Sampling Strategies in Evolving Cancer

From Mathematical Models to Clinical Application

Dissertation for Fulfillment
of the Requirements for the Doctoral Degree
Doctor rerum naturalium
the Faculty of Mathematics and Natural Sciences
Christian-Albrechts University of Kiel

Submitted by
Luka Opašić, dr.med.

Department of Evolutionary Theory,
Max Planck Institute for Evolutionary Biology, Plön

Kiel, 2020

First referee: Prof. dr. Arne Traulsen

Second referee: Prof. dr. Hinrich Schulenburg

Date of oral examination: 25.3.2020.

Approved for printing:

To my mother who lost her battle with cancer

Abstract

Despite tremendous resource investment in the fight against cancer over the last 50 years, prognosis of the late-stage malignant disease is almost inevitably unfavourable as most cancers are resilient against treatment. The source of this cancer resilience lies in rapid somatic evolution and consequential extensive intra-tumour genetic heterogeneity. Unravelling this complex genetic landscape of cancer is thus mandatory if we are ever to overcome the emergence of resistance. However, current sampling procedures allow only to investigate a small subset of malignant cells. Drawing conclusions about characteristics of the entire cancer from this partial information inevitably introduces a bias. To counter this bias, this thesis investigates sampling strategies in cancer genomic profiling by combining mathematical and computational modelling, and validating the models with cancer genomic data. In the first chapter, I introduce a mathematical model for calculating the number of samples needed to successfully identify mutations present in every cancer cell from multi-region genomic profiling of solid tumours. The clonality inference procedure is further tested in a spatial model of intratumour heterogeneity. Moreover, I show how the size of individual samples affects the probability for correct clonality estimation and how an optimized sampling pattern can improve detection accuracy to a great extent. In the next chapter, presented theoretical model was applied to genomic data derived from patients with gastric adenocarcinoma. We find that in three out of nine patients, the existing number of samples is sufficient to infer the clonality of detected mutations with a high degree of certainty. Additionally, an attempt to characterise the mode of evolution in primary tumours of each patient revealed a diversity of patterns characteristic for different evolutionary trajectories. The third chapter is dedicated to modelling the process of cancer genetic profiling in solid tumours using liquid biopsies. Here, I present the extent of sampling bias encountered in this type of diagnostics and show how it can lead to a distorted view of the tumour sub-clonal composition. I estimate the amount of genomic material necessary for detection and quantification of both individual and sets of genetic alterations present range of frequencies. Finally, in the last chapter, I present a software package for simulating spatial tumours developed in Python that provides an easily accessible resource for studying the evolutionary processes of cancer progression. The work presented in this thesis demonstrates how mathematical and computational modelling, combined with clinical data, can be used to support cancer diagnostics and assist clinicians in making better informed treatment decisions.

Kurzzusammenfassung

Trotz des außerordentlichen Einsatzes im Kampf gegen Krebs in den letzten 50 Jahren, sind die Prognosen für bösartige Erkrankungen aufgrund der Widerstandfähigkeit der Krebszellen gegen die Therapie schlecht. Die Ursache der Widerstandfähigkeit der Krebszellen liegt in ihrer schnellen somatischen Evolution und dem hohen Maße an intratumoraler genetischer Heterogenität. Um das Auftreten der Resistenz verhindern zu können, ist die Entschlüsselung der komplexen genetischen Vielfalt der Krebszellen notwendig. Derzeit lassen es die Prozeduren zur Probeentnahme von Tumorzellen nur zu, einen kleinen Anteil der bösartigen Zellen zu untersuchen. Schlussfolgerungen über die Charakteristik der gesamten Tumorzellen basierend auf diesen Teilinformationen sind oft problematisch. Um diesem Bias entgegen zu wirken, befasst sich diese Arbeit mit Stichprobenverfahren im Bereich des Genomic-Profiling von Krebszellen mithilfe der Kombination von mathematischen und rechnergestützten Modellierungen. Zur Validierung der Modelle werden Genomdaten von Krebszellen genutzt. Im ersten Kapitel stelle ich mathematische Modelle zur Berechnung der Anzahl von Stichproben, die zur vollständigen Identifikation aller in den Krebszellen vorkommenden Mutationen genutzt werden, vor. Diese basieren auf genetischen Profilen von soliden Tumoren. Die „clonality-inference“-Prozedur wird dann an einem räumlichen Modell der intratumoralen Heterogenität getestet. Ferner zeige ich, wie die Größe der individuellen Stichproben die Wahrscheinlichkeit der korrekten Schätzung der Klone beeinflusst und wie ein optimiertes Sampling-Muster die Erkennungspräzision in einem hohen Maße steigert. Im nächsten Kapitel wird der vorgestellte theoretische Ansatz auf Patientendaten von Patienten mit einem gastritischem Adenokarzinom angewandt. Wir können zeigen, dass für drei der neun Patienten die bestehende Anzahl von Stichproben ausreicht, um mit einem hohen Grad an Sicherheit die Klonalität der entdeckten Mutationen zu folgern. Zusätzlich wird durch den Versuch, die Evolutionsfaktoren in den primären Tumoren der Patienten zu charakterisieren, die Vielfalt der Muster verschiedenster evolutionärer Trajektorien aufgedeckt. Das dritte Kapitel befasst sich mit der Modellierung des Prozesses des genetischen Profiling von Krebszellen in soliden Tumoren mit Hilfe von Flüssigbiopsie. Hier stelle ich das Ausmaß der Auswahlverzerrung, welches in dieser Art der Diagnostik zu finden ist, dar und zeige wie diese Diagnostik zu einer verfälschten Darstellung der subklonalen Zusammensetzung des Tumors führt. Weiter berechne ich die Menge an genetischem Material, welche notwendig zur Detektion und Quantifizierung von sowohl individuellen, als auch Gruppen von genetischen Veränderungen ist. Abschließend stelle ich im letzten Kapitel ein Softwarepaket zur Simulierung von räumlichen Tumoren

vor. Dieses bietet eine leicht zugängliche Quelle, um die evolutionären Prozesse der Krebsprogression zu untersuchen. Die in dieser Arbeit vorgestellten Ergebnisse verdeutlichen, wie mathematische und rechnergestützte Modellierung in Kombination mit klinischen Daten genutzt werden kann, um die Diagnostik im Bereich der Krebstherapie zu unterstützen und den Ärzten zu helfen, Entscheidungen bezüglich der Therapie zu treffen.

Contents

Chapter 1

Introduction

1.1 Motivation

One in five males and one in six females will develop cancer during their lifetime, and one in eight males and one in eleven females will die from it according to the World Health Organisation's International Agency for Research on Cancer (IARC). IARC estimated the number of 18.1 million new cancer cases and 9.6 million cancer deaths in 2018 (Bray et al., 2018; Ferlay et al., 2019). This makes cancer one of the leading causes of death worldwide.

Such high prevalence and mortality of the malignant diseases possess great challenges for the society as a whole. Most significantly, cancer has a great negative psychosocial impact on the lives of patients, their friends and relatives (Singer, 2018). Also, the increasing cost of cancer care puts a high financial burden on societies (Meropol et al., 2009; Luengo-Fernandez et al., 2013). As an example, cancer medications were the leading class in hospital drug expenditures already before the year 2009 (Hoffman et al., 2009). Despite the great cost of newly developed medication, its curative potential in treatment of the late-stage malignant disease is extremely limited (Davis et al., 2017). Cancer is difficult to treat for several reasons: Standard chemotherapy is non-selective and is affecting all dividing cells in the body. Since neoplastic cells divide faster than the normal cells, they are more impacted by this type of treatment. Still, the effect of chemotherapy on non-cancer cells is severe, which leads to side effects that have great consequences on the patient well-being and their ability to tolerate sufficiently high doses of therapy (Schirrmacher, 2019). More modern therapy options, despite being designed do target tumour-specific alterations, are most often, unable to bring the entire cancer population to extinction and thus often fail to achieve full remission in the patient. Emergence of resistance

in the cancer cell population and wrong choice of target make targeted therapy ineffective after some time. Delayed time of diagnosis also contributes to the poor outcome. However, neoplasms detected in very early stages are often fully curable diseases and improvement in early detection showed a great decline in mortality rates for certain cancers (Neal et al., 2015; Pastorino et al., 2019; Hawkes, 2019).

Structure of the thesis

This thesis is a small contribution to the global effort that strives to improve the understanding of cancer and to devise more effective cancer treatments. Throughout the six chapters that constitutes the thesis, I will present topics that directly address some of the aforementioned causes of treatment failure. Since this thesis takes an interdisciplinary approach to study cancer I will start with Introduction where I will introduce three distinct aspects of cancer: biological–its development, clinical–its diagnostic and treatment and mathematical–using mathematics to describe cancer. Chapter 2 deals with the problems of spatial sampling in diagnostics of potential targets for personalized therapy. Here, I will present a theoretical model which can be used to identify the alterations present in every cancer cell with high confidence. In the following chapter, I will apply theoretical models on patient-derived genomic data in an attempt to bridge the gap between theory and clinical practice. Liquid biopsy, the method that shows promise for improving the early cancer detection, is the topic of Chapter 4. It investigates the extent of sampling error in liquid biopsy diagnostics and its consequences for clinical medicine. In chapter 5, I will present more closely a computational spatial model of intratumour heterogeneity used in earlier parts of this thesis. Finally, I will summarize my contributions and offer future perspectives to the presented research.

1.2 Cancer biology

Development and maintenance of tissue homeostasis relies on highly complex interactions between inter- and intracellular molecular mechanisms (Hall, 2016). This interaction ensures complete cell differentiation, cooperation between cells and ultimately programmed cell death (Pellettieri and Sánchez Alvarado, 2007). Disruption in any of these basic processes leads to various structural and functional abnormalities of which some can lead to cancer and ultimately be lethal to the organism. This disruption can be caused by environmental stimuli such as tobacco smoke or gastric acidity or by random somatic mutations. In the former case, an unfavourable environment provokes reactive adaptation of tissue by phenotypic–morphological transformation from one differentiated somatic cell type to another differentiated somatic cell type in the process of *metaplasia* (Giroux and Rustgi, 2017). In Barret's oesophagus squamous epithelium undergoes adaptive transdifferentiation to more acid-resistant columnar epithelium with clonally derived gland-like structures (Nicholson et al., 2012). Upon termination of stimulus, transformed tissue usually returns to physiological state. However, in some instances, a series of irreversible pathological changes within cells and tissue can occur. These changes are manifested as the absence of full cell differentiation and are known as epithelial *dysplasia*. Dysplasia is considered precancerous transformation and can lead to fully invasive cancer (Warnakulasuriya et al., 2008). In both metaplasia and dysplasia, growth pattern is affected but without excessive cell proliferation. Dysplastic lesions that do progress acquire early a number of mutations usually found in fully developed cancer (Wood et al., 2017). Still, in most cases, dysplastic lesions remain stable or are eliminated by the immune system (e.g. Cervical intraepithelial neoplasia (CIN) – cervical dysplasia) (Tainio et al., 2018). High grade dysplasia is known as carcinoma *in situ* (e.g. Dermatosis praecancerosa Bowen, Erythroplasia of Queyrat, Paget's disease of the breast). In this state, altered cells don not penetrate the basement membrane of the epithelium (Vinay Kumar, 2017). Penetration does occur with type-specific probabilities and marks the transition of the precancerous lesion into fully invasive cancer (Cox et al., 1999; Fanning and Flood, 2012; Bundred and Dixon, 2013).

Cancer (Latin word for Greek karkínos – καρκινος eng. crab – due to resemblance of a late-stage breast malignancy with a crustacean) is referring to malignant neoplasms whose main feature is the ability to metastasize and to invade and destroy surrounding tissue. These two features are absent in benign neoplasms and make a distinction between them. There are, however, cancers like basalioma, that generally do not seed metastases (Piva de Freitas et al., 2017). Still, they are invasive and considered full carcinomas. Tumour is a term used as a synonym for a neoplasm. It is not however synonymous with cancer as tumours do not have to be malignant. Tumour is referring to solid neoplasms, as excessive proliferation in neoplasms forms macroscopic tuber-like structure composed new altered cells. To ancient physicians, this swelling looked like a *tumour*, one of four cardinal signs of inflammation together with *rubor, calore, dolore* from so-called *Celsus tetrad*, and the name remained until this day. Even though each neoplasm is unique, they are classified into types of based on the features such as location or origin of the cell. Carcinoma, the most common malignancy, is a malignant neoplasm of epithelial cell origin regardless of the germ layer. Sarcomas are malignancies derived from cells of mesenchymal origin such as bone and muscle. They are rather rare representing 1% of all malignancies. Leukemia lymphoma and myeloma are neoplasms of the haematopoietic system and lymphoid tissues (Vinay Kumar, 2017). Leukaemia is usually considered as structureless as they do not form solid, spatially extended tumour, and as such are more tractable for mathematical modelling than solid tumours (see later paragraphs) Even though invasion and metastases are the distinctive features of cancer, work presented in this thesis focuses on cancer evolution and intratumour genetic heterogeneity as they are major factors related to therapy resistance.

Mechanisms of excessive proliferation

In the context of this thesis, it is important to get the idea of how mutations drive the increased cell proliferation as this is the feature is extensively modelled in the following chapters. Tissue homeostasis, cell processes and activities such as division, tissue repair or immunity are dependent on precise inter- and intracellular signalling (Hall, 2016). Disruption in some of the critical elements in signalling cascades can lead cells to proliferation and malignant transforma-

tion (Sanchez-Vega et al., 2018). Excessive proliferation of somatic cells can be caused by mutation, overexpression or amplification of specific genes that code protein for cell cycle regulation. These *proto-oncogenes* can be transcription factors, signal transduction proteins, growth factors and growth factor receptors (Croce, 2008). Also, in a malignant stage, cell proliferation in healthy tissue is usually mediated by paracrine signalling where cells stimulate proliferation of neighbouring cells by secretion of growth factors. In cancer, cells are self-stimulating their mitosis by secreting growth factors into their environment thus forming an autocrine loop (Heldin, 2012). Secretion is caused by overexpression or amplification of genes that code for growth factors (e.g. *TGF-β* and *PDGF* in breast cancers (Elenbaas and Weinberg, 2001). Growth factor receptors are trans-membrane proteins who transfer information from the outside to the inside of a cell. Being stimulated either by ligand or self-stimulation, they activate a complex cascade of intracellular signalling pathways. Same cell proliferation pathways can in cancer be activated even without the presence of growth factors in the tumour stroma. Mutations in growth factor receptors can make them constitutively active (e.g. *ALK* tyrosine kinase in lung cancer) which leads to constant cell proliferation and tumour growth (Shaw and Engelman, 2013). Downstream elements in these pathways can also be mutated and sustain proliferation without external stimulus. One of the most common mutations found in cancer is a mutation in the *RAS* gene family (*KRAS*, *HRAS*, *NRAS*) and *BRAF* (Hobbs et al., 2016). Because these mutations are so common, RAS proteins are being aimed as targets in precision therapy, although with only modest success (O'Bryan, 2019). Furthermore, amplifications and mutations in genes that code for transcription factors (*MYC*) and cell cycle regulators (cyclins and cyclin-dependent kinases) will also result in high mitotic activity, independent of external stimulus (Dang, 2012; Sante et al., 2019). As previously stated, non-physiological excessive cell divisions can often lead to malignancy. Therefore, to curb the unnecessary proliferation, cells use complex mechanism of growth inhibition checkpoints, proteins called tumour suppressors. In the case of expressed oncogene, if the tumour suppressor system is fully operational, the cell enters into cell cycle arrest, quiescence and ultimately apoptosis. Some parts of the mechanism are so essential that a mu-

tation in single tumour suppressor can lead to fully invasive cancer as in the example of *RB*, the first discovered tumour suppressor gene. RB protein acts as an inhibitor of G_1/S transition during the cell cycle in the presence of DNA damage in the cell (Weinberg, 1995). Loss of its function enables these cells to enter the S phase of the cell cycle, after which mitosis will proceed to its completion. Mutation in both alleles of the *RB* gene in cone precursor cells of retina causes retinoblastoma (Xu et al., 2014). This was first proposed by Knudson (1971) who used epidemiological data of retinoblastoma patients to get insight into underlying genetic events much long before allowed by molecular biology. However, there are recent indications that single mutations might be sufficient to initiate a genetic instability in cancer (Coelho et al., 2019). Active oncogenes and inactive tumour-suppressor genes are not the only faulty cellular mechanisms responsible for carcinogenesis. The cell life cycle of cell usually ends with controlled death, called apoptosis, which is triggered by the absence of survival factors or internally as a protective response to DNA damage. By committing suicide, cells eliminates the chance of neoplastic alterations. Inability of a cell to commit to apoptosis is crucial to the development of cancer (Brown and Attardi, 2005). There are other mechanisms of cancer inhibition that are also impaired in tumours.

Intratumour heterogeneity and evolution

Increased rate of cell division and impaired ability to repair DNA damage lead to constant accumulation of mutations in dividing cancer cells (figure 1.1). Per-cell point mutation rate is 10 to 30 times larger than in healthy tissue which translates to 1.14 mutations per genome per cell division (Werner et al., 2019). This means that every cell in the cancer is genetically unique and that the total number of mutations in the tumour is extremely large as 1 cm^3 of tumour contains $10^8 - 10^9$ cells (Del Monte, 2009). However, in solid tumours, sub-clones are not all equally distributed within the mass. Cells divide in their neighbourhood and are held together by spatial constrains of surrounding cells and the extracellular matrix. Therefore, genetic heterogeneity in solid cancers entails a specific spatial distribution. Spatial intratumour genetic heterogeneity was first described in renal cancer where sequencing of different parts of the tumour uncovered distinct mutations present in individual samples and not shared among

other samples (Gerlinger et al., 2012, 2014). Soon after that, spatial intra-tumour heterogeneity was characterised across cancer types (de Bruin et al., 2014; Rasche et al., 2017; McGranahan and Swanton, 2017). The main feature modelled in this thesis is the somatic evolution of tumour and its genetic heterogeneity. Spatial heterogeneity is modelled in the way that each cell, along with their distinctive mutations, carries a set of unique spatial coordinates. As previously described, cancer contains a great number of different mutations due to impaired DNA repair mechanisms. Mutation, one of the drivers of evolution, can create enormous genetic variation within individuals in the population (Lynch et al., 2016). Despite this great number of unique mutations, most of them are considered to be neutral or deleterious in treatment-naive tumour (Williams et al., 2016; Reiter et al., 2019). But then, medical intervention introduces an great evolutionary pressure on the cancer cell population. Some mutations, neutral or even deleterious in the absence of therapy, enable survival of carrier cells in this new environment. Natural selection, another force of evolution, then acts upon heterogeneous population of cancer cells by selecting for cells that contain advantageous mutations in this changed environment (Turajlic et al., 2019). This scenario of populations undergoing severe stress, adaptation and eventually avoiding extinction through the means of natural selection is called evolutionary rescue (Bell and Futuyma, 2017; Uecker, 2017) and it is an important paradigm in cancer treatment. Studying under which conditions populations go extinct is critical for the future development of therapy. Additionally, some mutations that may be beneficial under one condition may in turn be deleterious in others. This translates to a phenomenon in which an cancer cell that has developed resistance to one drug displays increased sensitivity to another drug, known as collateral sensitivity (Pál et al., 2015; Barbosa et al., 2017). Identification of sensitivity networks might allow us to steer the cancer population through genotype space, using sequential drug treatments, to avoid the emergence of resistance, which has similarly been shown in bacterial models of antibacterial treatment (Nichol et al., 2015). In conclusion, evolution is usually the ultimate cause of treatment failure and consideration of evolution in clinical management of cancer is imperative if we ever want to win The War on Cancer.

