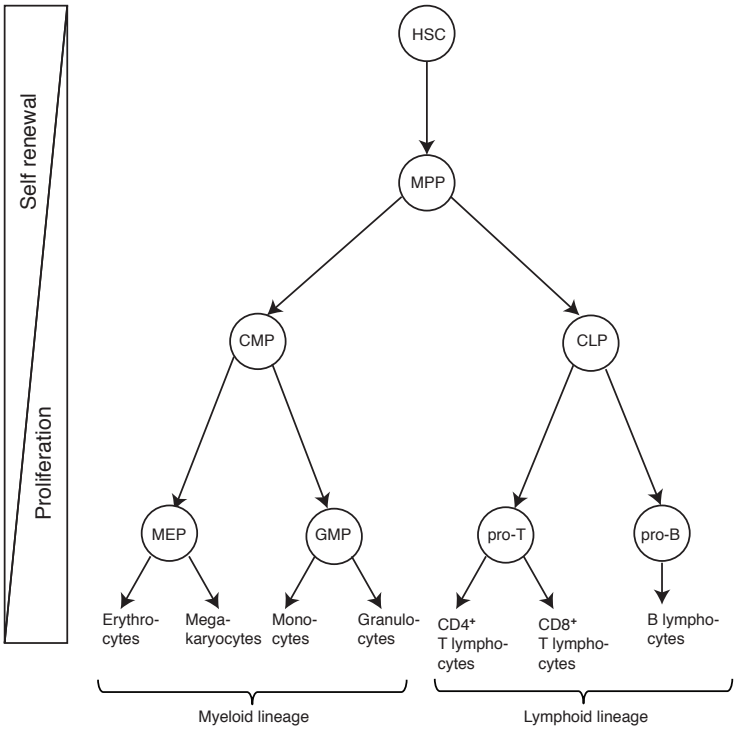


Figure 3.1: Hematopoietic lineage tree. Hematopoietic stem cells (HSC) reside at the top of the hematopoietic tree. Myeloid and lymphoid lineages split after differentiation of multipotent progenitors (MPP) into common myeloid progenitors (CMP) or common lymphoid progenitors (CLP). MEP, megakaryocyte-erythrocyte progenitor; GMP, granulocyte-monocyte progenitor; pro-B, pro-B cells; pro-T, pro T cells. Illustration based on Figure 1 in Graf, 2008.

et al., 2017). Whether clonal hematopoiesis in these individuals was due to neutral drift or due to a selective advantage provided by unknown driver mutations remains unknown, in particular, because the mutation dynamics in undisturbed hematopoiesis are not well understood (Lee-Six et al., 2018).

3.1.3 Mutagenesis in dynamic tissues

In the absence of external mutagens, most mutations are acquired during replication or from spontaneous deamination of methylated cytosines to thymidines. As replication is linked to cell division, accumulation of replication errors can be interpreted as a mitotic clock. By contrast, accumulation of mismatches due to spontaneous deamination is independent of cell divisions and thus represents a chronological rather than a mitotic clock (Gao et al., 2016). If, however, division rates are (approximately) constant over time, a detailed distinction of both contributions will not be needed and both processes can be added up to obtain an effective mutation rate per

cell division.

Estimates of the somatic mutation rate in human somatic tissues range from 0.27×10^{-9} (Lynch, 2010) to 2.66×10^{-9} (Milholland et al., 2017) mutations per nucleotide and cell division. This translates to one to ten mutations per cell division in the human genome of approximately 3.3×10^9 basepairs. Thus mutations are frequent enough to label most cell divisions and rare enough to be unique cellular markers (Kimura, 1969).

In an exponentially growing tissue of monoclonal origin, mutations present in the founder cell will be inherited to its entire progeny. In every subsequent division, new mutations will be acquired, but these will be present in a sub-tree of the total division tree only (Fig. 3.2A). On average, the tissue fraction sharing a newly acquired mutation halves with every generation, while the number of newly acquired mutations doubles with every generation. However, deviations from average are possible due to neutral drift or selection of a favourable mutation. Thus the dynamics of cell division and death imprint on the mutation frequency distribution of the tissue.

A straightforward approach to measure the mutation frequency distribution of a tissue sample is bulk whole genome sequencing (WGS). WGS identifies mutations present in a substantial fraction of the sample, albeit single cell information is lost. In addition, WGS reports the fraction of sequencing reads harboring a specific mutation, which provides a direct estimate of the variant allele frequency (VAF), the relative frequency of a mutation in the tissue. In diploid genomes, mutations are typically present on one of the two alleles only, so that the fraction of cells carrying a heterozygous mutation scales 2:1 with the VAF. Therefore, a typical VAF histogram measured in a monoclonal, diploid cell population consists of a clonal peak at 50% and a subclonal tail at small VAFs, as shown schematically in Fig. 3.2B.

3.2 Experimental motivation: Mutation frequency distribution in normal blood samples

We start with a qualitative analysis of the mutation frequency distribution in peripheral blood cells sampled from mice and humans of different age groups (Fig. 3.3). The data comprise WGS data of murine granulocytes from one eight-week-old and one two-year-old mouse as well as WGS data of unsorted human leukocytes from 68 children diagnosed with neuroblastoma and 39 adults diagnosed with glioblastoma.

3.2.1 Sequencing and mutation calling

Murine samples

Sample preparation and mutation calling of the murine samples were performed by Ruzhica Bogeska (experimental part) and Megan Druce (sequence alignment and mutation calling), both from the group of Michael Milsom (German Cancer Research Center, Heidelberg). Briefly,

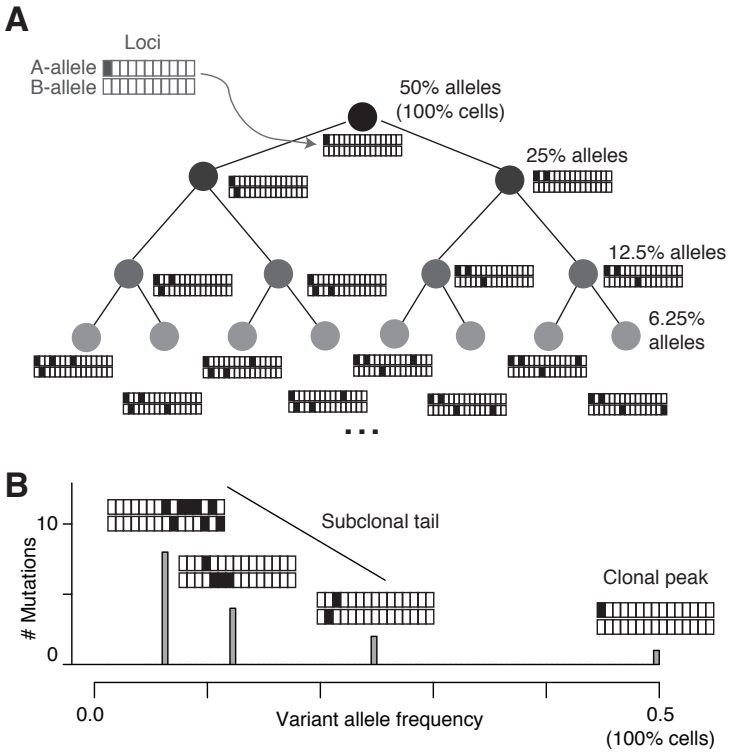


Figure 3.2: Accumulation of replication-coupled mutations in an exponentially growing tissue. **A** Mutation acquisition at a constant rate per cell division (barcodes represent DNA loci; black, mutated; white, wildtype). Percentages indicate the fraction of alleles harboring a specific mutational profile. **B** Variant allele frequency histogram of the example shown in (A; barcodes illustrate the mutations being present at a specific VAF).

$CD11b^+$ $Gr1^+$ granulocytes were sorted from the peripheral blood of one eight-week-old and one two-year-old C57/BL6 mouse, sequenced at 140x coverage and aligned to the murine reference genome mm10. Somatic mutations were called against control samples obtained from analogous preparation of tail cells of the same individuals.

Human samples

The human blood samples were taken from 68 neuroblastoma patients (median age = 4 ys, IQR=[2 ys, 6 ys]) and 39 glioblastoma patients (median age = 54 ys, IQR = [43 ys, 63 ys]), and originally served as germline controls when sequencing whole genomes of the tumors. Except for one glioblastoma patient, the samples were taken prior to treatment with chemoradiation therapy. The samples consisted of unsorted leukocytes, were sequenced at a target coverage of 80x and were aligned to the human reference genomes hg19 or hs37d5 by the DKFZ Omics IT and Data Management Core Facility. Somatic mutations were called with Strelka v2.8.4 (Saunders et al., 2012) at default parameters, using the corresponding tumor as a gemline

control, and annotated using annovar (Wang et al., 2010). Low-quality variants and mutations in repeat regions[1] were filtered upon mutation calling. Moreover, mutations at loci with copy number changes in the tumor (as determined with ACEseq; Kleinheinz et al., 2017; analysis performed by Jing Yang and Umut Toprak, German Cancer Research Center, Heidelberg) were excluded, since germline mutations on these loci may have been lost in the tumor. To achieve comparability between individual patients, the mutation counts were extrapolated from the analyzed genome fraction to the entire genome.

3.2.2 Mutation frequency distributions

Fig. 3.4A shows the mutation frequency distributions in murine granulocytes of the eight-week-old and the two-year-old mouse. Both histograms are consistent with a tissue expansion from a monoclonal origin, reflected in few clonal mutations at VAFs around 0.5 and increasing numbers of mutations at subclonal levels. Notably, the overall mutational burden approximately doubles from 235 mutations in the young mouse to 555 mutations in the old mouse. By contrast, the number of subclonal mutations (VAF < 0.25) in human samples is overall higher in children (median 342) than in adults (median 269), despite a high degree of inter-individual variability (Fig. 3.4B; p=0.017, Wilcoxon rank sum test with continuity correction). While these observations appear contradictory at a first glance, they may well be due to differences between the organisms, inter-individual heterogeneity and the type of data, as we will see later on. Without pre-empting the discussion, let us briefly summarize the available data basis:

- In mice, the mutational burden of granulocytes doubles with age. Murine data is under better control than human data, as both mice are of the same inbred strain and were kept in controlled environmental conditions. However, the data suffers from the lack of replicates.

- In humans, an increase in the mutational burden with age is not evident. Human data are available for more patients and have therefore more statistical power than the murine data. On the other hand, we can expect a higher variability among human samples as humans do not live in controlled environments, the blood samples were obtained from cancer patients (and the matching cancer samples were used to distinguish germline and somatic mutations), and the samples were obtained from unsorted leukocytes, so that a mixed cell population was analyzed.

[1] extracted from UCSC table browser, https://genome.ucsc.edu/cgi-bin/hgTables, selecting 'RepeatMasker' and 'Simple Repeats'; downloaded on December 5th, 2018.

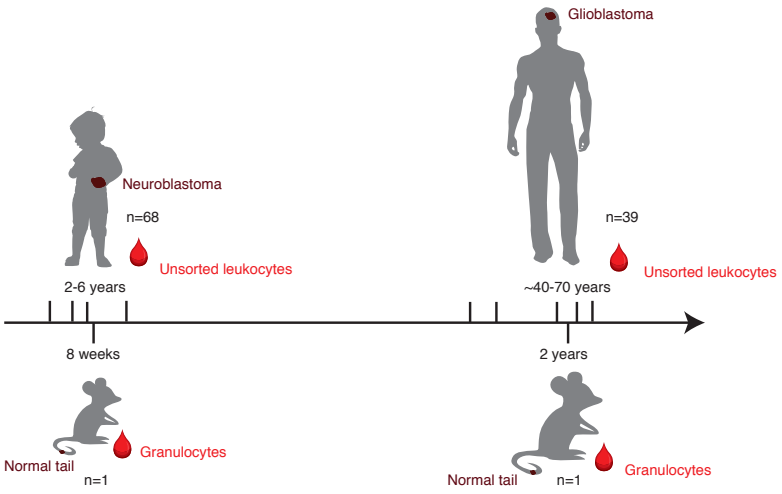


Figure 3.3: Whole genome sequencing of normal blood. Whole genomes of sorted granulocytes and unsorted leukocytes were sequenced from two mice (140x target sequencing coverage) and 107 humans (80x target sequencing coverage), respectively. For the murine samples, tail was sequenced as a germline control. The human blood samples were originally sequenced as germline controls for neuro- and glioblastomas. To call somatic variants in the blood, the samples were switched, i.e., the tumor was used as a germline control for the blood. Illustrations are modified from Vecteezy (www.vecteezy.com).

3.3 Neutral mutation accumulation in growing and homeostatic tissues

3.3.1 Exponential growth

To begin with, we recapitulate existing theory for an exponentially growing tissue with continuous accumulation of neutral mutations (Ohtsuki and Innan, 2017; Williams et al., 2016). Both approaches trace the fate of mutations that are acquired during cell divisions in a clonal expansion (Fig. 3.2). In the framework provided by Williams et al., 2016, cell death is neglected and all newly acquired mutations expand deterministically in the population. Ohtsuki and Innan refined this approach by accounting for the stochasticity in the clone size distributions. We will show that this refinement is necessary if cell loss during tissue expansion is not negligible.

Exponential growth with deterministic clonal expansions

We model tissue expansion from a single cell and assume that this expansion follows an exponential growth law at rate $\beta = \lambda - \delta$, where λ is the cellular division rate and δ the cellular death rate, with $\lambda > \delta$. Following Williams et al., 2016, we assume that all newly acquired mutations expand deterministically at rate β. The number of cells at time t, $N(t)$, then obeys

$$N(t) = e^{\beta t}. \tag{3.1}$$

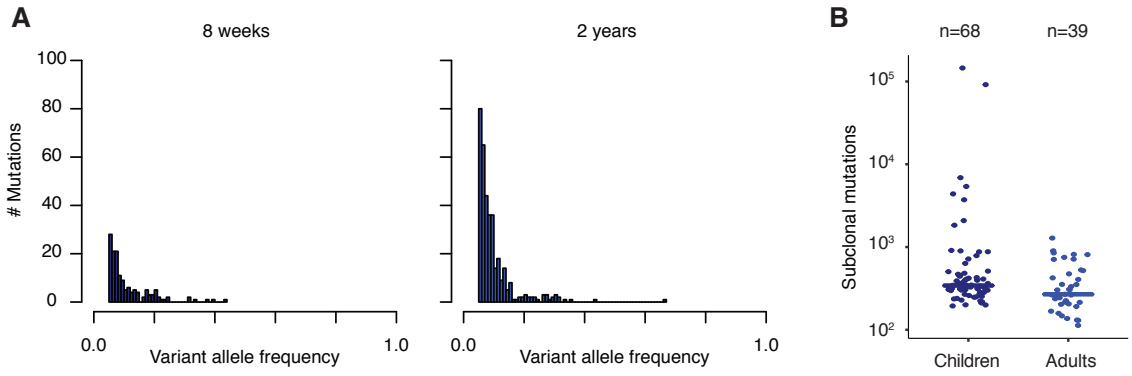


Figure 3.4: Mutation frequency distribution in the peripheral blood of mice and humans. **A** Variant allele frequencies of single nucleotide variants detected in peripheral granulocytes of an eight-week-old (left panel) and a two-year-old mouse (right panel). Mutations supported by less than five sequencing reads were excluded. **B** Subclonal mutation counts in unsorted leukocytes of human children and adults (requiring at least three reads supporting the variant and $0.05 < VAF < 0.25$; values are extrapolated from the analyzed genome fraction to the entire genome). Mouse data: Ruzhica Bogeska and Megan Druce; human data: Frank Westermann (children) and Peter Lichter & 'SysGlio' Consortium (adults).

We now assume that mutations are acquired at a constant rate per cell division, μ. The total number of mutations, M, is thus determined by

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \mu\lambda N(t), \tag{3.2}$$

which can be solved by integration to give

$$M(t) = \int_0^t \mu\lambda N(t')\mathrm{d}t' \tag{3.3}$$

$$= \frac{\mu\lambda}{\beta}\left(N(t) - 1\right). \tag{3.4}$$

As cells are assumed to divide deterministically, the tissue fraction containing a specific mutation will be constant over time. We introduce the tissue fraction of a single cell $f = \frac{1}{N(t)}$ and find the number of mutations with fraction f or larger:

$$M(f) = \frac{\mu\lambda}{\beta}\left(\frac{1}{f} - 1\right). \tag{3.5}$$

Thus the number of mutations present in a tissue fraction of f or larger grows linearly with $1/f$ (Fig. 3.5A, B).

We expect that treating all cells as proliferating exponentially in a deterministic setting has limitations. Therefore, we compare the theoretical prediction to stochastic simulations of

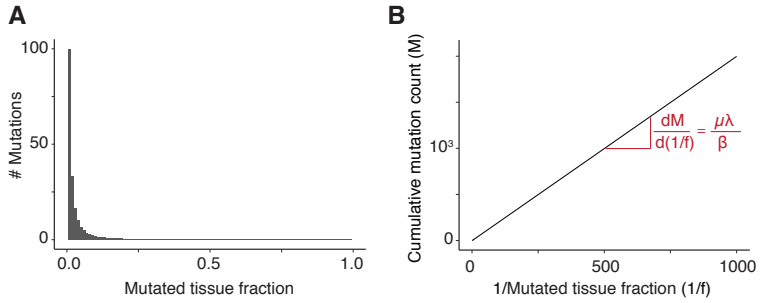


Figure 3.5: Predicted mutation accumulation in exponentially growing tissues following Williams et al., 2016. **A** Predicted mutation frequency histogram. **B** Predicted number of mutations present in at least the indicated tissue fraction (the slope of the curve scales with the mutation rate). The mutation rate was taken as 2 mutations per effective cell division (one per daughter cell).

mutation accumulation in a growing population at different death rates (recall that $\beta = \lambda - \delta$). To this end, we simulate the expansion of a single cell stochastically (Fig. 3.6A). The simulation works by tracking the mutational profiles of dividing cells in an agent-based fashion and is initiated with a single cell. At each time step, the cell to divide is randomly selected and inherits its mutational profile to both daughter cells. New mutations, acquired at division, are sampled from a binomial distribution for both daughter cells (using a sample size of 3×10^9, corresponding to the number of basepairs in the human genome, and a per-base substitution rate of 0.33×10^{-9} per division, corresponding to an average mutation rate of one mutation per daughter cell and division). The simulation is terminated once the population size reaches 1,000 cells.

In the absence of cell death, the theoretical prediction matches the simulated mutation frequency distribution accurately (Fig. 3.6B, black curve and red shading). However, with increasing cellular death rates, the simulated curve bends down from the theoretical prediction (Fig. 3.6B, orange shadings). Thus the linear prediction between the cumulative number of mutations and the inverse of the mutated tissue fraction will be accurate for neutral mutation accumulation only if cell death is negligible. By contrast, in tissues with high cellular turnover — due to death of or differentiation — a more refined approach is needed that properly accounts for the stochasticity of small cell numbers.

Exponential growth with stochastic clonal expansions

To account for neutral drift in mutation accumulation, we employ the theoretical framework developed by Ohtsuki and Innan, 2017. Ohtsuki and Innan extended the theoretical setting by Williams et al. to account for the stochasticity of clonal expansions. Like Williams et al., Ohtsuki and Innan model the growth of the total cell population deterministically:

$$N(t) = e^{(\lambda-\delta)t} \qquad \text{with} \qquad \lambda > \delta, \tag{3.6}$$

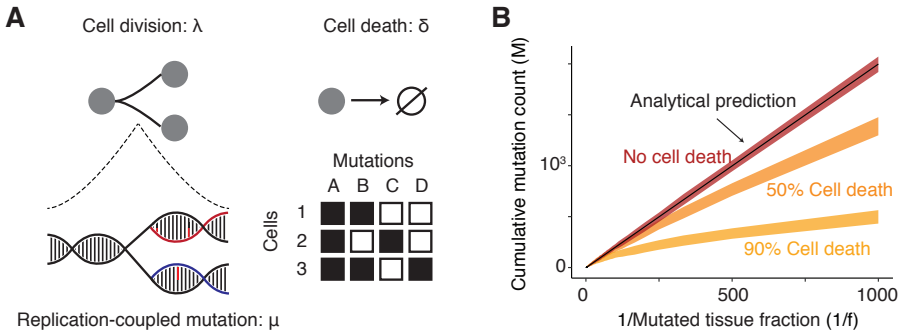


Figure 3.6: Stochastic simulation of mutation accumulation. **A** Simulation design. Cells either divide (rate λ) or die (rate δ). At cell divisions, a randomly selected cell inherits its mutational profile to both daughter cells. Newly acquired mutations are sampled according to the mutation rate μ. The mutational profile of each cell is stored in a mutation table. **B** Cumulative number of mutations according to the analytical prediction (black line; c.f. Fig. 3.5) and from stochastic simulations without cell death (darkred) and with $\delta = 0.5\lambda$ (darkorange) and $\delta = 0.9\lambda$ (orange). Shown are the 95% confidence bands of 100 simulations each. All simulations were terminated once a population size of 1,000 cells was reached; the mutation rate was taken as 2 mutations per effective cell division (one per daughter cell).

and rely on a deterministic description of mutation acquisition, as before:

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \mu\lambda N(t). \tag{3.7}$$

However, in contrast to Williams et al., Ohtsuki and Innan model clonal expansions of newly acquired mutations stochastically, using the transition probabilities of a supercritical linear birth-death process (i.e., $\lambda > \delta$). At constant birth and death rates, the probability that a mutation acquired at time t' will be present in a clone of size i at time t is given as (Bailey, 1964)

$$P_{\mathrm{exp},i}(t,t'|\lambda,\delta) = \begin{cases} x(t-t') & \text{if } i = 0 \\ (1-x(t-t'))\,(1-y(t-t'))\,y(t-t')^{i-1} & \text{if } i \geq 1 \end{cases}, \tag{3.8}$$

with

$$x(t) = \frac{\delta e^{(\lambda-\delta)t} - \delta}{\lambda e^{(\lambda-\delta)t} - \delta} \tag{3.9}$$

$$y(t) = \frac{\lambda e^{(\lambda-\delta)t} - \lambda}{\lambda e^{(\lambda-\delta)t} - \delta}. \tag{3.10}$$

Integration over time yields the total number of mutations present in i cells at time t, in the following denoted as $S_i(t|\mu,\lambda,\delta)$:

$$S_i(t,|\mu,\lambda,\delta) = \int_0^t \overbrace{P_{\mathrm{exp},i}(t,t'|\lambda,\delta)}^{\text{Neutral drift during expansion}} \times \underbrace{\mu\lambda N(t')}_{\text{Number of mutations acquired at } t'} \mathrm{d}t'. \tag{3.11}$$

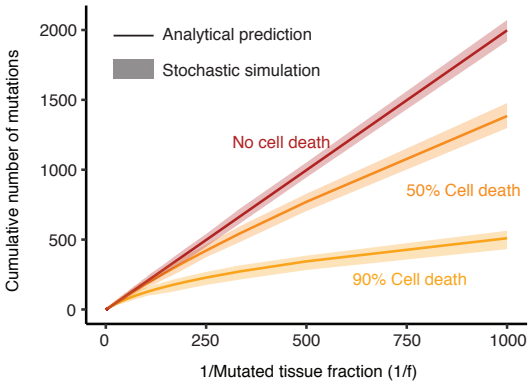


Figure 3.7: Predicted cumulative number of mutations (present in at least the indicated tissue fraction) if accounting for stochastic clonal expansions (following Ohtsuki and Innan, 2017). Shown are the theoretical predictions (lines) for expanding populations without cell death (darkred) and with $\delta = 0.5\lambda$ (darkorange) and $\delta = 0.9\lambda$ (orange), along with the 95% confidence bands of corresponding stochastic simulations (shaded areas, 100 simulations each). Simulations are the same as shown in Fig. 3.5 (terminated at a population size of 1,000 cells; μ =2 mutations per effective cell division).

Finally, expressing the tissue fractions harboring individual mutations as $f(t) = \frac{i}{N(t)}$, we obtain the number of mutations, $M(f)$, present in a tissue fraction of at least f at time t:

$$M(f, t|\mu, \lambda, \delta) = \sum_{i=fN(t)}^{N(t)} \int_0^t P_{\exp,i}(t, t'|\lambda, \delta)\mu\lambda N(t')\mathrm{d}t', \tag{3.12}$$

where $f = \frac{0}{N(t)}, \frac{1}{N(t)}, \ldots, 1$. Note that, in contrast to the theoretical framework proposed by Williams et al., 2016, the mutation frequency-distribution is time-dependent, which is due to stochastic drift.

When comparing the analytical prediction of Ohtsuki and Innan to stochastic simulations of mutation accumulation in growing cell populations (c.f. Section 3.3.1), we note that Eqn. 3.12 provides accurate predictions for scenarios with and without cellular death (Fig. 3.7). We note that this agreement of the theory with stochastic simulations was achieved by describing stochastic clonal drift while retaining a deterministic picture of mutation accumulation.