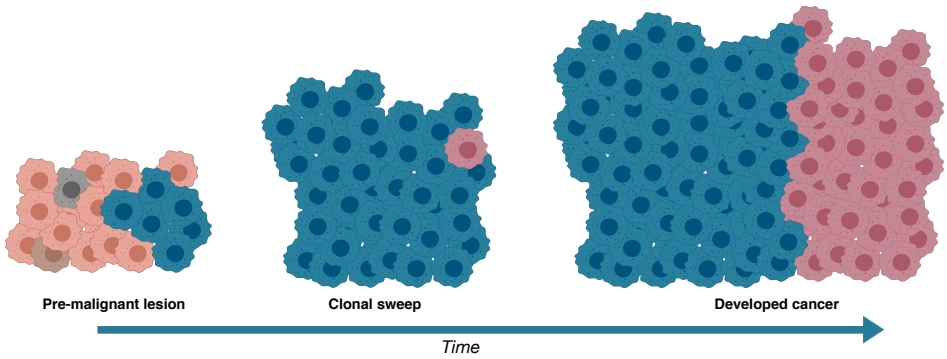

Figure 1.1: Somatic evolution of cancer. Cancer emerges from a pre-malignant lesion by accumulation of driver mutations. Acquisition of fully malignant phenotype with high fitness often leads to fixation of the mutant after the clonal sweep. Subsequent evolution in untreated cancer can be neutral (Adapted from (Reiter et al., 2019). Figure created with BioRender.com)

1.3 Clinical management of cancer

1.3.1 Laboratory diagnostics of malignant diseases

Laboratory diagnostics of neoplastic diseases is currently advancing at a rapid pace. While traditional methods for characterizing of neoplasms based on phenotype are still a standard, genomics has the potential to change the diagnostic process entirely (Wall and Tonellato, 2012).

Histopathology

Traditionally, and still up to this day, cancer diagnostics relies on direct observation of formalin-fixed paraffin-embedded (FFPE) tumour tissue under the optical microscope. Classification of cancer types is still largely based on cell and tissue morphology. In most of the cases, this phenotype-based method is sufficient to make a successful cancer diagnosis (Vinay Kumar, 2017). Diagnostics includes identification of the cancer cell origin and classification into one of histological types (Adenocarcinoma, squamous cell carcinoma, anaplastic carcinoma etc.). On top of that, neoplasms are further graded based on the level of differentiation into stages from (I to IV) (Edge and Compton, 2010). In general, the less differentiated cells are, the worse is the prognosis (Sun et al.,

2006). Finally, information on the extension of the primary tumour, its spread into regional lymph nodes and presence of distant metastasis is combined into the TNM score which serves as a key factor in determining appropriate treatment and estimation of prognosis for every cancer patient (American Joint Committee on Cancer, 2018).

Molecular pathology and genomics

Standard histologic evaluation is becoming increasingly complemented by techniques that utilize genetic distinctiveness of different neoplastic diseases to characterise their behaviour and treatment response (Imyanitov and Sokolenko, 2018). An early adopted molecular technique was immunohistochemistry (IHC) staining. Specific proteins can be made directly observable under the microscope after labelling with specific antibodies. Since numerous molecules are variably expressed in different cancer types or different differentiation stages, this serves as a powerful tool for both differential diagnosis and prognosis assessment (Kim et al., 2016). For example, a combination of myoepithelial markers is routinely used discriminate between invasive and non-invasive breast cancer ((Dewar et al., 2011)). Additional upgrades to the microscopy technique came with fluorescent in-situ hybridisation (FISH). This method enables detection of chromosomal abnormalities such as deletions, gains, translocations, amplifications and polysomy, which have great clinical implications for many cancer types (Cheng et al., 2017). It became part of routine pathological genetic analysis especially in diagnostics of soft tissue sarcoma (Sugita and Hasegawa, 2017). However, FISH is able to detect only large, catastrophic, chromosome level, genomic events. To detect specific genetic alteration on the nucleotide level, polymerase chain reaction (PCR) was introduced (Evans, 2009). Improved PCR methods such as Digital Droplet PCR (ddPCR) allows the detection of rare alleles from the analysed sample (Hindson et al., 2011). Common examples of PCR usage in tumour diagnostics include the detection of the EGFR mutation in Non-small-cell lung carcinoma (NSCLC) (Zhang et al., 2015) or an amplification of oncogenes such as NMYC in neuroblastoma (Brodeur et al., 1984) or HER2 in breast cancer (Mitri et al., 2012). However, both FISH and PCR come with their own specific limitations. For both methods, the examiner is testing specimens on a panel of genetic alterations that are expected to

be found in certain malignancies (Gozzetti and Le Beau, 2000; Imyanitov and Sokolenko, 2018). The number of detected genetic alterations on both chromosome level and nucleotide level can be increased by sequencing tumour DNA using the Next-Generation Sequencing (NGS) technology. This technique is currently used to sequence and analyse mutations in only a number of genes of interest. However, as the costs of NGS decreases, we can expect whole-genome sequencing becoming a routine diagnostic procedure in the near future (Kamps et al., 2017). With the introduction of targeted therapy, the identification of tumour-specific target alterations and resistance mutations became crucial for the success these therapies (Savage and Antman, 2002). Sequencing in clinical use is currently only limited to targeted sequencing. Sequencing data from cancer patients was used in the third chapter of this thesis in order to apply the theoretical method we developed in the preceding chapter.

1.3.2 Sampling of neoplasms

Sampling procedure is the first and most important step in cancer diagnostics. For that reason I decided to make sampling the main topic of my thesis. Every time we attempt to assess the characteristics of the whole of a collective from its subset, we encounter a sampling error. Problems with sampling are well known in pathology and many techniques are designed to reduce the amount of sampling error. For example, when assessing the tumour margins, the *bread loafing* technique relies on observing a number of sections of the specimen and has a much higher false-negative error rate compared to the complete circumferential peripheral and deep margin assessment (CCPDMA) or Mohs surgery, a technique which allows the complete examination of the surgical margin without statistical inference (extrapolation) (Kimyai-Asadi et al., 2005).

Every histopathological examination procedure starts with the collection of tissue or cells by a clinician. Samples can include surgical excision, needle biopsy, fine-needle aspiration or cytological smear, each with its own distinctive characteristics that determine the sampling efficiency (Vinay Kumar, 2017) (see figure 1.2). Specimens for histopathological diagnostics have to be sampled in a way that prevents total disruption of the relationship between the cells within

the sample. This is the case with biopsy sampling, done using a cylinder large enough to keep the tissue integrity complete. Examples of tissue biopsy sampling include punch biopsy, a procedure that includes sampling suspicious skin lesions using a hollow, circular scalpel (Vinay Kumar, 2017). For diagnostics of pleural mesothelioma, tissue sampling is performed using minimally invasive video-assisted thoracoscopic surgery which allows doctors to extract multiple samples from a lesion (Xu et al., 2018). Spatial biopsy sampling is the focus of the second and third chapter part of the thesis. There I use the data obtained from a tissue biopsy to find the minimal number of samples needed for correct classification of mutations. Since in invasive diagnostic sampling procedures, such as minimally invasive video-assisted thoracoscopic surgery, availability of samples is limited due to the spatial constraints, knowledge of the minimal number of samples is of great value. Finally, tissue specimens can be obtained by a surgical excision of the lesion. Having an entire tumour extracted, there are no spatial constraints on the number of samples that can be harvested both for histological and genetic analysis. It is imperative to conduct proper and well-thought sampling process as sampling error might lead to erroneous treatment and ultimately have great negative consequences for patients. While the information and integrity of the spatial structure in excision and needle biopsy mostly remains intact, they are completely absent in case of both aspiration and cytological smear. The type of collected specimen mostly depends on the location–accessibility of the lesion and the clinical stage of the disease. Naturally, the least invasive techniques are preferred when appropriate and some examples include: Cervical cytological smear, procedure used for screening of cervical precancerous and cancerous lesions called Papanicolaou test (Papanicolaou and Traut, 1997). Another one is fine needle aspiration, process of extracting cells through a needle and is used to sample bodily fluids like blood, pleural effusion, cerebrospinal fluid (CSF). This technique is also used to obtain tissue and fluid from solid or cystic lesions such as thyroid nodules and breast masses (Mazzaferri, 1993; Ahmadinejad et al., 2017). Cells obtained this way are further observed through the microscope or analysed with previously discussed molecular methods. The fourth chapter of this thesis is dealing with liquid biopsy, a type of needle aspiration sampling where genetic material is iso-

lated from bodily fluids and sequenced. Liquid biopsy holds great potential to become an indirect diagnostic and screening methods for cancer (Cohen et al., 2018). It is important to notice that aspiration techniques lose the original spatial properties of a tumour. That is the reason why cytological diagnostic procedures are usually inferior to the histopathological.

In summary, sampling of neoplastic disease for diagnostic purposes is key to infer appropriate and eventually personalized treatment plans. Therefore, I investigate both spatial (biopsy) and non-spatial (liquid biopsy) sampling. More closely, I focus on the role of the spatial tumour structure on the ability to classify mutations based on clonality and the effects of the sample size on the ability to detect cancer mutations from the genetic material obtained from bodily fluids using liquid biopsy.

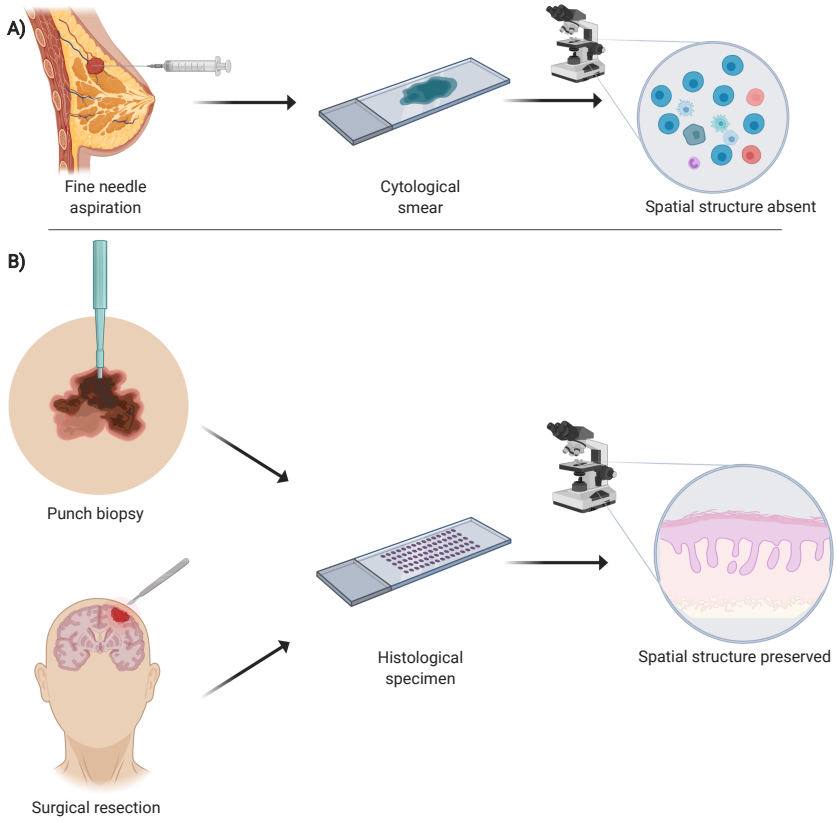

Figure 1.2: Sampling of neoplasms. Fine needle aspiration and cytological smears (A) lose spatial structure of the tissue but are less invasive procedures. Punch biopsy and whole tumour excision (B) are more invasive but provide more information about neoplasms. All procedures are subjected to sampling errors. Figure created with BioRender.com

1.4 Mathematical oncology

Mathematical oncology can be defined as the use of mathematics, modelling and simulation in cancer research (Rockne and Scott, 2019; Rockne et al., 2019). In general, mathematical modelling is applied when encountering a problem that is not experimentally tractable (Servedio et al., 2014). Cancer is a complex and dynamic set of diseases, inaccessible to observation and analysis during most of its existence, especially during early formation stages. Available experimental tools to study phenomena such as seeding of metastasis, cell dormancy

or somatic evolution are lacking so they are ideal examples of cases where mathematical models can have a great role in the investigation. Traditionally, mathematical methods have been used most extensively in the field of cancer epidemiology to discover factors that affect the incidence and health outcomes of oncological diseases. In their pioneering work, Armitage and Doll were able to use the cancer mortality statistics to get insight into the mechanisms of carcinogenesis (Armitage and Doll, 1954, 1957). They suggested carcinogenesis is a multi-step process by creating mathematical models which were able to explain the observed epidemiological data. Their work marked the beginning of mathematical modelling in cancer research. The field has been steadily growing since then and it successfully introduced a variety of mathematical tools to study cancer (Altrock et al., 2015a). There are many ways cancer can be modelled and it is always challenging to decide which approach is optimal. The majority of cancer models is based on specific assumptions about the tumour growth. Growth dynamics can be modelled by ordinary differential equations. The simplest model to describe tumour growth is the exponential law. Exponential growth has been identified in some cancers, most notably in those deriving from hematopoietic tissue (Hart et al., 1998; Rodriguez-Brenes et al., 2013). An important quantitative concept that was early established from exponential growth is the doubling time of the tumour (Collins, 1956). It has since been widely used for quantification of tumour growth rate (Mehrara et al., 2007). However, exponential growth is considered not applicable to many cancers over long periods of time, especially in solid tumours (Shackney et al., 1978; Wodarz and Komarova, 2014). Since cancer cannot sustain its exponential growth indefinitely due to the spatial constrains and nutrient limitations, at some point its growth reaches the saturation. For that reason, more refined models include a carrying capacity that sets an upper limit on the cancer cell population. Prominent examples here are the logistic growth model or the Gompertz model (Gerlee, 2013). It was recently shown that the same type of cancer can display diverse growth patterns, both exponential-like without and logistic-like with a carrying capacity (Gruber et al., 2019). This shows how often it is hard to decide which modelling approach to take for given experimental data. Complexity of malignant diseases is manifested in,

among other things, stochastic nature of cancer development and its response to treatment. This is evident in the process of accumulation of mutations and in effect these mutations have on fitness of mutants (Turajlic et al., 2019). Additionally, stochastic effects of small populations are important for in cancer phenomena such as cancer dormancy. Certain trends can be discovered despite cancer's stochastic nature using stochastic modelling (Altrock et al., 2015a). One example of stochastic modelling is the branching process where each cell in a single generation produces two cells with a certain probability. Daughter cell can be allowed to acquire mutation with a certain probability. Model can be upgraded by adding a variety of features to the model like the death process, different types of mutations etc. Another example of stochastic processes often used in modelling cancer is the Moran process (Moran, 1958). It describes the change of clone sizes in finite, fixed-sized populations. Each time step one cell chosen for reproduction with a probability proportional to its fitness replaces one cell randomly chosen to die with its offspring. Since the focus of this thesis is sampling bias in cancer diagnostics, I chose the type of model that takes into account spatial relations between cells in the tumour. Models that are best in describing properties emerging from local interaction among neighbouring cells are ones based on a spatial grid such as cellular automata models (Gerlee and Anderson, 2007). In this type of models each cell occupies a node on a lattice and is surrounded by a number of cells, usually four or eight cells. Cells can be of different genotype or phenotype and their behaviour is dictated by a set of rules for the interactions with neighbouring cells. There are other ways to study cancer by modelling the physical properties of the tumour tissue. For example, tumour tissue can be described as a viscous fluid using partial differential equations (Wodarz and Komarova, 2014). These models do not capture features I focus in this thesis.

The main challenge in cancer modelling, and modelling in general is the determination of the level of abstraction (Servedio et al., 2014). Models used to describe cancer can be simple as exponential growth and as complicated as real-size three-dimensional simulation of tumour with all its spatial complexity (Collins, 1956; Waclaw et al., 2015). Models containing complex cancer traits such as vasculature structure, extensive microenvironment or metabolic net-

works although more realistic, can be very difficult to interpret. On the other hand, overly simplistic models can fail in adequate representation of studied phenomenon which could lead to spurious scientific findings.

Chapter 2

Classification of cancer mutations based on clonality

This Chapter underwent peer-review and is published open access in BMC Cancer with a title: *How many samples are needed to infer truly clonal mutations from heterogenous tumours?* (Opasic, Zhou, Werner, Dingli, and Traulsen, 2019)

2.1 Abstract

Background Modern cancer treatment strategies aim to target tumour specific genetic (or epigenetic) alterations. Treatment response improves if these alterations are clonal, i.e. present in all cancer cells within tumours. However, the identification of truly clonal alterations is impaired by the tremendous intra-tumour genetic heterogeneity and unavoidable sampling biases.

Methods Here, we investigate the underlying causes of these spatial sampling biases and how the distribution and sizes of biopsies in sampling protocols can be optimized to minimize such biases.

Results We find that in the ideal case, less than a handful of samples can be enough to infer truly clonal mutations. The frequency of the largest sub-clone at diagnosis is the main factor determining the accuracy of truncal mutation estimation in structured tumours. If the first sub-clone is dominating the tumour, higher spatial dispersion of samples and larger sample size can increase the accuracy of the estimation. In such an improved sampling scheme, fewer samples will enable the detection of truly clonal alterations with the same probability.

Conclusions Taking spatial tumour structure into account will decrease the

probability to misclassify a sub-clonal mutation as clonal and promises better informed treatment decisions.

2.2 Background

In the past years, it has become increasingly clear that cancers are typically highly heterogeneous and characterised by a large degree of spatial diversity, which complicates cancer therapy (Gerlinger et al., 2012, 2014). Modern anticancer therapies aim at targeting tumour-specific genetic and epigenetic alterations, e.g. by specifically designed molecules (Cunningham et al., 2004) or immuno-therapy (Hodi et al., 2010; Vanneman and Dranoff, 2012; McGranahan et al., 2016). The paradigmatic example has been Chronic Myeloid Leukemia that is driven by the BCR-ABL oncogene. Tyrosine kinase inhibitors (TKI) such as Imatinib can inhibit the critical gene driving the disease leading to long lasting remissions, improved survival and perhaps even cure in some patients (Savage and Antman, 2002; Druker et al., 2006; Lenaerts et al., 2010; Fujimaki et al., 2018). With rare exceptions (Long et al., 2016), this goal to date has not materialized for other tumours since in many of them, the appropriate driver mutations(s) are either unknown, not targetable or treatment resistant. Although now tumour sequencing is available commercially and mutations within tumours can be identified routinely, the clinical benefit of such therapies has been limited, since it is likely that the identified mutations are not responsible for driving the tumour in that specific patient (Le Tourneau et al., 2015), or resistance emerges fast after an initial brief treatment response (Komarova and Wodarz, 2005; Sharma et al., 2017; Syn et al., 2017; Salgia and Kulkarni, 2018; Dagogo-Jack and Shaw, 2018). It is important to note that simply because a mutation is 'common' in a specific tumour type does not make it an appropriate target of therapy. Determining which mutations are targetable is not simple for a variety of reasons including (i) the identified mutations may not be drivers in that patient, (ii) more than one driver mutation may be present, (iii) genetic and spatial heterogeneity within the tumour make it difficult to be reasonably certain that the truncal/clonal mutations that could be targeted have been properly identified, which is the main focus of our work here.

The accumulation of mutations in a growing tumour leads to the presence of cells with different mutational profiles. Classical branching models predict that mutations will be increasingly present at lower frequency (Sottoriva et al.,

2015). More specifically, late alterations are typically found in a small proportions of cells, whereas early alterations are expected to be more abundant (Williams et al., 2016). For example, mutations present in the first tumour initiating cancer cell should, in principle, be clonal and consequently found in every cancer cell of the particular patient. In clinical protocols, it is often assumed that a mutation that is present in approximately 50% of the sequencing reads of a single tumour bulk sample is likely clonal (after adjusting for tumour ploidy and tumour purity). However, this reasoning is problematic, since it requires that the underlying sample is representative of the whole tumour. Multi-region profiling shows that this is not the case (Gerlinger et al., 2012, 2014; Werner et al., 2017). These multi-region sequencing studies have revealed a much more complicated picture of severe inter- and intra-tumour genetic heterogeneity (Gerlinger et al., 2012, 2014; de Bruin et al., 2014; Rasche et al., 2017; McGranahan and Swanton, 2017). Mutations that appear clonal in a single sample can be sub-clonal or even absent in other samples of the same tumour (Jamal-Hanjani et al., 2017). Therefore, targeting such mutations would not be expected to provide long term therapeutic benefit as we would at best treat only the part of the tumour that contains these sub-clonal mutations.

Indeed, determining which are the truly clonal alterations in a neoplasm that contains billions of cells distributed in complex spatial patterns is a challenging problem that has important implications for modern cancer therapies. Limitations of sequencing depth, genetic heterogeneity within single samples, contamination with healthy tissue, and loss of genetic elements due to genome instability all complicate the classification of these alterations (Meyerson et al., 2010). In this regard, multi-region sequencing has been shown to be more informative for the discrimination of mutations than single bulk sequencing (Gerlinger et al., 2012; Sottoriva et al., 2013, 2015; Ling et al., 2015; Siegmund and Shibata, 2016; Sun et al., 2017). As multi-region sequencing of tumours has become feasible, this has led to the development of a range of phylogenetic methods and tools to construct phylogenetic trees from cancer and infer truncal (clonal) mutations (Nik-Zainal et al., 2012; Schwartz and Schaeffer, 2017).