3.3.2 Homeostatic turnover

While the theoretical frameworks by Williams et al., 2016, and Ohtsuki and Innan, 2017, model the dynamics of mutation accumulation in exponential tissue expansions as occuring during embryogenesis or carcinogenesis, homeostatic turnover in adult tissues translates to a balanced birth-death process that yields a constant population size over time. Therefore, in order to predict the mutation frequency distribution in adult hematopoietic tissues, we will now couple an early

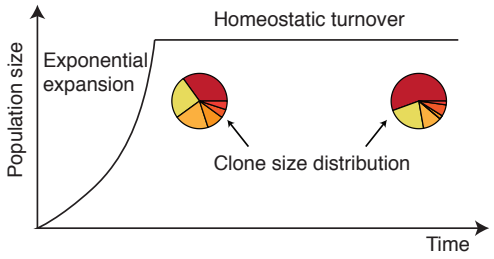


Figure 3.8: Two-stage tissue dynamics with an exponential expansion phase that is followed by a phase of homeostatic turnover. Piecharts represent clone size distributions if following the progeny of labeled cells over time. Due to neutral drift dynamics, the proportion of individual clones may change during homeostatic turnover.

exponential growth phase, translating to embryonic tissue expansion, to a subsequent steady-state phase, translating to homeostatic turnover in the adult (Fig. 3.8). To achieve this, we will first analyze neutral drift under balanced division and death rates and eventually link this to a preceding phase of exponential growth.

To model the clonal dynamics of mutations during homeostatic tissue turnover, we assume that cells divide at rate λ and die or differentiate at rate $\delta = \lambda$. We further assume a constant population size, N_{ss}, over time, which is a valid assumption if N is large. Analogous to Section 3.3.1, we model clonal expansions using the transition probabilities of a linear birth-death process. However, we now need to find the transition probabilities of a critical birth-death process, that is $\lambda = \delta$. An explicit solution for the probability that a cells at time t' transition into b cells at time t is obtained by expanding the probability generating function of the linear birth-death process (see Appendix B), which yields:

$$P_{\text{ss},a,b}(t',t|\lambda) = \begin{cases} 0 & a = 0, b \neq 0, \\ p^a & b = 0, \\ \sum_{k=1}^{b} \frac{k}{b} \binom{a}{k} p^{a-k}(1-p)^k \binom{b}{k} p^{b-k}(1-p)^k & a \neq 0, b \neq 0, \end{cases} \tag{3.13}$$

$$p = \frac{\lambda(t-t')}{1+\lambda(t-t')}. \tag{3.14}$$

In analogy to Eqn. 3.11, we can now determine the total number of mutations that were acquired during adulthood and are present in i cells at time t to

$$S_i(t|\mu,\lambda,N_{\text{ss}},T_{\text{ss}}) = \mu\lambda N_{\text{ss}} \int_{T_{\text{ss}}}^{t} P_{\text{ss},1,i}(t',t|\lambda)\mathrm{d}t', \tag{3.15}$$

where T_{ss} marks the beginning of the steady-state phase. Finally, in order to obtain the mutation frequency distribution of the two-stage system, we couple the expansion phase during embryogenesis with the steady-state phase during adulthood. Cells acquiring a mutation during embryogenesis may have grown to clones of any size between 0 and N_{ss} at the transition to

adulthood, and eventually reached their final clone size due to neutral drift in the homeostatic tissue. Labeling the parameters of the expansion phase with 'exp' and those of the steady-state phase with 'ss' we obtain the mutation frequency distribution of the coupled system:

$$S_i(t|\mu,\lambda_{\text{exp}},\delta_{\text{exp}},\lambda_{\text{ss}},N_{\text{ss}},T_{\text{ss}}) = \tag{3.16}$$

$$\underbrace{\int_0^{T_{\text{ss}}} \sum_{k=1}^{N_{\text{ss}}} \overbrace{P_{\text{exp},k}(t'|T_{\text{ss}},\lambda,\delta)}^{\text{Neutral drift during expansion}} \quad \mu\lambda e^{(\lambda_{\text{exp}}-\delta_{\text{exp}})t'} \times \overbrace{P_{\text{ss},k,i}(T_{\text{ss}},t|\lambda_{\text{ss}})}^{\text{Neutral drift during steady-state}} \quad \mathrm{d}t'}_{\text{Mutations acquired during expansion}} +$$

$$+ \underbrace{\int_{T_{\text{ss}}}^{t} P_{\text{ss},1,i}(t',t|\lambda_{\text{ss}})\mu\lambda N_{\text{ss}}\mathrm{d}t'}_{\text{Mutations acquired during steady-state}}, \qquad t \geq T_{\text{ss}}.$$

As before, we obtain the number of mutations present in a tissue fraction of at least f by setting $f = \frac{i}{N}$:

$$M(f,t|\mu,\lambda_{\text{exp}},\delta_{\text{exp}},\lambda_{ss},N_{\text{ss}},T_{\text{ss}}) = \sum_{i=fN(t)}^{N(t)} S_i(t|\mu,\lambda_{\text{exp}},\delta_{\text{exp}},\lambda_{\text{ss}},N_{\text{ss}},T_{\text{ss}}). \tag{3.17}$$

Again, we assess the accuracy of our analytical prediction with stochastic simulations. To this end, we first simulate clonal expansions until the population size reaches 1,000 cells, using the agent-based simulation approach, as before (c.f. Section 3.3.1). Keeping in mind our ultimate goal to model embryonic and adult hematopoiesis, we simulate the expansion phase with a cellular death rate of $\delta_{\text{exp}} = 0.5\lambda_{\text{exp}}$, thus assuming fast expansion while accounting for some cell death or differentiation. Subsequently, we assume that the cell population remains constant, i.e., division and death rates are balanced. Following previous estimates of cell division rates in murine hematopoietic stem cells, we choose $\lambda_{\text{ss}} = \delta_{\text{ss}} = 0.009\,\text{d}^{-1}$ (Busch et al., 2015). To assess the accuracy of our theoretical prediction at different time points, we report the simulation result five weeks, two years and four years after the steady-state was reached (the five-weeks time point corresponds roughly to eight-week-old mice, as hematopoietic stem cells transition to an adult proliferation phenotype three weeks after birth; Bowie et al., 2007). We find that the analytical prediction slightly overestimates the average from stochastic simulations at all timepoints (Fig. 3.9), which may, in part, be due to random fluctuations of the population size in the simulations and rare outliers that were not captured by the simulation. This is corroborated by the large variance that is suggested by the stochastic simulations. Interestingly, the model predicts that the overall mutational burden increases with age, and that mutations drift to high tissue fractions with time (red arrows in Fig. 3.9). Thus neutral drift dynamics in homeostatic tissues changes the mutation frequency distribution distinctly from exponentially growing tissues.

Taken together, our theoretical considerations show that different tissue dynamics, such as growth versus homeostasis, but also the degree of cellular turnover, affect mutation accumulation

in distinct ways. This provides a theoretical framework based on which deviations from neutrality, as expected during tumor initiation, may become identifiable.

3.4 Model applications: Mutation accumulation in the hematopoietic system

Having a theoretical expectation of the mutation frequency distribution at hand, we now exemplify potential applications of our approach by revisiting the mutational burden in murine and human blood samples. To this end, we need to ensure that mutation accumulation in hematopoietic stem cells and peripheral blood cells correlate sufficiently well. This is a necessary prerequisite, since our theoretical framework describes mutation acquisition and clonal expansions in an influx-free population, comparable to stem cells. However, additional cell divisions during differentiation likely contribute to the mutation frequency distribution in downstream compartments. Accurate prediction of mutation accumulation in granulocytes therefore requires the coupling of clonal drift dynamics within granulocytes with stochastic influx from upstream compartments. However, this becomes computationally cumbersome, as the cell counts in downstream compartments are large. Yet, mutation accumulation in the measurable range (VAFs>1%, irrespective of the read coverage; Alioto et al., 2015) correlates sufficiently well between stem cells and downstream compartments, as we will see in the following. Thus mutation accumulation in peripheral granulocytes can be approximated by an influx-free cell population.

3.4.1 Mutation spectrum in peripheral granulocytes as a readout for hematopoietic stem cell dynamics

To show that the measurable mutation frequency distribution is only marginally affected by cell divisions during progressive differentiation, we analyze the clonal dynamics between long-term hematopoietic stem cells (LT-HSCs), short-term hematopoietic stem cells (ST-HSCs) and multipotent progenitors (MPPs), corresonding to the three most upstream compartments in the hematopoietic system as defined by Busch et al., 2015 (Fig. 3.10A). Hierarchical tissue organization may affect the mutation frequency distribution in downstream compartments by (i) acquisition of new mutations during divisions in transiently amplifying compartments and (ii) neutral drift of mutations acquired in LT-HSCS (Fig. 3.10B). However, our WGS data reports mutations only if present in at least 10% of the sample, and hence the relevant question for our purpose is not, whether hierarchical tissue organization changes the mutation frequency distribution at all, but whether it does so within the measurable window.

Previous studies estimated a total number of 500×10^6 bone marrow cells (Colvin et al., 2004) and a LT-HSC frequency of 3×10^{-5} cells (Höfer and Rodewald, 2016), yielding approximately 10,000 LT-HSCs in the mouse. MPPs are about ten times more frequent than LT-HSCs (Busch

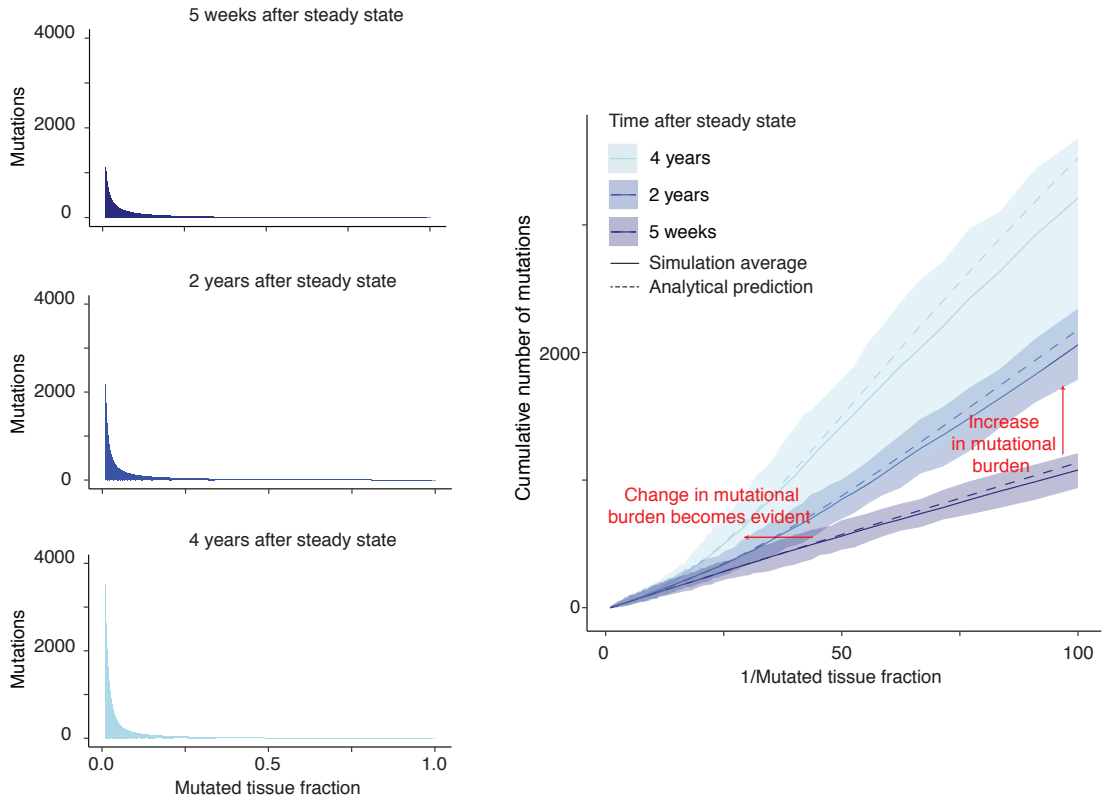


Figure 3.9: Predicted mutation frequency distribution if coupling an exponential expansion phase to a phase of homeostatic turnover. Shown are the theoretical predictions of the mutation frequency distribution at different time points post steady-state (left) and the corresponding cumulative number of mutations (right, dashed lines; shown is the number of mutations present in at least the indicated tissue fraction). The cumulative distribution is compared to the result of analogous stochastic simulations (solid lines, means; shaded areas, 95% confidence intervals from 100 simulations each). Simulations were run with $\delta_{exp} = 0.5\lambda_{exp}$ and $\lambda_{exp} = 1\ d^{-1}$ until a population size of 1,000 cells was reached, and then continued with $\lambda_{ss} = \delta_{ss} = 0.009\ d^{-1}$ for the indicated time. Mutation rate in all predictions/simulations was taken as six mutations per cell division (three per daughter cell).

et al., 2015), corresponding to roughly 100,000 cells per mouse. Thus a mutation acquired in a downstream compartment becomes detectable with WGS only if the cell of origin expands to a clone size of several thousand cells.

To assess the clone size distribution of expanding LT-HSCs, ST-HSCs and MPPs, we simulate the three most upstream compartments stochastically, adopting the model structure and parametrization from Busch et al., 2015. Briefly, we assume a linear differentiation path between LT-HSCs, ST-HSCs and MPPs, as shown in Fig. 3.10A. Moreover, we assume that cells divide and differentiate at constant rates λ and α, and neglect cellular loss due to death. Indexing LT-HSCs, ST-HSCs and MPPs with 0, 1, and 2, respectively, we define the state vector $\boldsymbol{n} = (n_0, n_1, n_2)$ and formulate the Master equation that describes the temporal evolution of the state probabilities in the linear differentiation path:

$$\frac{\mathrm{d}P_{\boldsymbol{n}}(t)}{\mathrm{d}t} = \sum_{i=1}^{2} \Bigg[\underbrace{\overbrace{\lambda_i(n_i-1)P_{\boldsymbol{n}-e_i}(t)}^{\text{Division of cell type i}} + \overbrace{\alpha_i(n_i+1)P_{\boldsymbol{n}+e_i-e_{i+1}}(t)}^{\text{Differentiation of cell type i into cell type i + 1}}}_{\text{Stochastic processes running into state n}} \underbrace{-\lambda_i n_i P_{\boldsymbol{n}}(t) - \alpha_i n_i P_{\boldsymbol{n}}(t)}_{\text{Sochastic processes running out of state n}} \Bigg], \quad (3.18)$$

where e_i is the $(i+1)$-th unit vector and e_3 is set to $(0,0,0)$ by convenience.

From the Master equation one can extract the stochastic moments (Appendix C.1), allowing us to simulate the temporal evolution of the mean and the variance of single cell expansions starting in different compartments (cf. Eqns. C.11 and C.12 in the Appendix). With the parametrization by Busch et al., 2015 (c.f. Table 3.1), the capacity of robust clonal expansion declines with progressive differentiation, consistent with a decreasing capacity to self-renew (Fig. 3.10C). While a single LT-HSC can produce long-term surviving progeny, ST-HSCs may only transiently expand and MPPs become rapidly extinct. Accordingly, LT-HSCs can produce the largest clones, consisting of up to 50 LT-HSCs, 100 ST-HSCs and 1000 MPPs after two years (Fig. 3.10D; assessed with stochastic simulations of the reactions shown in Fig. 3.10A and using the steady-state parametrization listed in Table 3.1; c.f. Appendix C.2). Since we estimate the number of MPPs per mouse to approximately 10^5 (Table 3.1), measurable mutations must be present in at least 1,000 MPPs and are thus most likely acquired in the stem cell compartment. Moreover, as large stem cell clones are rare (Fig. 3.10D), most measurable mutations will be acquired during embryogenesis. Thus it is very unlikely that mutations acquired during adulthood contribute significantly to the measurable range of the mutation frequency distribution in granulocytes.

Next, we show that neutral drift dynamics downstream of ST-HSCs do not measurably alter the mutation frequency distribution in granulocytes either. To this end, we once again employ stochastic simulations of murine hematopoiesis. As we are now interested in the fate of individual mutations, we simulate mutation frequencies with an agent-based model (Fig. 3.11

Table 3.1: Parameters used to simulate the mutation frequency distributions in LT-HSCs, ST-HSCs and MPPs.

	Parameter	Value	Comment
Expansion phase	$\lambda_{S,0}$ (d^{-1})	1	chosen to reflect a high proliferation rate in the fetus
	$\lambda_{A,0}$ (d^{-1})	0.6	chosen according to the model fit in Section 3.4.2
	$\lambda_{S,1}$ (d^{-1})	0.7940611	tailored to obtain ~ 29,000 ST-HSCs
	$\lambda_{A,1}$ (d^{-1})	0.6	
	$\lambda_{S,2}$ (d^{-1})	0	MPP expansion is not modeled
	$\lambda_{A,2}$ (d^{-1})	0	
Steady-state phase	$\lambda_{S,0}$ (d^{-1})	0.009	from Busch et al., 2015
	α_0 (d^{-1})	0.009	
	$\lambda_{S,1}$ (d^{-1})	0.042	
	α_1 (d^{-1})	0.045	
	$\lambda_{S,1}$ (d^{-1})	4	
	$\alpha_{S,2}$ (d^{-1})	4.014	
Population size at steady-state	N_{LT-HSC}	10,000	from Colvin et al., 2004; Höfer and Rodewald, 2016
	$N_{ST-HSC} : N_{LT-HSC}$	2.9	from Busch et al., 2015
	$N_{MPP} : N_{LT-HSC}$	9	

Abbreviations used: $\lambda_{S,0}$, rate of symmetric LT-HSC division; $\lambda_{A,0}$, rate of asymmetric LT-HSC division; $\lambda_{S,1}$, rate of symmetric ST-HSC division; $\lambda_{A,1}$, rate of asymmetric ST-HSC division; $\lambda_{S,2}$, rate of symmetric MPP division; $\lambda_{A,2}$, rate of asymmetric MPP division; $\alpha_{S,0}$, rate of LT-HSC differentiation; $\alpha_{S,1}$, rate of ST-HSC differentiation; $\alpha_{S,2}$, rate of MPP differentiation; μ, mutation rate; N_{LT-HSC}, number of LT-HSCs at steady state; N_{ST-HSC}, number of ST-HSCs at steady state; N_{MPP}, number of MPPs at steady state.

and Appendix C.3). Briefly, the model simulates the embryonic expansion of a single LT-HSC on the macroscopic scale, and mutation acquisition and inheritance during cell divisions on the microscopic scale. Embryonic expansion is run until the population size of LT-HSCs reaches 10,000 cells, corresponding to the stem cell population size at homeostasis. Subsequently, the simulation is approximated by sampling the cellular fate of each LT-HSC and ST-HSC from the simulated clone size distributions after two years (c.f. Fig. 3.10D and Appendix C.2). This approximation accelerates the simulation by neglecting any mutations newly acquired during adulthood. Importantly, this does not impair the model prediction within the measurable window, since adult mutations are unlikely to drift to measurable frequencies, as discussed above (c.f. Fig. 3.10C, D).

Simulations of the agent-based model suggest that the variance of the mutation frequency distribution indeed increases downstream of LT-HSCs (Fig. 3.12A). However, this is primarily due to the relatively slow proliferating ST-HSCs, while the fast turnover of MPPs barely changes the distribution (Fig. 3.12B). As the potential to self-renew decreases with progressive differentiation in the hematopoietic system (Busch et al., 2015), compartments downstream of MPPs are unlikely to affect the mutation frequency distribution beyond the impact of MPPs. Thus mutation accumulation in ST-HSCs approximately models mutation accumulation in more downstream states within the measurable window.

3.4.2 Mutation frequency distribution in murine granulocytes

With a setup to simulate mutation accumulation in undisturbed hematopoiesis at hand, we now test whether neutral drift explains the measured increase in the mutational burden of murine granulocytes during ageing (c.f. Fig. 3.4). To begin with, we treat the measurement in peripheral granulocytes as a direct readout of the mutational profile in LT-HSCs, following the discussion of Section 3.4.1. This reduces our system to a single, self-sustaining compartment, for which the theory of Section 3.3.2 applies. To compare theory and data, we fit the free parameters of Eqn. 3.17 to the measured mutation frequency distribution in granulocytes of the eight-week-old mouse. Using the calibrated model, we then predict the mutation frequency distribution after two years and compare it to the measured data in the two-year-old mouse.

In principle, Eqn. 3.17 has six free parameters: the cell division and loss rate during expansion, λ_{exp} and δ_{exp}, the cell division rate during steady-state, λ_{ss}, the number of stem cells at steady-state, N_{ss}, the time point at which expansion is completed, T_{ss}, and the mutation rate per cell division, μ. Adopting previous estimates from the literature, we reduce the parameter estimation to two unknown parameters:

- We fix N_{ss} to 10,000 LT-HSCs and λ_{ss} to 0.009 d^{-1}, as before (Table 3.11).
- As hematopoietic stem cells have been shown to switch from an embryonic to an adult proliferation phenotype within the first three weeks of newborn mice, we fix T_{ss} to three

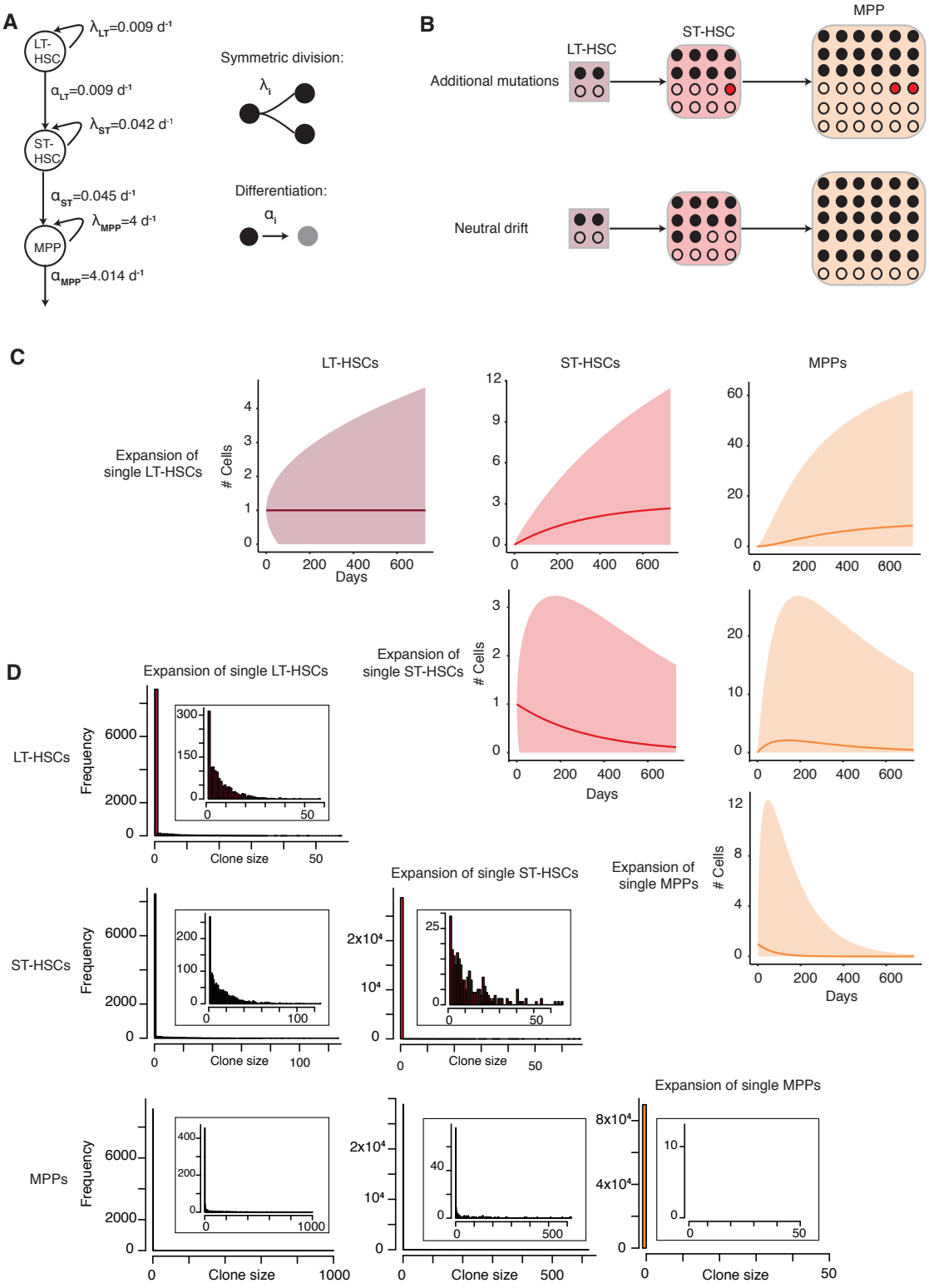

A
LT-HSC
ST-HSC
MPP
λLT=0.009 d-1
αLT=0.009 d-1
λST=0.042 d-1
αST=0.045 d-1
λMPP=4 d-1
αMPP=4.014 d-1
Symmetric division:
Differentiation:
B
LT-HSC
ST-HSC
MPP
Additional mutations
Neutral drift
C
LT-HSCs
ST-HSCs
MPPs
Expansion of single LT-HSCs
Expansion of single ST-HSCs
Expansion of single MPPs
Cells
Days
D
Expansion of single LT-HSCs
Expansion of single ST-HSCs
Expansion of single MPPs
LT-HSCs
ST-HSCs
MPPs
Frequency
Clone size

Figure 3.10 ***(previous page)*****:** Single cell expansions of LT-HSCs, ST-HSCs and MPPs. **A** Model scheme and parametrization of adult hematopoiesis according to Busch et al., 2015. Cells divide at rate λ and differentiate at rate α. **B** The mutation frequency distribution may change over the course of progressive differentiation due to newly acquired mutations and neutral drift. **C** Mean (solid lines) and variance (shaded areas) of clonal expansions of a single LT-HSC (three upmost panels), a single ST-HSC (middle panels) or a single MPP (bottom panel) over two years. Mean and variance were computed from the master equation of the stochastic system shown in (A). **D** Clone size distributions of LT-HSCs (three leftmost panels, 10,000 simulations), ST-HSCs (middle panels, 29,000 simulations) and MPPs (right panel, 90,000 simulations) after two years assessed with stochastic simulations (c.f. Appendix C.2). Insets show the clone size distribution of surviving cells only. The number of simulations was scaled with the relative compartment sizes according to Busch et al., 2015 and assuming 10,000 LT-HSCs per mouse. Parametrization for all simulations was taken according to Table 3.1, 'Steady-state phase'.

weeks after birth (Bowie et al., 2007). Thus we assume that the blood sample of the eight-week-old mouse was taken five weeks after the steady-state was reached.

- We arbitrarily fix $\lambda_{\text{exp}} = 1\ \text{d}^{-1}$, since the absolute duration of the the expansion phase (consisting of embryogenesis and the first three weeks after birth) is irrelevant for our analysis.

This leaves the extent of cellular loss, δ_{exp}, and the division-coupled mutation rate, μ, to be determined. To do this, we perform a uniform scan over the range $1 \leq \mu \leq 10$ and $0\ \text{d}^{-1} \leq \delta_{\text{exp}} \leq 0.9\ \text{d}^{-1}$ (10 values each) and compute the residual sum of squares between the measured and predicted mutation frequency distribution in the blood of the eight-week-old mouse. We find the best fit at a mutation rate of seven mutations per cell division and a relative loss rate (due to death or differentiation) of $\delta_{\text{exp}}/\lambda_{\text{exp}} = 0.6$, indicating that approximately every second stem cell division during embryogenesis is a symmetric division (Fig. 3.13A). With this parametrization we capture the mutation frequency distribution in the granulocytes of the eight-week-old mouse accurately (Fig. 3.13B, left panel).

Next, we analyze the mutation frequency distribution in the granulocytes of the two-year-old mouse, and find that the number of mutations shared by large tissue fractions is comparable to the young mouse, whereas the number of mutations shared by small tissue fractions is elevated (Fig. 3.13B, blue points in right panel). This observation resembles the mutation frequency distribution predicted under neutral dynamics in ageing mice qualitatively (c.f. Fig. 3.9). Quantitatively, however, an increase in the mutational burden is not expected within the measurable window if using the same parametrization as before (Fig. 3.13B). Moreover, simulations accounting for additional divisions of ST-HSCs do not fill the gap between theory and data (assessed with the agent-based model introduced in Section 3.4.1; Fig. 3.13C). This raises the question of whether (i) the experimental observation is due to non-neutral dynamics, (ii) additional parameters need to be adjusted to reconcile the measured data with the theoretical prediction of neutral drift or (iii) the observation represents a rare outlier.

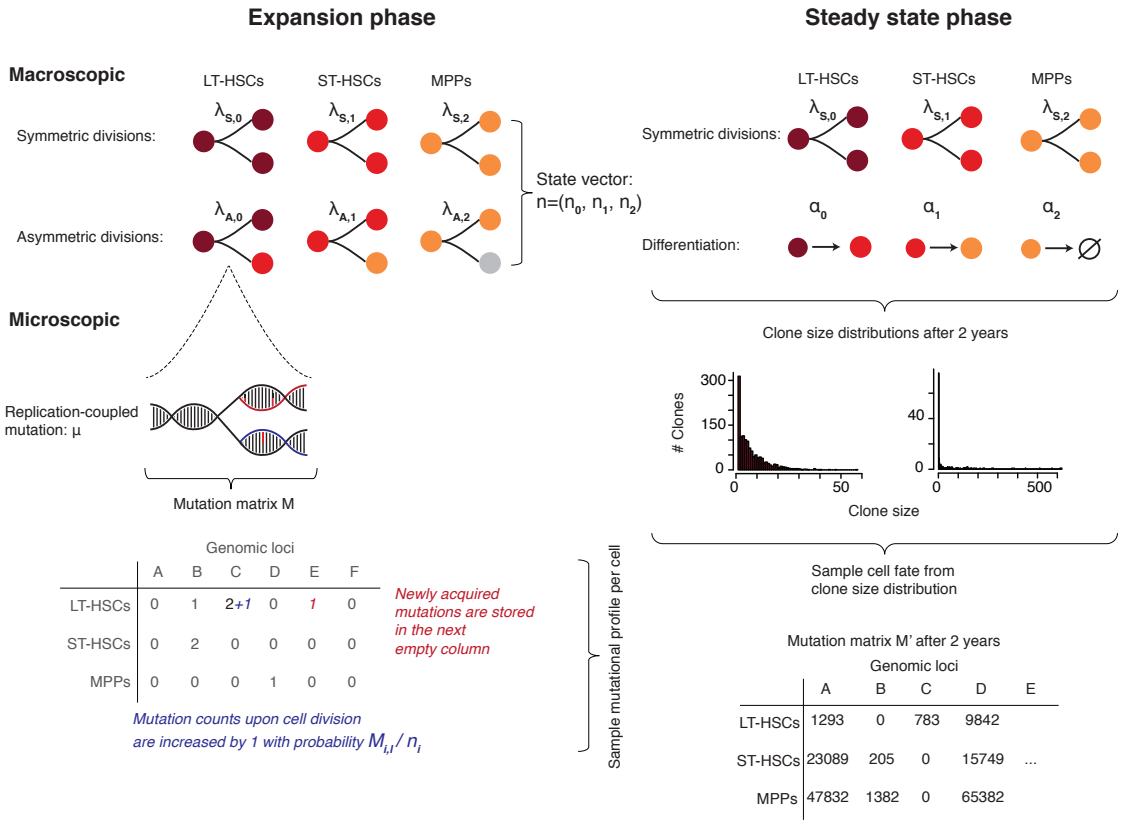


Figure 3.11: Agent-based simulation of mutation accumulation in adult hematopoiesis. The expansion phase is simulated with symmetric and asymmetric divisions on the macroscopic scale. On the microscopic scale, the model simulates acquisition and inheritance of mutations during divisions. The steady-state phase is approximated by sampling the mutational profiles of individual cells from the mutation frequency distribution at the end of the expansion phase, and expanding them according to a random sample from the simulated clone size distributions after two years. Details of the simulation algorithm are explained in Appendix C.3.

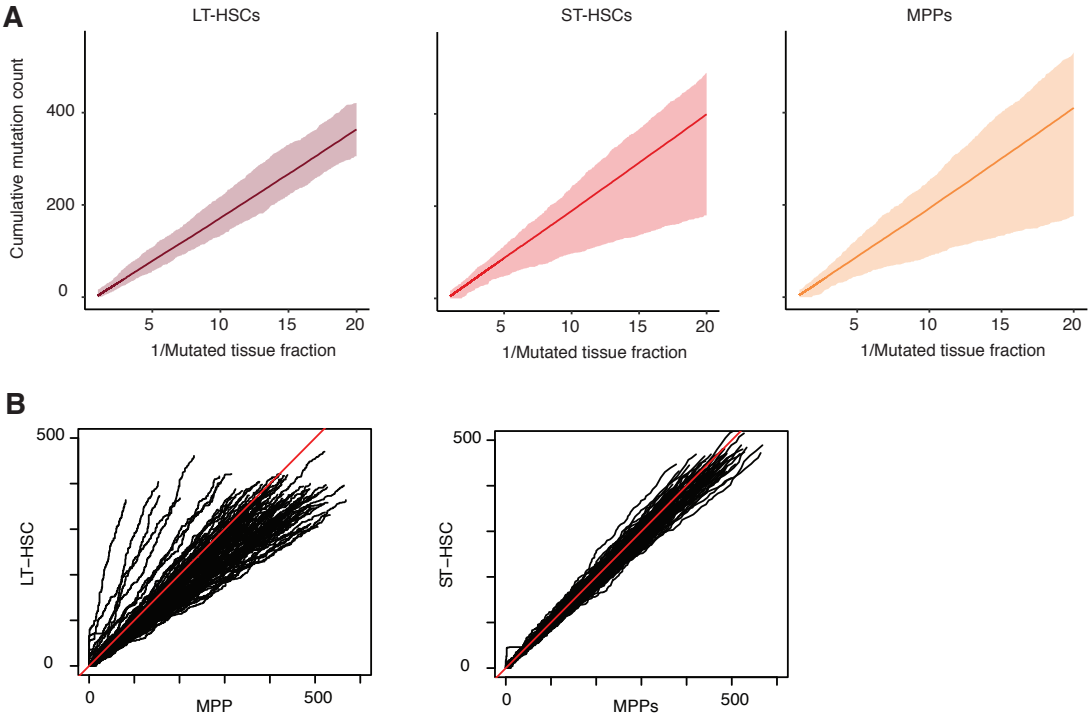


Figure 3.12: Mutation frequency distribution in LT-HSC, ST-HSC and MPPs simulated with the agent-based model. **A** Mean (solid lines) and 95% confidence interval of 100 simulated mutation frequency distributions in LT-HSCs, ST-HSCs and MPPs. **B** Comparison of mutation frequencies between MPPs and LT-HSCs (left), and between MPPs and ST-HSCs (right) for the 100 simulations shown in (A; red lines, bisetrices). All simulations were performed with the agent-based model illustrated in Fig. 3.11, using the parameters of Table 3.1.

As non-neutral dynamics can be caused by driver mutations, we screen the mutational profile of the two-year-old mouse for exonic mutations. Interestingly, we find a single exonic mutation only, which is a non-synonymous substitution in exon 13 of the *Janus kinase 3* gene (*Jak3*), generating the constitutively active mutant Jak3^{A568V} that is homologous to the human variant JAK3^{A572V} (Fig. 3.13D; Walters et al., 2006). This mutation is found at a variant allele frequency of 0.09 and has previously been associated with T-cell lymphoma (Koo et al., 2012). Consistent with the role of JAK3 in promoting lymphopoiesis (Springuel et al., 2015), activating mutants in Jak3 have been shown to induce lymphoproliferative syndromes in bone marrow transplantation models (Cornejo et al., 2009). Moreover, hyperproliferative T lymphocytes with the Jak3^{A568V} mutation colonize the bone marrow upon expansion in the periphery (Rivera-Munoz et al., 2018). A possible effect of the acquired Jak3^{A568V} mutation in the two-year-old mouse could hence involve a reduction of the stem cell pool due to competition with invading CD8^{+} T lymphocytes. Since blood counts of myeloid cells remain unaffected in a Jak3^{A568V} mouse model (Cornejo et al.,

2009), LT-HSCs would need to compensate their reduced cell counts by increased proliferation. Indeed, decreasing the pool of LT-HSCs from 10,000 to 1,000, while increasing the division rate during homeostasis from 0.009 d^{-1} to 0.05 d^{-1} improves model predictions at two years (Fig. 3.13E). Thus the unexpectedly high mutational burden in the two-year-old mouse may have been induced by the $Jak3^{A568V}$ mutation and hence may mirror a pathological rather than a normal situation. However, as the data lack a replicate, additional sequencing is necessary to conclusively disentangle the dynamics of undisturbed hematopoiesis in ageing mice.

3.4.3 Mutation frequency distribution in human leukocytes

A potential application of our theoretical framework is the identification of pre-cancerous lesions due to deviations from the mutation frequency distribution expected under neutral tissue dynamics. As an outlook, we therefore anecdotally revisit the mutation frequency distributions in human leukocytes introduced at the beginning of this chapter. As these patients were diagnosed with neuro- or glioblastoma, but not with a hematological disorder, we expect that the mutational burden conforms to the theoretical expectation under neutral stem cell dynamics. To test this hypothesis quantitatively with model simulations, we would need to simulate a larger stem cell pool than in mice, as a recent study estimated that humans have approximately ten times as many LT-HSCs as mice, while the rate of stem cell turnover is comparable (10^5 LT-HSCs that divide at most once in two months; Lee-Six et al., 2018). However, simulations of large stem cell pools become computationally costly, wherefore we compare the experimental data to simulations of a smaller pool of 10^4 cells, as before. As changes in the mutation frequency distribution become measurable earlier if the stem cell pool is small, this simplification may wrongly classify non-neutral dynamics as neutral, but not vice versa.

As shown in Fig. 3.14A and B, the mutation frequency distributions in the blood of most children conform to the expectation under neutral dynamics, whereas approximately 40% of the samples have more mutations than expected (Fig. 3.14A, C, D). Deviation from neutrality is primarily evident at small VAFs and of a mild extent in some samples (Fig. 3.14C), while striking in others (Fig. 3.14D). The unexpectedly high mutational burden in the latter raises the question whether oncogenic mutations drove a pre-leukemic phenotype in these patients. To address this question, we scan the mutational profiles for driver mutations associated with clonal hematopoiesis (Steensma et al., 2015) and leukemia (www.intogen.org; Gonzalez-Perez et al., 2013). Note that the true number of driver mutations in the blood samples might be higher, since loci with a copy number change in the matching tumor sample (which was used as a germline control) had to be excluded from mutation calling. In line with the conjunction that the hypermutation genotypes may have evolved non-neutrally, mutations in leukemic driver genes, such as *NOTCH1*, *JAK3* or *RUNX1*, are frequent in hypermutated samples, while rare in the rest (Fig. 3.14 and Fig. 3.15). Moreover, in two, but not all of the hypermutated samples, we find non-synonymous and stop-gain mutations in the mismatch repair genes *MSH5* and *MLH3*, respectively. Thus the strong deviation of some blood samples from the theoretical expectation

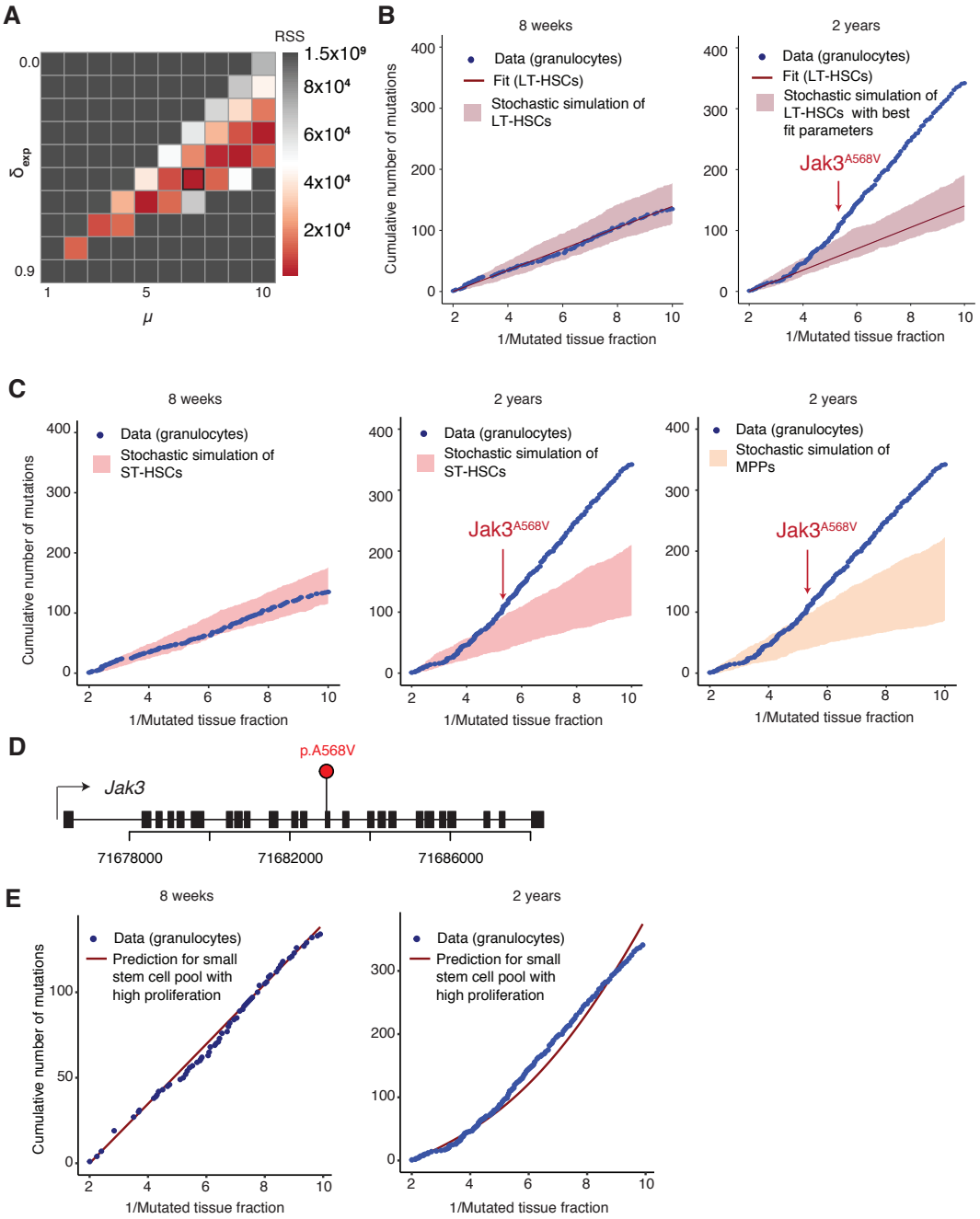


Figure 3.13: Mutation accumulation in undisturbed hematopoiesis. **A** Parameter scan over the mutation rate per cell division, μ, and the differentiation/death rate, δ_{exp}, during expansion of the hematopoietic system in the murine embryo. Shown are the residual sums of squares when comparing the theoretical expectation in the stem cell compartment according to Eqn. 3.17 to the measured mutation frequency distribution in granulocytes of the eight-week-old mouse (c.f. Fig. 3.4). The best fit is marked by a black square.

Caption continues on next page.

Figure 3.13 ***(continued)*****: B, C** Expected mutation frequency distribution in murine LT-HSCs (B), ST-HSCs and MPPs (C) of eight-week-old and two-year-old mice along with measured data in peripheral granulocytes. Theoretical predictions are based on Eqn. 3.17 (B, lines) and 100 stochastic simulations with the agent-based model outlined in Section 3.4.1 (B, C, shaded areas). The parametrization is based on the best fit marked in (A) (see also Table 3.1; lines represent the analytical prediction according to Eqn. 3.17; shaded areas represent 95% confidence intervals of the stochastic simulations; red arrows indicate the measured tissue fraction of the $Jak3^{A568V}$ mutation in the two-year-old mouse). **D** Genomic locus of the *Jak3* gene with A568V mutation in exon 13. **E** Expected mutation frequency distribution in murine LT-HSCs if assuming a compartment size of 1,000 LT-HSCs, a division rate 0.05 d^{-1} during adult hematopoiesis and all other parameters as in (B, C) (lines, theoretical prediction according to Eqn. 3.17). Tissue fractions were obtained from the measured mutation frequency distribution by multiplying variant allele frequencies at heterozygous loci by two, thus accounting for diploid genomes. Mutations with VAFs<0.05 and VAFs>0.25, as well as mutations supported by less than five sequencing reads were excluded. Experimental data: Ruzhica Bogeska. Bioinformatic pre-processing: Megan Druce.

under neutral tissue dynamics may indeed be due to pre-cancerous lesions.

Finally, we analyze the adult blood samples and ask whether putatively pre-cancerous lesions are also present in the glioblastoma patients. In contrast to the neuroblastoma patients, the mutation frequency distributions in the peripheral leukocytes of all glioblastoma patients look qualitatively similar, and all individuals have subclonal mutations in the same order of magnitude (100 - 1000; Fig. 3.16A). Moreover, leukemic driver mutations are rarely found in the adult blood samples, indicating that the mutation frequency distributions were predominantly shaped by neutral stem cell dynamics (Fig. 3.15). To test whether a single model of mutation accumulation captures the measured mutation frequency distributions of all adult blood samples, we simulate adult hematopoiesis as before, but extend the simulation over 50 years. We find that a single model of mutation accumulation explains the data of approximately 60% of the patients only (Fig. 3.16B-D), suggesting that the dynamics of mutation acquisition and stem cell turnover is not homogenous among individuals. Interestingly, comparison of the measured mutation frequency distribution in children and adults with the model predictions suggests that the mutation rate in children is approximately twice as high as the rate in adults (14 versus six mutations per cell division in children and adults, respectively; Fig. 3.14, 3.16). However, further measurements will be needed in order to assess whether this observation is true or due to differences in sample acquisition and processing.

In summary, unsorted blood samples of cancer patients with no diagnosed leukemias suggest that undisturbed hematopoiesis and putatively pre-cancerous lesions shape the mutation frequency distribution in peripheral blood cells distinctly. Moreover, the shape of the mutation frequency distribution in non-hypermutated samples matches our theoretical expectation qualitatively. Thus our theoretical framework may facilitate the interpretation of deep whole genome sequencing data. Nevertheless, a better understanding of the dynamics and the inter-

individual heterogeneity in the human hematopoietic system will be necessary to exploit the predictive potential of our theoretical framework in a more quantitative way.

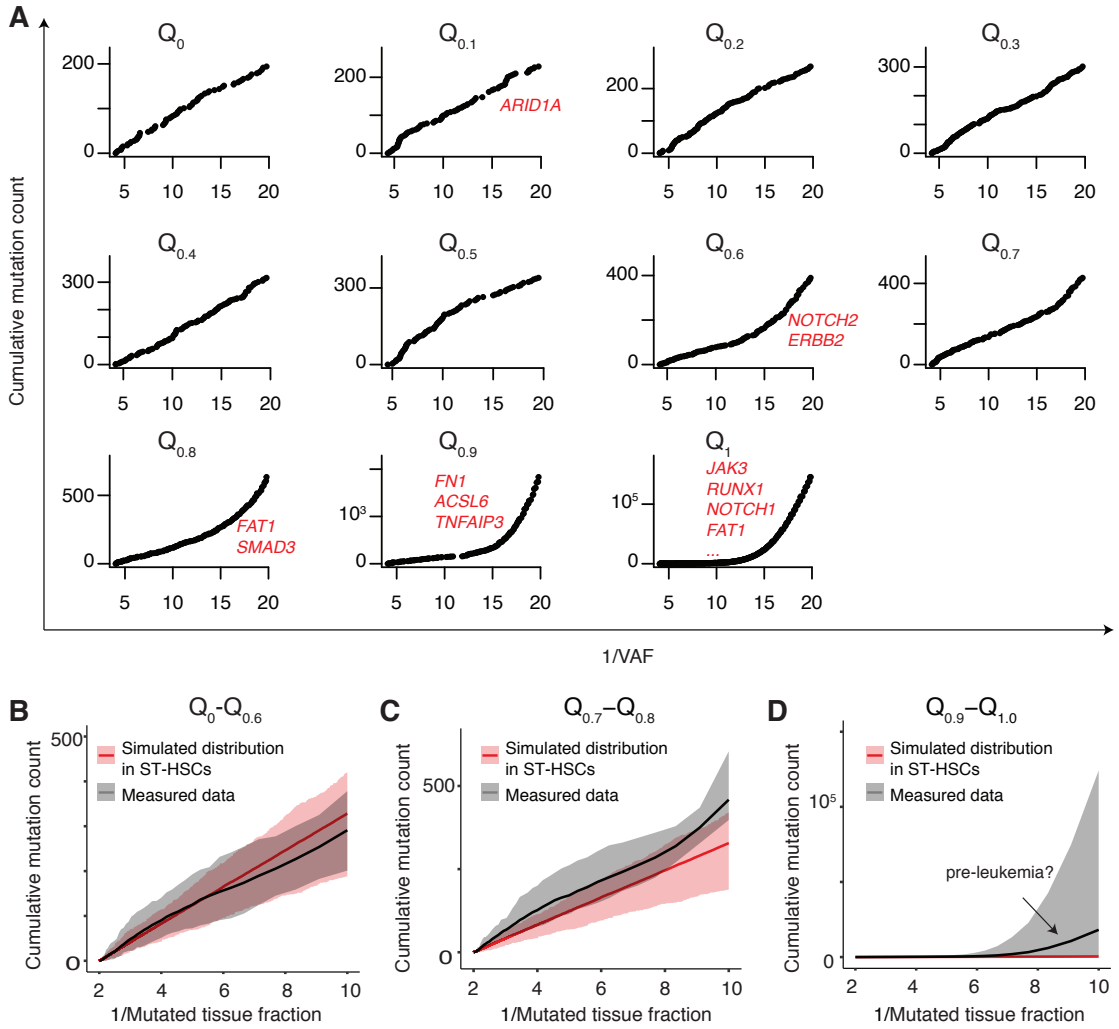


Figure 3.14: Cumulative mutation frequency distributions measured in blood samples from children diagnosed with neuroblastoma (shown are mutations supported by at least three sequencing reads, excluding mutations with VAFs≤0.05 and VAFs≥0.25; values are extrapolated from the analyzed genome fraction to the entire genome). **A** Ten exemplary samples, corresponding to the deciles of the measured subclonal mutation count among the cohort of in total 68 samples. Leukemic driver mutations are indicated in red. **B-D** Mean (lines) and 95% confidence intervals (shaded areas) of 100 simulated mutation frequency distributions in ST-HSCs. Simulations are the same as shown in Fig. 3.13C (middle panel) and as described in Section 3.4.1, but mutation counts were multiplied by two to account for the steeper curve measured in human samples (thus, yielding a mutation rate of 14 mutations per cell division). The measured mean and 95% confidence intervals of the samples corresponding to the lower six deciles (B), the seventh and the eigth deciles (C) and the ninth and tenth deciles (D) are shown alongside and are colored in black (mutated tissue fractions were obtained by multiplying the measured VAFs of heterozygous mutations with two to account for diploid genomes). Experimental data: Frank Westermann.