The detailed architecture of any tumour is likely unique and driven by the

complex interactions between microevolution, the immune response as well as the presence of physical barriers to growth of the tumour population in each specific patient. The latter depend on the location of the tumour within the body. This complexity makes it difficult to reconstruct the branching process that underlies the growth of the tumour population. In the absence of such knowledge, what would be the optimal sampling approach for each individual tumour and how can we maximize our probability to identify truly clonal (and hopefully driver) mutations within these tumours. This is the focus of our work here.

We have previously shown that it is not necessary to reconstruct the complete phylogenetic tree of a tumour to estimate the probability to identify all clonal alterations correctly (Werner et al., 2017). It is easier and sufficient to identify only the earliest branching events, which then allows the detection of all truly clonal genetic alterations within individual tumours. We consider the earliest branching event that separates sub-clonal mutations from those mutations present in the ancestral population of cancer cells. We refer to these branch-defining sub-clonal mutations on the ancestral population background as first-tier mutations (Figure 2.1(d)). Taking more samples will naturally increase the chance to exclude misclassified sub-clonal mutations. However, this obviously implies a cost-benefit tradeoff and a proper understanding of the scaling of these probabilities with increasing tumour sample numbers can better inform treatment strategies.

In our previous work, we showed how the probability to correctly classify clonal mutations scales with additional samples and how many samples are needed to identify the truncal mutations with a certain level of confidence given the life histories of a tumour (Werner et al., 2017). But critically, this initial study did not consider successive branchings from the ancestral population, the spatial structures of both the tumour and the samples as well as the influence of the sizes of the biopsies taken.

Here, we quantify the probability to correctly classify the clonal mutations of individual tumours growing in space with sufficiently low mutation rates per cell division (as is for example the case for clinical targeted or exome sequencing protocols). We also show how the size of tumour samples influences our ability

to identify truly clonal alterations and how we can increase the accuracy of the detection by exclusion of low frequency mutations. We address the problem of spatially structured tumours, which can have great repercussions on clonality inferences. Finally, we compare different sampling protocols by comparing standardized spatial sampling patterns against random sampling. By applying standardized sampling patterns one can further increase the probability to correctly classify truly clonal mutations.

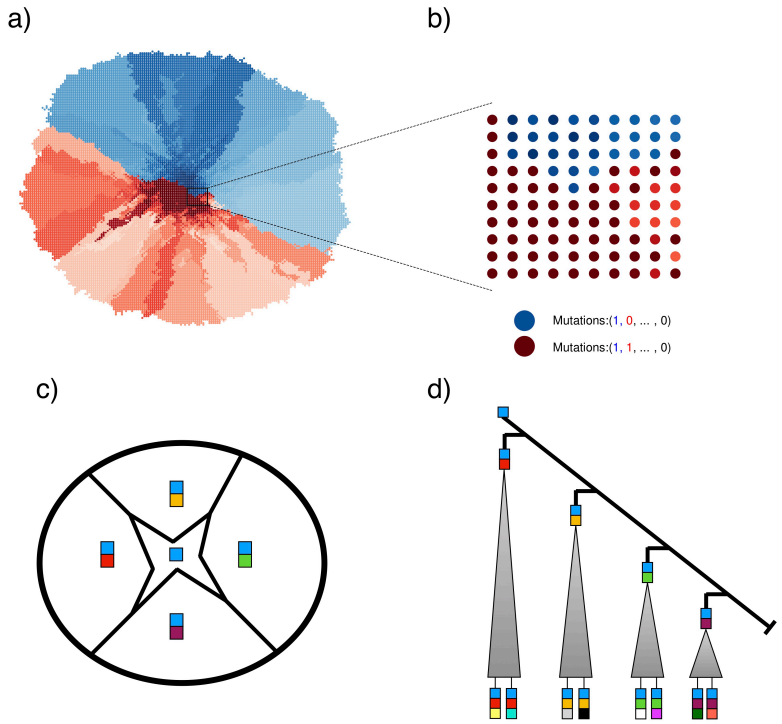

Figure 2.1: **Spatial model of intratumour heterogeneity. a)** Schematic illustration of our spatial cancer model. Tumour cells are represented as nodes on a two-dimensional lattice. Each cell has a mutational profile. With each cell division, a cell can mutate with probability $\mu = 0.5$ and thus intratumour heterogeneity is generated. While we typically think of each node as a cell, it could in principle also represent a small subpopulation of cells. **b)** The mutational profiles of cells within a bulk biopsy are then combined to give the mutational profile of the biopsy sample. Red colored subpopulations are cells that carry the first sub-clonal mutation (red in the mutational profile), while blue cells do not contain it. All cells carry one truncal mutation (blue in the mutational profile). Cells acquire new mutations and are presented with a different nuances of the original colour. In this example, because the red and blue subpopulations are approximately equal in size, we call the tumour well-balanced. **c)** The different first-tier sub-clones are spatially represented within the tumour, placed around the centrally positioned ancestral population. **d)** An ancestral population that contains a set of truncal mutations (blue square) branches multiple times into first-tier sub-clones with their own private sets of mutations. Having only samples from one of the major branches will result in a misclassification of the sub-clonal mutation that founded that branch.

2.3 Methods

2.3.1 Computational model of tumour heterogeneity:

We simulate tumours on a lattice where filled nodes represent the presence of individual tumour cells. For the structured case, the neighbouring cells are thought to represent a real – but arguably highly idealized – tumour architecture. For the unstructured case – often refereed to as the "well mixed" case – the neighborhood on the lattice has no relevance for simulations, as any cell can divide and place its offspring to any empty site. Simulations start by transforming one cell into a tumour cell through the introduction of the first mutation (in principal many mutations could be introduced, however this does not matter for our purpose here). Simulations run in discrete time steps. During each time step, one cell is chosen randomly for reproduction. Once chosen, it divides into one adjacent empty space, if the tumour is spatially structured, or in any random available place chosen from non-occupied spaces, for a well-mixed tumour. With each division, one daughter cell accumulates a novel mutation proportional with probability μ. After division, one randomly chosen tumour cell will die with a probability $d = 0.1$. To simulate a non-aggressive tumour, we have chosen a lower death rate than suggested by Waclaw et al. (Waclaw et al., 2015) ($d = 0.5b$) for the simulation of highly aggressive tumours.

Computationally, the tumour is represented by a sparse matrix, wherein the position of a cell, the ID of its parent cell and the signature identifier of each new mutation is stored. This information allows us to reconstruct the mutational profiles of any cell at any given time point. We assume that each mutation can arise only once during division and can only be lost when the a cell dies (corresponding to the infinite allele assumption). Moreover, we assume all mutations to be neutral – they do not affect the fitness of the carrier cell. Our assumption of neutrality should not impact the generalizability of our results. After a full sub-clonal sweep, the dominant sub-clone would appear as ancestral population, thus thus leading to a tumour population with similar underlying branching structure. The nature of our simulation makes the structured tumour grow mostly at its periphery. Once the centre of the

tumour is densely populated cells can only divide if neighbouring space becomes available after a random cell death. This pattern is supported by observations of similar peripheral growth patterns in some real tumours (Lloyd et al., 2016).

In our analysis only the presence of new mutations is important and not the number of new mutations in each cell. We therefore assume that during each division daughter cells receive a new mutation with probability $\mu = 0.5$. This assumption is supported by the estimated high mutation rates in neoplasms. With effective mutation rates (mutation per surviving lineage) of up to 10^{-7} mutations per base per division, we can expect a mutation occurring during almost every division within the exome of cancer cell (Williams et al., 2016). With this mutation rate we achieve early branching and extensive intra-tumour heterogeneity from the early stages of tumour growth. This feature of the model is compatible with the fact that only early mutations are likely to spread sufficiently to be detected by Next Generation Sequencing (Sun et al., 2017). Early branching provides a broad spectrum of sizes of the first sub-clone in our simulations due to the stochastic nature of sub-clonal growth and mutation accumulation. At the end of the simulation we calculate the frequency of each mutation within the tumour. We specifically denote frequencies of first tier sub-clonal mutations and the most frequent sub-clonal mutation. In the evolutionary history of the tumour, we define the first-tier branch as the subpopulation of cells that diverged directly from the ancestral population of tumour cells. To reconstruct the truncal mutations from multi-region tumour sampling, we need to either sample from two different first-tier subpopulations that emerged from the ancestral cancer population or from one first-tier subpopulation and the ancestral subpopulation, because sampling from the same first-tier branch will falsely identify branch-defining mutations as clonal mutations.

2.3.2 Sampling and clonality inference

A single simulated biopsy is composed of a group of cells in close proximity (Figure 2.4), or a single cell (Figure 2.2) initially taken from a random location of the lattice. In a well-mixed tumour, due to the absence of spatial structure, a

sample is a number of randomly pooled cells (Figure 2.4), or a single cell (Figure 2.2). We reconstruct the mutational profiles of the sampled cells and calculate the frequencies of the mutations within each sample. As we are unable to detect low frequency mutations with current sequencing technology (sequencing depth threshold), we vary the threshold ε to detect a mutation within the sample. Mutations that appear clonal across a tumour are those mutations present in all taken samples. However, in our simulations we know the ground truth and we can test how often these mutations actually represent truly clonal mutations present in the first cancer initiating cell. If no mutations were wrongly classified as clonal we mark our sampling as correct. Otherwise, if there is at least one sub-clonal mutation misclassified as clonal, we consider our sampling incorrect. To get the proportion of correct estimations for single tumours, we repeat the sampling process 10 000 times with n samples (shown as dots in Figures 2.2, 2.4).

For pattern sampling we chose four single-cell samples from the tumour edge located at opposite directions and assessed clonality as previously described. To calculate the proportion of correct estimations for each individual tumour using our sampling pattern, we rotated the samples stepwise and assessed clonality on each step until we covered the whole tumour circumference (Figure 2.6(a)). Rotations of the samples allowed us to make multiple repetitions of sampling on a single tumour using the same pattern. The proportion of correct estimations was then compared with random single-cell sampling and our mathematical model quantification.

2.4 Mathematical model

Let us first consider a simple model with only a single bifurcation representing the entire phylogenetic tree of the tumour. This bifurcation generates a branching subpopulations of cells that diverged directly from the ancestral population of initiating tumour cells. This branch contains a new sub-clonal mutation compared to the ancestral population. Here we define a balancing factor f as the proportion of the subpopulation within this branch, while the proportion of the other branch of the ancestral population is $1 - f$. If we take n independent

samples at random, the probability $p_f(n)$ of finding the true clonal (truncal) alteration is the probability that not all n samples come from the branch with the new sub-clonal mutation, in our case this is

$$p_f(n) = 1 - f^n. \tag{2.1}$$

We now generalize the expression of $p_f(n)$ for a phylogenetic tree with a large number of bifurcations. Among all bifurcations, we are specially interested in the branches that diverged directly from the ancestral population. These branches are defined as first-tier branches, each of which contains a distinct sub-clonal mutation compared to the ancestral population (Figure 2.1d). Within each of these first-tier subpopulations, subsequent sub-clonal mutations would happen constantly (Figure 2.1d). However, these subsequent events are unnecessary for identifying the truly truncal alteration, so in what follows we ignore them and focus on the first-tier branches. We assume that the mutation rate of cells in the ancestral population does not change over time, so the first-tier branching mutations arrive after equidistant intervals. As a result, the balance factor f is unchanged and applies to all the first-tier branching mutations. Suppose there are M first-tier branches, which are ordered by their time of occurrence. The proportion of cells in the kth first-tier subpopulation is f times $(1-f)^{k-1}$ (the fraction of cells that did not carry any of the previous $k-1$ first-tier sub-clonal mutations). In this way, the proportions of cells in these first-tier branches are given by f, $(1-f)f$, $(1-f)^2 f$,..., $(1-f)^{M-1}f$. To identify the true truncal alteration with n independent samples, these samples should not come from one single first-tier subpopulation. Thus, the probability $p_f(n)$ is given by

$$\begin{aligned} p_f(n) &= 1 - f^n - [(1-f)f]^n - [(1-f)^2 f]^n - \\ &\quad \ldots - [(1-f)^{M-2} f]^n - [(1-f)^{M-1} f]^n. \\ &= 1 - f^n \sum_{i=0}^{M-1} (1-f)^{ni} \end{aligned} \tag{2.2}$$

If M is sufficiently large, the geometric series can be used to approximate $p_f(n)$

by the simplified expression

$$p_f(n) \approx \frac{1 - f^n - (1-f)^n}{1 - (1-f)^n}. \tag{2.3}$$

The estimation of f for specific tumour from multi-region sequencing data has been The parameter f, necessary for the calculation of the probability $p_f(n)$ can be estimated from data in the following way: The idea is to compare the sets of clonal mutations identified by all permutations of tumour samples. More specifically, first the intersection of all alterations of all n biopsy samples has to be determined. After that, the intersection of all possible combinations of $i = 2$ biopsy samples is generated. The frequency at which both intersections coincide is an estimate for the probability of a correct classification, $p_f(i)/p_f(n)$. The same procedure is performed for all possible combinations of $i = 3, 4, ...n - 1$ biopsy samples. Using these probabilities, one can estimate f for a given cancer by fitting the estimated probabilities to

$$\frac{p_f(i)}{p_f(n)} = \frac{1 - f^i - (1-f)^i}{1 - (1-f)^i} \frac{1 - (1-f)^n}{1 - f^n - (1-f)^n}. \tag{2.4}$$

2.5 Results

2.5.1 Expectation from our mathematical model

In a highly homogeneous cancer with low mutational burden, even the largest sub-clonal mutation is only present in a small proportion of the tumour. Thus, our mathematical model predicts a fairly small chance to get false positive clonal mutations, see Fig. 2.2a with $f = 0.1$. Already with $n = 3$ samples the probability to correctly classify truly clonal mutations is $> 98\%$. For the case where the first branching event leads to a tumour with two roughly equally-sized populations $f = 0.5$ (both of which will carry a tremendous amount of private mutations and many subsequent branchings) we reach a probability of $> 98\%$ already with $n = 6$ samples. Finally, in tumours where specific sub-clonal mutations undergo great expansion, it is highly likely that this expanding mutation and its sub-clonal mutations will be categorized as clonal (Fig. 2.2a with $f = 0.9$). This is because with increasing f it becomes less probable to sample from the part of the cancer without that abundant sub-clonal mutation. To reach the same level of confidence $> 98\%$, as in shown tumours with lower f, we need > 38 samples for $f = 0.9$.

2.5.2 Validation of the mathematical model using single cell sampling in simulated tumours

The proportion of the largest sub-clonal mutation f has a great effect on the clonality analysis. The probability to correctly classify clonal mutations with $n = 2$, $n = 5$ and $n = 10$ independent samples changes substantially with the value of f (Figure 2.2 b.) In principle, it is possible to correctly estimate clonality with only two samples, in particular if the largest sub-clone is sufficiently small. Theoretically, two samples give a correct estimation with probability $> 98\%$ for tumours where the proportion of the largest sub-clonal mutation is below $f = 0.1$. However, this probability drops rapidly with increasing f. Using more samples can substantially increase the probability of correct clonality assessments. With $n = 10$ samples, we cover most of the range of f and only for $f > 0.8$, our estimates become less reliable. For most values of f, 10

random tumour samples are sufficient to reach a high probability of a correct clonality assessment. We validate our mathematical model by comparing it with the results from stochastic spatial simulations of cancer growth (Figure 2.2b for $n = 2, 5, 10$.). Each point represents the proportion of correct estimations of clonality inferred from 10 000 iterations of n independent and random samples from a single tumour, in which the proportion of the largest sub-clonal mutation is f. Results obtained from simulations are in good agreement with our theoretical expectation, in particular results with more than two samples show almost perfect agreement between the mathematical model and the simulated tumours, despite the spatial correlations between clones arising in our computational model – which become crucial if we sample more than one cell.

2.5.3 Effect of the sample size on clonality estimation

Biological samples from tumours typically contain a variety of different subpopulations of cancer cells, healthy surrounding and 'supporting' tissue as well as leukocytes that infiltrate the tumour. All of these cells can influence the interpretation of the sequencing data and the correct assignment of mutations. In current clinical applications, sequencing a group of cancer cells is standard – with single-cell genomic profiling so far an approach of the future. For that reason, we investigated the clonality inference by using multiple large samples, each containing 1% of the total number of cells in the tumor. In the previous analysis of single cell samples, there was no mutational frequency component – every mutation was of equal value for the clonality estimation. With multiple cells, we gain additional information of the frequency for each mutation within the sample.

2.5.3.1 Well-mixed tumours

As we previously stated, in our analysis a sub-clonal mutation is misclassified as clonal if it is present in all samples, therefore to classify it correctly there should be at least one sample where that mutation is absent. In a well-mixed tumour, mutational frequencies within large single samples already represent the spectrum of frequencies within the whole tumour. That makes them un-

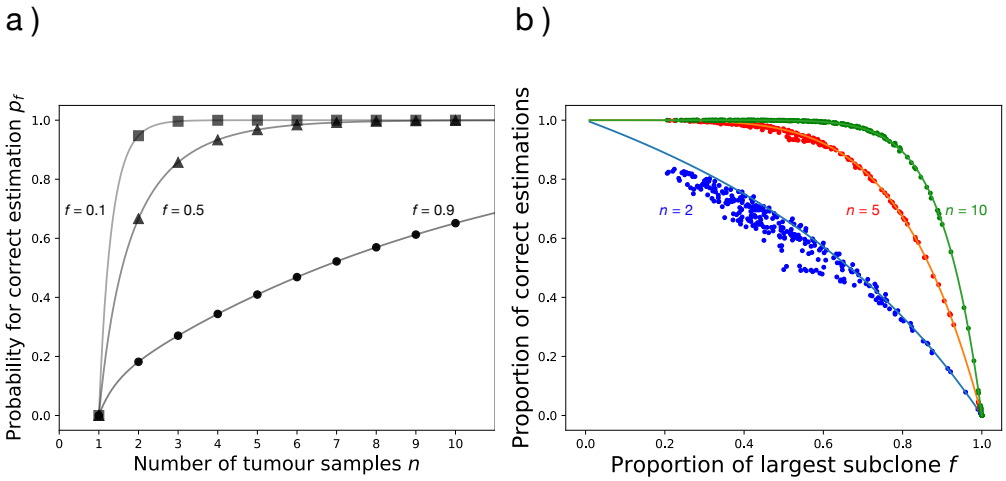

Figure 2.2: **Comparison of clonality inferences in structured and unstructured models of tumours with a different proportion of the largest sub-clone. a)** The probability to correctly identify set of truly clonal mutations with n tumour samples in our model. In tumours where the size of the largest sub-clone f is small ($f = 0.1$) the probability to correctly identify truly clonal mutations is already sufficiently ($> 98\%$) high with two samples. In balanced tumours with $f = 0.5$, five samples give the same probability. **b)** The quality of our clonality estimation in dependence of the proportion of the first sub-clone. Lines represent solutions of the mathematical model, while dots represent results from clonality inferences of spatial computer simulations. A number of randomly distributed single-cell samples n ($n = 2$ shown in blue , $n = 5$ shown in red and $n = 10$ shown in green) was taken from each simulated tumour and clonality of present mutations was estimated. Each single dot represents the proportion of correct estimations for one tumour by sampling n tumour samples after 10 000 repetitions. Results from simulations are in agreement with model predictions for the full range of f.

usable for the classification of truly clonal mutations from multiple samples by means of exclusion from multiple sampling, as a large number of mutations appear clonal if we do not consider the frequency of each mutation within the sample. We can reduce the number of candidate mutations by excluding mutations with frequency below a certain threshold ε. By doing so, we remove sub-clonal mutations with low frequencies from the analysis and get a high proportion of correct estimates (Figure 2.3). Bringing the threshold ε above the frequency of the most abundant sub-clonal mutation (f) leads to a cor-

rect clonality assessment regardless of the number of samples. However, such clear demarcation likely is the result of our idealized scenario. In reality copy number changes and limited sequencing depth shift frequencies of mutations and introduce additional errors, leaving some uncertainty for the minimal list of clonal mutations.