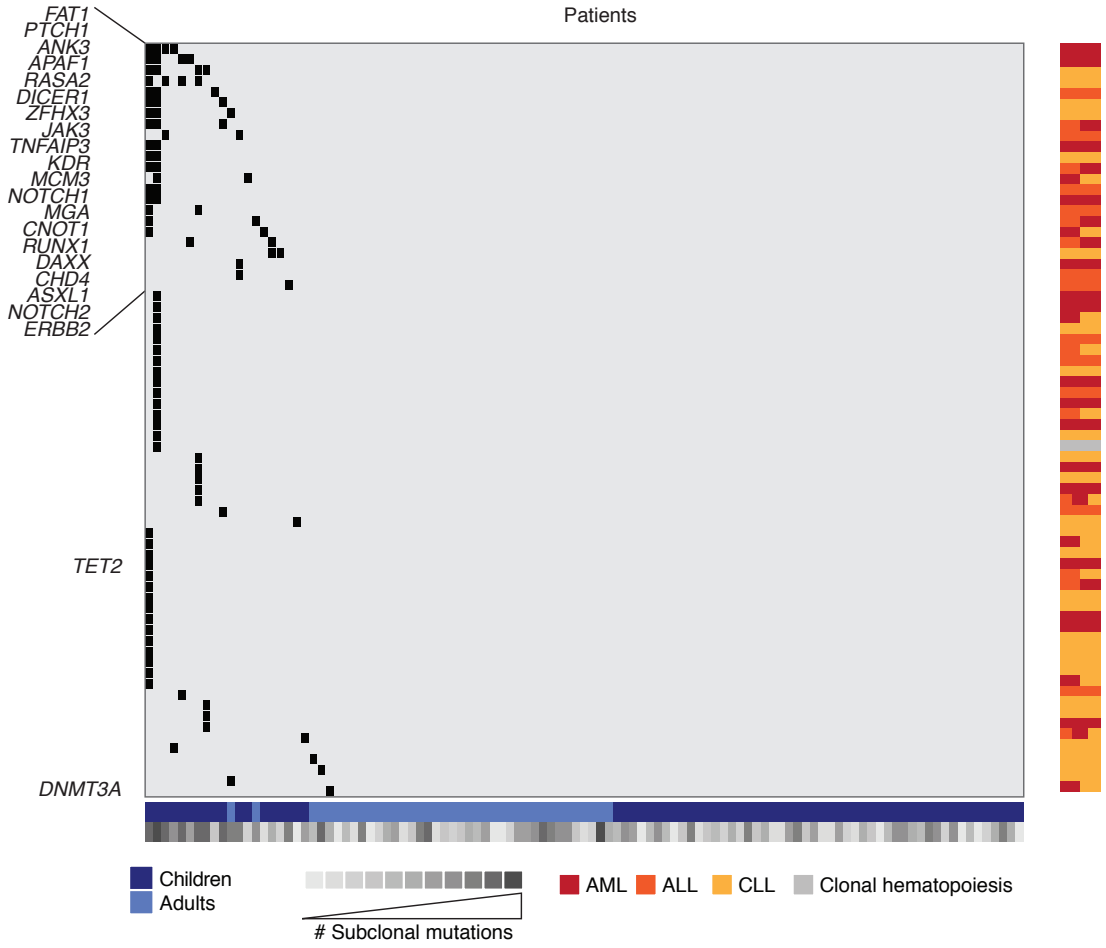


Figure 3.15: Leukemic driver mutations in blood leukocytes from neuro- and glioblastoma patients. Shown are non-synonymous mutations in leukemic driver genes (www.intogen.org) and genes associated with clonal hematopoiesis (Steensma et al., 2015).

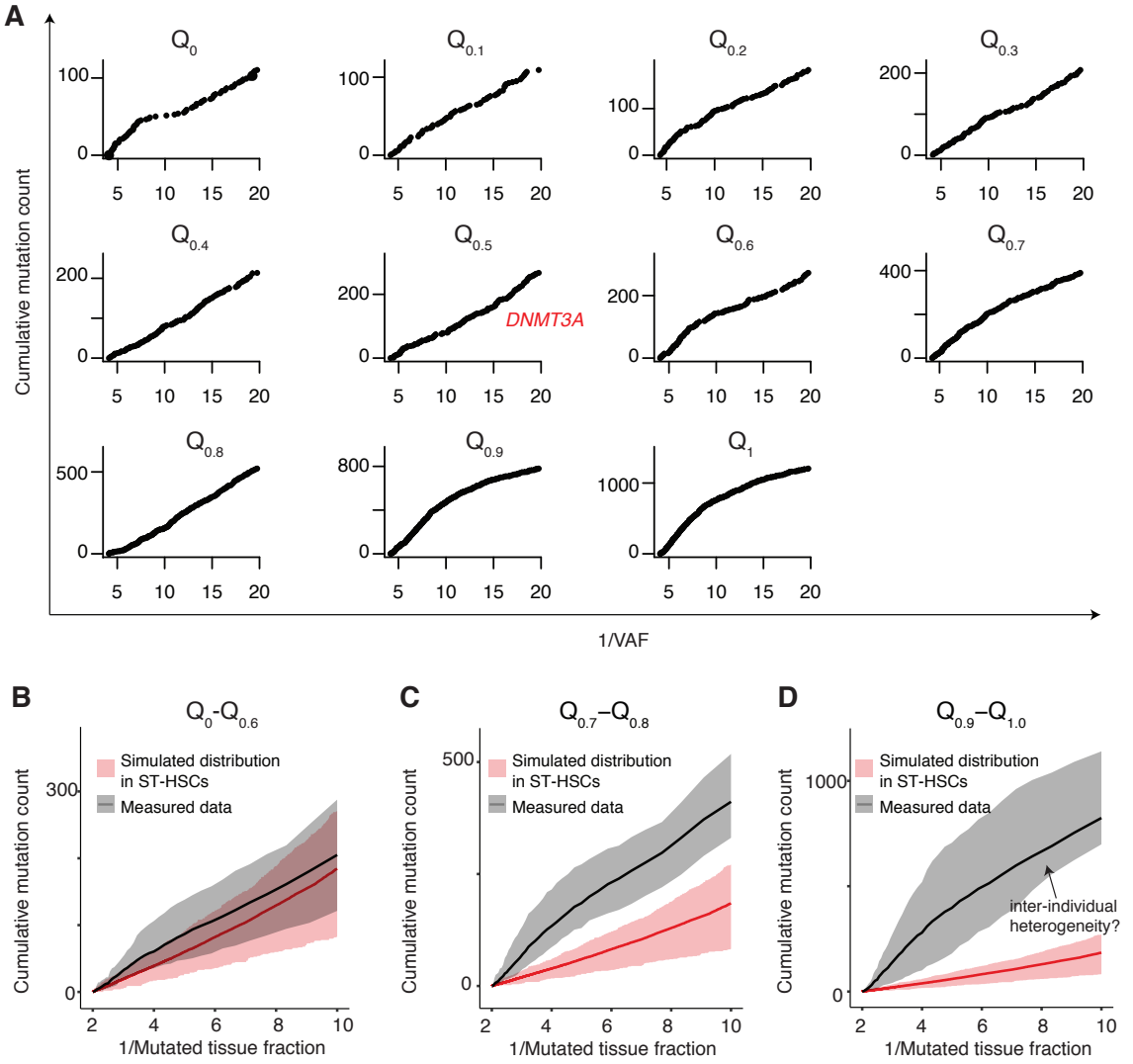


Figure 3.16: Cumulative mutation frequency distributions measured in blood samples from adults diagnosed with glioblastoma (shown are mutations supported by at least three sequencing reads, excluding mutations with VAFs≤0.05 and VAFs≥0.25; values are extrapolated from the analyzed genome fraction to the entire genome). **A** Ten exemplary samples, corresponding to the deciles of the measured subclonal mutation count among the cohort of in total 39 samples. Leukemic driver mutations are indicated in red. **B-D** Mean (lines) and 95% confidence intervals (shaded areas) of 100 simulated mutation frequency distributions in ST-HSCs. Simulations were done as in Fig. 3.13C (middle panel) and as described in Section 3.4.1, but for 50 years and at a mutation rate of six mutations per cell division (all other parameters were taken as before, Table 3.1). The measured mean and 95% confidence intervals of the samples corresponding to the lower six deciles (B), the seventh and the eigth deciles (C) and the ninth and tenth deciles (D) are shown alongside and are colored in black (mutated tissue fractions were obtained by multiplying the measured VAFs of heterozygous mutations with two to account for diploid genomes). Experimental data: Peter Lichter & 'SysGlio' Consortium.

3.5 Discussion

The transition of normal tissues to pre-cancerous lesions is of particular interest to understand the cellular processes governing tumor initiation and to guide early diagnosis. In the current chapter, we analyzed how clonal dynamics in normal tissues imprint on the mutation frequency distribution measured with whole genome sequencing and how this knowledge may be used to differentiate between normal and pre-cancerous situations.

The purpose of this chapter is two-fold. On the one hand, we developed a theoretical framework to describe mutation accumulation in neutrally evolving somatic tissues. On the other hand, we analyzed how this framework may be used to identify pre-leukemic lesions with whole genome sequencing. To this end, we first assessed the validity of existing theory on mutation accumulation in exponentially growing tissues by comparing the model predictions to stochastic simulations. This revealed that a deterministic view on mutation accumulation, as suggested by Williams et al., 2016, relies on simplifying assumptions that are valid in the limit of negligible cell loss only. By contrast, we showed that, in general, reliable predictions of the mutation frequency distribution are only obtained if accounting for the stochasticity of clonal drift, as suggested by Ohtsuki and Innan, 2017. It is noteworthy that cell loss, due to death or differentiation, is non-negligible in many physiological, but also pathological tissue expansions. For example, embryonic expansion of the hematopoietic stem cell compartment is accompanied by the production of mature blood cells via differentiation of early hematopoietic stem cells (Golub and Cumano, 2013). Likewise, many tumors retain a functional hierarchy, in which only a minor fraction of the daughter cells produced by 'cancer stem cell' divisions survives lastingly and contributes to tumor growth, while most of the daughter cells eventually die. To give an example, we showed in Chapter 2 that high cell death rates are a characteristic of glioblastoma growth, which may be due to functional heterogeneity (see also Lan et al., 2017) or other growth impediments. Thus a deterministic view on mutation accumulation in expanding tissues likely oversimplifies the mutational dynamics in many cases.

While existing theory on mutation accumulation in somatic tissues has primarily focused on exponentially growing cell populations, such as tumors, physiological tissues typically undergo homeostatic turnover upon embryonic expansion to their definite size. Thus, in order to predict mutation accumulation in adult tissues over time, we extended the model by Ohtsuki and Innan, 2017, to a two-stage system of initial expansion and subsequent homeostasis. This extension is necessary to understand the expected mutation frequency distribution in normal somatic tissues and thus to distinguish neutral mutation accumulation from pre-cancerous situations. Moreover, coupling an expansion phase with subsequent homeostasis may also be useful when sequencing tumors that grow according to the Gompertzian law and have approached the carrying capacity. We validated our theoretical framework with stochastic simulations and showed that the analytical solution is on average accurate and may hence aid a mechanistic interpretation of bulk whole genome sequencing data. However, stochastic simulations of the analogous process showed that individual measurements may deviate from the average

prediction as the mutation frequeny distribution varies according to the stochasticity of mutation accumulation. Whether an analytical prediction of the variance inherent to the stochastic process can be found analytically is subject to future work. Until then model predictions should ideally be complemented with stochastic simulations and compared to an ensemble of measurements.

The advantage of stochastic simulations to validate our model predictions becomes especially evident when measuring mutations in a downstream compartment of a hierarchical system as discussed in Section 3.4.1. Here, a complete analytical description of the clonal drift dynamics would become computationally cumbersome as multiple summation and integration steps are required. Nevertheless, using the example of murine hematopoiesis, we showed that divisions during progressive differentiation affect the measurable mutation frequency distribution only if the downstream compartments have a high capacity to self-renew (e.g., ST-HSCs). By contrast, transiently amplifying compartments that rapidly turn over (e.g., MPPs) do not change the mutation frequencies in the mesaurable window. Thus WGS of downstream compartments can serve as a reliable readout for the mutation frequency distribution in the stem cell compartment, but this needs to be examined for each tissue individually.

A future application of our theoretical framework could lie in the identification of pre-cancerous lesions with deep whole genome sequencing. The idea behind this approach is that non-neutral growth dynamics associated with tumor initiation cause deviations of the mutation frequency distribution from the neutral expectation. Thus, in order to develop this approach in the future, it is necessary to understand how whole genome sequencing data of normal and pre-cancerous tissue samples conform to our model predictions. As an outlook to this application, we analyzed available sequencing data of undisturbed hematopoiesis in the second part of this chapter. First, we used whole genome sequencing data from granulocytes of two healthy mice of different ages. Unexpectedly, we observed that the mutational burden at small variant allele frequencies is markedly increased in the granulocytes of the two-year-old mouse as compared to the eight-week-old counterpart, and that this increase is not expected under the current understanding of the size and the dynamics of the hematopoietic system in mice (Bowie et al., 2007; Busch et al., 2015; Colvin et al., 2004; Höfer and Rodewald, 2016). However, model predictions can be tailored to match the experimental observation if decreasing the number of stem cells, while increasing the rate of cellular turnover. If this is true, it either implies a different parametrization of the hematopoietic system than currently agreed on, or suggests that non-neutral dynamics affected the mutational burden in the two-year-old mouse. Indeed, we found a mutation in the *Jak3* gene that may account for the observed increase in the mutation count by altering the dynamics of the hematopoietic system (Cornejo et al., 2009; Rivera-Munoz et al., 2018). However, as the current dataset lacks replicates, robust conclusions on the hematopoietic dynamics during ageing cannot be drawn at this stage. It is hence unclear whether the present observation reflected a rare outlier, a pathological situation or normal hematopoietic dynamics during ageing. The presented data should thus be seen as a motivating example for further sequencing studies on normal blood.

The advantage of inferring clonal dynamics with next generation sequencing, as compared to fate mapping or cellular barcoding, lies in the minimally invasive character of data acquisition by which the system can be studied in an unperturbed way. Moreover, this approach can be easily applied to humans, for whom the dynamics of tissue maintenance are known in much less detail. To illustrate this potential, we analyzed the mutation frequency distribution in blood samples from children and adults with diagnosed neuro- or glioblastoma, but without diagnosed leukemia. The human dataset, in contrast to the murine dataset, had the advantage of a large sample size, which increases the statistical power. On the downside, however, a mixed cell population of unsorted leukocytes was sequenced, and much less is known on the dynamics of human hematopoiesis than on murine hematopoiesis. As expected for human data, the mutational burden varied between individual patients, which may in part be due to stochastic drift, but also due to differences in the tissue dynamics, the individual disease history and genetic predisposition. Nevertheless, a subgroup of the neuroblastoma patients stood out by having many more mutations than the rest, reminiscent of a 'hypermutation genotype'. Patients of this subgroup accumulated several leukemic driver mutations associated with both myeloid and lymphoid leukemias and in two cases also mutations in mismatch repair genes. Interestingly, the blood samples of these patients had been taken prior to chemoradiation therapy, so that a therapy-induced hypermutation genotype is unlikely. Although the cause of these hypermutation genotypes remains unknown, this observation indicates that pathological states may be identifiable in the mutation frequency distribution measured with whole genome sequencing.

Apart from the hypermutated cases, the majority of human blood samples were qualitatively reconciled with the theoretical expectation of neutral mutation accumulation. Interestingly, the mutation rate appeared to be higher among children (approximately fourteen base pair substitutions per division) than among adults (approximately six base pair substitutions per division). This difference might be due to the different lengths of the genomic regions that could be used for mutation calling (on average 2.2×10^9 basepairs in glioblastoma patients and 1.8×10^9 basepairs in neuroblastoma patients). Although we attempted to account for this difference by extrapolating the mutation count to the entire genome, more robust results are expected with a proper germline control, consisting of a pure sample of normal non-hematopoietic tissue. Moreover, measuring mutations in sorted granulocytes may provide more accurate insights in the dynamics of mutation acquisition than measuring a mixed population of leukocytes whose composition varies both between individuals and between children and adults.

Interestingly, our estimate of six to fourteen mutations per cell division contradicts two previous studies that analyzed the mutation frequency distribution in normal blood samples from 241 adults (Ju et al., 2017) and in single cell expansions of hematopoietic stem cells from a single individual with whole genome sequencing (Lee-Six et al., 2018). These studies estimated that the mutation rate during early embryogenesis is at most three base pair substitutions per division. A possible reason for this difference may be the lower sequencing coverage used in these studies

(15x and 30x) compared to our data (80x), as well as the different algorithms used to call somatic variants. Here, a re-analysis of our data with different variant callers would be instructive, but goes beyond the scope of this thesis.

In summary, the project presented in this chapter highlights the potentials and limitations of whole genome sequencing studies to analyze clonal dynamics in tissues *in vivo*. We learned from theoretical considerations how different clonal dynamics manifest on the measurable mutation frequency distribution in stem cells and more downstream compartments. These thought experiments showed that — putting aside the mutation rate — the relevant parameters influencing the measured distribution are the compartment size, the degree of self renewal and the speed of cellular turnover. Comparison of the model predictions with exploratory datasets qualitatively confirmed our theoretical predictions in most cases, but revealed unexpected phenotypes in a smaller subset. Thus combining our theoretical framework with sequencing data of defined cell populations may contribute to the development of new tools for early cancer diagnosis, and also to a better understanding of human tissue dynamics *in vivo*. It will be interesting how these questions can be tackled within the emerging field of next generation sequencing in the future.

CHAPTER 4

Final discussion & conclusions

According to the 'cancer evolution hypothesis' (Nowell, 1976), tumors progress towards more malignant stages by iterative cycles of mutation and selection. In this view, genetic evolution and tumor growth are mutually dependent, since the rate of cell divisions is determined by the genotype, and, conversely, the acquisition of genetic mutations is driven by cell divisions. Next generation sequencing has facilitated the measurement of genetic heterogeneity within tissue samples, providing a glimpse on the somatic evolution in normal tissues and cancer.

In this thesis, I analyzed the evolutionary dynamics during glioblastoma growth and hematopoiesis. Both projects relied on whole genome sequencing data and thus reveal the potentials and limitations of deep sequencing to understand tissue dynamics in health and disease. In homeostatic tissues, such as the blood or the skin, cells are continuously replaced by influx from a stem cell compartment, and constant population sizes are achieved by a detailed balance between cell divisions and death. Thus in the absence of functional mutations, one may assume that the probabilities for cell division, death and mutation are homogenous among all cells of a particular differentiation stage. We saw in Chapter 3 that full analytical predictions for the mutation frequency distribution in normal tissues can be obtained with few simplifying assumptions, and that these predictions depend on the cellular turnover rate, the mutation rate and the size of the tissue. Moreover, a measured deviation from the expected mutation frequency distribution may serve to identify pre-cancerous lesions. Thus sequencing normal tissues may indeed reveal characteristics of tissue homeostasis and tumor initiation, although prior knowledge on the architecture and the dynamics of the specific tissue facilitates data interpretation. With decreasing costs of high-throughput sequencing studies, it will be interesting to trace the mutational profiles in the blood of large patient cohorts longitudinally. As some patients may eventually progress to leukemia, studies of this kind have the potential to

not only identify early markers of tumor initiation, but also the time scale over which the tumor develops. Eventually, this may help to develop new tools for early diagnosis, the most promising approach to combat cancer.

In contrast to normal tissues, cancer grows from unbalanced cell divisions, due to one or multiple functional mutations. As individual cancer cells may acquire additional driver mutations over the course of tumorigenesis, it is unclear whether they share equal probabilities to divide, die and mutate. Measuring evolutionary dynamics in cancer is therefore more challenging than in normal tissues and predicting the full mutation frequency distribution is practically impossible. However, insights into the evolutionary history of cancers can still be gained from whole genome sequencing data if adjusting the research question to the explanatory power of the data. Moreover, the predictive power increases with the size of the studied cohort and with prior knowledge on the functional effect of individual driver mutations. We saw in Chapter 2 that the measured mutation frequency distribution in cancer reflects major subclonal branchings and thus reveals the temporal order of driver mutations over the course of tumorigenesis. In addition, coarse dynamical estimates, such as the extent of cell death during tumor growth or the selective advantage provided by pervasive driver mutations are contained in deep sequencing data from cancer samples. Future studies, combining bulk whole genome sequencing with genomic, epigenomic and transcriptomic information on the single cell level, will help to gain a more comprehensive picture of the evolutionary dynamics during tumor growth. In this way, the contribution of different levels of heterogeneity to cancer progression and treatment failure may be disentangled. Moreover, with genomic editing becoming increasingly available, it will be interesting to recapitulate the genetic history of human tumors in mouse models more realistically. In that regard, it may also become possible to test if an early, tumor initiating mutation becomes dispensable at a later stage of tumorigenesis, or whether the entire tumor can be treated by efficient therapeutic targeting of this mutation. The road to take should therefore complement the more phenomenological insights into cancer evolution, obtained from high-throughput sequencing studies of human cancers, with functional studies in adequate model systems, which may support the design of novel treatment approaches.

Taken together, apart from biological insights into the evolution of glioblastomas and normal blood cells, this thesis also provides general insights into the potential of deep whole genome sequencing data to uncover the evolutionary dynamics underlying tumor initiation and progression. It remains to be seen whether these insights can contribute to therapeutic approaches or diagnostic tools in the future.

Bibliography

Akashi, K., Traver, D., Miyamoto, T., and Weissman, I. L. (2000). A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*, 404(6774):193–197.

Al-Kaabi, A., Van Bockel, L., Pothen, A., and Willems, S. (2014). p16INK4A and p14ARF gene promoter hypermethylation as prognostic biomarker in oral and oropharyngeal squamous cell carcinoma: a review. *Dis. Markers*, 2014:260549.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*. Garland, New York.

Alexander, B. M. and Cloughesy, T. F. (2017). Adult glioblastoma. *J. Clin. Oncol.*, 35(21):2402–2409.

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., and Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. *Nat. Genet.*, 47(12):1402.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., 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., Eyfjord, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Imielinsk, M., Jager, N., Jones, D. T., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., Lopez-Otin, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., 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., Valdes-Mas, R., van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., 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., Claviez, A., Rosenwald, A., Rosenwald, A., Borkhardt, A., Brors, B., Radlwimmer, B., Lawerenz, C., Lopez, C., Langenberger, D., Karsch, D., Lenze, D., Kube, D., Leich, E., Richter, G., Korbel, J., Hoell, J., Eils, J., Hezaveh, K., Trumper,

L., Rosolowski, M., Weniger, M., Rohde, M., Kreuz, M., Loeffler, M., Schilhabel, M., Dreyling, M., Hansmann, M. L., Hummel, M., Szczepanowski, M., Ammerpohl, O., Stadler, P. F., Moller, P., Kuppers, R., Haas, S., Eberth, S., Schreiber, S., Bernhart, S. H., Hoffmann, S., Radomski, S., Kostezka, U., Klapper, W., Sotiriou, C., Larsimont, D., Vincent, D., Maetens, M., Mariani, O., Sieuwerts, A. M., Martens, J. W., Jonasson, J. G., Treilleux, I., Thomas, E., Mac Grogan, G., Mannina, C., Arnould, L., Burillier, L., Merlin, J. L., Lefebvre, M., Bibeau, F., Massemin, B., Penault-Llorca, F., Lopez, Q., Mathieu, M. C., Lonning, P. E., Schlooz-Vries, M., Tol, J., van Laarhoven, H., Sweep, F., and Bult, P. (2013a). Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., and Stratton, M. R. (2013b). Deciphering signatures of mutational processes operative in human cancer. *Cell reports*, 3(1):246–259.

Alioto, T. S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M. D., Hovig, E., Heisler, L. E., Beck, T. A., Simpson, J. T., Tonon, L., Sertier, A. S., Patch, A. M., Jager, N., Ginsbach, P., Drews, R., Paramasivam, N., Kabbe, R., Chotewutmontri, S., Diessl, N., Previti, C., Schmidt, S., Brors, B., Feuerbach, L., Heinold, M., Grobner, S., Korshunov, A., Tarpey, P. S., Butler, A. P., Hinton, J., Jones, D., Menzies, A., Raine, K., Shepherd, R., Stebbings, L., Teague, J. W., Ribeca, P., Giner, F. C., Beltran, S., Raineri, E., Dabad, M., Heath, S. C., Gut, M., Denroche, R. E., Harding, N. J., Yamaguchi, T. N., Fujimoto, A., Nakagawa, H., Quesada, V., Valdes-Mas, R., Nakken, S., Vodak, D., Bower, L., Lynch, A. G., Anderson, C. L., Waddell, N., Pearson, J. V., Grimmond, S. M., Peto, M., Spellman, P., He, M., Kandoth, C., Lee, S., Zhang, J., Letourneau, L., Ma, S., Seth, S., Torrents, D., Xi, L., Wheeler, D. A., Lopez-Otin, C., Campo, E., Campbell, P. J., Boutros, P. C., Puente, X. S., Gerhard, D. S., Pfister, S. M., McPherson, J. D., Hudson, T. J., Schlesner, M., Lichter, P., Eils, R., Jones, D. T., and Gut, I. G. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat Commun*, 6:10001.

Amorim, J. P., Santos, G., Vinagre, J., and Soares, P. (2016). The Role of ATRX in the Alternative Lengthening of Telomeres (ALT) Phenotype. *Genes*, 7(9):66.

Arita, H., Yamasaki, K., Matsushita, Y., Nakamura, T., Shimokawa, A., Takami, H., Tanaka, S., Mukasa, A., Shirahata, M., Shimizu, S., Suzuki, K., Saito, K., Kobayashi, K., Higuchi, F., Uzuka, T., Otani, R., Tamura, K., Sumita, K., Ohno, M., Miyakita, Y., Kagawa, N., Hashimoto, N., Hatae, R., Yoshimoto, K., Shinojima, N., Nakamura, H., Kanemura, Y., Okita, Y., Kinoshita, M., Ishibashi, K., Shofuda, T., Kodama, Y., Mori, K., Tomogane, Y., Fukai, J., Fujita, K., Terakawa, Y., Tsuyuguchi, N., Moriuchi, S., Nonaka, M., Suzuki, H., Shibuya, M., Maehara, T., Saito, N., Nagane, M., Kawahara, N., Ueki, K., Yoshimine, T., Miyaoka, E., Nishikawa, R., Komori, T., Narita, Y., and Ichimura, K. (2016). A combination of TERT promoter mutation and MGMT methylation status predicts clinically relevant subgroups of newly diagnosed glioblastomas. *Acta Neuropathol Commun*, 4(1):79.

Bachoo, R. M., Maher, E. A., Ligon, K. L., Sharpless, N. E., Chan, S. S., You, M. J., Tang, Y., DeFrances, J., Stover, E., Weissleder, R., Rowitch, D. H., Louis, D. N., and DePinho, R. A. (2002). Epidermal growth factor receptor and Ink4a/Arf: convergent mechanisms governing terminal differentiation and transformation along the neural stem cell to astrocyte axis. *Cancer Cell*, 1(3):269–277.

Bailey, N. (1964). *The Elements of Stochastic Processes with Applications to the Natural Sciences*. John Wiley.

Balss, J., Meyer, J., Mueller, W., Korshunov, A., Hartmann, C., and von Deimling, A. (2008). Analysis of the IDH1 codon 132 mutation in brain tumors. *Acta Neuropathol.*, 116(6):597–602.

Barthel, F. P., Wesseling, P., and Verhaak, R. G. (2018). Reconstructing the molecular life history of gliomas. *Acta Neuropathol.*, 135(5):649–670.

Batlle, E. and Clevers, H. (2017). Cancer stem cells revisited. *Nat. Med.*, 23(10):1124–1134.

Beerenwinkel, N., Schwarz, R. F., Gerstung, M., and Markowetz, F. (2014). Cancer evolution: mathematical models and computational inference. *Syst. Biol.*, 64(1):1–25.

Bernstein, J. J. and Woodard, C. A. (1995). Glioblastoma cells do not intravasate into blood vessels. *Neurosurgery*, 36(1):124–132.

Bhat, K. P. L., Balasubramaniyan, V., Vaillant, B., Ezhilarasan, R., Hummelink, K., Hollingsworth, F., Wani, K., Heathcock, L., James, J. D., Goodman, L. D., Conroy, S., Long, L., Lelic, N., Wang, S., Gumin, J., Raj, D., Kodama, Y., Raghunathan, A., Olar, A., Joshi, K., Pelloski, C. E., Heimberger, A., Kim, S. H., Cahill, D. P., Rao, G., Den Dunnen, W. F. A., Boddeke, H. W. G. M., Phillips, H. S., Nakano, I., Lang, F. F., Colman, H., Sulman, E. P., and Aldape, K. (2013). Mesenchymal differentiation mediated by NF-κB promotes radiation resistance in glioblastoma. *Cancer Cell*, 24(3):331–346.

Bigner, S. H., Mark, J., Bullard, D. E., Mahaley Jr, M. S., and Bigner, D. D. (1986). Chromosomal evolution in malignant human gliomas starts with specific and usually numerical deviations. *Cancer Genet. Cytogenet.*, 22(2):121–135.

Bigner, S. H., Mark, J., Mahaley, M. S., and Bigner, D. D. (1984). Patterns of the early, gross chromosomal changes in malignant human gliomas. *Hereditas*, 101(1):103–113.

Bigner, S. H., Wong, A. J., Mark, J., Muhlbaier, L. H., Kinzler, K. W., Vogelstein, B., and Bigner, D. D. (1987). Relationship between gene amplification and chromosomal deviations in malignant human gliomas. *Cancer Genet. Cytogenet.*, 29(1):165–170.

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., Verstegen, M. M., van der Laan, L. J., de Jonge, J., IJzermans, J. N., Vries, R. G., van de Wetering, M., Stratton, M. R., Clevers, H., Cuppen, E., and van Boxtel, R. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, 538(7624):260–264.

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2011). Control-freec: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, 28(3):423–425.

Bowie, M. B., Kent, D. G., Dykstra, B., McKnight, K. D., McCaffrey, L., Hoodless, P. A., and Eaves, C. J. (2007). Identification of a new intrinsically timed developmental checkpoint that reprograms key hematopoietic stem cell properties. *Proc. Natl. Acad. Sci. U.S.A.*, 104(14):5878–5882.

Bozic, I., Gerold, J. M., and Nowak, M. A. (2016). Quantifying clonal and subclonal passenger mutations in cancer evolution. *PLoS Comput. Biol.*, 12(2):e1004731.

Brastianos, P. K., Nayyar, N., Rosebrock, D., Leshchiner, I., Gill, C. M., Livitz, D., Bertalan, M. S., D'Andrea, M., Hoang, K., Aquilanti, E., Chukwueke, U. N., Kaneb, A., Chi, A., Plotkin, S., Gerstner, E. R., Frosch, M. P., Suva, M. L., Cahill, D. P., Getz, G., and Batchelor, T. T. (2017). Resolving the phylogenetic origin of glioblastoma via multifocal genomic analysis of pre-treatment and treatment-resistant autopsy specimens. *NPJ Precis Oncol*, 1(1):33.

Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., Zheng, S., Chakravarty, D., Sanborn, J. Z., Berman, S. H., Beroukhim, R., Bernard, B., Wu, C. J., Genovese, G., Shmulevich, I., Barnholtz-Sloan, J., Zou, L., Vegesna, R., Shukla, S. A., Ciriello, G., Yung, W. K., Zhang, W., Sougnez, C., Mikkelsen, T., Aldape, K., Bigner, D. D., Van Meir, E. G., Prados, M., Sloan, A., Black, K. L., Eschbacher, J., Finocchiaro, G., Friedman, W., Andrews, D. W., Guha, A., Iacocca, M., O'Neill, B. P., Foltz, G., Myers, J., Weisenberger, D. J., Penny, R., Kucherlapati, R., Perou, C. M., Hayes, D. N., Gibbs, R., Marra, M., Mills, G. B., Lander, E., Spellman, P., Wilson, R., Sander, C., Weinstein, J., Meyerson, M., Gabriel, S., Laird, P. W., Haussler, D., Getz, G., Chin, L., Benz, C., Barnholtz-Sloan, J., Barrett, W., Ostrom, Q., Wolinsky, Y., Black, K. L., Bose, B., Boulos, P. T., Boulos, M., Brown, J., Czerinski, C., Eppley, M., Iacocca, M., Kempista, T., Kitko, T., Koyfman, Y., Rabeno, B., Rastogi, P., Sugarman, M., Swanson, P., Yalamanchii, K., Otey, I. P., Liu, Y. S., Xiao, Y., Auman, J. T., Chen, P. C., Hadjipanayis, A., Lee, E., Lee, S., Park, P. J., Seidman, J., Yang, L., Kucherlapati, R., Kalkanis, S., Mikkelsen, T., Poisson, L. M., Raghunathan, A., Scarpace, L., Bernard, B., Bressler, R., Eakin, A., Iype, L., Kreisberg, R. B., Leinonen, K., Reynolds, S., Rovira, H., Thorsson, V., Shmulevich, I., Annala, M. J., Penny, R., Paulauskis, J., Curley, E., Hatfield, M., Mallery, D., Morris, S., Shelton, T., Shelton, C., Sherman, M., Yena, P., Cuppini, L., DiMeco, F., Eoli, M., Finocchiaro, G., Maderna, E., Pollo, B., Saini, M., Balu, S., Hoadley, K. A., Li, L., Miller, C. R., Shi, Y., Topal, M. D., Wu, J., Dunn, G., Giannini, C., O'Neill, B. P., Aksoy, B. A., Antipin, Y., Borsu, L., Berman, S. H., Brennan, C. W., Cerami,

E., Chakravarty, D., Ciriello, G., Gao, J., Gross, B., Jacobsen, A., Ladanyi, M., Lash, A., Liang, Y., Reva, B., Sander, C., Schultz, N., Shen, R., Socci, N. D., Viale, A., Ferguson, M. L., Chen, Q. R., Demchok, J. A., Dillon, L. A., Shaw, K. R., Sheth, M., Tarnuzzer, R., Wang, Z., Yang, L., Davidsen, T., Guyer, M. S., Ozenberger, B. A., Sofia, H. J., Bergsten, J., Eckman, J., Harr, J., Myers, J., Smith, C., Tucker, K., Winemiller, C., Zach, L. A., Ljubimova, J. Y., Eley, G., Ayala, B., Jensen, M. A., Kahn, A., Pihl, T. D., Pot, D. A., Wan, Y., Eschbacher, J., Foltz, G., Hansen, N., Hothi, P., Lin, B., Shah, N., Yoon, J. G., Lau, C., Berens, M., Ardlie, K., Beroukhim, R., Carter, S. L., Cherniack, A. D., Noble, M., Cho, J., Cibulskis, K., DiCara, D., Frazer, S., Gabriel, S. B., Gehlenborg, N., Gentry, J., Heiman, D., Kim, J., Jing, R., Lander, E. S., Lawrence, M., Lin, P., Mallard, W., Meyerson, M., Onofrio, R. C., Saksena, G., Schumacher, S., Sougnez, C., Stojanov, P., Tabak, B., Voet, D., Zhang, H., Zou, L., Getz, G., Dees, N. N., Ding, L., Fulton, L. L., Fulton, R. S., Kanchi, K. L., Mardis, E. R., Wilson, R. K., Baylin, S. B., Andrews, D. W., Harshyne, L., Cohen, M. L., Devine, K., Sloan, A. E., VandenBerg, S. R., Berger, M. S., Prados, M., Carlin, D., Craft, B., Ellrott, K., Goldman, M., Goldstein, T., Grifford, M., Haussler, D., Ma, S., Ng, S., Salama, S. R., Sanborn, J. Z., Stuart, J., Swatloski, T., Waltman, P., Zhu, J., Foss, R., Frentzen, B., Friedman, W., McTiernan, R., Yachnis, A., Hayes, D. N., Perou, C. M., Zheng, S., Vegesna, R., Mao, Y., Akbani, R., Aldape, K., Bogler, O., Fuller, G. N., Liu, W., Liu, Y., Lu, Y., Mills, G., Protopopov, A., Ren, X., Sun, Y., Wu, C. J., Yung, W. K., Zhang, W., Zhang, J., Chen, K., Weinstein, J. N., Chin, L., Verhaak, R. G., Noushmehr, H., Weisenberger, D. J., Bootwalla, M. S., Lai, P. H., Triche, T. J., Van Den Berg, D. J., Laird, P. W., Gutmann, D. H., Lehman, N. L., VanMeir, E. G., Brat, D., Olson, J. J., Mastrogianakis, G. M., Devi, N. S., Zhang, Z., Bigner, D., Lipp, E., and McLendon, R. (2013). The somatic genomic landscape of glioblastoma. *Cell*, 155(2):462–477.

Brosnan-Cashman, J. A., Yuan, M., Graham, M. K., Rizzo, A. J., Myers, K. M., Davis, C., Zhang, R., Esopi, D. M., Raabe, E. H., Eberhart, C. G., Heaphy, C. M., and Meeker, A. K. (2018). ATRX loss induces multiple hallmarks of the alternative lengthening of telomeres (ALT) phenotype in human glioma cell lines in a cell line-specific manner. *PLoS ONE*, 13(9):e0204159.

Busch, K., Klapproth, K., Barile, M., Flossdorf, M., Holland-Letz, T., Schlenner, S. M., Reth, M., Höfer, T., and Rodewald, H.-R. (2015). Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature*, 518(7540):542.

Busque, L., Mio, R., Mattioli, J., Brais, E., Blais, N., Lalonde, Y., Maragh, M., and Gilliland, D. G. (1996). Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood*, 88(1):59–65.

Cadet, J. and Wagner, J. R. (2013). DNA base damage by reactive oxygen species, oxidizing agents, and UV radiation. *Cold Spring Harb Perspect Biol*, 5(2):a012559.

Cahill, D. P., Levine, K. K., Betensky, R. A., Codd, P. J., Romany, C. A., Reavie, L. B., Batchelor, T. T., Futreal, P. A., Stratton, M. R., Curry, W. T., Iafrate, A. J., and Louis, D. N. (2007). Loss of the

mismatch repair protein MSH6 in human glioblastomas is associated with tumor progression during temozolomide treatment. *Clin. Cancer Res.*, 13(7):2038–2045.

Capper, D., Jones, D. T. W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D. E., Kratz, A., Wefers, A. K., Huang, K., Pajtler, K. W., Schweizer, L., Stichel, D., Olar, A., Engel, N. W., Lindenberg, K., Harter, P. N., Braczynski, A. K., Plate, K. H., Dohmen, H., Garvalov, B. K., Coras, R., Holsken, A., Hewer, E., Bewerunge-Hudler, M., Schick, M., Fischer, R., Beschorner, R., Schittenhelm, J., Staszewski, O., Wani, K., Varlet, P., Pages, M., Temming, P., Lohmann, D., Selt, F., Witt, H., Milde, T., Witt, O., Aronica, E., Giangaspero, F., Rushing, E., Scheurlen, W., Geisenberger, C., Rodriguez, F. J., Becker, A., Preusser, M., Haberler, C., Bjerkvig, R., Cryan, J., Farrell, M., Deckert, M., Hench, J., Frank, S., Serrano, J., Kannan, K., Tsirigos, A., Bruck, W., Hofer, S., Brehmer, S., Seiz-Rosenhagen, M., Hanggi, D., Hans, V., Rozsnoki, S., Hansford, J. R., Kohlhof, P., Kristensen, B. W., Lechner, M., Lopes, B., Mawrin, C., Ketter, R., Kulozik, A., Khatib, Z., Heppner, F., Koch, A., Jouvet, A., Keohane, C., Muhleisen, H., Mueller, W., Pohl, U., Prinz, M., Benner, A., Zapatka, M., Gottardo, N. G., Driever, P. H., Kramm, C. M., Muller, H. L., Rutkowski, S., von Hoff, K., Fruhwald, M. C., Gnekow, A., Fleischhack, G., Tippelt, S., Calaminus, G., Monoranu, C. M., Perry, A., Jones, C., Jacques, T. S., Radlwimmer, B., Gessi, M., Pietsch, T., Schramm, J., Schackert, G., Westphal, M., Reifenberger, G., Wesseling, P., Weller, M., Collins, V. P., Blumcke, I., Bendszus, M., Debus, J., Huang, A., Jabado, N., Northcott, P. A., Paulus, W., Gajjar, A., Robinson, G. W., Taylor, M. D., Jaunmuktane, Z., Ryzhova, M., Platten, M., Unterberg, A., Wick, W., Karajannis, M. A., Mittelbronn, M., Acker, T., Hartmann, C., Aldape, K., Schuller, U., Buslei, R., Lichter, P., Kool, M., Herold-Mende, C., Ellison, D. W., Hasselblatt, M., Snuderl, M., Brandner, S., Korshunov, A., von Deimling, A., and Pfister, S. M. (2018). DNA methylation-based classification of central nervous system tumours. *Nature*, 555(7697):469–474.

Chatterjee, N. and Walker, G. C. (2017). Mechanisms of DNA damage, repair, and mutagenesis. *Environ. Mol. Mutagen.*, 58(5):235–263.

Cheng, Y.-K., Beroukhim, R., Levine, R. L., Mellinghoff, I. K., and Michor, F. (2011). Reply to Parsons: Many tumor types follow the monoclonal model of tumor initiation. *Proc. Natl. Acad. Sci. U.S.A.*, page 201018584.

Chiba, K., Johnson, J. Z., Vogan, J. M., Wagner, T., Boyle, J. M., and Hockemeyer, D. (2015). Cancer-associated TERT promoter mutations abrogate telomerase silencing. *Elife*, 4:e07918.

Chiba, K., Lorbeer, F. K., Shain, A. H., McSwiggen, D. T., Schruf, E., Oh, A., Ryu, J., Darzacq, X., Bastian, B. C., and Hockemeyer, D. (2017). Mutations in the promoter of the telomerase gene TERT contribute to tumorigenesis by a two-step mechanism. *Science*, 357(6358):1416–1420.

Christensen, J. L., Wright, D. E., Wagers, A. J., and Weissman, I. L. (2004). Circulation and chemotaxis of fetal hematopoietic stem cells. *PLoS Biol.*, 2(3):e75.

Colvin, G., Lambert, J., Abedi, M., Hsieh, C., Carlson, J., Stewart, F., and Quesenberry, P. (2004). Murine marrow cellularity and the concept of stem cell competition: geographic and quantitative determinants in stem cell biology. *Leukemia*, 18(3):575–583.

Cornejo, M. G., Kharas, M. G., Werneck, M. B., Le Bras, S., Moore, S. A., Ball, B., Beylot-Barry, M., Rodig, S. J., Aster, J. C., Lee, B. H., Cantor, H., Merlio, J. P., Gilliland, D. G., and Mercher, T. (2009). Constitutive JAK3 activation induces lymphoproliferative syndromes in murine bone marrow transplantation models. *Blood*, 113(12):2746–2754.

Dahlback, H.-S. S., Brandal, P., Meling, T. R., Gorunova, L., Scheie, D., and Heim, S. (2009). Genomic aberrations in 80 cases of primary glioblastoma multiforme: Pathogenetic heterogeneity and putative cytogenetic pathways. *Genes Chromosomes Cancer*, 48(10):908–924.

Davis, A., Gao, R., and Navin, N. (2017). Tumor evolution: Linear, branching, neutral or punctuated? *Biochim Biophys Acta Rev Cancer*, 1867(2):151–161.

Davis, A. J. and Chen, D. J. (2013). DNA double strand break repair via non-homologous end-joining. *Transl Cancer Res*, 2(3):130–143.

De Witt Hamer, P. C. (2010). Small molecule kinase inhibitors in glioblastoma: a systematic review of clinical studies. *Neuro-oncology*, 12(3):304–316.

Del Monte, U. (2009). Does the cell number 10(9) still really fit one gram of tumor tissue? *Cell Cycle*, 8(3):505–506.

Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L., and Morris, Q. (2015). PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.*, 16(1):35.

Desouky, O., Ding, N., and Zhou, G. (2015). Targeted and non-targeted effects of ionizing radiation. *J Radiat Res Appl Sci*, 8(2):247–254.

Devita Jr, V. T., Young, R. C., and Canellos, G. P. (1975). Combination versus single agent chemotherapy: a review of the basis for selection of drug treatment of cancer. *Cancer*, 35(1):98–110.

Dick, J. E., Magli, M. C., Huszar, D., Phillips, R. A., and Bernstein, A. (1985). Introduction of a selectable gene into primitive stem cells capable of long-term reconstitution of the hemopoietic system of W/Wv mice. *Cell*, 42(1):71–79.

Drabløs, F., Feyzi, E., Aas, P. A., Vaagbø, C. B., Kavli, B., Bratlie, M. S., Peña-Diaz, J., Otterlei, M., Slupphaug, G., and Krokan, H. E. (2004). Alkylation damage in DNA and RNA–repair mechanisms and medical significance. *DNA Repair (Amst.)*, 3(11):1389–1407.

Drachman, D. A. (2005). Do we have brain to spare? *Neurology*, 64(12):2004–2005.

Engler, J. R., Robinson, A. E., Smirnov, I., Hodgson, J. G., Berger, M. S., Gupta, N., James, C. D., Molinaro, A., and Phillips, J. J. (2012). Increased microglia/macrophage gene expression in a subset of adult and pediatric astrocytomas. *PLoS ONE*, 7(8):e43339.

Fearon, E. R. and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767.

Fisher, R. A. (1999). *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press.

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., Stefancsik, R., Harsha, B., Kok, C. Y., Jia, M., Jubb, H., Sondka, Z., Thompson, S., De, T., and Campbell, P. J. (2016). COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, 45(D1):D777–D783.

Ford, C., Hamerton, J., Barnes, D., and Loutit, J. (1956). Cytological identification of radiation-chimaeras. *Nature*, 177(4506):452–454.

Francis, J. M., Zhang, C. Z., Maire, C. L., Jung, J., Manzo, V. E., Adalsteinsson, V. A., Homer, H., Haidar, S., Blumenstiel, B., Pedamallu, C. S., Ligon, A. H., Love, J. C., Meyerson, M., and Ligon, K. L. (2014). EGFR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing. *Cancer Discov*, 4(8):956–971.

Friedmann-Morvinski, D., Bushong, E. A., Ke, E., Soda, Y., Marumoto, T., Singer, O., Ellisman, M. H., and Verma, I. M. (2012). Dedifferentiation of neurons and astrocytes by oncogenes can induce gliomas in mice. *Science*, 338(6110):1080–1084.

Fu, D., Calvo, J. A., and Samson, L. D. (2012). Balancing repair and tolerance of DNA damage caused by alkylating agents. *Nat. Rev. Cancer*, 12(2):104–120.

Fujii, S., Akiyama, M., Aoki, K., Sugaya, Y., Higuchi, K., Hiraoka, M., Miki, Y., Saitoh, N., Yoshiyama, K., Ihara, K., Seki, M., Ohtsubo, E., and Maki, H. (1999). DNA replication errors produced by the replicative apparatus of Escherichia coli. *J. Mol. Biol.*, 289(4):835–850.

Fujisawa, H., Reis, R. M., Nakamura, M., Colella, S., Yonekawa, Y., Kleihues, P., and Ohgaki, H. (2000). Loss of heterozygosity on chromosome 10 is more extensive in primary (de novo) than in secondary glioblastomas. *Laboratory investigation*, 80(1):65–72.

Gao, Z., Wyman, M. J., Sella, G., and Przeworski, M. (2016). Interpreting the dependence of mutation rates on age and time. *PLoS Biol.*, 14(1):e1002355.

Genovese, G., Kahler, A. K., Handsaker, R. E., Lindberg, J., Rose, S. A., Bakhoum, S. F., Chambert, K., Mick, E., Neale, B. M., Fromer, M., Purcell, S. M., Svantesson, O., Landen, M., Hoglund, M., Lehmann, S., Gabriel, S. B., Moran, J. L., Lander, E. S., Sullivan, P. F., Sklar, P., Gronberg, H., Hultman, C. M., and McCarroll, S. A. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.*, 371(26):2477–2487.