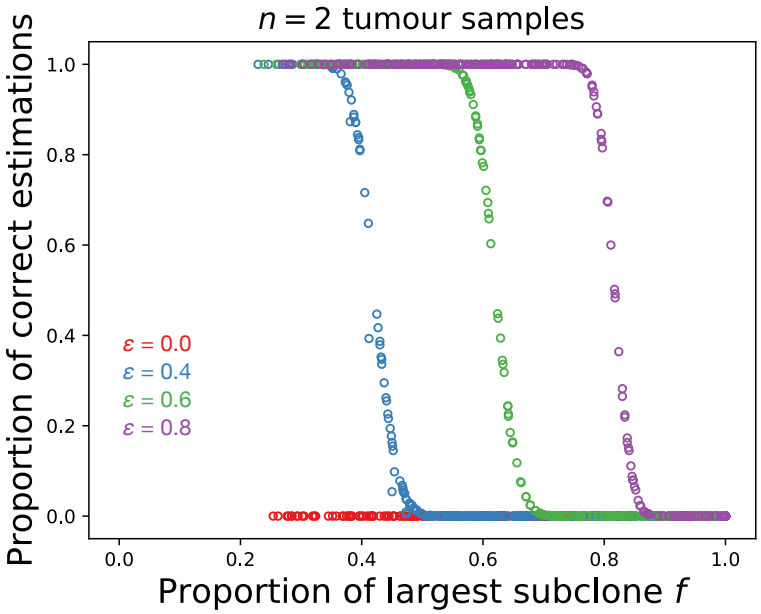

Figure 2.3: **Effect of sample size on clonality inference in well-mixed tumour.** The proportion of correct clonality estimates for $n = 2$ samples, each containing 1% of total number of cancer cells. ε represents the frequency bellow which we discard mutations from the clonality analysis. The value of ε where identification of clonal mutations becomes impossible corresponds to the proportion of the largest sub-clone f, as each sample is representative of the whole tumour.

2.5.3.2 Structured tumours

In structured tumours the clonality inference with multiple large samples is less accurate than the equivalent single-cell sample when including all mutations in the analysis, even though with large samples much more cells are included in the analysis than with single-cell sampling (Figure 2.4 for $\varepsilon = 0$). Large

samples (1% of tumour size) contain more sub-clonal mutations that might be considered clonal. Having the frequency of each mutation within the sample, we can consider all mutations present at sufficiently low frequency within the sample as sub-clonal. By doing so we stop considering all mutations present in every sample as clonal. By raising the mutation detection threshold ε, low frequency sub-clonal mutations are removed from the analysis which increases the accuracy of our classification, ultimately surpassing the probabilities predicted by our model for single-cell sampling (Figure 2.4). We would get the best results by considering only mutations that appear clonal within the sample. Yet, some of clonal mutations might appear sub-clonal within the samples due to contamination with healthy tissue, copy number variation or sequencing noise, and would be wrongly excluded from the analysis.

To further test our approach, we repeated the same inference on a spatial model of tumour growth originally developed by Waclaw et al. (Waclaw et al., 2015), shown in Fig. 2.5. This model is very different from ours not only in the dimensionality, but also in many other details of the stochastic implementation. For example, the model is not based on a spatial lattice, allowing more complex configurations of cells in space. Nonetheless, the results on this three-dimensional model show the same qualitative features we observed in our two-dimensional scenario. The structure of the tumour has the same effect on the probability to correctly detect clones, furthermore both size of the samples and removal of sub-clonal mutations within samples from the analysis are showing similar trends as in our original computational implementation.

2.5.4 Clonality inference using non-random spatial sampling

Until now, we only considered a random sampling process. In reality, this approach is not applicable and it would be useful to have clearly defined spatial relations between individual samples. Thus, we now compare the clonality inference using four samples arranged in a circular spatial pattern (Figure 2.6a) against four randomly distributed samples. We use samples from the (circular) edge of the tumour with the greatest distance between samples. In order to

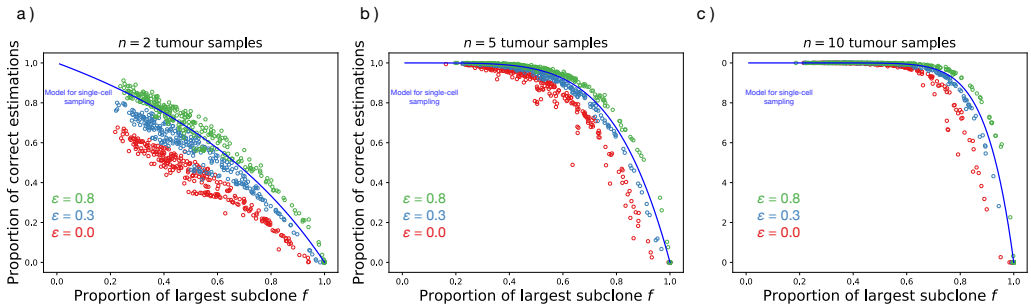

Figure 2.4: **Effect of sample size on clonality inference in spatial tumour.** To test the effect of biopsy size, we tested the accuracy of clonality estimations by sampling groups of cells (one sample = 200 cells, corresponding to 1% of tumour size). Batches of tumour samples (two for a), five for b) and ten for c)) were taken at random locations and used to estimate the clonality of the present mutations. We repeated sampling process 10 000 times for on each tumour and calculated proportion of correct estimations – single point on figure. We exclude mutations below a certain frequency ε from the analysis ($\varepsilon = 0.0$ or no exclusion for red, $\varepsilon = 0.3$ for blue, $\varepsilon = 0.8$ for green), which increases the accuracy of clonality estimations.

calculate the probability for a correct estimation of clonal mutations for a single tumour, we repeat the sampling process on the same tumour after we rotate all samples (Figure 2.6a) while maintaining the distance between them. We find that by sampling in this pattern, we increase the probability to correctly classify clonal mutations in tumours across most range of sub-clone proportions, f, compared to random sampling (Figure 2.6b). This is especially pronounced for intermediate number of samples. Interestingly, for $n > 1/(1-f)$, the classification of clonal mutations is correct almost with certainty. Only in cases where the proportion of the largest sub-clone f is close to 1, random sampling can be superior to pattern sampling.

These results can be generated to any number of samples positioned in an equidistant pattern. If we keep the same distance between samples, to have at least one sample not containing the largest sub-clonal mutation $(1-f)$, on average we need to sample $n = 1/(1-f)$ samples. This translates to at least 10 samples needed for a correct classification of clonal mutations for tumours with $f > 0.9$.

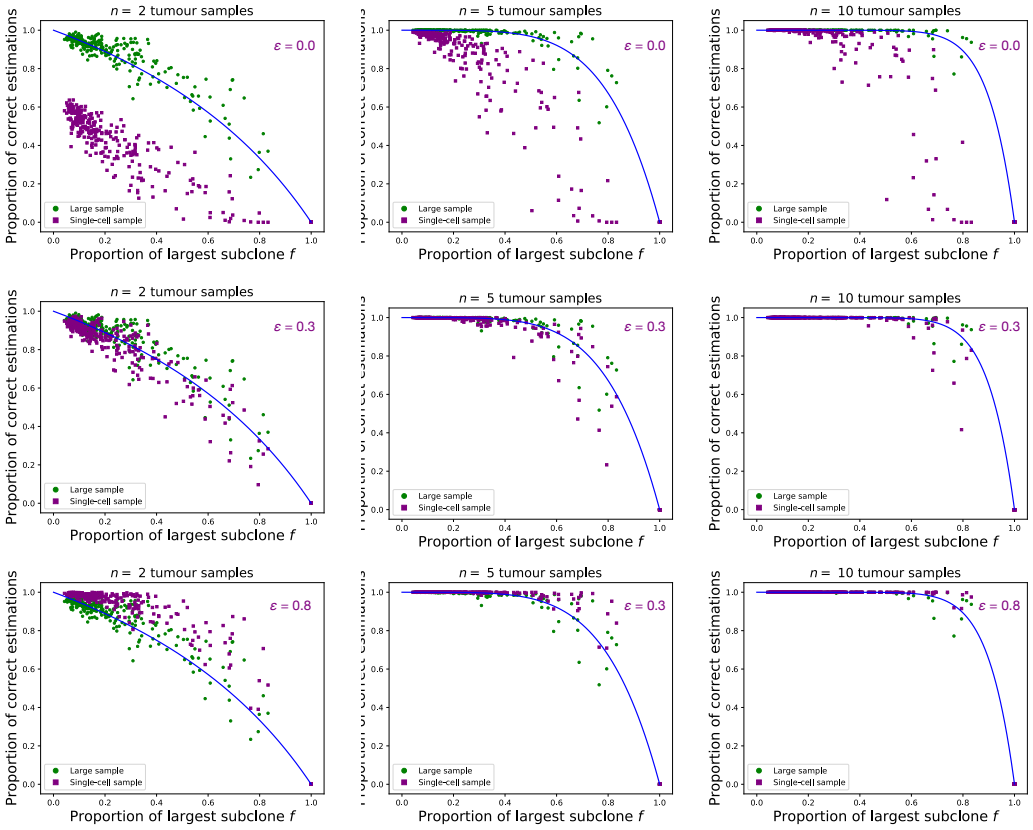

Figure 2.5: **Clonality inference on the three-dimensional, spatial tumour model.** To test the robustness of our results, we repeated the clonality inference process in the previously published spatial cancer model of Waclaw et al. (Waclaw et al., 2015), which is very different from our model. Small biopsies (green; one biopsy = 1 node) have a much greater probability to correctly classify clonal mutations than large biopsies (purple; one biopsy =8% of the total tumour size) if we include all mutations present in each sample $\varepsilon = 0.0$. As we increase the threshold of the mutation frequency ε (middle panel $\varepsilon = 0.3$, bottom panel $\varepsilon = 0.8$), the accuracy of larger samples is increasing, and goes beyond single-cell samples and model predictions, same as in our simpler model. model. Number of tumours = 300 with maximum size of $5 \cdot 10^6$ cells were simulated with a death rate of $d = 0.8$. Mutation rate $\gamma = 0.02$ mutations per division.

To test the possibility that improvement in classification is caused by specific (unknown) properties of cells from the edge of tumour and not due to pattern sampling, we took a series of random samples from the edge of the tumour and

estimated clonality. These results match those from purely random sampling.

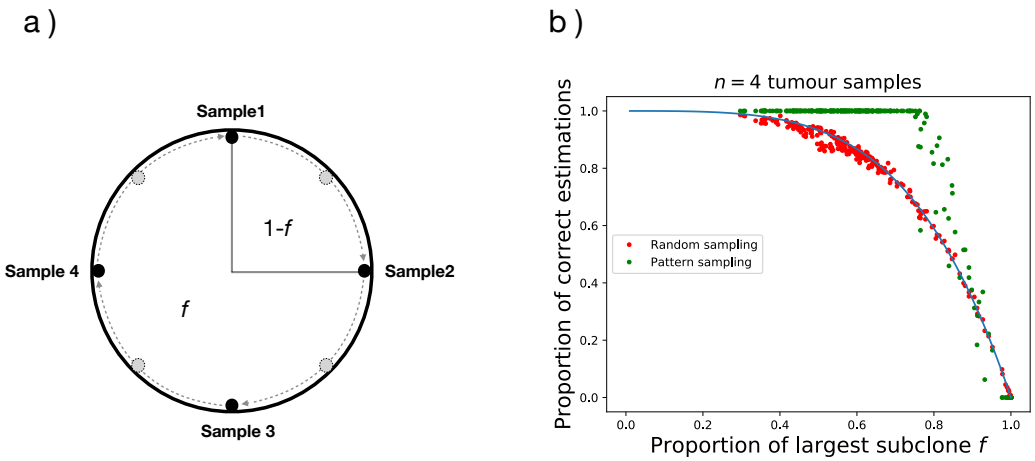

Figure 2.6: **Pattern spatial sampling and identification of clonal mutations.** We compared our model model (blue line) to simulation results using random single-cell sampling (red points) and to simulation results using sampling in specific pattern (blue points). Sampling pattern in panel a). Each red point represents the proportion of correct estimates with $n = 4$ samples after 10 000 sampling repetitions. For the green points, the template pattern was rotated to obtain multiple sampling repetitions from each tumour, while maintaining the distance between the cells and the peripheral location of cells. The proportion of correct estimations was calculated from a number of possible sampling repetitions for each tumour. Sampling in pattern appears superior to random sampling throughout most of the range of proportion of largest subclone f (Fig b)).

2.6 Discussion

Targeting "driver mutations" in cancer is considered an important new approach to therapy that takes into consideration a varied mutational landscape present in tumours, even when arising in the same tissue (Bailey et al., 2018; Sanchez-Vega et al., 2018; Groisberg et al., 2018). The paradigm set by TKI therapy

of CML and ALK inhibition in a small subset of patients with non-small cell lung cancer is quite compelling. In these two tumours, the mechanistic understanding of how the mutation drives the tumour is clear and therefore the term "driver" mutation is justified. Similarly, c-Kit expression on gastrointestinal stromal tumours (GIST) renders these tumours sensitive to imatinib as are the rare cases of mastocytosis with eosinophilia due to PDGFRA expression or mutant c-Kit expression (Gotlib, 2017; Evans et al., 2017; Dufresne et al., 2018). $BRAF^{V600E}$ mutations in malignant melanoma render the cells sensitive to vemurafenib (Chapman et al., 2011). Sequencing of other rare tumours has also led to the discovery of mutations that can be meaningfully targeted (Drilon et al., 2013). However, in the majority of cases, the identification of a mutation by itself does not imply that it is a driver – even if this was shown to be the case in a similar tumour and it is present in a significant fraction of the sample. Clonality needs to be proven with a reasonable certainty if there is any hope that targeted therapy will be effective. Sequencing a single sample and inferring that a mutation is "actionable" is fraught with problems, since the sample of the tumour sequenced may not be representative of the whole tumour, and in addition sampling has also to contend with the problems of false positive and negatives, high background noise due to the potential presence of not fully malignant cells that may still harbor normal copies of important genes such as TP53 (Krimmel et al., 2016) as well as contamination with normal tissue. It is therefore not surprising that despite major efforts, the practical benefit of NGS sequencing for the individual patient to date has been limited. A recent example illustrates this case: In a series of 95 patients with cancer seen at MD Anderson Cancer Center, NGS sequencing identified at least one mutation in 92 % of patients. The most common were in TP53 (25%) and KRAS (10%). In principle, 36% of the tumours sequenced had an actionable mutation and 13 patients received therapy based on this sequencing data to target the presumed driver mutation. Four patients had a partial response, six had stable disease while three progressed (Groisberg et al., 2018). It is difficult to justify the current clinical approach with these results. Proving that a mutation is truncal and therefore clonal should lead to better identification of driver mutations and proper targeting of such mutations is more likely to give meaningful

results. It appears that a proper sampling strategy for multi-region sequencing of a tumour is a key component in the process for the correct identification of truly clonal mutations. Such a list that is developed for every unique patient sequenced will likely be enriched for 'driver' mutations. In this work, we discuss how to improve the strategy determining this list of clonal mutations with a high level of certainty. The future introduction of multi-region tumour profiling into clinical practice requires a better understanding of the underlying mechanisms of intratumour heterogeneity and a more standardized approach to tumour sampling. We are still unable to steer the biological processes within a tumour to affect its heterogeneity (Nichol et al., 2015) , but we can optimize the way we collect and analyse tumour samples.

Our study provides insights into both intrinsic and extrinsic factors that influence the probability to detect truly clonal mutations. As the construction of the complete tumour phylogeny is not necessary for clonality inference, we focused on the reconstruction of the series of branching events coming from an ancestral population. We have developed a mathematical model for the calculation of the probability for correct identification of truly clonal mutations from multi-region sampling of cancer with a large number of bifurcations. In that process, the largest sub-clone is the most relevant factor in the identification of truly clonal mutations. Its proportion is a consequence of the time since the emergence of the first sub-clone. The earlier the first sub-clonal mutation occurs, the more likely we are to misclassify it as truncal. A large abundance of this first sub-clonal mutation requires more samples to ensure that at least one sample does NOT contain that mutation. In addition, if the first sub-clone is only present in a small proportion of the tumour, there is a low probability of it being misclassified as clonal. Our results show that considering multiple branching events we now see that the probability to correctly classify mutations is much greater than previously thought (Werner et al., 2017).

In solid neoplasms, where cells grow in space, we have shown that larger samples are more likely to overestimate clonality of some mutations if the analysis include all mutations present in each sample. It is necessary to exclude mutations present in low and medium frequency from the analysis. Doing so we not just reach the same accuracy as we would by single-cell sampling, but

we even increase it substantially. However, exclusion of lower frequency mutations from the analysis might cause false negative classification of some clonal mutations whose frequencies are variable within a sample due to contamination with healthy tissue, duplication of genetic material within some cancer cells, or sequencing error. In the presence of other detected clonal mutations, a false negative error is of less concern than a false positive, as false positives would deprive the patient of effective therapy. Decision on the cutoff for the mutation exclusion must be individually chosen based on the number of available clonal mutation candidates.

Finally, we bring a theoretical rationale for sampling in non-random patterns. We showed that placing biopsies in pattern equally distant from each other, might substantially increase the probability to correctly classify truly clonal mutations when compared with random sampling. We are aware that our computational simulations of tumour growth are simplified and lack many features of living systems, such as cell migration in tumours with more complex growth pattern. There, a different spatial sampling strategy might be more successful. Our results provide a rationale for pathologists when taking samples for multi region tumour sequencing and clinicians during endoscopic sampling of neoplasms of e.g. gastrointestinal tract. By choosing samples with maximum spread in suggested pattern one maximizes the chances to correctly classify clonal mutations. We also offer a way to estimate a level of certainty for a list of detected clonal mutations which can serve as a guidance to oncologists in their choice of appropriate target. Our results provide some considerations for the improved clinical assessment of targetable mutations in treatment naive tumours.

2.7 Conclusions

In conclusion, the correct classification of clonal (truncal) mutations is of great importance for the success of anti-tumour therapy. We have shown how the probability to identify truly clonal mutations depends on sub-clonal composition of tumour and how many samples one must take to be able to discern mutation clonality with great confidence. Furthermore using a computational

model of cancer heterogeneity, we have shown how the size of biopsies affects the probability to correctly identify clonal mutations. Finally, we showed that our suggested spatial sampling pattern is superior to random sampling.

2.8 Abbreviations

TKI: Tyrosine kinase inhibitors
GIST: Gastrointestinal stromal tumour
CML: Chronic myeloid leukemia
ALK: Anaplastic lymphoma kinase
NGS: Next generation sequencing

2.9 Keywords

Intratumour heterogeneity
Clonal mutations
Spatial model
Truncal mutations
Targeted therapy
Somatic evolution

2.10 Declarations

Ethics approval and consent to participate

This work does not involve human subjects, including human material or human data, therefore no approval or consent was required.

Availability of data and materials

Code available from corresponding author on request.

Competing interests

The authors declare that they have no competing interests.

Funding

B.W. is supported by the Geoffrey W. Lewis Post-Doctoral Training Fellowship.

2.10. Declarations

D.Z. is supported by China Scholarship Council Grant (No.201806315038), the Fundamental Research Funds for the Central Universities in China (Grant No. 20720180005) and Max-Planck stipend. The funding bodies did not have a role in any aspect of this study.

Author's contributions

All authors developed the concept, L.O. implemented simulations and analysed the model with input from all authors. L.O. and A.T. wrote the paper with input from D.D., B.W and D.Z. . All authors read and approved the final manuscript.

Acknowledgments

We thank Bartlomiej Waclaw for inspiring discussions and explanations on his model and Jordi Arranz and Marvin Böttcher for programming support.

Chapter 3

Application of theoretical models on data from multi-region genomic profiling of gastric cancer

---

This Chapter is a part of a joint collaboration project led by Prof. Dr. med. Christoph Roecken from the Department of Pathology of the Christian-Albrechts-University in Kiel, Germany. Results presented in this thesis are included in the draft of a joint manuscript currently prepared for submission for peer review.

3.1 Abstract

In this chapter, we apply the theoretical model for identification of truly clonal alterations from multi-region genomic data, presented in the previous chapter of this thesis, on data from a gastric adenocarcinoma study cohort. We estimated the robustness of clonality analysis using an already existing number of samples to find that for 3/9 patients we can be fairly sure that mutations classified as clonal are truly clonal. For the remaining six patients, more sampling is required to reach high confidence of the clonality estimation, as additional samples would likely truncate the existing list of mutations classified as clonal. Further, we tested whether different parts of a tumour exhibit different modes of evolution by comparing the distribution of variants allele frequencies from each sample to the neutral model of cancer growth. We found variability in distributions of variant allele frequencies in samples within the same tumour. These variabilities lead to various degrees of heterogeneity in respect to agreement with the neutral model. Our estimation of robustness of the clonality analysis holds a direct

clinical value as it assigns a confidence value to the list of mutations considered clonal in the tumour of each patient. Our method identified patients who may likely benefit from sequencing of additional samples. We suggest that decisions regarding the target of choice for targeted therapy should take into account the number of samples in order to maximize therapeutic potential.

3.2 Introduction

In recent years, the field of cancer genomics is continuously producing large amounts of data that describe changes in genomic structure and function of cancer cells. Despite an accumulation of knowledge, quantitative methods for understanding generated cancer data are often not established (Gatenby and Maini, 2003). For that reason, mathematical modelling has recently become common in cancer research as it proved itself to be useful in providing explanation of the observed data (Altrock et al., 2015b). First time mathematical modelling was used to explain observed cancer data goes back to the mid-20^{th} century. By examining age-specific mortality rates it was suggested that cancer is the result of successive cellular changes (Armitage and Doll, 1954, 1957). Recently, mathematical modelling has allowed insight into the early development of malignancies during the time when cancer is inaccessible to direct observation (Sottoriva et al., 2015; Hu et al., 2019). Even though the number of research efforts that include mathematical models in the study of cancer is steadily growing, the translation of theoretical research to the clinic is still difficult. Currently obtainable data are often not appropriate for coupling with mathematical models as they commonly lack important features such as temporal and spatial information (Anderson and Maini, 2018). There are promising clinical studies based on theoretical models of cancer evolution which actively affect the treatment schedule and dosage for patients with metastatic castrate-resistant prostate cancer (Zhang et al., 2017). However, these dynamic approaches to therapy require consistent feedback on the state of a tumour and its composition which can be done using liquid biopsy of circulating tumour DNA (Stanková et al., 2019). In this chapter we try to bridge the gap between theoretical models and their clinical utility by applying previously developed

models to patient-derived cancer data.

In his seminal publications, Gerlinger et al. (2012, 2014) used the multi-region genomic profiling of renal cancer to show that different parts of the tumour carry different mutations and quantified the extent of intratumour genetic heterogeneity. Genetic heterogeneity is a consequence of accumulation of mutations within the cells over time (Nowell, 1976). This occurs naturally in every somatic cell due to the imperfection of DNA replication and DNA repair mechanisms. In healthy cells, fidelity of DNA replication is high with less than one mistake for every 10^9 nucleotides added (McCulloch and Kunkel, 2008). If DNA repair apparatus is impaired, cells accumulate more mutations than normal, which greatly increases the chances of malignant alteration. Impairment of these mechanisms can be hereditary as in xeroderma pigmentosum, Lynch syndrome or Muir–Torre syndrome. Additionally, random somatic mutation of the gene responsible for DNA repair can make healthy cells hypermutable and thus more susceptible to malignancy (Norgauer et al., 2003); (Ang et al., 2011); (Sehgal et al., 2014; Martincorena and Campbell, 2015). Knowledge about shared, clonal, mutations is imperative for effective treatment, as modern cancer therapy targets tumour-specific alterations (McGranahan et al., 2016). To achieve maximum effectiveness, the therapeutic target must ideally be clonal – present in every cancer cell. Otherwise, a part of the tumour would be innately unresponsive to the therapy. Sub-clonal mutations, within a tumour, serve as a reservoir of mutations from which resistance mutations are drafted. They are the driving force behind the evolutionary force of adaptation and awareness of the presence of resistance mutations is crucial for the emerging field of adaptive therapy. Furthermore, sub-clonal mutations can give us insight into the evolutionary history of tumours (Ling et al., 2015). Information on the frequencies of sub-clonal mutations present in the tumour can be used to distinguish tumours driven by positive sub-clonal selection from neutrally growing tumours (Williams et al., 2016; Sun et al., 2017; Williams et al., 2018). Here, we apply cancer genomic data on to theoretical models to gain valuable information about both clonal and sub-clonal mutations in the context of cancer evolution from gastric cancer genomic data. In the previous chapter of this thesis, we presented a model for calculating the sample number necessary to identify

clonal mutations from multi-region sampling of spatial tumours. Here, we use a previously presented theoretical model to estimate the number of samples necessary for the identification of truly clonal genetic alterations from gastric cancer data. We measured the robustness of existing clonality analysis for each patient using the currently available number of samples. We also investigated variability in the mode of evolution among different samples of individual tumours. By examination and comparison of variant allelic frequencies between individual samples within the tumour, we were able to gain insight into evolutionary dynamics of the tumour on a regional level.

3.3 Materials and methods

3.3.1 Study cohort and dataset

The dataset was produced and shared by Prof.Dr. Roecken from the Department of Pathology of the University Hospital Schleswig-Holstein, Kiel, Germany. The study included a total of 9 patients with gastric cancer who underwent surgical resection of their primary disease (see Figure 3.1. Resected primary tumours > 3 cm in diameter were histologically classified and subjected to multi-region sampling using a core needle biopsy. The number of harvested tumour samples for each patient ranged from four to ten samples. Matched adjacent normal tissue was sequenced to provide a reference for the analysis of tumour samples. Whole exome sequencing for each individual sample was done with average sequencing coverage for both tumour and normal tissue of $\sim 115x$. The exome was annotated using Annovar (Wang et al., 2010) and variant calling was done using VarDict software (Lai et al., 2016) by our collaborators from Avera Cancer Institute Center for Precision Oncology in Sioux Falls, S.D., USA. The provided dataset contained name, class, position on the chromosome and frequencies of allelic variants in each sequenced tumour sample. To calculate the variant allele frequencies in the whole tumour, allelic frequencies from individual patients were pooled by combining the number of detected variants from sequencing data of each sample.

3.3.2 Clonality

Our theoretical framework for the measurement of the robustness of the clonality analysis is described in the previous chapter of this thesis and in (Werner et al., 2017; Opasic et al., 2019). In contrast with clonality estimations on simulated tumours in the previous chapter we used information from phylogenetic trees our collaborators reconstructed using LICHeE software (Popic et al., 2015) (Supplementary figure 6.1). Since our method showed great sensitivity to false-negative variant calls, we decided to perform a clonality estimation using only branch-defining mutation events and not including all detected mutations in each sample. Additionally, we decided to use branch defining alterations as the

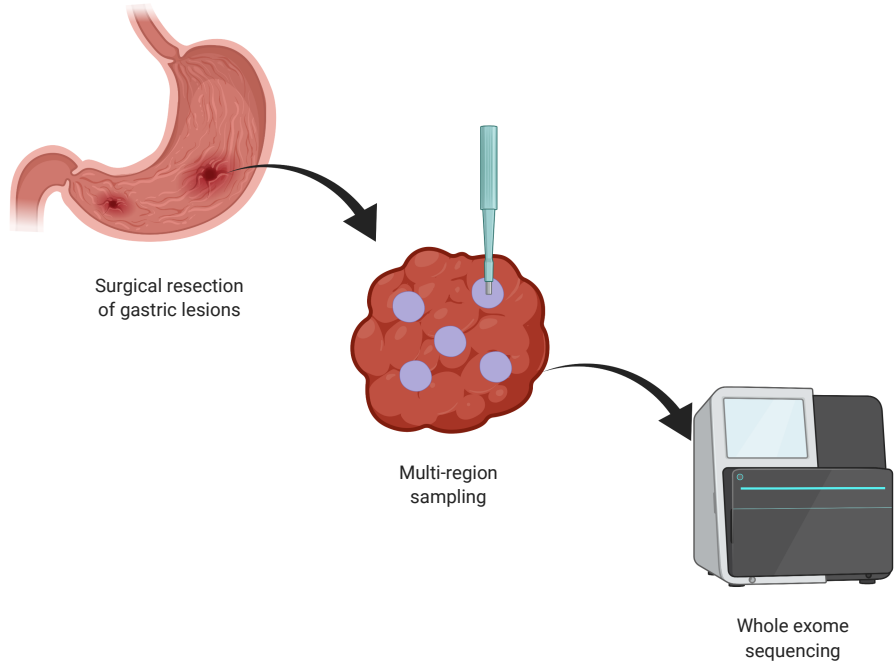

Figure 3.1: Patients with gastric adenocarcinoma underwent surgical resection of their primary disease. A number of biopsies was sampled and sequenced. Figure created with BioRender.com

sample itself is heterogeneous, composed of many different sub-clones, which can interfere with our analysis (Alves et al., 2017). Finally, it appears that the majority of sub-clonal mutations in treatment naive tumours are neutral as alterations that drive tumour progression are mostly clonal (Reiter et al., 2019). Therefore, we focused largely on more significant branch defining mutations. We assessed the robustness of clonality analysis in the following way: First, for each patient, we estimated the proportion of the largest sub-clonal alteration, a parameter f needed for the estimation of the probability for correct classification of clonal mutations. For that, we determined the intersection of all branch defining alterations of all n biopsy samples. Next, we generated the intersection of all possible combinations of $i = 2$ biopsy samples. The frequency at which both intersections coincide is an estimate of the probability of a correct classification, $pf(i)/pf(n)$. The same procedure was performed for all possible combinations of $i = 3, 4, ..., n$ biopsy samples. Using these probabilities, we

estimated f for a given patient by fitting the estimated probabilities to:

$$\frac{p_f(i)}{p_f(n)} = \frac{1 - f^i - (1-f)^i}{1-(1-f)^i} \frac{1-(1-f)^n}{1-f^n-(1-f)^n}. \tag{3.1}$$

After the estimation of factor f we calculate the number of samples needed for detection of truly clonal mutations with $> 90\%$ confidence using the following equation:

$$p_f(n) \approx \frac{1-f^n-(1-f)^n}{1-(1-f)^n}. \tag{3.2}$$

3.3.3 Evolutionary trajectory

To asses the agreement of the mode of evolution of gastric cancer with the neutral model, we used R package neutralitytestr developed by Williams et al. (2016). To reduce the probability that apparent deviation from neutrality is caused by increase in allelic frequency due to gene duplication events, all variant alleles that had likely undergone gene doubling were removed, as shown by Williams et al. (2019b). Exclusion was done based on the logR value calculated for each variant using variant calling software CNVkit 0.9.5 (Talevich et al., 2016) by our project collaborators and tumour purity c. First, logR was used to calculated copy number (CN) for each variant assuming tumour purity c using the following equation:

$$\mathrm{CN} = \frac{2(2^{log(R)} - 1) + c}{c}. \tag{3.3}$$

Next mutation copy number (MCN) was calculated for every given mutation i as follows:

$$\mathrm{MCN_i} = \frac{\mathrm{CN_i} \times \mathrm{VAF_i}}{c}. \tag{3.4}$$

Next, the sequencing data was represented as a histogram of mutant allele frequencies (VAF histogram) to validate our exclusion of duplicated segments by comparison of VAF histograms prior and posterior to the exclusion step. We found that distributions remained mostly unaltered after the exclusion step as expected (figures 3.4 and 6.2). Further, we used VAF histograms to select for variants that are sub-clonal in the sample and to locate the frequency range

of accumulated sub-clonal mutations. This was done by identification of the distinctive clonal peak in each sample. In the absence of healthy tissue contamination, a clonal peak is expected to be around a frequency of approximately 0.5. In the case of contamination, the same peak is to be found at a lower frequency. Contamination with healthy tissue was measured using two approaches. First, a board-certified surgical pathologist visually assessed the tumour content of each sample. Second, purity was computationally estimated by our collaborators using Sequenza software package (Favero et al., 2015). Since the two modes of purity estimations showed quite different results, we decided to use the raw variant frequencies unadjusted for the purity for our neutrality analysis. Next, we represent the subset of data as the cumulative distribution of sub-clonal mutations in a range of frequencies between the lower edge of the clonal peak and the highest histogram bar (see supplementary figures 6.3 to 6.11). The cumulative distribution of mutations from each sample was then compared with the neutral model as presented in (Williams et al., 2016). In a neutral growing tumour, the number of sub-clonal mutations is expected to accumulate linearly with the inverse of their frequency (1/f). Using the neutralitytestr package (Williams et al., 2016), we calculated the effective mutation rate for each sample from the slope of the cumulative distribution curve. To quantify the degree of deviation from neutrality, we calculated the Kolmogorov distance between the normalized cumulative distribution of mutations and the expectations of the neutral model. Finally, we compared Kolmogorov values for each sample within the tumour.

3.4 Results and Discussion

3.4.1 Robustness of clonality analysis

Although the usage of genome sequencing in clinical practice is currently limited (Schwarze et al., 2018), we can expect genomics to enter clinical medicine in many specialities (Manolio et al., 2019). Clinical oncologists are already using sequencing to devise individual treatment regimes for their patients (Zhao et al., 2019). Currently, the mere presence of an actionable alteration in the tumour is sufficient to proceed with the targeted therapy. At the same time, information about abundance of the mutation is usually not considered (Wang et al., 2012).

In order to discriminate public (clonal) and private (sub-clonal) mutations it is required to have more than one tumour sample sequenced (Siegmund and Shibata, 2016). We estimated the reliability of identification of clonal mutations using an approach described in the previous chapter of this thesis (Werner et al., 2017; Opasic et al., 2019) by estimating the number of samples required for the correct identification of clonal mutations within the cancer with $> 90\%$ certainty. First, we assessed the "balance factor" f for every individual tumour by fitting a theoretical curve to the information gain with each multi-region sample as in (Opasic et al., 2019). In this study, we based our f estimation on fully reconstructed phylogenetic trees and used the branch-defining sub-clonal alterations to estimate the information gain with each additional multi-region sample (Supplementary figure 6.1). We found that six out of nine tumours are highly unbalanced, which implies that the number of truly clonal mutations is probably lower than currently considered (Figure 3.2). In these cases, sequencing of many additional multi-region samples would be necessary to identify truly clonal mutations with certainty. In one patient (E4526) with an almost balanced phylogenetic tree (f=0.56), five samples were sufficient for the identification of truly clonal mutations with probability $> 90\%$. For two other patients (E3350 and E5095) with unbalanced phylogenetic trees (f=0.2 and f=0.01), we can be almost certain that the list of clonal mutations is correct using the existing number samples, which can be as low as three (Fig. 3.3). Previously Werner et al. (2017) found that 6/7 unbalanced analysed tumours had received

treatment before resection and 2/3 balanced tumours were treatment-naive.

Their results imply that the selective pressure of therapy leads to unbalanced phylogenetic structures, therefore more samples are required for correct classification of clonal mutations in treated tumours. In our case, 3/6 unbalanced tumours had received neoadjuvant therapy and 2/3 balanced tumours were untreated. One of the explanations for balanced treated tumours is that the selective forces of therapy lead to sub-clonal sweep of the entire tumour population. In that case, all future growth following the sweep would be neutral which would appear as a balanced phylogenetic tree. Furthermore, the expected frequency of mutations emerging after the sweep is extremely low. This is because in neutral growth, expected frequency is $1/N$, where N is number of cells at the time of mutation emergence. A highly unbalanced tree in an untreated tumour can be a consequence of stochastic effects in both cancer growth dynamics and stochastic sampling effect. Further, selection other than the one caused by therapy might indeed be present in tumours, what would lead to non-neutral growth patterns. This is quite common as Williams et al. (2016) found that a majority of cancers, around two thirds of 904, could not be explained by a neutral model.

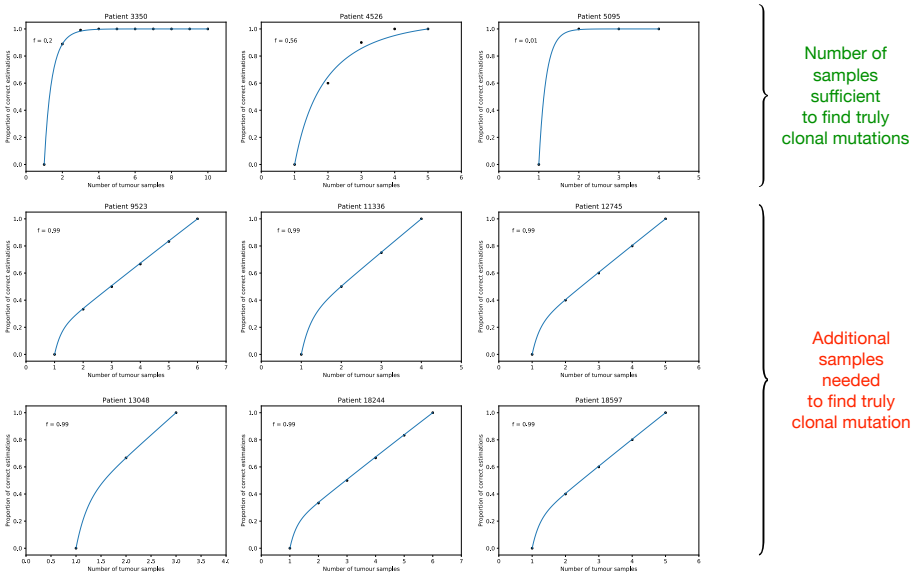

Figure 3.2: The balancing factor f was estimated for each patient using branch defining sets of mutations. For every combination of sample number from 2 to the total number of samples, the intersection of mutations was compared with the intersection using the total number of samples. Each dot represents the proportion of coincidences for a given number of samples. The blue line represents the best fit of the model E.q. 3.1 to the information gained with each additional sample. For patients in the top row, the existing number of samples is sufficient for correct clonality estimation.

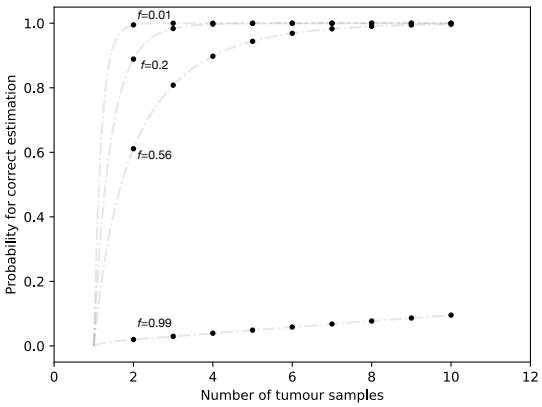

Figure 3.3: Probability for correct clonality estimation with n samples calculated using eq. 3.2 for tumours with balancing factor f assessed prior for each individual patient (see Fig. 3.2).

3.4.2 Evolutionary trajectory

In this segment, we attempt to quantify the variability in the mode of evolution within gastric cancers. Tumour growth can be driven by competition between emerging sub-clones where each acquires more and more beneficial mutations (Reiter et al., 2019). In contrast, tumour growth can be predominantly neutral; mutations responsible for malignant phenotype appeared in the first malignant cell and all subsequent mutations are effectively neutral (Williams et al., 2016). Finally, due to the different microenvironments, different parts of the tumour might exhibit different modes of evolution. To distinguish these two modes of evolution in tumours from the patients, we analysed the variant allele frequencies from each sample of an individual tumour. Specifically, we are interested in testing whether the distribution of variant alleles frequencies in our data can be explained by the neutral model as shown by Williams et al. (2016).