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S. C., Gonzalez Rosado, S., Rosebrock, D., Mitchell, T. J., Rubanova, Y., Anur, P., Yu, K., Tarabichi, M., Deshwar, A., Wintersinger, J., Kleinheinz, K., Vazquez-Garcia, I., Haase, K., Jerman, L., Sengupta, S., Macintyre, G., Malikic, S., Donmez, N., Livitz, D. G., Cmero, M., Demeulemeester, J., Schumacher, S., Fan, Y., Yao, X., Lee, J., Schlesner, M., Boutros, P. C., Bowtell, D. D., Zhu, H., Getz, G., Imielinski, M., Beroukhim, R., Sahinalp, S. C. C., Ji, Y., Peifer, M., Markowetz, F., Mustonen, V., Yuan, K., Wang, W., Morris, Q. D., Spellman, P. T., Wedge, D. C., Van Loo, P., Evolution, P., Group, H. W., and network, P. (2017). The evolutionary history of 2,658 cancers. *bioRxiv*, page 161562.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361.

Goldberg-Zimring, D., Talos, I.-F., Bhagwat, J. G., Haker, S. J., Black, P. M., and Zou, K. H. (2005). Statistical validation of brain tumor shape approximation via spherical harmonics for image-guided neurosurgery. *Acad Radiol*, 12(4):459–466.

Golub, R. and Cumano, A. (2013). Embryonic hematopoiesis. *Blood Cells Mol. Dis.*, 51(4):226–231.

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., Santos, A., and Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods*, 10(11):1081–1082.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, 17(6):333–351.

Graf, T. (2008). Immunology: blood lines redrawn. *Nature*, 452(7188):702–703.

Greaves, M. and Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, 481(7381):306–313.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674.

Hecht, S. S. (2003). Tobacco carcinogens, their biomarkers and tobacco-induced cancer. *Nat. Rev. Cancer*, 3(10):733–744.

Hegi, M. E., Diserens, A. C., Gorlia, T., Hamou, M. F., de Tribolet, N., Weller, M., Kros, J. M., Hainfellner, J. A., Mason, W., Mariani, L., Bromberg, J. E., Hau, P., Mirimanoff, R. O., Cairncross, J. G., Janzer, R. C., and Stupp, R. (2005). MGMT gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.*, 352(10):997–1003.

Heide, T., Zapata, L., Williams, M. J., Werner, B., Barnes, C. P., Graham, T., and Sottoriva, A. (2018). Reply: Neutral tumor evolution? *bioRxiv*, page 274142.

Held, L. (2008). *Methoden der statistischen Inferenz*. Spektrum Akademischer Verlag Heidelberg.

Herculano-Houzel, S., Mota, B., and Lent, R. (2006). Cellular scaling rules for rodent brains. *Proc. Natl. Acad. Sci. U.S.A.*, 103(32):12138–12143.

Höfer, T. and Rodewald, H.-R. (2016). Output without input: the lifelong productivity of hematopoietic stem cells. *Curr. Opin. Cell Biol.*, 43:69–77.

Holland, E. C., Hively, W. P., DePinho, R. A., and Varmus, H. E. (1998). A constitutively active epidermal growth factor receptor cooperates with disruption of G1 cell-cycle arrest pathways to induce glioma-like lesions in mice. *Genes Dev.*, 12(23):3675–3685.

Huang, X., Liu, S., Wu, L., Jiang, M., and Hou, Y. (2018). High Throughput Single Cell RNA Sequencing, Bioinformatics Analysis and Applications. *Adv. Exp. Med. Biol.*, pages 33–43.

Hustedt, N. and Durocher, D. (2017). The control of DNA repair by the cell cycle. *Nat. Cell Biol.*, 19(1):1–9.

Jackson, S. P. and Bartek, J. (2009). The DNA-damage response in human biology and disease. *Nature*, 461(7267):1071–1078.

Jahn, K., Kuipers, J., and Beerenwinkel, N. (2016). Tree inference for single-cell data. *Genome Biol.*, 17(1):86.

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P. V., Mar, B. G., Lindsley, R. C., Mermel, C. H., Burtt, N., Chavez, A., Higgins, J. M., Moltchanov, V., Kuo, F. C., Kluk, M. J., Henderson, B., Kinnunen, L., Koistinen, H. A., Ladenvall, C., Getz, G., Correa, A., Banahan, B. F., Gabriel, S., Kathiresan, S., Stringham, H. M., McCarthy, M. I., Boehnke, M., Tuomilehto, J., Haiman, C., Groop, L., Atzmon, G., Wilson, J. G., Neuberg, D., Altshuler, D., and Ebert, B. L. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.*, 371(26):2488–2498.

Johnson, B. E., Mazor, T., Hong, C., Barnes, M., Aihara, K., McLean, C. Y., Fouse, S. D., Yamamoto, S., Ueda, H., Tatsuno, K., Asthana, S., Jalbert, L. E., Nelson, S. J., Bollen, A. W., Gustafson, W. C., Charron, E., Weiss, W. A., Smirnov, I. V., Song, J. S., Olshen, A. B., Cha, S., Zhao, Y., Moore, R. A., Mungall, A. J., Jones, S. J. M., Hirst, M., Marra, M. A., Saito, N., Aburatani, H., Mukasa, A.,

Berger, M. S., Chang, S. M., Taylor, B. S., and Costello, J. F. (2014). Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science*, 343(6167):189–193.

Joo, K. M., Kim, S. Y., Jin, X., Song, S. Y., Kong, D. S., Lee, J. I., Jeon, J. W., Kim, M. H., Kang, B. G., Jung, Y., Jin, J., Hong, S. C., Park, W. Y., Lee, D. S., Kim, H., and Nam, D. H. (2008). Clinical and biological implications of CD133-positive and CD133-negative cells in glioblastomas. *Laboratory investigation*, 88(8):808–815.

Ju, Y. S., Martincorena, I., Gerstung, M., Petljak, M., Alexandrov, L. B., Rahbari, R., Wedge, D. C., Davies, H. R., Ramakrishna, M., Fullam, A., Martin, S., Alder, C., Patel, N., Gamble, S., O'Meara, S., Giri, D. D., Sauer, T., Pinder, S. E., Purdie, C. A., Borg, A., Stunnenberg, H., van de Vijver, M., Tan, B. K., Caldas, C., Tutt, A., Ueno, N. T., van 't Veer, L. J., Martens, J. W., Sotiriou, C., Knappskog, S., Span, P. N., Lakhani, S. R., Eyfjord, J. E., Børresen-Dale, A. L., Richardson, A., Thompson, A. M., Viari, A., Hurles, M. E., Nik-Zainal, S., Campbell, P. J., and Stratton, M. R. (2017). Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature*, 543(7647):714–718.

Juratli, T. A., Thiede, C., Koerner, M. V. A., Tummala, S. S., Daubner, D., Shankar, G. M., Williams, E. A., Martinez-Lage, M., Soucek, S., Robel, K., Penson, T., Krause, M., Appold, S., Meinhardt, M., Pinzer, T., Miller, J. J., Krex, D., Ely, H. A., Silverman, I. M., Christiansen, J., Schackert, G., Wakimoto, H., Kirsch, M., Brastianos, P. K., and Cahill, D. P. (2017). Intratumoral heterogeneity and TERT promoter mutations in progressive/higher-grade meningiomas. *Oncotarget*, 8(65):109228–109237.

Keller, G., Paige, C., Gilboa, E., and Wagner, E. F. (1985). Expression of a foreign gene in myeloid and lymphoid cells derived from multipotent haematopoietic precursors. *Nature*, 318(6042):149–154.

Khan, A. and Zhang, X. (2015). dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.*, 44(D1):D164–D171.

Kim, J., Lee, I. H., Cho, H. J., Park, C. K., Jung, Y. S., Kim, Y., Nam, S. H., Kim, B. S., Johnson, M. D., Kong, D. S., Seol, H. J., Lee, J. I., Joo, K. M., Yoon, Y., Park, W. Y., Lee, J., Park, P. J., and Nam, D. H. (2015). Spatiotemporal Evolution of the Primary Glioblastoma Genome. *Cancer Cell*, 28(3):318–328.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903.

Kimura, M. (1991). The neutral theory of molecular evolution: a review of recent evidence. *Jpn. J. Genet.*, 66(4):367–386.

Kleinheinz, K., Bludau, I., Huebschmann, D., Heinold, M., Kensche, P., Gu, Z., Lopez, C., Hummel, M., Klapper, W., Moeller, P., Vater, I., Wagener, R., Brors, B., Siebert, R., Eils, R., and Schlesner,

M. (2017). Aceseq-allele specific copy number estimation from whole genome sequencing. *bioRxiv*, page 210807.

Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U.S.A.*, 68(4):820–823.

Kondo, M., Weissman, I. L., and Akashi, K. (1997). Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell*, 91(5):661–672.

Koo, G. C., Tan, S. Y., Tang, T., Poon, S. L., Allen, G. E., Tan, L., Chong, S. C., Ong, W. S., Tay, K., Tao, M., Quek, R., Loong, S., Yeoh, K., Kim, L., Tan, D., Goh, C., Cutcutache, I., Yu, W., Ng, C., Rajasegaran, V., Heng, H., Gan, A., Ong, C., Rozen, S., Tan, P., Teh, B., and Lim, S. (2012). Janus kinase 3–activating mutations identified in natural killer/T-cell Lymphoma. *Cancer Discov.*, 2(7):591–597.

Körber, V., Yang, J., Barah, P., Wu, Y., Stichel, D., Gu, Z., Fletcher, M. N. C., Jones, D., Hentschel, B., Lamszus, K., Tonn, J. C., Schackert, G., Sabel, M., Felsberg, J., Zacher, A., Kaulich, K., Hübschmann, D., Herold-Mende, C., von Deimling, A., Weller, M., Radlwimmer, B., Schlesner, M., Reifenberger, G., Höfer, T., and Lichter, P. (2019). Evolutionary trajectories of IDH-wildtype glioblastomas reveal a common path of early tumorigenesis instigated years ahead of initial diagnosis. *Cancer Cell (accepted)*.

Kosakovsky Pond, S. L. and Frost, S. D. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.*, 22(5):1208–1222.

Kow, Y. W. (2002). Repair of deaminated bases in DNA. *Free Radic. Biol. Med.*, 33(7):886–893.

Krokan, H. E. and Bjørås, M. (2013). Base excision repair. *Cold Spring Harb Perspect Biol*, 5(4):a012583.

Kunkel, T. A. (2004). DNA replication fidelity. *J. Biol. Chem.*, 279(17):16895–16898.

Kunkel, T. A. (2009). Evolving views of DNA replication (in)fidelity. In *Cold Spring Harb. Symp. Quant. Biol.*, volume 74, pages 91–101. Cold Spring Harbor Laboratory Press.

Kunkel, T. A. and Erie, D. A. (2015). Eukaryotic Mismatch Repair in Relation to DNA Replication. *Annu. Rev. Genet.*, 49:291–313.

Lai, A., Kharbanda, S., Pope, W. B., Tran, A., Solis, O. E., Peale, F., Forrest, W. F., Pujara, K., Carrillo, J. A., Pandita, A., Ellingson, B. M., Bowers, C. W., Soriano, R. H., Schmidt, N. O., Mohan, S., Yong, W. H., Seshagiri, S., Modrusan, Z., Jiang, Z., Aldape, K. D., Mischel, P. S., Liau, L. M., Escovedo, C. J., Chen, W., Nghiemphu, P. L., James, C. D., Prados, M. D., Westphal, M., Lamszus, K., Cloughesy, T., and Phillips, H. S. (2011). Evidence for sequenced molecular evolution of IDH1 mutant glioblastoma from a distinct cell of origin. *J. Clin. Oncol.*, 29(34):4482–4490.

Lan, X., Jorg, D. J., Cavalli, F. M. G., Richards, L. M., Nguyen, L. V., Vanner, R. J., Guilhamon, P., Lee, L., Kushida, M. M., Pellacani, D., Park, N. I., Coutinho, F. J., Whetstone, H., Selvadurai, H. J., Che, C., Luu, B., Carles, A., Moksa, M., Rastegar, N., Head, R., Dolma, S., Prinos, P., Cusimano, M. D., Das, S., Bernstein, M., Arrowsmith, C. H., Mungall, A. J., Moore, R. A., Ma, Y., Gallo, M., Lupien, M., Pugh, T. J., Taylor, M. D., Hirst, M., Eaves, C. J., Simons, B. D., and Dirks, P. B. (2017). Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. *Nature*, 549(7671):227–232.

Landa, I., Ibrahimpasic, T., Boucai, L., Sinha, R., Knauf, J. A., Shah, R. H., Dogan, S., Ricarte-Filho, J. C., Krishnamoorthy, G. P., Xu, B., Schultz, N., Berger, M. F., Sander, C., Taylor, B. S., Ghossein, R., Ganly, I., and Fagin, J. A. (2016). Genomic and transcriptomic hallmarks of poorly differentiated and anaplastic thyroid cancers. *J. Clin. Invest.*, 126(3):1052–1066.

Lander, A. D., Gokoffski, K. K., Wan, F. Y., Nie, Q., and Calof, A. L. (2009). Cell lineages and the logic of proliferative control. *PLoS biology*, 7(1):e1000015.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K.,

Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Landy, H. J., Lee, T. T., Potter, P., Feun, L., and Markoe, A. (2000). Early MRI findings in high grade glioma. *J. Neurooncol.*, 47(1):65–72.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., 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., Cortes, 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., 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., and Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.

Lee, E., Yong, R. L., Paddison, P., and Zhu, J. (2018a). Comparison of glioblastoma (GBM) molecular classification methods. In *Semin. Cancer Biol.*, volume 53, pages 201–211. Elsevier.

Lee, J. H., Lee, J. E., Kahng, J. Y., Kim, S. H., Park, J. S., Yoon, S. J., Um, J. Y., Kim, W. K., Lee, J. K., Park, J., Kim, E. H., Lee, J. H., Lee, J. H., Chung, W. S., Ju, Y. S., Park, S. H., Chang, J. H., Kang, S. G., and Lee, J. H. (2018b). Human glioblastoma arises from subventricular zone cells with low-level driver mutations. *Nature*, 560(7717):243–247.

Lee-Six, H., Øbro, N. F., Shepherd, M. S., Grossmann, S., Dawson, K., Belmonte, M., Osborne, R. J., Huntly, B. J., Martincorena, I., Anderson, E., Green, A. R., Kent, D. G., and Campbell, P. J. (2018). Population dynamics of normal human blood inferred from somatic mutations. *Nature*, 561(7724):473–478.

Lefranc, F., Brotchi, J., and Kiss, R. (2005). Possible future issues in the treatment of glioblastomas: special emphasis on cell migration and the resistance of migrating glioblastoma cells to apoptosis. *J. Clin. Oncol.*, 23(10):2411–2422.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, X. and Heyer, W.-D. (2008). Homologous recombination in DNA repair and DNA damage tolerance. *Cell Res.*, 18(1):99–113.

Llaguno, S. A., Chen, J., Kwon, C.-H., Jackson, E. L., Li, Y., Burns, D. K., Alvarez-Buylla, A., and Parada, L. F. (2009). Malignant astrocytomas originate from neural stem/progenitor cells in a somatic tumor suppressor mouse model. *Cancer Cell*, 15(1):45–56.

Loeb, L. A. and Kunkel, T. A. (1982). Fidelity of DNA synthesis. *Annu. Rev. Biochem.*, 51(1):429–457.

Lottaz, C., Beier, D., Meyer, K., Kumar, P., Hermann, A., Schwarz, J., Junker, M., Oefner, P. J., Bogdahn, U., Wischhusen, J., Spang, R., Storch, A., and Beier, C. P. (2010). Transcriptional profiles of CD133+ and CD133- glioblastoma-derived cancer stem cell lines suggest different cells of origin. *Cancer Res.*, 70(5):2030–2040.

Louis, D. N., Perry, A., Reifenberger, G., Von Deimling, A., Figarella-Branger, D., Cavenee, W. K., Ohgaki, H., Wiestler, O. D., Kleihues, P., and Ellison, D. W. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol.*, 131(6):803–820.

Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. *"Proc. Natl. Acad. Sci. U.S.A.*, 107(3):961–968.

Maciejowski, J. and de Lange, T. (2017). Telomeres in cancer: tumour suppression and genome instability. *Nat. Rev. Mol. Cell Biol.*, 18(3):175–186.

Martincorena, I. and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489.

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., and Campbell, P. J. (2018). Universal patterns of selection in cancer and somatic tissues. *Cell*, 173(7):1823.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., Stebbings, L., Menzies, A., Widaa, S., Stratton, M. R., Jones, P. H., and Campbell, P. J. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–886.

McGregor, W. G. (1999). DNA repair, DNA replication, and UV mutagenesis. In *J. Investig. Dermatol. Symp. Proc.*, volume 4, pages 1–5. Elsevier.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20(9):1297–1303.

Medvinsky, A. L., Samoylina, N. L., Müller, A. M., and Dzierzak, E. A. (1993). An early pre-liver intraembryonic source of CFU-S in the developing mouse. *Nature*, 364(6432):64–67.

Mellinghoff, I. K., Cloughesy, T. F., and Mischel, P. S. (2007). PTEN-mediated resistance to epidermal growth factor receptor kinase inhibitors. *Clin. Cancer Res.*, 13(2):378–381.

Meyerson, M., Gabriel, S., and Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.*, 11(10):685–696.

Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y., and Vijg, J. (2017). Differences between germline and somatic mutation rates in humans and mice. *Nat Commun*, 8:15183.

Milo, R., Jorgensen, P., Moran, U., Weber, G., and Springer, M. (2010). BioNumbers–the database of key numbers in molecular and cell biology. *Nucleic Acids Res.*, 38(Database issue):D750–D753.

Mjelle, R., Hegre, S. A., Aas, P. A., Slupphaug, G., Drabløs, F., Sætrom, P., and Krokan, H. E. (2015). Cell cycle regulation of human DNA repair and chromatin remodeling genes. *DNA Repair (Amst.)*, 30:53–67.

Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A., and López-Bigas, N. (2016). OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.*, 17(1):128.

Nakamura, M., Yang, F., Fujisawa, H., Yonekawa, Y., Kleihues, P., and Ohgaki, H. (2000). Loss of heterozygosity on chromosome 19 in secondary glioblastomas. *J. Neuropathol. Exp. Neurol.*, 59(6):539–543.

Nei, M. (1975). *Molecular population genetics and evolution.* North-Holland Publishing Company.

Nguyen, H. N., Lie, A., Li, T., Chowdhury, R., Liu, F., Ozer, B., Wei, B., Green, R. M., Ellingson, B. M., Wang, H. J., Elashoff, R., Liau, L. M., Yong, W. H., Nghiemphu, P. L., Cloughesy, T., and Lai, A. (2016). Human TERT promoter mutation enables survival advantage from MGMT promoter methylation in IDH1 wild-type primary glioblastoma treated by standard chemoradiotherapy. *Neuro-oncology*, 19(3):394–404.

Nishi, N., Kawai, S., Yonezawa, T., Fujimoto, K., and Masui, K. (2009). Early appearance of high grade glioma on magnetic resonance imaging. *Neurol. Med. Chir. (Tokyo)*, 49(1):8–12.

Nordling, C. (1953). A new theory on the cancer-inducing mechanism. *Br. J. Cancer*, 7(1):68–72.

Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., Pan, F., Pelloski, C. E., Sulman, E. P., Bhat, K. P., Verhaak, R. G., Hoadley, K. A., Hayes, D. N., Perou, C. M., Schmidt, H. K., Ding, L., Wilson, R. K., Van Den Berg, D., Shen, H., Bengtsson, H., Neuvial, P., Cope, L. M., Buckley, J., Herman, J. G., Baylin, S. B., Laird, P. W., and Aldape, K. (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell*, 17(5):510–522.

Nowak, M. A. (2006). *Evolutionary dynamics*. Harvard University Press.

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28.

Oesper, L., Mahmoody, A., and Raphael, B. J. (2013). THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.*, 14(7):R80.

Ohgaki, H., Dessen, P., Jourde, B., Horstmann, S., Nishikawa, T., Di Patre, P. L., Burkhard, C., Schuler, D., Probst-Hensch, N. M., Maiorka, P. C., Baeza, N., Pisani, P., Yonekawa, Y., Yasargil, M. G., Lutolf, U. M., and Kleihues, P. (2004). Genetic pathways to glioblastoma: a population-based study. *Cancer Res.*, 64(19):6892–6899.

Ohgaki, H. and Kleihues, P. (2013). The definition of primary and secondary glioblastoma. *Clin. Cancer Res.*, 19(4):764–772.

Ohtsuki, H. and Innan, H. (2017). Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population. *Theor Popul Biol*, 117:43–50.

O'Huallachain, M., Karczewski, K. J., Weissman, S. M., Urban, A. E., and Snyder, M. P. (2012). Extensive genetic variation in somatic human tissues. *Proc. Natl. Acad. Sci. U.S.A.*, 109(44):18018–18023.

Ostrom, Q. T., Gittleman, H., Liao, P., Vecchione-Koval, T., Wolinsky, Y., Kruchko, C., and Barnholtz-Sloan, J. S. (2017). CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2010-2014. *Neuro-oncology*, 19(suppl_5):v1–v88.

Ozawa, T., Riester, M., Cheng, Y.-K., Huse, J. T., Squatrito, M., Helmy, K., Charles, N., Michor, F., and Holland, E. C. (2014). Most human non-GCIMP glioblastoma subtypes evolve from a common proneural-like precursor glioma. *Cancer Cell*, 26(2):288–300.

Pakkenberg, B. and Gundersen, H. J. G. (1997). Neocortical neuron number in humans: effect of sex and age. *J. Comp. Neurol.*, 384(2):312–320.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suva, M. L., Regev, A., and Bernstein, B. E. (2014). Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401.

Pei, W., Feyerabend, T. B., Rossler, J., Wang, X., Postrach, D., Busch, K., Rode, I., Klapproth, K., Dietlein, N., Quedenau, C., Chen, W., Sauer, S., Wolf, S., Hofer, T., and Rodewald, H. R. (2017). Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature*, 548(7668):456–460.

Phillips, H. S., Kharbanda, S., Chen, R., Forrest, W. F., Soriano, R. H., Wu, T. D., Misra, A., Nigro, J. M., Colman, H., Soroceanu, L., Williams, P. M., Modrusan, Z., Feuerstein, B. G., and Aldape, K. (2006). Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell*, 9(3):157–173.

Prados, M. D., Byron, S. A., Tran, N. L., Phillips, J. J., Molinaro, A. M., Ligon, K. L., Wen, P. Y., Kuhn, J. G., Mellinghoff, I. K., de Groot, J. F., Colman, H., Cloughesy, T. F., Chang, S. M., Ryken, T. C., Tembe, W. D., Kiefer, J. A., Berens, M. E., Craig, D. W., Carpten, J. D., and Trent, J. M. (2015). Toward precision medicine in glioblastoma: the promise and the challenges. *Neuro-oncology*, 17(8):1051–1063.

Qazi, M., Vora, P., Venugopal, C., Sidhu, S., Moffat, J., Swanton, C., and Singh, S. (2017). Intratumoral heterogeneity: pathways to treatment resistance and relapse in human glioblastoma. *Ann. Oncol.*, 28(7):1448–1456.

Rivera-Munoz, P., Laurent, A. P., Siret, A., Lopez, C. K., Ignacimouttou, C., Cornejo, M. G., Bawa, O., Rameau, P., Bernard, O. A., Dessen, P., Gilliland, G. D., Mercher, T., and Malinge, S. (2018). Partial trisomy 21 contributes to T-cell malignancies induced by JAK3-activating mutations in murine models. *Blood Adv*, 2(13):1616–1627.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., and Shah, S. P. (2014). PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, 11(4):396–398.

Samokhvalov, I. M., Samokhvalova, N. I., and Nishikawa, S.-i. (2007). Cell tracing shows the contribution of the yolk sac to adult haematopoiesis. *Nature*, 446(7139):1056–1061.

Sanai, N., Polley, M.-Y., McDermott, M. W., Parsa, A. T., and Berger, M. S. (2011). An extent of resection threshold for newly diagnosed glioblastomas. *J. Neurosurg.*, 115(1):3–8.

Saunders, C. T., Wong, W. S., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817.

Savage, D. C. (1977). Microbial ecology of the gastrointestinal tract. *Annu. Rev. Microbiol.*, 31(1):107–133.

Schärer, O. D. (2013). Nucleotide excision repair in eukaryotes. *Cold Spring Harb Perspect Biol*, 5(10):a012609.

Schroering, A. G., Edelbrock, M. A., Richards, T. J., and Williams, K. J. (2007). The cell cycle and DNA mismatch repair. *Exp. Cell Res.*, 313(2):292–304.

Schwartz, R. and Schäffer, A. A. (2017). The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.*, 18(4):213–229.

Schweitzer, T., Vince, G., Herbold, C., Roosen, K., and Tonn, J.-C. (2001). Extraneural metastases of primary brain tumors. *J. Neurooncol.*, 53(2):107–114.