To compare the accumulation of mutations with the neutral model, we presented sequencing data from each sample as histograms of variant allelic frequencies (VAF) (figure 3.4 and supplementary figure 6.2). VAF histograms of neutrally growing tumours display a specific pattern in which sub-clonal alterations accumulate following a $1/f$ power-law distribution (Williams et al.,

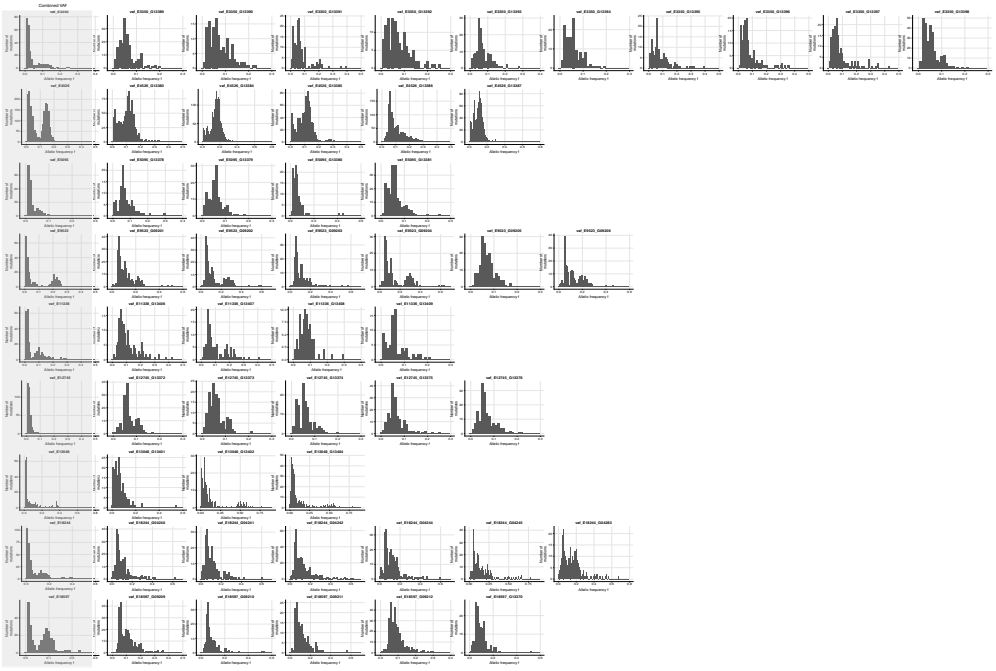

Figure 3.4: Genomic data of sampled tumours presented as variant allelic frequency histograms. Genomic regions that likely underwent duplication events were removed from the analysis.

2016). We tested whether the distribution of variant allelic frequencies from our patients obey this distribution. The distribution of variant frequencies within the sample can be greatly affected by genetic events such as copy number variation (Williams et al., 2016). Loss of a single chromosome can increase the perceived frequency of alterations present in the remaining chromosome even though the fraction of cells carrying this variant remains the same. For that reason, for each variant we calculated the cancer cell fraction (CCF) based on the tumour purity and local copy numbers. Unfortunately, the distributions of CCFs were extremely different from both VAF histograms, and distributions proposed by theoretical models (Williams et al., 2016; Sun et al., 2017; Williams et al., 2018). Consequentially, we decided to analyse raw allele frequencies only on variants present in regions of the genome we know are diploid. We did not take into consideration contamination with healthy tissue as there were great discrepancies between pathological and computational estimations of tumour

content. We are, however, aware that contamination with healthy tissue lowers the frequency of each variant (Williams et al., 2016). Finally, to perform the analysis we had to identify and exclude the mutations which are clonal in each sample. We identify clonal mutations as alterations present in the peak with the highest frequency in the VAF histogram. Data was then represented as the cumulative distribution of sub-clonal alterations for each sample and a neutral model was fitted to the data as in Williams et al. (2016). Frequency range of mutations used to compare against the neutral model is marked grey on each VAF histogram in the panel A of supplementary figures 6.3 to 6.11. If the coefficient of determination of goodness of fit is $R^2 \geq 0.98$ we can consider the data compatible with the model.

We found that the distribution of variant frequencies in 13/16 samples from treated tumours can be explained by the neutral model (see supplementary figures 6.3 to 6.11). This is surprising as chemotherapy induces strong selective pressure in cancer (Venkatesan et al., 2017). We expected to observe signs of sub-clonal selection in the form of additional sub-clonal peaks in the VAF distribution (Williams et al., 2018). This was not the case, as the majority of samples follow the neutral growth dynamics. In other samples that did not agree with the neutral model, no additional sub-clonal peak were observed. One of the reasons for apparent neutrality in treated tumours can be a complete clonal sweep and the fixation of adaptive mutations which would revert the tumour dynamic back to neutral (Turajlic et al., 2019).

Among five treatment-naive patients, 20/32 samples show agreement with the neutral model (supplementary figures 6.3 to 6.11). No patient displayed exclusively neutral dynamics across all samples. The absence of apparent selective advantage of sub-clones was also observed in oral pre-cancerous and cancerous lesions (Wood et al., 2017). Interestingly, in patient E4526, which is the only case with microsatellite instability (MSI), there is clear observable bump below $f < 0.2$ (supplementary figure 6.7). We, suggest that this prominent peak is caused by a set of clonal mutations whose frequency is lowered by contamination with healthy tissue. Such a high number of clonal mutations in such low frequencies obfuscates the distribution of sub-clonal mutations which are crucial for the assessment of the tumour dynamics. Additionally, we estimated

the effective mutation rate (μ/β, where μ is the per-base per-division mutation rate and β is effective division rate) from the slope of the fitted curve. Inferred values hold only for neutral tumours ($R^2 \geq 0.98$), since in non-neutral tumours mutation rates cannot be estimated as the model does not fit the dynamics of these tumours (Williams et al., 2016). After we estimated which samples concur with the neutral model, we quantified the deviation from neutrality within different parts of the tumour. For each sample, the Kolmogorov distance was calculated between the empirical and theoretical normalized cumulative number of mutations for given frequency range (panel C in supplementary figures 6.3 to 6.11) and used as a measure of deviation from neutrality. Finally, we compared the degrees of intratumour heterogeneity in respect to deviation from a neutral model between the patients (Fig. 3.5). To compare the heterogeneity in deviation for a neutral model, for every patient we calculated the mean and standard deviation of Kolmogorov distances of each sample. We found great variability in the deviation from neutrality within each patient and among patients (Fig. 3.5). In some cases (E13048 and E12745), all parts of the tumour have a similar distribution of mutations. Others display variance in compatibility with the neutral model

This method has so far been used in several studies to infer the evolutionary growth of cancer (Wood et al., 2017). Sun et al. (2017) presented a way to distinguish tumours undergoing strong positive selection from those evolving neutrally by analysing VAF histograms from multiple tumour samples. Using a computational model of tumour heterogeneity, they showed that different spatial samples in neutrally growing tumours have similar looking VAF histograms. In contrast, two distant spatial samples of non-neutral tumours produce different VAF histograms. They showed that detectable genetic divergence between regions is caused by selection when it fails to result in a complete sweep. This between-region genetic divergence can be used to evaluate patterns of intratumour heterogeneity. The detection of clonal selection in cancer is rather difficult and all currently available approaches have some drawbacks (Williams et al., 2019a). For instance, VAF histogram of non-neutral growing tumours can still appear neutral, as the driver frequency is biased towards 0 and 1. According to Bozic et al. (2019) there is a narrow window of parameters where a driver

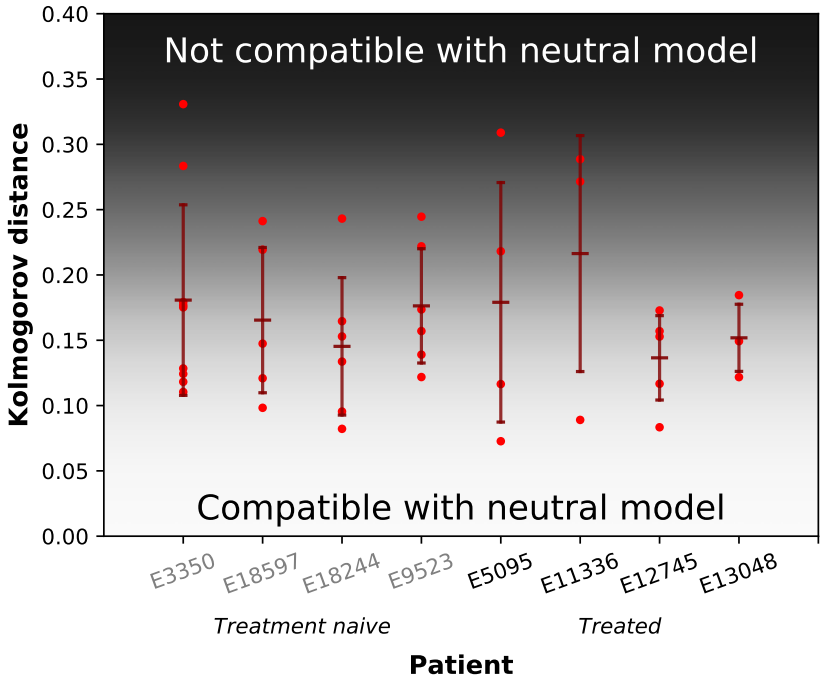

Figure 3.5: Compliance of each tumour sample with the neutral model is measured as Kolmogorov distance between empirical and theoretical distribution of normalized cumulative mutation number in defined frequency range. Red lines represent standard deviation of deviations from neutrality for each patient.

sub-clonal cluster can be visible before it fixates within the sample. Deviation from neutrality in the region ($VAF > 0.12$) can be caused by false sub-clonal clusters in the VAF spectrum due to the over-representation of the lineage in the sample as a result of the spatial effects of tumour growth (Chkhaidze et al., 2019). Another method, the dN/dS—the ratio of nonsynonymous mutations to synonymous mutations, can be used to infer selection (Williams et al., 2019b). However, synonymous mutations often are not neutral (Supek et al., 2014) and the method relies on many assumptions not applicable to cancer such as constant population size (Williams et al., 2019a).

3.5 Conclusions

In this chapter, we applied existing theoretical models on patient-derived sequencing data in an attempt to bring the results of mathematical modelling closer to clinical practice. This has shown to be challenging a task mainly due to the limitations of current sequencing technology and the noise associated with it. We estimated robustness of the clonality analysis in nine patients with gastric adenocarcinoma and found that in only 3/9 patients the number of samples is sufficiently large to identify the list of truly clonal genetic alteration with high confidence. For the rest, the list of clonal mutations is likely to be shorter and sequencing of more samples would provide additional information on the clonality status of each detected mutation. Our results show that multi-region tumour sequencing is necessary for clonality estimations and that it is beneficial to perform the robustness measure on every candidate for targeted therapy procedures. Aside from diagnostic value, the method used in this chapter is of economic significance as it provides information on the minimal number of samples needed to infer the clonality of mutations.

Apart from the robustness analysis, in this chapter we attempted to characterise the modes of evolution within each tumour. By fitting a neutral model to a cumulative number of sub-clonal mutations in a specific interval, we found that data from 33/48 of samples could be explained by neutral growth dynamics. We were, however, unable to detect the difference in the mode of evolution between treated and treatment-naive tumours. We also showed variability in the deviation from neutrality within each patient. Our analysis of the evolutionary trajectory has no direct clinical utility at the time of writing. This might change as our knowledge of sub-clonal composition and tumour evolutionary history develops. Of direct clinical utility is that we showed is that multiple samples are necessary when evaluating evolutionary aspects of tumour, as different parts show different sub-clonal distributions. Finally, detection of clonal selection in tumours is difficult, especially when looking at a single snapshot in the tumour evolution, as static VAF histograms are likely to capture only a narrow window of tumour evolution. Late and/or weak selective mutations remain at a low frequency while early and/or strong selective mutations fixate

and appear clonal (Bozic et al., 2019) which might explain a high degree of perceived neutrality in our samples.

3.6 Contributions

This chapter is a part of larger project led by Prof. Dr. Christoph Roecken from the Department of Pathology, Christian-Albrechts-University, Kiel, Germany. Prof. Roecken was responsible for the acquisition of the samples and histological characterisation of neoplasms. Sequencing was performed by Philip Rosenstiel at the Institute for Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany. Bioinformatic analysis of the sequencing data was done by collaborating group led by Tobias Meissner from the Department of Molecular and Experimental Medicine, Avera Cancer Institute, Sioux Falls, USA. That includes: somatic mutation calling, copy number profiling, purity estimation, phylogenetic tree reconstruction.

3.7 Ethics statement

Ethical approval was obtained from the local ethical review board (D 453/10 and D 525/15) of the University Hospital Schleswig-Holstein, Kiel, Germany, which permitted us to use the samples from those patients who had also given written informed consent for a prospective scientific use of their patient material.

Chapter 4

The extent of sampling bias in liquid biopsy cancer diagnostics

---

This chapter is a working draft of a manuscript (*The extent of sampling bias in liquid biopsy cancer diagnostics* Opasic, Zhou, Scott, Traulsen, *in prep.*) of a project conceived during my short research visit to Theory Division led by Jacob Scott, MD. PhD. in the Department of Translational Hematology and Oncology Research in the Cleveland Clinic, Cleveland, Ohio, USA. Next to dr. Scott, this project is co-supervised by Prof. Dr. Arne Traulsen and Prof. Dr. Da Zhou.

4.1 Abstract

Background Liquid biopsy holds great potential for the tracking of evolutionary dynamics of cancer in patients over time. However, the amounts of tumour DNA we are able to extract and sequence is small such that stochastic sampling effects distort the obtained results. These effects distort our view on the status of solid tumour as mutations might appear under- or overrepresented in the sample.

Methods In this paper, we investigate the extent of the bias in liquid biopsy sampling. We apply a set of statistical methods to calculate the sample size necessary to detect individual tumour mutations and to estimate error in the mutation frequency inference for a given sample size.

Results We show that with a small amount of genomic material we are likely to detect only the most common mutations within the cancer. Further, we provide a method to calculate the error of mutation frequency estimation obtained

from liquid biopsy sampling.

Conclusions Our result highlights the importance of sample size consideration when drawing conclusions from liquid biopsy genome profiling. Additionally, we provide to researchers and clinicians a tool to estimate the extent of the sampling error in their results and to calculate the sample size required to accurately measure mutations of a specific frequency.

4.2 Introduction

Solid tumours release proteins, individual or groups of circulating tumour cells (CTC), circulating cell-free tumour DNA (ctDNA) and metabolites into the bloodstream (Lindblom and Liljegren, 2000). Obtainable by minimally invasive procedures, these analytes can provide valuable insight in the status of malignant processes over time (Diehl et al., 2008; Heitzer et al., 2019). Despite easy access, currently only protein tumour markers isolated from peripheral blood such as CEA, Ca-19-9, CA 125, PSA, are being used routinely to monitor tumour burden and disease recurrence (Duffy, 2013). Even though the presence of circulating cancer cells (CTC) in serum has been known for long time (Ashworth, 1869; Mandel and Métais, 1948; Engell, 1955; Leon et al., 1977; Stroun et al., 1989) and there are many promising studies, ongoing clinical trails and FDA-approved diagnostic products in the US, CTCs are not widely utilized in clinical practice due to the lack of clinical utility (Riethdorf et al., 2007; de Bono et al., 2008; Zhang et al., 2012; Rack et al., 2014; Alix-Panabieres and Pantel, 2014; Heitzer et al., 2019).

Circulating cell-free tumour DNA (ctDNA) is genetic material released from necrotic and apoptotic cells of primary tumour, metastases, or from CTC's (Alix-Panabières and Pantel, 2017). With the advancement of isolation and sequencing technologies, it is possible to sequence minute amounts of ctDNA from blood with great accuracy and the opportunities for utilization of ctDNA are vast (Sonnenberg et al., 2014; Heitzer et al., 2015; Gale et al., 2018). ctDNA holds potential to provide us with great insight into the extensive genetic diversity within cancer and elucidate different sub-clonal compositions that vRY across cancer types and between individual patients (Gerlinger et al., 2012, 2014; de Bruin et al., 2014; Jamal-Hanjani et al., 2016; Rasche et al., 2017; McGranahan and Swanton, 2017). The extent of genetic heterogeneity has repercussions on cancer resistance to therapy and patient's final outcome (Dagogo-Jack and Shaw, 2018). Furthermore, knowledge of the state of sub-clonal composition over time could unravel complex phylogenetic relations between cancer populations and the mode of cancer clonal evolution (Abbosh et al., 2017; Siravegna et al., 2017; Scott et al., 2019)

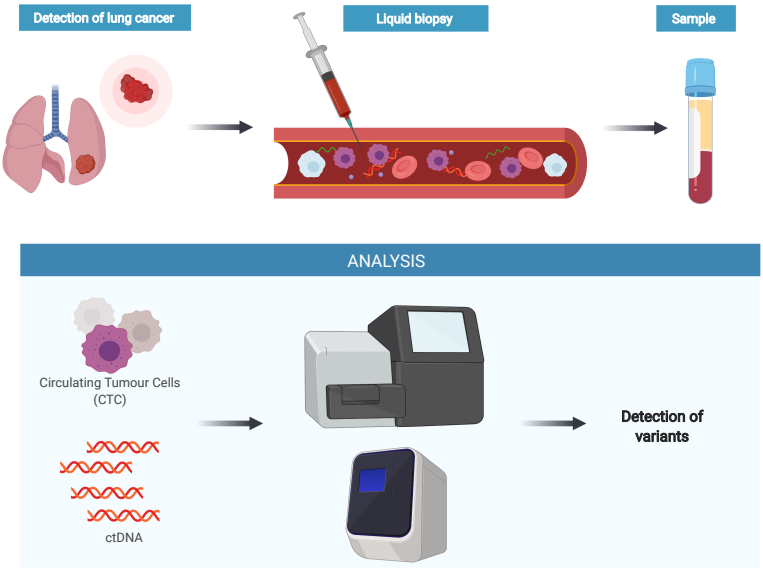

Figure 4.1: Tumour continuously sheds genetic material in form of circulating tumour cells and cell-free tumour DNA. Small amounts of material can be isolated and sequenced. Figure created with BioRender.com

More specifically, ctDNA has the potential to be used as an non-invasive biomarker for early detection of cancer (Abbosh et al., 2017). Currently, the CancerSEEK test offers more than 69% of sensitivity for five types of cancer for which there is no available screening test (Cohen et al., 2018). Another still ongoing and yet to be published study for the development of a non-invasive, multi-cancer detection assay, The Circulating Cell-free Genome Atlas (CCGA), shows great promise (Klein et al., 2018). Those tests might bring great benefit in cancers where early detection is critical for positive outcome and where patients are avoiding standard diagnostic procedures due to their invasiveness. Currently however, clinical usage of ctDNA is limited to detection of certain targetable mutations e.g cEGFR in NSCLC and SEPT9 and KRAS in colorectal cancer (Sacher et al., 2016; Hench et al., 2018; Khan et al., 2018). The ability of liquid biopsies to predict the evolution of resistance have yet to fulfil its potential. So far, a single prospective clinical trial that demonstrated predictive power using sequential profiling of ctDNA coupled with solid tis-

sue biopsies could predict waiting time to relapse in patients with metastatic colorectal cancer (Khan et al., 2018). ctDNA was also shown to be useful in detecting minimal residual disease (Tie et al., 2016). Success of adaptive therapy is expected to rely heavily on the ability of liquid biopsy to provide real-time information on cancer dynamics. Switching between different therapy regimes, required by adaptive therapy, must be done to respond to changes in sub-clonal architecture of the cancer and the proportion of resistant sub-clones (Gatenby et al., 2009). The potential of this emerging technology is great as it enables us to gain insight into temporal dynamics of the ever-changing genomic landscape of cancer. As it releases genomic material from different parts of the tumour (Jamal-Hanjani et al., 2016), ctDNA should provide comprehensive view of the tumour genomic landscape with large enough sample. However, Heitzer et al. (2019) rightly emphasizes that low levels of ctDNA in plasma bring considerable challenges in getting reliable and accurate representations of true variant allele frequencies (VAFs) from mutations within tumour due to stochastic sampling effects. Some mutations might be overrepresented in ctDNA, then again, certain sub-clonal driver alterations and resistance-causing mutations, might be underrepresented or even completely absent from the liquid biopsy sample. Detected mutational frequencies therefore greatly depend on the amount of available genomic material. The quantity of DNA able to be extracted from the blood of a cancer patient varies with the status of the disease. In one patient, the fraction of mutant DNA fragments from total circulating DNA can drop from 13.4% before surgery, to 0.015% three days after surgery, and then to 0.66% 244 days after surgery at the time of relapse (Diehl et al., 2008). In addition, extremely short half-life of ctDNA (16 minutes (Lo et al., 1999)) contributes to small amount of tumour DNA being isolated and successfully sequenced. We are able, however, to detect mutations at low allele fractions in ctDNA down to 2% (Gale et al., 2018). In this work we address the issue of small sample size and provide a theoretical framework to measure the extent of the sampling bias in liquid biopsy diagnostics. We use a computational model of tumour heterogeneity and a simulation of liquid biopsy to gain insight into the ability to detect the presence of mutations and estimate their true frequencies from liquid biopsy samples.

4.3 Methods

In this work we combine computational modeling of cancer with statistical inference to investigate the theoretical limitations in accuracy of liquid biopsy. First, we present the theoretical framework for the calcualation of the extent of the sampling error in liquid biopsy sampling. Next, we present a spatial model of tumour heterogeneity that is used to test the mathematical framework. Further, we simulate liquid biopy by randomly sampling a variable number of cells from simulated tumours. The sample size of a liquid biopsy is presented as the number of single circulating tumour cells (CTCs). Further, sample size can be viewed as a number of whole genome equivalents – full genomes composed out of ctDNA fragments originating from different cancer cells. Those two measures are equivalent if ctDNA offers a good representation of cancer and CTCs in the bloodstream.

4.3.1 Detecting single mutations

Consider a tumour composed of N cells, K of which carry a certain mutation of interest. Then the true proportion of the mutant cells in the tumour is $p = K/N$. Now we randomly sample n cells from the tumour and assume that the sample size n is much smaller than the whole population size N, $n \ll N$. Let X_n be the number of the mutant cells within these samples of size n, then X_n approximately obeys a binomial distribution

$$\Pr(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{4.1}$$

We are interested in the probability of correctly detecting the mutation by using random samples with size n, i.e. the probability that at least one mutant sample is present in these samples of size n. This is given by

$$\Pr(X_n \geq 1) = 1 - \Pr(X_n = 0) = 1 - (1-p)^n. \tag{4.2}$$

For example, if $\Pr(X_n \geq 1)$ is required to be larger than 90%, the we have

$$1 - (1 - p)^n \geq 90\%, \tag{4.3}$$

that is,

$$n \geq \frac{\ln(0.1)}{\ln(1 - p)} \tag{4.4}$$

gives the threshold for the sample size.

4.3.2 Inferring mutation frequencies

We first consider the single mutation case. Still, X_n represents the number of cells carrying a particular mutation of interest in a sample of size n. Let $Z_n = X_n/n$ be the sampled proportion of mutant cells. Here we would like to quantify the difference between the sampled proportion Z_n and the true proportion p by using the mean squared error as follows

$$E[(Z_n - p)^2] = E[(\frac{X_n}{n} - E(\frac{X_n}{n}))^2] = \frac{1}{n^2}\mathrm{Var}(X_n). \tag{4.5}$$

As X_n follows a binomial distribution parameterized by n and p, we have

$$E(Z_n - p)^2 = \frac{1}{n}p(1-p). \tag{4.6}$$

We now calculate the sample size n needed to reach a squared error lower than ε^2 with a given confidence level, e.g. 90%.

$$\begin{aligned} \Pr\left((Z_n - p)^2 \leq \varepsilon^2\right) &= \Pr\left(n(p-\varepsilon) \leq X_n \leq n(\varepsilon + p)\right) \\ &\approx \sum_{\lceil n(p-\varepsilon)\rceil \leq i \leq \lceil n(p+\varepsilon)\rceil} \binom{n}{i} p^i (1-p)^{n-i} \\ &\geq 90\%. \end{aligned} \tag{4.7}$$

We can get the lower bound of sample size n by solving the above inequality.

We next consider the multiple mutations case, in which there are M different types of mutations in the tumour. Let p_i be the true proportion of the mutant i and Z_n^i be the sampled proportion of the mutant i. Then, the mean squared error between p_i and Z_n^i is given by

$$E(Z_n^i - p_i)^2 = \frac{1}{n}p_i(1-p_i). \tag{4.8}$$

To get the overall difference between the sampled and whole tumour, we take

the average of all the mean squared errors for M different mutations

$$\begin{aligned} \text{MSE}_L &= \frac{1}{M} E[(Z_n^1 - p_1)^2 + (Z_n^2 - p_2)^2 + ... + (Z_n^M - p_M)^2] \\ &= \frac{1}{nM} \sum_{i=1}^{M} p_i(1-p_i) \approx \frac{1}{n} \int_0^1 p(1-p) f(p) dp, \end{aligned} \tag{4.9}$$

where $f(p)$ is the probability density of mutational frequency p across all the mutations. Eq. (4.9) defines the overall mean squared error of liquid biopsy. We can calculate MSE_L from the distribution $f(p)$.

$f(p)$ is a distribution of the mutant allele frequencies within the tumour. Naturally with accumulation of mutations in tumour over time, low frequency alterations will be present in much higher number then the high frequency ones (Williams et al., 2016). This can be estimated from the distribution of frequencies within the liquid biopsy sample, provided that the liquid biopsy is a good representation of the whole tumour population.

We here show some results of MSE_L for different distributions of $f(p)$.

- If $f(p)$ is a truncated exponential distribution, i.e.

$$f(p) = \frac{\lambda e^{-\lambda p}}{1 - e^{-\lambda}} \quad \text{for } \lambda > 0 \text{ and } 0 \leq p \leq 1, \tag{4.10}$$

  then

$$\text{MSE}_L = \frac{1}{n} \frac{(\lambda - 2) + (\lambda + 2)e^{-}\lambda}{\lambda^2(1 - e^{-\lambda})}. \tag{4.11}$$

- If $f(p)$ is truncated power-law distribution, i.e.

$$f(p) = \begin{cases} \dfrac{1}{\ln p_L^{-1}} \dfrac{1}{p} & (\text{ for } \alpha = 1 \text{ and } \ p_L \leq p \leq 1) \\ \dfrac{1-\alpha}{1 - p_L^{1-\alpha}} p^{-\alpha} & (\text{ for } \alpha > 1 \text{ and } \ p_L \leq p \leq 1) \end{cases} \tag{4.12}$$

then

$$\mathrm{MSE}_L = \begin{cases} \dfrac{1}{n}\dfrac{(1-p_L)_L^2}{2\ln(p_L^{-1})} & (\text{ for } \alpha = 1) \\ \dfrac{1}{n}\dfrac{1-\alpha}{1-p_L^{1-\alpha}}\left(\dfrac{1}{(2-\alpha)(3-\alpha)} - \dfrac{1}{2-\alpha}p_L^{2-\alpha} + \dfrac{1}{3-\alpha}p_L^{3-\alpha}\right) & (\text{ for } \alpha = 1) \end{cases} \tag{4.13}$$

We call the case with $\alpha = 1$ simple power-law distribution and it is parameter-free. Note that both exponential and power-law distributions have their own parameters to be determined by data fitting. We can use variant allele frequencies (VAFs) (Williams et al., 2016, 2018) from liquid biopsy samples to estimate these parameters. More specifically, we can check if the probability density function of normalized VAF histogram follows either an exponential or a power-law distribution. Using linear regression on log-transformed density values, we test if the distribution of mutations follows an exponential distribution. If that is the case, MSE_L is calculated with Eq. (4.11) using λ obtained from slope by linear regression. If the probability density function of sampled mutations agrees with power-law distribution, we proceed to calculate MSE_L using Eq.(4.12) for power-law distribution of mutation frequencies.

4.3.3 Liquid biopsy sampling from a spatial cancer model

We model the spatial expansion of the tumour on a two dimensional lattice where each position in matrix represents one tumour cell (Opasic et al., 2019). Empty sites represent healthy cells or an intracellular matrix and they are excluded from the analysis. Every time step, each cell within tumour divides into the surrounding site (Moore neighborhood) not occupied by a cancer cell. One time step corresponds to the average division time of a cancer cell. If the cell chosen for division does not have an available spot for division, it will not divide. With every division, the daughter cell receives a neutral mutation. We adopt the infinite alleles model (Kimura, 1969), i.e. the same mutation cannot arise twice and cannot be lost. Initial cancer cells hold a set of mutations that will be passed on every subsequent cell, these are considered to be clonal

or public. All other mutations acquired during tumour development will be sub-clonal or private.

To simulate a liquid biopsy, we randomly select a group of cells from the simulated tumour, reconstruct the mutational profile of each sampled cell and calculate the frequencies (Z_n^i) for each mutation detected in the sample. We then compare the frequencies of mutations within the sample with the true frequency of the same mutations in the whole tumour by calculating $(Z_n^i - p_i)^2$. To get the overall agreement between sample and whole tumour, we use

$$\text{MSE}_S = \frac{1}{M}\sum_{i=1}^{M}(Z_n^i - p_i)^2. \tag{4.14}$$

MSE_S is the true mean squared error in our simulation study. In clinical practice however, we do not know the true value of p_i, and thus we cannot calculate MSE_S directly. In this case, MSE_L in Eq. (4.9) can be used to estimate MSE_S.

4.4 Results

We investigate the effect of the sample size on the correct estimation of the tumour mutational burden. We identify the limitations of genomic data obtained by liquid biopsy sampling and from multi-region sampling of a tumour and provide a theoretical rationale for drawing conclusions from genomic data obtained by liquid biopsy. We further investigate the possibilities for utilization of liquid biopsies as a proxy for measuring intratumour sub-clonal composition.

4.4.1 Detection and quantification of individual mutations

Detection of individual mutations. Detection of the presence of a mutation within a tumour is the main goal of liquid biopsy diagnostics. Let us consider a simple scenario where a single mutation is present in half of the tumour ($p = 0.5$). By sampling one cell, we have a 50% probability to have the specific mutation present in our sample. Now, we can calculate the number of cells needed to detect any mutant present in frequency p using Eq. (4.4). To reach a probability of $> 90\%$ for having at least one mutant with frequency $p = 0.5$ present in the sample, the liquid biopsy must contain at least six CTCs or whole genome equivalents (Werner et al., 2017). The amount of genetic material needed to detect rarer variants increases dramatically depending on the mutational frequency. For less abundant variants, it climbs to > 40 cells needed for $p = 0.1$ and > 458 for $p = 0.01$ (Figure 4.2), an amount far beyond the currently available genomic material and our ability of isolating and sequencing it from the bloodstream.

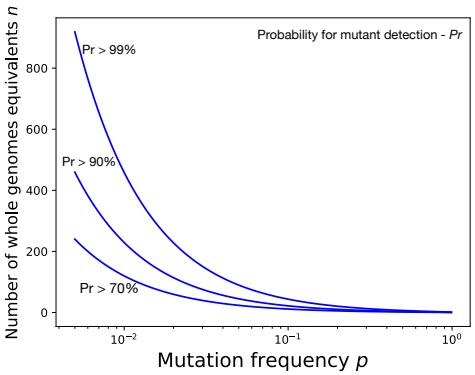

Figure 4.2: Number of cells or whole genome equivalents needed for detection of mutations with frequency p with $> 90\%$ confidence level. Calculated using Eq. (4.4)

Frequency estimation of individual mutations. The sampled frequency Z_n from a sample of a certain size n is an unbiased estimation of the exact frequency p of the mutation within the tumour. If we are interested in knowing the effect of this estimation, we can consider the mean squared error between the sampled and exact frequencies. Note that the number of mutants in the sample approximately obeys a binomial distribution Eq. (4.1). We can thus calculate the theoretical estimation error using Eq. (4.6) (Figure 4.3 left panel, blue line). We show that with a realistic number of cells in the sample, $n < 10$, the average difference between detected and true frequency of the most common sub-clonal mutation stays larger than $\varepsilon = 0.2$. The estimated accuracy of the frequency of a single mutation is increasing as we include more cells into the analysis (Figure 4.3). To validate the theoretical result in Eq. (4.6), we have also repeatedly sampled simulated tumours using samples containing n cells and calculated the average difference between observed frequency (Z_n) and true frequency (p) of the most common sub-clonal mutation (Figure 4.3 left panel, discrete symbols). The theoretical and simulated results show good agreement.

Next, we estimated the minimum sample size n needed to estimate the frequency of common individual mutations ($0.01 < p < 0.9$) with an error lower than ε with a 90% confidence level. We calculated the minimum sample size using Eq. (4.7) for a range of ε (Figure 4.3 right panel). Since the true frequency

of detected mutations in a tumour in reality is unknown, we calculated the n for a range of frequencies ($0.01 < p < 0.9$) and chose the highest value of n to ensure all detected mutations with frequency p would be estimated within the set error ε. As expected, higher precision (lower ε) generally requires more samples.

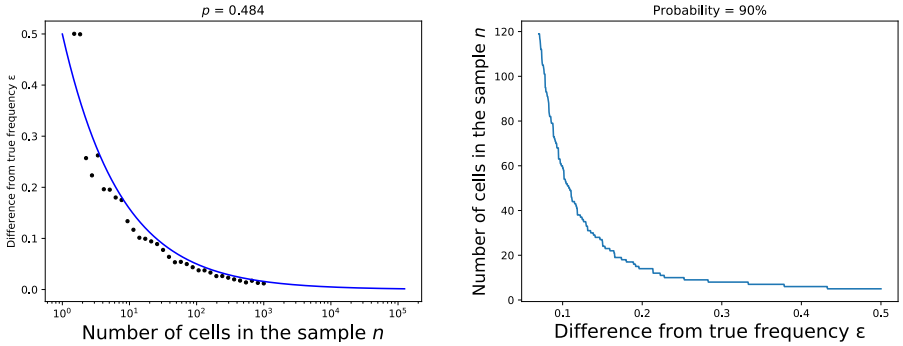

Figure 4.3: The left panel shows the decrease of frequency estimation error with increasing size of the liquid biopsy sample n. Each dot represents the average difference between the sampled (Z_n) and the true mutation frequency (p) of the largest sub-clonal mutation after 100 sampling repetitions on the same simulated tumour. The blue line represents the accuracy calculated for the biggest sub-clonal mutation with the frequency p as a squared root of the Eq. (4.6). The right panel represents the sample size (n) necessary to reach accuracy greater than ε with 90% probability for any mutation with frequency ($0.01 < p < 0.9$). Calculated with Eq (4.7) for the range of ε.

4.4.2 Frequency estimation of multiple mutations

In clinical practice, it is quite possible to obtain frequencies for a set of different mutations from liquid biopsy samples. Due to the small amount of sequenced genetic material, however, the inferred frequencies and their relationship will very likely be distorted. Here, we investigate this issue by estimating the average error of the measured frequencies for a given sample size. We calculated true mean squared error MSE_S (Eq. (4.14)) by comparing frequencies detected in the sample with true frequencies of the same mutations in the whole simulated tumour. MSE_S is further used as a ground truth for testing error estimation from the sample without knowing mutations' true frequencies (MSE_L (Eq. (4.9))))

If the empirical distribution of mutation frequencies within the sample follows an exponential distribution, we can calculate MSE_L for each sample size n using equation Eq. (4.11). The parameter λ necessary for the calculation was estimated by fitting the linear regression model to log-transformed probability density function obtained from the sample VAF histogram. At the same time for each individual sample, we compared the frequencies of mutations within the sample with corresponding frequencies in the whole tumour and calculated MSE_S. We used this value as a reference to test the accuracy of our MSE estimation. Figure 4.5 shows estimated MSE (MSE_L) and observed MSE (MSE_S) with a sample of size n for a single simulated spatial tumour. The MSE estimated using our approach is comparable to the observed MSE from the same sample. The average difference between observed and estimated MSE from the same sampling event are shown in the right panel of Figure 4.5. The difference for sample size $n > 100$ becomes negligible and the additional information gain with increasing sample size becomes smaller.

If the empirical distribution of mutation frequencies in the tumour follows a power-law distribution, we can calculate MSE_L for each sample size n using Eq.(4.12). The factor α is estimated from the distribution of VAF histogram obtained from the sample and fitting linear regression to the normalized histogram values. We test the accuracy of (MSE_L) (Eq. (4.9)) estimation against (MSE_S) (Eq. (4.14)) directly measured by comparing frequencies of sampled

mutations with simulated tumours (see Figure 4.5).

Williams et al. (2016) showed that sub-clonal mutant allele frequencies of roughly a third of analysed tumours follow a simple power-law distribution ($\alpha = 1$). In that case, we get a parameter free way for estimating MSE_L. To test how this parameter-free assumption affects the MSE estimation, we simulated a neutrally growing spatial tumour and calculated MSE_S using all the mutations present in the tumour with frequencies larger than 0.05 ($p_L > 0.05$) for different sizes of samples (n) (blue line on Figure 4.4). Next we performed a repeated number of sampling events (1000 repetitions) with a sample size n and calculated MSE_S (violin plot on Figure 4.5). Finally we calculated MSE_L using Eq. (4.13) with the assumption of truncated power-law distribution of the probability density of mutations for $p_L > 0.05$ (black line on Figure 4.5). It shows that MSE_L with a simple power law assumption was consistently larger than MSE_S. When the sample size reaches $n > 100$, the difference in error becomes negligible. Note that the simple power-law distribution is parameter-free and thus more convenient for utilization in clinical practice.

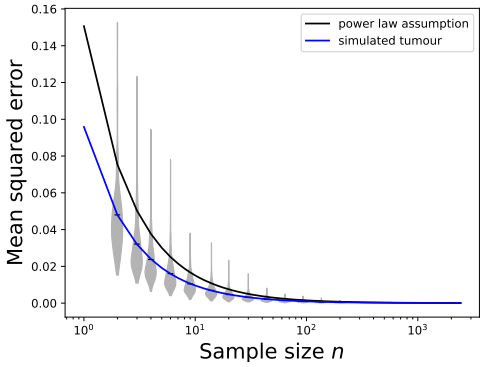

Figure 4.4: Error in mutation frequency estimation with sample size n on a single simulated tumour. The violin plot show individual MSE values for each sampling iteration. The blue line represents MSE calculated using Eq. 4.14 and the black line represents an MSE approximation calculated using simple power-law distribution Eq. 4.12 with $\alpha = 1$ with lower boundary of detectable mutations of $p_L = 0.05$.

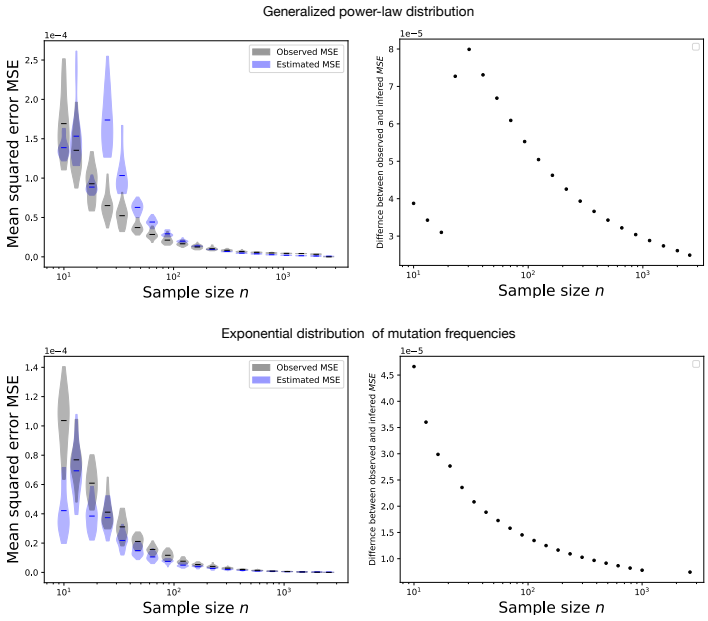

Figure 4.5: Violin plots on left panels represent individual MSE. Grey represent MSE directly observed by comparison of detected frequency with actual frequency of the mutation in the tumour. Blue violin plots contain the estimated MSE using Eq. 4.13 for the top panel and Eq. 4.11 for bottom panel. The right panels show average difference between observed and estimated MSE from the same sample.

4.5 Discussion

Liquid biopsies have shown potential to greatly improve medical diagnostics. It is expected to provide a wide range of clinical applications from non-invasive prenatal diagnostics by analysis of fetal cfDNA in maternal circulation (Mackie et al., 2017) to transplantation medicine – characterisation of graft-derived cell-free DNA (GcfDNA) as markers of transplant rejection (Schütz et al., 2017), to ultimately cancer diagnostics. Up to some extent, diagnostics using liquid biopsy is already part of routine clinical procedures e.g. detection of EGFR mutations in NSCLC where it is known that presence of some mutations induces sensitivity to EGFR tyrosine kinase inhibitor (TKI) therapy (Saarenheimo et al., 2019; Hench et al., 2018). The detection of single resistance mutations from a liquid biopsy is of great clinical importance, as they might serve as prognostic marker and a call for a change of the therapeutic option. Acquired

resistance to tyrosine kinase inhibitor was detected from a liquid biopsy of a single patient with NSCLC with EGFR mutation and initial sensitivity to TKI (Thress et al., 2015). In this case, the sample was collected at the time of systemic progression when amount of cfDNA is greater than before the time of relapse (Diehl et al., 2008). Ideally, resistance would be detected from liquid biopsy long before clinically observable progression. Knowledge about the amount of genomic material needed for successful detection of mutations would be beneficial, as additional samples could be harvested, pooled and analysed. In this work, we present a simple procedure to calculate a number of CTCs or ctDNA whole genome equivalents necessary to have a mutation present in an extracted sample with a desired probability. We show that individual low frequency mutations $p < 0.1$ are not likely to be detected in single sampling attempt with the amount of yield currently used (< 10 on average for 7.5ml of blood (Li et al., 2018)). We offer a way to calculate the probability for a false negative results when detecting potentially resistance-driving mutations. Information on the presence or absence of specific mutations is, for itself, of great clinical value as it can be used for making informed personalised therapeutical decisions. However, additional quantitative information on the mutation frequency tracked over time can help us to assess tumour evolutionary dynamics. Tracking the change of allele frequencies over time would open the possibility to predict future tumour growth trajectories and therapeutic outcome.