Sidow, A. and Spies, N. (2015). Concepts in solid tumor evolution. *Trends Genet.*, 31(4):208–214.

Simons, B. D. (2016). Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis. *Proc. Natl. Acad. Sci. U.S.A.*, 113(1):128–133.

Singh, S. K., Clarke, I. D., Terasaki, M., Bonn, V. E., Hawkins, C., Squire, J., and Dirks, P. B. (2003). Identification of a cancer stem cell in human brain tumors. *Cancer Res.*, 63(18):5821–5828.

Singh, S. K., Hawkins, C., Clarke, I. D., Squire, J. A., Bayani, J., Hide, T., Henkelman, R. M., Cusimano, M. D., and Dirks, P. B. (2004). Identification of human brain tumour initiating cells. *Nature*, 432(7015):396–401.

Snuderl, M., Fazlollahi, L., Le, L. P., Nitta, M., Zhelyazkova, B. H., Davidson, C. J., Akhavanfard, S., Cahill, D. P., Aldape, K. D., Betensky, R. A., Louis, D. N., and Iafrate, A. J. (2011). Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma. *Cancer Cell*, 20(6):810–817.

Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., Marjoram, P., Siegmund, K., Press, M. F., Shibata, D., and Curtis, C. (2015). A Big Bang model of human colorectal tumor growth. *Nat. Genet.*, 47(3):209–216.

Sottoriva, A., Spiteri, I., Piccirillo, S. G., Touloumis, A., Collins, V. P., Marioni, J. C., Curtis, C., Watts, C., and Tavaré, S. (2013). Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, 110(10):4009–4014.

Spiteri, I., Caravagna, G., Cresswell, G. D., Vatsiou, A., Nichol, D., Acar, A., Ermini, L., Chkhaidze, K., Werner, B., Mair, R., Brognaro, E., Verhaak, R. G. W., Sanguinetti, G., Piccirillo, S. G. M., Watts, C., and Sottoriva, A. (2018). Evolutionary dynamics of residual disease in human glioblastoma. *Ann. Oncol.*

Springuel, L., Renauld, J.-C., and Knoops, L. (2015). JAK kinase targeting in hematologic malignancies: a sinuous pathway from identification of genetic alterations towards clinical indications. *Haematologica*, 100(10):1240–1253.

Steensma, D. P., Bejar, R., Jaiswal, S., Lindsley, R. C., Sekeres, M. A., Hasserjian, R. P., and Ebert, B. L. (2015). Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood*, 126(1):9–16.

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature*, 458(7239):719–724.

Strom, S. P. (2016). Current practices and guidelines for clinical next-generation sequencing oncology testing. *Cancer Biol Med*, 13(1):3–11.

Stupp, R., Mason, W. P., van den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J., Belanger, K., Brandes, A. A., Marosi, C., Bogdahn, U., Curschmann, J., Janzer, R. C., Ludwin, S. K., Gorlia, T., Allgeier, A., Lacombe, D., Cairncross, J. G., Eisenhauer, E., and Mirimanoff, R. O. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N. Engl. J. Med.*, 352(10):987–996.

Tarabichi, M., Martincorena, I., Gerstung, M., Leroi, A. M., Markowetz, F., Spellman, P. T., Morris, Q. D., Lingjaerde, O. C., Wedge, D. C., Van Loo, P., Dentro, S. C., Leshchiner, I., Gerstung, M., Jolly, C., Haase, K., Tarabichi, M., Wintersinger, J., Deshwar, A. G., Yu, K., Gonzalez, S., Rubanova, Y., Macintyre, G., Adams, D. J., Anur, P., Beroukhim, R., Boutros, P. C., Bowtell, D. D., Campbell, P. J., Cao, S., Christie, E. L., Cmero, M., Cun, Y., Dawson, K. J., Demeulemeester, J., Donmez, N., Drews, R. M., Eils, R., Fan, Y., Fittall, M., Garsed, D. W., Getz, G., Ha, G., Imielinski, M., Jerman, L., Ji, Y., Kleinheinz, K., Lee, J., Lee-Six, H., Livitz, D. G., Malikic, S., Markowetz, F., Martincorena, I., Mitchell, T. J., Mustonen, V., Oesper, L., Peifer, M., Peto, M., Raphael, B. J., Rosebrock, D., Sahinalp, S. C., Salcedo, A., Schlesner, M., Schumacher, S., Sengupta, S., Shi, R., Shin, S. J., Stein, L. D., Vazquez-Garcia, I., Vembu, S., Wheeler, D. A., Yang, T. P., Yao, X., Yuan, K., Zhu, H., Wang, W., Morris, Q. D., Spellman, P. T., Wedge, D. C., and Van Loo, P. (2018). Neutral tumor evolution? *Nat. Genet.*, 50(12):1630–1633.

The Cancer Genome Atlas Research Network, McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., Olson, J. J., Mikkelsen, T., Lehman, N., Aldape, K., Yung, W. K., Bogler, O., Weinstein, J. N., VandenBerg, S., Berger, M., Prados, M., Muzny, D., Morgan, M., Scherer, S., Sabo, A., Nazareth, L., Lewis, L., Hall, O., Zhu, Y., Ren, Y., Alvi, O., Yao, J., Hawes, A., Jhangiani, S., Fowler, G., San Lucas, A., Kovar, C., Cree, A., Dinh, H., Santibanez, J., Joshi, V., Gonzalez-Garay, M. L., Miller, C. A., Milosavljevic, A., Donehower, L., Wheeler, D. A., Gibbs, R. A., Cibulskis, K., Sougnez, C., Fennell, T., Mahan, S., Wilkinson, J., Ziaugra, L., Onofrio, R., Bloom, T., Nicol, R., Ardlie, K., Baldwin, J., Gabriel, S., Lander, E. S., Ding, L., Fulton, R. S., McLellan, M. D., Wallis, J., Larson, D. E., Shi, X., Abbott, R., Fulton, L., Chen, K., Koboldt, D. C.,

Wendl, M. C., Meyer, R., Tang, Y., Lin, L., Osborne, J. R., Dunford-Shore, B. H., Miner, T. L., Delehaunty, K., Markovic, C., Swift, G., Courtney, W., Pohl, C., Abbott, S., Hawkins, A., Leong, S., Haipek, C., Schmidt, H., Wiechert, M., Vickery, T., Scott, S., Dooling, D. J., Chinwalla, A., Weinstock, G. M., Mardis, E. R., Wilson, R. K., Getz, G., Winckler, W., Verhaak, R. G., Lawrence, M. S., O'Kelly, M., Robinson, J., Alexe, G., Beroukhim, R., Carter, S., Chiang, D., Gould, J., Gupta, S., Korn, J., Mermel, C., Mesirov, J., Monti, S., Nguyen, H., Parkin, M., Reich, M., Stransky, N., Weir, B. A., Garraway, L., Golub, T., Meyerson, M., Chin, L., Protopopov, A., Zhang, J., Perna, I., Aronson, S., Sathiamoorthy, N., Ren, G., Yao, J., Wiedemeyer, W. R., Kim, H., Kong, S. W., Xiao, Y., Kohane, I. S., Seidman, J., Park, P. J., Kucherlapati, R., Laird, P. W., Cope, L., Herman, J. G., Weisenberger, D. J., Pan, F., Van den Berg, D., Van Neste, L., Yi, J. M., Schuebel, K. E., Baylin, S. B., Absher, D. M., Li, J. Z., Southwick, A., Brady, S., Aggarwal, A., Chung, T., Sherlock, G., Brooks, J. D., Myers, R. M., Spellman, P. T., Purdom, E., Jakkula, L. R., Lapuk, A. V., Marr, H., Dorton, S., Choi, Y. G., Han, J., Ray, A., Wang, V., Durinck, S., Robinson, M., Wang, N. J., Vranizan, K., Peng, V., Van Name, E., Fontenay, G. V., Ngai, J., Conboy, J. G., Parvin, B., Feiler, H. S., Speed, T. P., Gray, J. W., Brennan, C., Socci, N. D., Olshen, A., Taylor, B. S., Lash, A., Schultz, N., Reva, B., Antipin, Y., Stukalov, A., Gross, B., Cerami, E., Wang, W. Q., Qin, L. X., Seshan, V. E., Villafania, L., Cavatore, M., Borsu, L., Viale, A., Gerald, W., Sander, C., Ladanyi, M., Perou, C. M., Hayes, D. N., Topal, M. D., Hoadley, K. A., Qi, Y., Balu, S., Shi, Y., Wu, J., Penny, R., Bittner, M., Shelton, T., Lenkiewicz, E., Morris, S., Beasley, D., Sanders, S., Kahn, A., Sfeir, R., Chen, J., Nassau, D., Feng, L., Hickey, E., Barker, A., Gerhard, D. S., Vockley, J., Compton, C., Vaught, J., Fielding, P., Ferguson, M. L., Schaefer, C., Zhang, J., Madhavan, S., Buetow, K. H., Collins, F., Good, P., Guyer, M., Ozenberger, B., Peterson, J., and Thomson, E. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068.

Tohma, Y., Gratas, C., Biernat, W., Peraud, A., et al. (1998). PTEN (MMAC1) mutations are frequent in primary glioblastomas (de novo) but not in secondary glioblastomas. *J. Neuropathol. Exp. Neurol.*, 57(7):684–689.

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

Turke, A. B., Zejnullahu, K., Wu, Y. L., Song, Y., Dias-Santagata, D., Lifshits, E., Toschi, L., Rogers, A., Mok, T., Sequist, L., Lindeman, N. I., Murphy, C., Akhavanfard, S., Yeap, B. Y., Xiao, Y., Capelletti, M., Iafrate, A. J., Lee, C., Christensen, J. G., Engelman, J. A., and Janne, P. A. (2010). Preexistence and clonal selection of MET amplification in EGFR mutant NSCLC. *Cancer Cell*, 17(1):77–88.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B. A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H. S., Hodgson, J. G., James,

C. D., Sarkaria, J. N., Brennan, C., Kahn, A., Spellman, P. T., Wilson, R. K., Speed, T. P., Gray, J. W., Meyerson, M., Getz, G., Perou, C. M., and Hayes, D. N. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17(1):98–110.

Vinagre, J., Almeida, A., Populo, H., Batista, R., Lyra, J., Pinto, V., Coelho, R., Celestino, R., Prazeres, H., Lima, L., Melo, M., da Rocha, A. G., Preto, A., Castro, P., Castro, L., Pardal, F., Lopes, J. M., Santos, L. L., Reis, R. M., Cameselle-Teijeiro, J., Sobrinho-Simoes, M., Lima, J., Maximo, V., and Soares, P. (2013). Frequency of TERT promoter mutations in human cancers. *Nat Commun*, 4:2185.

Vranová, V., NeCesalová, E., Kuglík, P., Cejpek, P., PeSáková, M., Budínská, E., Relichová, J., and Veselská, R. (2007). Screening of genomic imbalances in glioblastoma multiforme using high-resolution comparative genomic hybridization. *Oncol. Rep.*, 17(2):457–464.

Walters, D. K., Mercher, T., Gu, T.-L., O'Hare, T., Tyner, J. W., Loriaux, M., Goss, V. L., Lee, K. A., Eide, C. A., Wong, M. J., et al. (2006). Activating alleles of JAK3 in acute megakaryoblastic leukemia. *Cancer Cell*, 10(1):65–75.

Wang, J., Cazzato, E., Ladewig, E., Frattini, V., Rosenbloom, D. I., Zairis, S., Abate, F., Liu, Z., Elliott, O., Shin, Y. J., Lee, J. K., Lee, I. H., Park, W. Y., Eoli, M., Blumberg, A. J., Lasorella, A., Nam, D. H., Finocchiaro, G., Iavarone, A., and Rabadan, R. (2016). Clonal evolution of glioblastoma under therapy. *Nat. Genet.*, 48(7):768–776.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, 38(16):e164–e164.

Wang, Q., Hu, B., Hu, X., Kim, H., Squatrito, M., Scarpace, L., deCarvalho, A. C., Lyu, S., Li, P., Li, Y., Barthel, F., Cho, H. J., Lin, Y. H., Satani, N., Martinez-Ledesma, E., Zheng, S., Chang, E., Sauve, C. G., Olar, A., Lan, Z. D., Finocchiaro, G., Phillips, J. J., Berger, M. S., Gabrusiewicz, K. R., Wang, G., Eskilsson, E., Hu, J., Mikkelsen, T., DePinho, R. A., Muller, F., Heimberger, A. B., Sulman, E. P., Nam, D. H., and Verhaak, R. G. W. (2017). Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell*, 32(1):42–56.

Waters, T. R. and Swann, P. F. (2000). Thymine-DNA glycosylase and G to A transition mutations at CpG sites. *Mutat. Res. Rev. Mutat. Res.*, 462(2):137–147.

Wee, P. and Wang, Z. (2017). Epidermal growth factor receptor cell proliferation signaling pathways. *Cancers*, 9(5):52.

Weeden, C. E. and Asselin-Labat, M.-L. (2018). Mechanisms of DNA damage repair in adult stem cells and implications for cancer formation. *Biochim Biophys Acta Mol Basis Dis*, 1864(1):89–101.

Westphal, M., Maire, C. L., and Lamszus, K. (2017). EGFR as a Target for Glioblastoma Treatment: An Unfulfilled Promise. *CNS drugs*, 31(9):723–735.

Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A., and Sottoriva, A. (2016). Identification of neutral tumor evolution across cancer types. *Nat. Genet.*, 48(3):238–244.

Wippold, F., Lämmle, M., Anatelli, F., Lennerz, J., and Perry, A. (2006). Neuropathology for the neuroradiologist: palisades and pseudopalisades. *AJNR Am J Neuroradiol*, 27(10):2037–2041.

Zhang, J., FG Stevens, M., and D Bradshaw, T. (2012). Temozolomide: mechanisms of action, repair and resistance. *Curr Mol Pharmacol*, 5(1):102–114.

Zink, F., Stacey, S. N., Norddahl, G. L., Frigge, M. L., Magnusson, O. T., Jonsdottir, I., Thorgeirsson, T. E., Sigurdsson, A., Gudjonsson, S. A., Gudmundsson, J., Jonasson, J. G., Tryggvadottir, L., Jonsson, T., Helgason, A., Gylfason, A., Sulem, P., Rafnar, T., Thorsteinsdottir, U., Gudbjartsson, D. F., Masson, G., Kong, A., and Stefansson, K. (2017). Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood*, 130(6):742–752.

Zovein, A. C., Hofmann, J. J., Lynch, M., French, W. J., Turlo, K. A., Yang, Y., Becker, M. S., Zanetta, L., Dejana, E., Gasson, J. C., Tallquist, M. D., and Iruela-Arispe, M. L. (2008). Fate tracing reveals the endothelial origin of hematopoietic stem cells. *Cell Stem Cell*, 3(6):625–636.

Acknowledgments

I genuinely thank my supervisor Thomas Höfer, who gave me the opportunity and, much more than this, encouraged me to pursue this PhD project. Thank you for accompanying me during my PhD by letting me develop my own ideas, supporting and improving them, by teaching me how to write and to publish, by generating an inspiring and intellectual, but at the same time humourous working environment and by backing me up in difficult times.

I profoundly thank the entire Höfer group for the pleasant working environment — thank you for the many good conversations, the beautiful hikes and the good time we spent together. I especially thank my office mates, Adrien, Carsten and Matthias for the shared time, small conversations over coffee and their open ear, and Lisa, Nils, Christoph and Ines for the good discussions we had.

My deep gratitude goes to my second assessor Frank Westermann for always taking time to dive into my projects, supporting them with experimental data whenever possible, for inspiring discussions and for a great collaboration. I thank his group members for their support with experiments and discussions; I would like to point out Sabine Hartlieb, Selina Jansky and especially Moritz Gartlgruber, who has been a reliable project partner in every respect, but also a great dialog partner for small chats in between.

I thank the members of the 'SysGlio' consortium for the fruitful and pleasant collaboration. I especially thank Jing Yang and Matthias Schlesner for the joined forces in data analysis, and Bernhard Radlwimmer, Guido Reifenberger and Peter Lichter for promoting our project with the many detailed discussions we had.

I thank Ruzhica Bogeska, Megan Druce and Michael Milsom for sharing their interesting data with me, which led me to a new way of thinking about somatic mutagenesis.

I thank Frederik Graw for being a member of my TAC-committee and for carefully reading my reports.

I thank Robert Russell and Hai-Kun Liu for agreeing to be part of my defense committee.

I thank Matthias, Pete and Dina for carefully reading my thesis and their valuable input.

I thank Nick and Diana for their patient support with technical and bureaucratic challenges of all kind.

I thank my parents and my brothers, Michael and Thomas, for their unconditional support throughout my thesis; I thank my mother for patiently correcting my English, Thomas for counselling me on correct mathematical annotation, and Michael for helping me with LaTex formatting.

Appendices

APPENDIX A

Supplementary information to phylogenetic inference

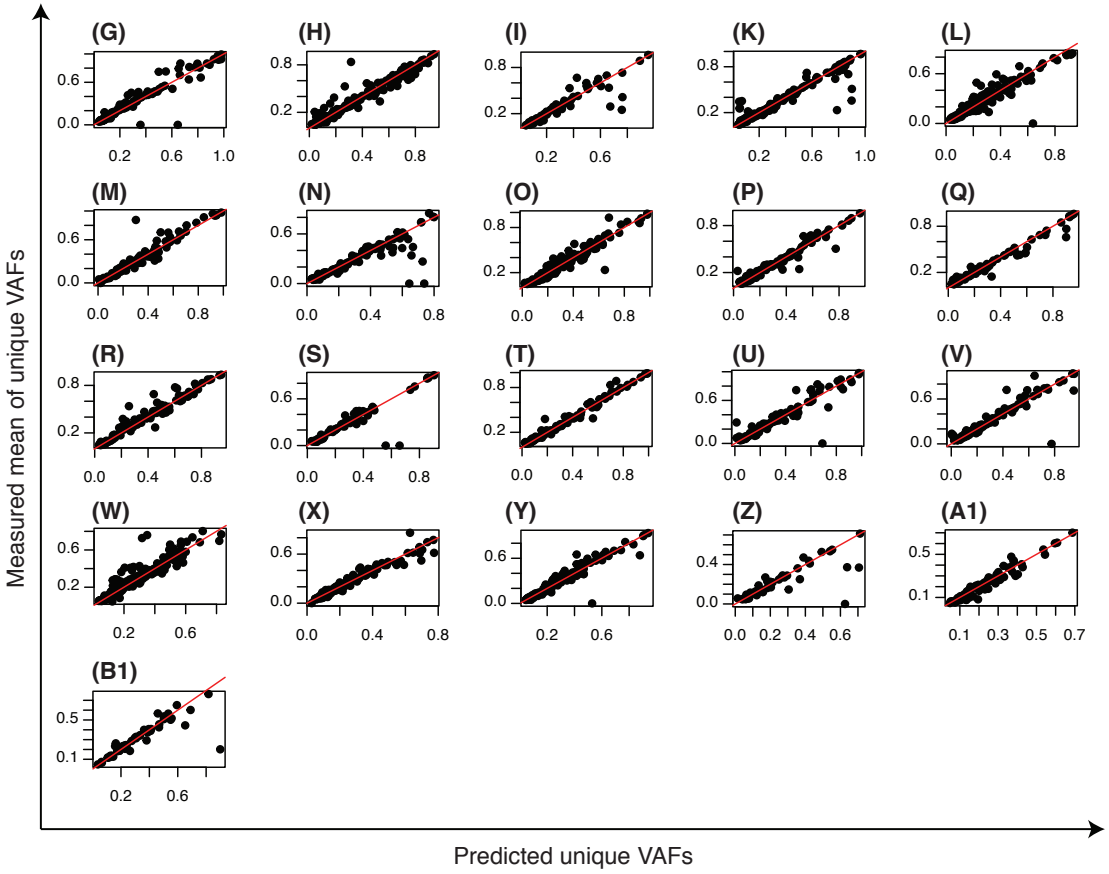


Figure A.1: Goodness of fit of inferred phylogenetic trees. At each fit, the predicted unique variant allele frequencies are plotted against the measured means of all variants expected at this VAF (the order of the plots corresponds to the order of the trees in Fig. 2.13, as indicated by capital letters in brackets; red lines, bisetrices).

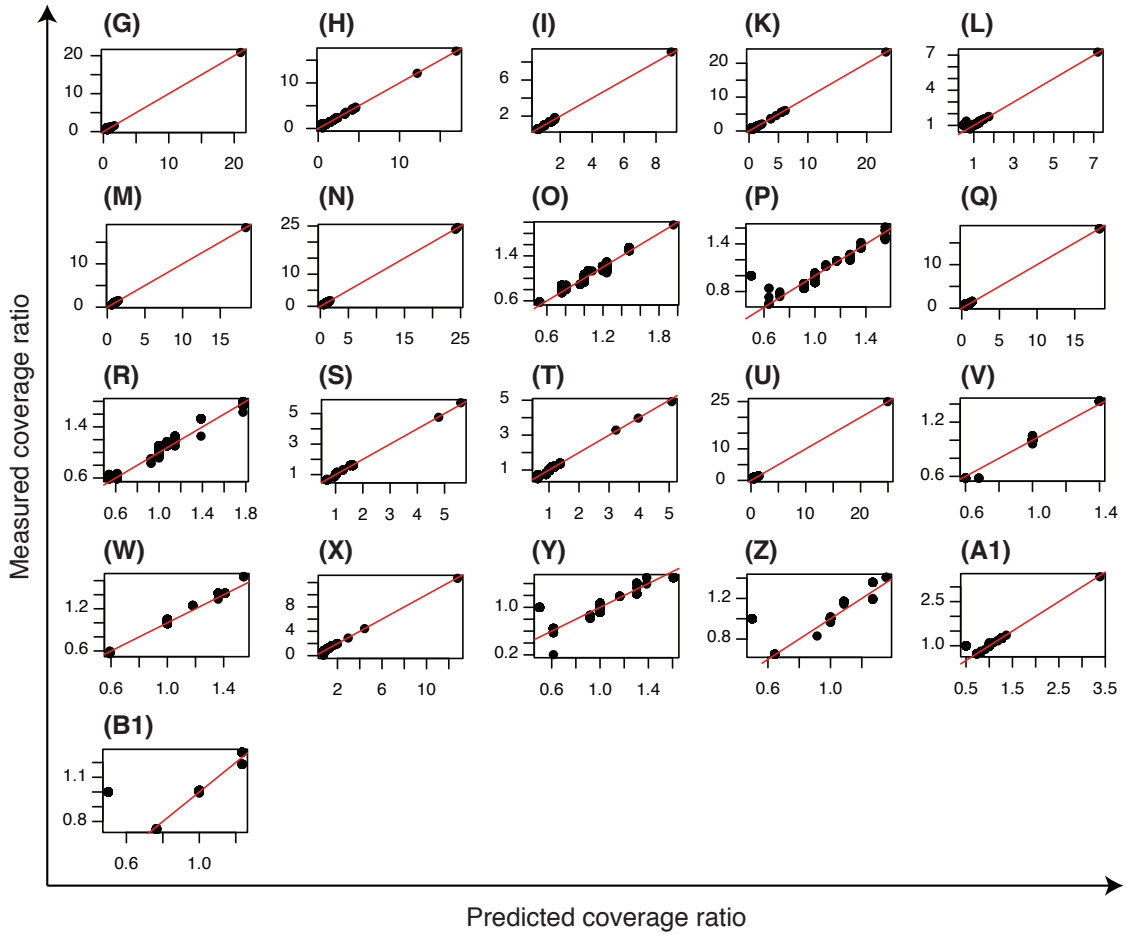


Figure A.2: Goodness of fit of inferred phylogenetic trees. At each fit, the predicted coverage ratios are plotted against their measured counterparts (the order of the plots corresponds to the order of the trees in Fig. 2.13, as indicated by capital letters in brackets; 1000 datapoints were downsampled per tumor for better visualization; red lines, bisetrices).

APPENDIX B

Transition probabilities in a critical birth-death process

In the following, we will derive the transition probabilities in a 'critical' birth-death process (linear birth-death process with equal birth and death probabilities). To this end, we will first derive the probability generating function following Bailey, 1964 and then derive the transition probabilities by Taylor expansion.

We study a critical birth-death process with division and death rate λ and population size $X(t)$. In an infinitesimal time interval $\mathrm{d}t$ the population can either increase by one (birth), decrease by one (death) or remain constant, yielding the following transtion probabililties:

$$\text{birth: } \lambda X(t)\mathrm{d}t \tag{B.1a}$$

$$\text{death: } \lambda X(t)\mathrm{d}t \tag{B.1b}$$

$$\text{neither birth nor death: } 1 - 2\lambda X(t)\mathrm{d}t. \tag{B.1c}$$

With this, the Master equation for the probabililty of having n individuals at time t is directly obtained:

$$\frac{\mathrm{d}P_n}{\mathrm{d}t} = \lambda(n-1)P_{n-1} - 2\lambda n P_n + \lambda(n+1)P_{n+1}, \qquad n \geq 1, \tag{B.2a}$$

$$P_{-1} = 0. \tag{B.2b}$$

To solve the Master equation we define the generating function $F(z,t)$,

$$F(z,t) = \sum_{n=0}^{\infty} z^n P_n(t) \tag{B.3a}$$

$$\frac{\mathrm{d}F(z,t)}{\mathrm{d}t} = \sum_{n=0}^{\infty} z^n \frac{\mathrm{d}P_n(t)}{\mathrm{d}t} \tag{B.3b}$$

$$= \sum_{n=0}^{\infty} z^n \left[\lambda(n-1)P_{n-1} - 2\lambda n P_n + \lambda(n+1)P_{n+1}\right] \tag{B.3c}$$

$$= \lambda \left[\sum_{n=1}^{\infty} z^n (n-1)P_{n-1} - 2\sum_{n=0}^{\infty} z^n n P_n + \sum_{n=0}^{\infty} z^n (n+1)P_{n+1}\right]. \tag{B.3d}$$

Rearranging the indices yields

$$\frac{\mathrm{d}F(z,t)}{\mathrm{d}t} = \lambda \left[\sum_{n=1}^{\infty} z^{n+1} n P_n - 2\sum_{n=1}^{\infty} z^n n P_n + \sum_{n=1}^{\infty} z^{n-1} n P_n\right]. \tag{B.4}$$

We note that $\frac{\partial F}{\partial z} = \sum_{n=1}^{\infty} n z^{n-1} P_n$ and obtain the partial differential equation

$$\frac{\partial F}{\partial t} = \lambda \left[z^2 - 2z + 1\right] \frac{\partial F}{\partial z}, \tag{B.5}$$

with initial condition

$$F(z, t=0) = \Psi(z). \tag{B.6}$$

Eqn. B.5 is a linear partial differential equation, which can be solved with the method of characteristics. To this end, we first derive with respect to x, using the chain rule:

$$\frac{\partial F}{\partial x} = \frac{\partial F}{\partial z}\frac{\partial z}{\partial x} + \frac{\partial F}{\partial t}\frac{\partial t}{\partial x}. \tag{B.7}$$

Next, we define the characteristic curves

$$\frac{\mathrm{d}z}{\mathrm{d}x} = P(z,t,x) = -\lambda(z-1)^2, \tag{B.8a}$$

$$\frac{\mathrm{d}t}{\mathrm{d}x} = Q(z,t,x) = 1, \tag{B.8b}$$

for

$$\frac{\mathrm{d}F}{\mathrm{d}x} = \frac{\partial F}{\partial z}\frac{\partial z}{\partial x} + \frac{\partial F}{\partial t}\frac{\partial t}{\partial x} = R(z,t,x) = 0. \tag{B.9}$$

and thus obtain a system of ordinary differential equations. From Eqns. B.8a and B.8b we conclude that

$$\frac{\mathrm{d}t}{\mathrm{d}x} = -\frac{1}{\lambda(z-1)^2}\frac{\mathrm{d}z}{\mathrm{d}x}. \tag{B.10}$$

Integration with respect to x yields

$$-\frac{\mathrm{d}z}{\lambda(z-1)^2} = \frac{\mathrm{d}t}{1}, \tag{B.11}$$

and thus

$$-\int \frac{\mathrm{d}z}{\lambda(z-1)^2} = \int \frac{\mathrm{d}t}{1} \tag{B.12a}$$

$$t = \frac{1}{\lambda(z-1)} + c_1. \tag{B.12b}$$

On the other hand, from Eqn. B.9 we conclude that

$$F(z,t) = c_2. \tag{B.13}$$

Thus we find the general solution:

$$F(z,t) = \Psi\left(\lambda t - \frac{1}{z-1}\right). \tag{B.14}$$

If we start with a cells at $t = 0$, Eqn. B.3a gives the initial condition $F(z,0) = \sum_{n=0}^{\infty} P_n(0) = z^a$ and thus

$$F(z,0) = z^a = \Psi\left(-\frac{1}{z-1}\right). \tag{B.15}$$

Moreover, Eqn. B.12b yields at $t = 0$

$$c_1 = -\frac{1}{z-1} \qquad \Leftrightarrow \qquad z = 1 - \frac{1}{c_1}. \tag{B.16}$$

From the initial condition, we thus have

$$c_2 = F(z,0) = z^a = \left(1 - \frac{1}{c_1}\right)^a = \Psi(c_1). \tag{B.17}$$

If $t \neq 0$, we recall $c_1 = t - \frac{1}{\lambda(z-1)}$ and find

$$1 - \frac{1}{c_1} = \frac{1-(z-1)(\lambda t-1)}{1-(z-1)\lambda t}. \tag{B.18}$$

Thus,

$$F(z,t) = \left(\frac{1-(z-1)(\lambda t-1)}{1-(z-1)\lambda t}\right)^a, \tag{B.19}$$

where we used Eqn. B.17. Finally, we rearrange by Taylor expansion around 0:

$$F(z,t) = \sum_i \frac{1}{i!}\frac{\partial^i F}{\partial z^i}(z-0)^i, \tag{B.20a}$$

$$\frac{\partial^i F}{\partial z^i} = \sum_{k=1}^{i} \binom{a}{k}\binom{i}{k} A^{a-k} B^{i+k} (\lambda t)^{n-k} k i!, \tag{B.20b}$$

$$A = \frac{1-(\lambda t-1)(z-1)}{1-\lambda t(z-1)}, \qquad B = \frac{1}{1-\lambda t(x-1)}, \tag{B.20c}$$

and obtain the transition probabilities from a to b:

$$P_{a,b} = \left.\frac{\partial^b F}{\partial z^b}\right|_{z=0} \tag{B.21a}$$

$$P_{a,b}(\lambda,t) = \begin{cases} 0 & a = 0 \\ p^a & a \neq 0, b = 0 \\ \sum_{k=1}^{b} \frac{k}{b}\binom{a}{k} p^{a-k}(1-p)^k \binom{b}{k} p^{b-k}(1-p)^k & a \neq 0, b \neq 0 \end{cases} \tag{B.21b}$$

$$p = \frac{\lambda t}{1+\lambda t}. \tag{B.21c}$$

APPENDIX C

Clonal dynamics in adult hematopoiesis

The stochasticity of clonal expansions in adult hematopoiesis were assessed explicitly by solving the master equation and, additionally, with stochastic simulations using Gillespie's algorithm (Gillespie, 1977). Both approaches are outlined in the following.

Indexing LT-HSCs, ST-HSCs and MPPs with 0, 1, and 2, respectively, we define the state vector $\boldsymbol{n} = (n_0, n_1, n_2)$ for the respective cell population counts. The model in Fig.3.10A is then described by the following transition probabilities:

$$\begin{aligned}
\text{division of LT-HSCs:} &\quad \lambda_0 n_0(t)\mathrm{d}t \\
\text{differentiation of LT-HSCs into ST-HSCs:} &\quad \alpha_0 n_0(t)\mathrm{d}t \\
\text{division of ST-HSCs:} &\quad \lambda_1 n_1(t)\mathrm{d}t \\
\text{differentiation of ST-HSCs into MPPs:} &\quad \alpha_1 n_1(t)\mathrm{d}t \\
\text{division of MPPs:} &\quad \lambda_2 n_2(t)\mathrm{d}t \\
\text{differentiation of MPPs into CMPs/CLPs:} &\quad \alpha_2 n_2(t)\mathrm{d}t,
\end{aligned} \tag{C.1}$$

Derivation and solution of the Master equation for this set of stochastic processes is provided in Section C.1, while stochastic simulations with Gillespie's algorithm are explained in Section C.2.

C.1 Solving the master equation

We will show the derivation of the master equation (Eqn. 3.18) and the solution for the mean and variance of the linear differentiation model shown in Fig.3.10A. The stochastic processes according to Eqns. C.1 translate to the following Master equation:

$$\frac{\mathrm{d}P_{\boldsymbol{n}}(t)}{\mathrm{d}t} = \sum_{i=0}^{2} \Bigg[\underbrace{\overbrace{\lambda_i(n_i-1)P_{\boldsymbol{n}-e_i}(t)}^{\text{Division of cell type i}} + \overbrace{\alpha_i(n_i+1)P_{\boldsymbol{n}+e_i-e_{i+1}}(t)}^{\text{Differentiation of cell type i into cell type i + 1}}}_{\text{Stochastic processes running into state n}} \underbrace{-\lambda_i n_i P_{\boldsymbol{n}}(t) - \alpha_i n_i P_{\boldsymbol{n}}(t)}_{\text{Sochastic processes running out of state n}} \Bigg],$$

$$P_{n_i} = 0, \text{ if any } n_i = -1, \tag{C.2}$$

where e_i is the $(i+1)$-th unit vector and e_3 is set to $(0,0,0)$ by convenience. From the Master equation we can derive the probability generating function, $F(z_0, z_1, z_2, t)$:

$$F(z_0, z_1, z_2, t) = \sum_n z_0^{n_0} z_1^{n_1} z_2^{n_2} P_n(t) \tag{C.3}$$

and, consequently,

$$\begin{aligned}\frac{\partial F}{\partial t} = \sum_n \Bigg[z_0^{n_0} z_1^{n_1} z_2^{n_2} \sum_{i=0}^{2} & \lambda_i(n_i-1)P_{\boldsymbol{n}-e_i}(t) + \\ & + \alpha_i(n_i+1)P_{\boldsymbol{n}+e_i-e_{i+1}}(t) + \\ & - \lambda_i n_i P_{\boldsymbol{n}}(t) - \alpha_i n_i P_{\boldsymbol{n}}(t) \Big]. \end{aligned} \tag{C.4}$$

Rearranging the indices yields

$$\begin{aligned}\frac{\partial F}{\partial t} = \sum_n \sum_{i=0}^{2} \Big[& z_i z_0^{n_0} z_1^{n_1} z_2^{n_2} \lambda_i n_i P_{\boldsymbol{n}}(t) + \\ & + \frac{z_{i+1}}{z_i} z_0^{n_0} z_1^{n_1} z_2^{n_2} \alpha_i n_i P_{\boldsymbol{n}}(t) + \\ & - \prod_{j=0}^{2} z_j^{n_j} \lambda_i n_i P_{\boldsymbol{n}}(t) - \prod_{j=0}^{2} z_j^{n_j} \alpha_i n_i P_{\boldsymbol{n}}(t) \Big]. \end{aligned} \tag{C.5}$$

Noting that

$$\frac{\partial F}{\partial z_i} = \sum_n n_i \frac{1}{z_i} z_0^{n_0} z_1^{n_1} z_2^{n_2} P_{\boldsymbol{n}}(t) \tag{C.6}$$

we can substitute

$$\begin{aligned}
\sum_n z_i z_0^{n_0} z_1^{n_1} z_2^{n_2} \lambda_i n_i P_{\boldsymbol{n}}(t) &= \lambda_i z_i^2 \frac{\partial F}{\partial z_i} \\
\sum_n \frac{z_{i+1}}{z_i} z_0^{n_0} z_1^{n_1} z_2^{n_2} \alpha_i n_i P_{\boldsymbol{n}}(t) &= \alpha_i z_{i+1} \frac{\partial F}{\partial z_i} \\
\sum_n z_0^{n_0} z_1^{n_1} z_2^{n_2} \lambda_i n_i P_{\boldsymbol{n}}(t) &= \lambda_i z_i \frac{\partial F}{\partial z_i} \\
\sum_n z_0^{n_0} z_1^{n_1} z_2^{n_2} \alpha_i n_i P_{\boldsymbol{n}}(t) &= \alpha_i z_i \frac{\partial F}{\partial z_i},
\end{aligned} \tag{C.7}$$

yielding

$$\frac{\partial F(z,t)}{\partial t} = \sum_{i=0}^{2} \left[\lambda_i z_i (z_i - 1) + \alpha_i (z_{i+1} - z_i)\right] \frac{\partial F(z,t)}{\partial z_i}, \tag{C.8}$$

where $z_3 = 1$. From the probability generating function we can determine the expected value of n_i by differentiation w.r.t. z_i and evaluation at $z = 1$:

$$\mathrm{E}(n_i) = \left.\frac{\partial F}{\partial z_i}\right|_{z=1} = \sum_n n_i P_{\boldsymbol{n}}(t). \tag{C.9}$$

Accordingly, we obtain the variance and covariance as:

$$\begin{aligned}
\mathrm{Var}(n_i) &= \mathrm{E}(n_i^2) - \mathrm{E}(n_i)^2 \\
&= \mathrm{E}(n_i(n_i - 1)) + \mathrm{E}(n_i) - \mathrm{E}(n_i)^2 \\
&= \left.\frac{\partial^2 F}{\partial z_i^2}\right|_{z=1} + \left.\frac{\partial F}{\partial z_i}\right|_{z=1} - \left(\left.\frac{\partial F}{\partial z_i}\right|_{z=1}\right)^2 \\
\mathrm{Cov}(n_i, n_j) &= \mathrm{E}(n_i n_j) - \mathrm{E}(n_i)\mathrm{E}(n_j) \\
&= \left.\frac{\partial^2 F}{\partial z_i z_j}\right|_{z=1} - \left.\frac{\partial F}{\partial z_i}\right|_{z=1} \left.\frac{\partial F}{\partial z_j}\right|_{z=1}.
\end{aligned} \tag{C.10}$$

Defining $x_i = \left.\frac{\partial F}{\partial z_i}\right|_{z=1}$, $y_i = \left.\frac{\partial^2 F}{\partial z_i^2}\right|_{z=1}$, and $c_{i+j} = \left.\frac{\partial F}{\partial z_i}\right|_{z=1} \left.\frac{\partial F}{\partial z_j}\right|_{z=1}$ we rewrite Eqns. C.9 and C.10 and obtain a system of linear differential equations for the temporal evolution of the mean, variance and covariance in the linear differentiation path:

$$\begin{aligned}
\mathrm{E}[n_{\mathrm{LT-HSC}}] &= \mathrm{E}[n_0] = x_0(t) \\
\mathrm{E}[n_{\mathrm{ST-HSC}}] &= \mathrm{E}[n_1] = x_1(t) \\
\mathrm{E}[n_{\mathrm{MPP}}] &= \mathrm{E}[n_2] = x_2(t) \\
\mathrm{Var}[n_{\mathrm{LT-HSC}}] &= \mathrm{Var}[n_0] = x_0(t) + y_0(t) - x_0(t)^2 \\
\mathrm{Var}[n_{\mathrm{ST-HSC}}] &= \mathrm{Var}[n_1] = x_1(t) + y_1(t) - x_1(t)^2 \\
\mathrm{Var}[n_{\mathrm{MPP}}] &= \mathrm{Var}[n_2] = x_2(t) + y_2(t) - x_2(t)^2,
\end{aligned} \tag{C.11}$$

with

$$\begin{aligned}
\frac{\mathrm{d}x_0}{\mathrm{d}t} &= (\lambda_0 - \alpha_0)\, x_0 \\
\frac{\mathrm{d}x_1}{\mathrm{d}t} &= (\lambda_1 - \alpha_1)\, x_1 + \alpha_0 x_0 \\
\frac{\mathrm{d}x_2}{\mathrm{d}t} &= (\lambda_2 - \alpha_2)\, x_2 + \alpha_1 x_1 \\
\frac{\mathrm{d}y_0}{\mathrm{d}t} &= 2\lambda_0 x_0 + (\lambda_0 - \alpha_0)\, y_0 \\
\frac{\mathrm{d}y_1}{\mathrm{d}t} &= 2\lambda_1 x_1 + 2\,(\lambda_1 - \alpha_1)\, y_1 + 2\alpha_0 c_1 \\
\frac{\mathrm{d}y_2}{\mathrm{d}t} &= 2\lambda_2 x_2 + 2\,(\lambda_2 - \alpha_2)\, y_2 + 2\alpha_1 c_3 \\
\frac{\mathrm{d}c_1}{\mathrm{d}t} &= (\lambda_0 - \alpha_0 + \lambda_1 - \alpha_1) c_1 + \alpha_0 y_0 \\
\frac{\mathrm{d}c_2}{\mathrm{d}t} &= \alpha_1 c_1 + (\lambda_0 - \alpha_0 + \lambda_2 - \alpha_2) c_2 \\
\frac{\mathrm{d}c_3}{\mathrm{d}t} &= \alpha_0 c_2 + (\lambda_1 - \alpha_1 + \lambda_2 - \alpha_2) c_3 + \alpha_1 y_1
\end{aligned} \tag{C.12}$$

Requiring $\mathrm{E}(n_i|t) = X_i, \mathrm{Var}(n_i|t_0) = 0$ and $\mathrm{Cov}(n_i, n_j|t_0) = 0, i = 0, 1, 2, \quad j = i + 1$ and $j \leq 2$, the initial conditions $y_i = Y_i$ and $c_i = C_i$ can be determined to

$$\begin{aligned}
\mathrm{E}(n_i|t_0) &= x_i = X_i \\
\mathrm{Var}(n_i|t_0) &= \mathrm{E}(n_i^2) - \mathrm{E}(n_i)^2 = Y_i + X_i - X_i^2 = 0 \qquad \leftrightarrow \qquad Y_i = X_i^2 - X_i \\
\mathrm{Cov}(n_i, n_j|t_0) &= C_{i+j} - X_i X_j = 0 \qquad \leftrightarrow \qquad C_i = X_i X_j,
\end{aligned} \tag{C.13}$$

with which the system can be solved numerically.

C.2 Stochastic simulations of clone size distributions in adult hematopoiesis

To assess the clone size distributions in adult hematopoiesis with stochastic simulations, we initialize the simulation with the state vector $\boldsymbol{n}^0$ and $t_0 = 0$, and at each simulated step choose the next event based on the probabilities of cell division and differentiation according to the reaction rates (Eqns. C.1):

$$\text{Symmetric division of cell type i:} \qquad P_{\mathrm{S,i}} = \frac{\lambda_{\mathrm{i}} n_{\mathrm{i}}}{\sum_{\mathrm{j}} \lambda_{\mathrm{j}} n_{\mathrm{j}} + \alpha_{\mathrm{j}} n_{\mathrm{j}}} \tag{C.14}$$

$$\text{Differentiation of cell type i to type i+1:} \qquad P_{\mathrm{D,i}} = \frac{\alpha_{\mathrm{i}} n_{\mathrm{i}}}{\sum_{\mathrm{j}} \lambda_{\mathrm{j}} n_{\mathrm{j}} + \alpha_{\mathrm{j}} n_{\mathrm{j}}}. \tag{C.15}$$

We then sample the time step Δt from an exponential distribution with rate $\sum_j \lambda_j n_j + \alpha_j n_j$ and update the system according to

$$n_i^{j+1} = \begin{cases} n_i^j + 1, & \text{if event = 'Symmetric division of cell type i',} \\ n_i^j - 1, & \text{if event = 'Differentiation of cell type i to i+1',} \\ n_i^j + 1, & \text{if event = 'Differentiation of cell type i-1 to i',} \\ n_i^j, & \text{else.} \end{cases}$$

$$t_{j+1} = t_j + \Delta t.$$

These steps are iteratively repeated until $t \geq t_{\max}$.

C.3 Stochastic simulations of neutral drift during adult hematopoiesis

To assess the effect of neutral drift during adult hematopoiesis on the cellular frequencies of mutations acquired during embryonic expansion, we simulate mutation frequencies with an agent-based model (Fig. 3.11). On the macroscopic scale, the agent-based model traces the number of LT-HSCs, ST-HSCs and MPPs in the state vector $\boldsymbol{n} = (n_0, n_1, n_2)$. On the microscopic scale, the mutation frequencies within each compartment (i.e., the number of cells harboring a specific mutation) are stored in a 3 x l mutation matrix M, whose rows correspond to the three compartments and whose l columns correspond to distinct loci in the genome. The first locus in the genome without a mutation is stored in the indicator variable k and is initialized to $k = 1$. During tissue expansion, cells divide symmetrically at rates $\lambda_{S,0}, \lambda_{S,1}, \lambda_{S,2}$ or asymmetrically at rates $\lambda_{A,0}, \lambda_{A,1}, \lambda_{A,2}$. Thus cellular loss due to death or direct differentiation is neglected during the expansion phase. New mutations are acquired at a single-base-pair substitution rate μ per cell division.

The simulation is initiated with $\boldsymbol{n}^0 = (1, 0, 0)$ and $M^0 = 0_{3,l}$, mimicking embryonic expansion from a single LT-HSC. On the macroscopic scale, the next event — symmetric or asymmetric cell division — is randomly sampled with probabilities determined by the number of cells and the reaction rates, according to:

$$\text{Symmetric division of cell type i:} \qquad P_{S,i} = \frac{\lambda_{S,i} n_i}{\sum_j \lambda_{S,j} n_j + \lambda_{A,j} n_j} \tag{C.16}$$

$$\text{Asymmetric division of cell type i:} \qquad P_{A,i} = \frac{\lambda_{A,i} n_i}{\sum_j \lambda_{S,j} n_j + \lambda_{A,j} n_j}. \tag{C.17}$$

On the microscopic scale, the number of new mutations per daughter cell, $n_{\text{mut},1}$ and $n_{\text{mut},2}$, are sampled from a binomial distribution with sample size 3×10^9 (corresponding to the size of

the genome) and success probability μ.

At each simulation step j, the state vector is updated according to the selected event:

$$n_{\mathrm{i}}^{\mathrm{j+1}} = \begin{cases} n_{\mathrm{i}}^{\mathrm{j}} + 1, & \text{if event = 'Symmetric division of cell type i',} \\ n_{\mathrm{i}}^{\mathrm{j}} + 1, & \text{if event = 'Asymmetric division of cell type i-1',} \\ n_{\mathrm{i}}^{\mathrm{j}}, & \text{else.} \end{cases}$$

Accordingly, the mutation matrix is updated by (i) modeling the inheritance of mutations during cell divisions and (ii) modeling the acquisition of new mutations:

If event = 'Symmetric division of cell type i':

$$M_{\mathrm{i,l}}^{\mathrm{j+1}} = \begin{cases} M_{\mathrm{i,l}}^{\mathrm{j}} + 1, & \text{with probability } \frac{M_{i,l}^{\mathrm{j}}}{n_{\mathrm{i}}}, \\ M_{\mathrm{i,l}}^{\mathrm{j}}, & \text{with probability } 1 - \frac{M_{i,l}^{\mathrm{j}}}{n_{\mathrm{i}}.} \end{cases}$$

While $k \leq n_{\mathrm{mut},1}$:

$$M_{\mathrm{i,k}}^{\mathrm{j+1}} = \begin{cases} 1, & \text{if event = 'Symmetric division of cell type i',} \\ 1, & \text{if event = 'Asymmetric division of cell type i',} \\ 0, & \text{else,} \end{cases}$$

$k = k + 1.$

While $k \leq n_{\mathrm{mut},2}$:

$$M_{\mathrm{i,k}}^{\mathrm{j+1}} = \begin{cases} 1, & \text{if event = 'Symmetric division of cell type i',} \\ 1, & \text{if event = 'Asymmetric division of cell type i-1',} \\ 0, & \text{else,} \end{cases}$$

$k = k + 1.$

Embryonic expansion is simulated until $n_0 = 10,000$, using the parameters provided in Table 3.1. As MPPs are rapidly turned over, it is sufficient to simulate neutral drift in MPPs during the homeostatic phase, so that $\lambda_{\mathrm{S},2} = \lambda_{\mathrm{A},2} = 0$ during embryonic expansion. Once the hematopoietic system is established (i.e., at $n_0 = 10,000$), all three compartments are simulated. However, due to the large cell number and rapid turnover of MPPs, stochastic simulations of the full system are computationally costly. We therefore approximate the homeostatic phase by sampling the fate of an individual mutation from the simulated clone size distribution after two years (c.f. Fig. 3.10D and Appendix C.2). Thus mutations newly acquired during adulthood are neglected by the simulation, in agreement with our previous argument that these mutations do not influence the measured mutation frequency distribution.

The final mutational profile is stored in the mutation matrix M', which is initialized to $M'^{0}_{3,l} = 0$. To simulate the fate of mutations present in LT-HSCs during adult hematopoieisis, we sample for each cell a mutational profile m, according to

$$m_l = \begin{cases} 1, & \text{with probability } \frac{M_{0,l}}{n_0}, \\ 0, & \text{with probability } 1 - \frac{M_{0,l}}{n_0}. \end{cases}$$

Next, we sample from the simulated clone size distributions the instances c_0, c_1, c_2, corresponding to the number of LT-HSCs, ST-HSCs and MPPs grown by this single cell after two years and update M':

$$\begin{aligned} M'^{j+1}_{0,.} &= M'^{j}_{0,.} + c_0 \times m \\ M'^{j+1}_{1,.} &= M'^{j}_{1,.} + c_1 \times m \\ M'^{j+1}_{2,.} &= M'^{j}_{2,.} + c_2 \times m. \end{aligned}$$

The fate of embryonic mutations present in ST-HSCs is analogously simulated.