This knowledge would also be integral for improvement scheduling therapeutic regimes in adaptive therapy which so far relies on the feedback from indirect indicators of tumour dynamics such as PSA (Zhang et al., 2017; West et al., 2019). In order to have the correct information on the frequency of the specific mutation, one must have a sample representative of the whole population. Here we calculated the sample size necessary for reaching a desired level accuracy of frequency estimation of single individual mutation. Error estimated from samples that contain $n < 50$ individual cells or whole genome equivalents in $> 90\%$ will be greater than 0.1. Further, we presented a method to calculate average error for a full range of individual mutations within the sample of certain size.

Our results emphasise the importance of sampling bias when using a liq-

uid biopsy for genetic profiling of neoplastic disease. We provide a theoretical rationale for any results obtained from the analysis of cancer genomic material obtained from bodily fluids. Moreover, we determine a way to calculate quantitative threshold for the identification and determination of a quantity of specific genetic alterations within the cancer using samples of specific size.

4.6 Additional information

Availability of data and materials

Code available from Luka Opasic on request.

Conflict of Interest

The authors declare no conflict of interest.

Funding

D.Z. is supported by China Scholarship Council Grant (No.201806315038), NSFC (Grant No.11971405), and Max-Planck stipend. The funding bodies did not have a role in any aspect of this study.

Author's contributions

All authors developed the concept, L.O. and D.Z implemented the model, conducted the simulations and analysed the model with input from all authors. All participated in writing the manuscript.

Chapter 5

CancerSim software package for python3

This chapter is composed of software documentation of the cancer model used in previous chapters of the thesis. For this chapter, an earlier has been rewritten entirely with emphasis on usability and upgradeability. This includes facilitated distribution of the software, extensive documentation containing reference manual with explanation of each function, quickstart examples with detailed explanation of the simulation setup, automated test for correctness checking. CancerSim software together with here presented documentation can be found at the git repository: `https://github.com/mpievolbio-scicomp/cancer_sim.git`.

CancerSim will be submitted for a peer-review to the Journal of Open Source Software.

CancerSim: A Cancer Simulation Package for python3

Luka Opasic, Jacob Scott, Arne Traulsen, Carsten Fortman-Grotte

Background

Cancer is a group of complex diseases characterized by excessive cell proliferation, invasion, and destruction of the surrounding tissue (Vinay Kumar, 2017). Its high division and mutation rates lead to excessive intratumour genetic heterogeneity which makes cancer highly adaptable to environmental pressures such as therapy (Turajlic et al., 2019). Throughout most of its existence tumour is inaccessible to direct observation and experimental evaluation. Therefore, computational modelling can be useful to study many aspects of cancer. Some examples where theoretical models can be of great use include early carcinogenesis as lesions are clinically observable when they already contain millions of cells, seeding of metastases, and cancer cell dormancy (Altrock et al., 2015).

Here, we present CancerSim, a software that simulates somatic evolution of tumours. The software produces two-dimensional, virtual spatial tumours with variable extent of intratumour genetic heterogeneity and realistic mutational profiles. Simulated tumours can be subjected to multi-region sampling to obtain mutation profiles that are realistic representation of sequencing data. This makes the software useful for studying various sampling strategies in clinical cancer diagnostics. An early version of this cancer evolution model was used to simulate tumours subjected to sampling for classification of mutations based on their abundance (Opasic et al., 2019). Target users are scientists working in the field of mathematical oncology and students with interest in studying somatic evolution of cancer.

Our model is abstract, not specific to any neoplasm type and does not consider a variety of biological features commonly found in neoplasm such as vasculature, immune contexture, availability of nutrients, and architecture of the tumour surroundings. It resembles the most to superficially spreading tumours like carcinoma in situ, skin cancers, or gastric cancers, but it can be used to model any tumour on this abstract level.

The tumour is simulated using a two-dimensional, on-lattice, agent-based model. The tumour lattice structure is established by a sparse matrix whose non-zero elements correspond to the individual cells. Each cell is surrounded by eight

neighbouring cells (Moore neighbourhood). The value of the matrix element is an index pointing to the last mutation cell acquired in a list of mutations which is updated in each simulation step.

The simulation advances in discrete time-steps. In each simulation step, every tumour cell in the tumour that has an unoccupied neighbour can divide with a certain probability (params.div___probability). The daughter cell resulting from a cell division inherits all mutations from the parent cell and acquires a new mutation with a given probability (params.mut_prob). Different division probabilities can be introduced for some cells in order to simulate variability in fitness of cells that acquired a beneficial or deleterious mutation. The simulation allows the acquisition of more than one mutational event per cell (params.mut_per_division). In that case variable amounts of sequencing noise (Williams et al., 2016) can be added to make the output data more biologically realistic.

Throughout the cancer growth phase, CancerSim stores information about the parent cell and a designation of newly acquired mutations for every cell. Complete mutational profiles of cells are reconstructed a posteriori based on the stored lineage information.

The division rules which allow only cells with empty neighbouring nodes to divide, cause exclusively peripheral growth and complete absence of dynamics in the tumour centre. To allow for variable degree of growth inside the tumour, we introduced a death process. At every time step, after all cells attempt their division, a number of random cells die and yield their position to host a new cancer cell in a subsequent time step.

After the simulation, the tumour matrix, and the lists of lineages and frequencies of each mutation in the tumour are exported to files. Furthermore, the virtual tumour can be sampled and a histogram over the frequency of mutations will be visualised. Alternatively, a saved tumour can be loaded from file and then subjected to the sampling process.

Installation

CancerSim is written in Python (version >3.5). We recommend to install it directly from the source code. To download the code:

EITHER clone the repository:

```
$> git clone https://github.com/mpievolbio-scicomp/cancer_sim.git
```

OR download the source code archive:

```
$> wget https://github.com/mpievolbio-scicomp/cancer_sim/archive/master.zip
$> unzip master.zip
$> mv cancer_sim-master cancer_sim
```

Change into the source code directory

```
$> cd cancer_sim
```

We provide for two alternatives to install the software after it was downloaded:

Alternative 1: Conda

New conda environment

We provide an `environment.yml` to be consumed by `conda`. To create a fully self-contained conda environment (named `casim`):

```
$> conda env create -n casim --file environment.yml
```

This will also install the cancer simulation code into the new environment.

To activate the new conda environment:

```
$> source activate casim
```

or

```
$> conda activate casim
```

if you have set up conda appropriately.

Install into existing and activated conda environment

To install the software into an already existing environment:

```
$> conda activate <name_of_existing_conda_environment>
$> conda env update --file environment.yml
```

Alternative 2: Using pip

The file `requirements.txt` is meant to be consumed by `pip`:

```
$> pip install -r requirements.txt [--user]
```

The option `--user` is needed to install without admin privileges.

Testing

Although not strictly required, we recommend to run the test suite after installation. Simply execute the `run_tests.sh` shell script:

```
$> ./run_tests.sh
```

This will generate a test log named `casim_test@<timestamp>.log` with `<timestamp>` being the date and time when the test was run. You should see an `OK` at the bottom of the log. If instead errors or failures are reported, something is wrong with the installation or the code itself. Feel free to open a github issue at https://github.com/mpievolbio-scicomp/cancer_sim/issues and attach the test log plus any information that may be useful to reproduce the error (version hash, computer hardware, operating system, python version, a dump of `conda env export` if applicable, ...).

The test suite is automatically run after each commit to the code base. Results are published on travis-ci.org.

High–level functionality

The parameters of the cancer simulation are given via a python module or programmatically via the `CancerSimulationParameters` class. A documented example `params.py` is included in the source code (under `test/params.py`) and reproduced here:

```
$> cat test/params.py
# Number of mesh points in each dimension
matrix_size                      = 100

# Number of generations to simulate.
num_of_generations               = 20

# Number of divisions per generation.
div_probability                  = 1

# Number of division for cells with mutation.
fittnes_advantage_div_prob       = 1

# Fraction of cells that die per generation.
dying_fraction                   = 0.1

# Fraction of cells with mutation that die per generation.
fitness_advantage_death_prob     = 0.0

# Rate of mutations.
mut_prob                         = 1

# Mutation probability for the adv. cells.
advantageous_mut_prob            = 1

# Number of mutations per cell division.
mut_per_division                 = 10
```

```
# Time after which adv. mutations occur.
time_of_adv_mut                 = 2

# Number of mutations present in first cancer cell.
num_of_clonal                   = 15

# Tumour multiplicity.
tumour_multiplicity             = None

# Sequencing read depth.
read_depth                      = 100

# Fraction of cells to be sampled.
# sampling_fraction             = 0.1
```

The simulation is started from the command line. The syntax is

```
$> python -m casim.casim [-h] [-o DIR] seed
```

The mandatory command line argument `seed` is the random seed. Using the same seed on two simulation runs with identical parameters results in identical results, this may be used for testing and debugging. The optional argument `DIR` specifies the directory where to store the simulation log and output data. If not given, output will be stored in the directory `casim_out` in the current directory. For each seed, a subdirectory `cancer_SEED` will be created. If that subdirectory already exists because an earlier run used the same seed, the run will abort. This is a safety catch to avoid overwriting data from previous runs.

Example 1

```
$> python -m casim.casim 1
```

Example 2

```
$> mkdir sim_out
$> python -m casim.casim -o sim_out 2
```

Results will be stored in the newly created directory `sim_out/`.

Reference Manual

The API reference manual is available at https://cancer-sim.readthedocs.io.

Examples

See our quickstart example in `docs/source/include/notebooks/quickstart_example.ipynb`.

Chapter 6

General discussion

Despite the tremendous investments into in cancer research during the last 50 years, advanced malignant disease is still incurable. Even though in some forms of cancer, such as acute lymphoblastic leukaemia, progress has been tremendous (Inaba et al., 2013), in general, survival rates are still disappointing (Lu et al., 2019). As a result, pressure for bringing novel therapeutic options to the market is immense. Many new therapeutic drugs indeed entered the market in recent years. European Medicines Agency approved the use of 48 cancer drugs for 68 indication in four years period between 2009 and 2013. Shockingly, no prolonged survival or improvement in quality of life was observed for most of released drugs (Davis et al., 2017). Shockingly, none of these drugs either prolonged or improved the quality of live of patients (Davis et al., 2017). Further, targeting druggable cancer alterations, identified in patients, using already available medication is still no better than the conventional treatment (Le Tourneau et al., 2015). There are two reasons for such a bad outcome of modern therapy, first is the wrong target of choice and second is the onset of resistance. To address the first prerequisite, it is imperative to estimate the abundance of the molecular target in the cancer cell population before the start of treatment. Next, the ability to detect the presence of potentially resistant sub-clones, not just at the beginning of the treatment, but also over time, is needed to actively respond to changes in cancer composition. This thesis is addressing both causes of therapeutic failure. In this thesis, I presented a series of small contributions which could improve the diagnostics of cancer-specific alterations and assist in making better predictions of the therapy response in every individual patient.

In the second chapter, we provided an answer to a concrete question: *How many samples are needed to infer truly clonal mutations from heterogenous*

tumours? We discovered that the answer depends directly on the frequency of the largest sub-clonal mutation. For each patient, the number of samples required for correct classification of mutations with high certainty can be calculated using our mathematical model, see Eq. 2.4 and 2.3. In a tumours with low mutational burden, already with $n > 3$ tumour samples, the probability to correctly classify truly clonal mutations is $> 98\%$. For the case where the first branching event leads to a tumour with two roughly equally-sized populations we reach the same probability already with $n = 6$ samples. Finally, in tumours where specific sub-clonal mutations undergo great expansion, it is highly likely that this expanding mutation will be falsely categorized as clonal, see Figure (2.2). Our mathematical theory was further tested on a computational model of tumour heterogeneity. Using the same computational model, we further explored the effects of sample size on the accuracy of mutation classification in both spatial and well-mixed tumours. We showed that if all mutations are included, the clonality inference with multiple large samples is less accurate than using the equivalent single-cell samples. To increase the accuracy, we proposed removing the low frequency mutations from the analysis, which did improve the classification, see Figure 2.4. Finally, we proposed a spatial sampling pattern which improves the accuracy of clonality estimation in comparison to random sampling, see Figure 2.6. Previous research focused mostly on identification of clonal mutations from a single bulk sample (Bozic et al., 2016). However, employing multiple samples improves the identification of both private and public mutations (Siegmund and Shibata, 2016). Our consideration of successive branching events, spatial distribution and size of sample lowered the estimated number of samples needed for identification of truly clonal mutations than previously proposed by Werner et al. (2017).

To bring our theoretical method closer to clinical application, we applied the model to multi-region sequencing data of gastric adenocarcinoma. For each patient we estimated whether the number of tumour samples is sufficient for identification of clonal mutations with high degree of certainty. We found that in six out of nine tumours, the number of truly clonal mutations is probably lower than the currently considered, see Fig. 3.2. In these cases, sequencing of additional multi-region samples would be necessary to identify truly clonal

mutations with certainty. In one patient with an almost balanced phylogenetic tree, five samples are sufficient for the identification of truly clonal mutations with probability $> 90\%$. For the other two patients with unbalanced phylogenetic trees, we can be almost certain that the list of clonal mutations is correct using the existing number samples, which can be as low as three (Fig. 3.3). In our theoretical approach, all the mutations were of similar significance regardless of their fitness. In reality, the majority of mutations is tipically neutral with only few present in the driver genes (Reiter et al., 2019). Therefore, most have no impact on the biology of cancer and its response to therapy. It is possible that the most abundant sub-clonal mutations are neutral and missclassifying them as clonal will not affect the patient survival. This would reduce the number of necessary samples in even some cases to as much a single sample (Reiter et al., 2019). In order to estimate the probability for correct identification of clonal mutations, the clonality analysis in patients, we employed a fully reconstructed phylogenetic trees and major branch defining alterations with biological significance. In the same chapter we characterised the extent of intratumour heterogeneity in respect to the mode of evolution. By testing the compatibility of variant allele frequency histograms of each cancer sample with the neutral growth model, we found variability in the degree of deviation from neutrality within each patient. Interestingly, there was no difference between treatment naive and treated cancers in that regard.

As previously stated, every late-stage cancer is likely to develop resistance to the therapy. A resistant mutation can be present in the tumour from its beginning and thus make it innately unresponsive to the treatment. Resistance can also emerge as a random mutation during the tumour growth which becomes adaptive after the introduction of therapy. Selective pressure changes the frequency of adaptive this allele over time. This change in allele frequency in the cancer population can be tracked through the profiling of tumoral genetic material from blood. Liquid biopsy is one of the most promising methods for tracking of evolutionary dynamics and response to therapy (Heitzer et al., 2019). It has the potential to detect the presence of resistance-causing mutations prior to the relapse. (Khan et al., 2018) This information would be of great value to oncologists as it would enable them to react much earlier. How-

ever, in chapter four of this thesis, we show that with the amount of genomic material currently available with liquid biopsy, we are likely to detect only the most common mutations within the cancer. Liquid biopsy must contain at least six CTCs or whole genome equivalents to detect the mutation with frequency $p = 0.5$ with a probability of $> 90\%$. For less abundant variants, > 40 cells needed for $p = 0.1$ and > 458 for $p = 0.01$, see Figure 4.2. This amount is far beyond the current ability of isolating and sequencing of cancer DNA from the bloodstream. Further, we provided a method to calculate the error of mutation frequency estimation from liquid biopsy sampling for both single mutation and a set of mutations. We tested the method on simulated tumours, see Figures 4.3, 4.4 and 4.5. Our results show that increase in yield of ctDNA is necessary to utilize liquid biopsy as a diagnostic tool. Utilization of continuous extra-corporeal blood separation techniques, apheresis, might be useful for isolation of larger quantities of ctDNA. One potential extension of this study could test how probabilities for detection of sub-clonal mutations from liquid biopsy sampling change with different modes of non-random DNA shedding patterns. This might enable us to use mutational profiles to test whether the liquid biopsy sample is well-mixed representation of tumour or it came from its specific specific segment. Experimental verification of our method could be done by isolating ctDNA from patients prior to surgical treatment. Genomic profiling of primary tumour tissue could be compared with mutational profile of ctDNA to check if our predictions for detection probability and frequency estimations are correct. Interestingly, deviations from theoretical expectations might provide us with insight into ctDNA shedding patterns and tumour behaviour.

In fifth chapter, I presented CancerSim, a software designed to simulate cancer growth and evolution. The software package was based and created from existing models used throughout the thesis in various forms. The main features of the model include spatial structure, variable mutation rates, addition of mutation with variable fitness effects, spatial sampling and simulation of sequencing. Software package presented in this state has a lot place for improvement. One of its major drawbacks is inability to simulate tumours with realistic cell numbers high turnover rate. It could be optimized for faster execution time by implementation in more efficient programming languages such

as C++ or by utilizing graphics processing unit (GPU) as it is faster in handling matrix and vector operations in comparison to the central processing unit (CPU). Optimization of the way mutations are stored in the cells would enable simulation of more division generations and greater turnover of cells. Simulation could be upgraded with addition of intratumour cellular migration and invasive behaviour against surrounding healthy tissue. Simulation of metastases seeding would provide good framework to study evolutionary relationship between primary tumour and metastases, or between different metastases. There are already several models that emphasise the importance of simulating migration features (Waclaw et al., 2015; Paterson et al., 2016).

The results presented in this thesis open many other interesting questions. From a theoretical perspective, it would be interesting to investigate the role of intratumour cell migration on the identification of truly clonal mutations. Migrating sub-clones would likely be present in more samples than in the absence of migration. This would likely introduce phylogenetic inconsistencies and make classification of mutations more difficult. Another possible extension of our model could include different non-random sampling patterns and sampling of samples of different sizes. Combination of single-cell samples and larger bulk samples might be more favourable option for identification of clonal mutations. Our method for estimation of the factor f, the most important parameter in our calculations, should be experimentally tested. By *in situ* labelling of the largest sub-clonal mutation one could estimate its abundance and spatial distribution in the tumour. This could further be used to measure the accuracy of our proposed method. More generally, broader experimental evaluation of computational spatial models is required. Since the distribution of sub-clones in simulated tumours is greatly determined by the implemented set of rules, it can turn out that the spatial genetic heterogeneity of simulated tumours is not an accurate representation of a real tumour. Unravelling this problem would be very informative for the modelling community. It would either further support or bring into question conclusions drawn from the spatial modelling. An improved understanding of the model fidelity would enable finer parametrization and smarter choice of rules for individual cell behaviour.

Finally and most importantly, the theoretical methods proposed in this

thesis could be tested in a clinical setting. One of our main assumptions is that targeting clonal alterations leads to better outcomes than targeting sub-clonal ones. However, it is still unknown whether patients whose molecular target we identify as clonal have better clinical outcome and improved quality of life in comparison to standard therapeutic protocols.

The goal of this thesis was to make a contribution towards translation of theoretical research to clinical medicine. This thesis tries to integrate mathematical methods, computational modelling and cancer genomics. Hopefully some of the work presented in this thesis will find a way into future clinical practice.

Bibliography

Abbosh, C., Birkbak, N.J., Wilson, G.A., Jamal-Hanjani, M., Constantin, T., Salari, R., Le Quesne, J., Moore, D.A., Veeriah, S., et al. Phylogenetic ctdna analysis depicts early-stage lung cancer evolution. *Nature*, 545(7655):446–+, 2017. (Cited on pages 63 and 64.)

Ahmadinejad, M., Hajimaghsoudi, L., Pouryaghobi, S.M., Ahmadinejad, I., Ahmadi, K. Diagnostic value of fine-needle aspiration biopsies and pathologic methods for benign and malignant breast masses and axillary node assessment. *Asian Pac J Cancer Prev*, 18(2):541–548, 2017. (Cited on page 11.)

Alix-Panabieres, C., Pantel, K. Opinion challenges in circulating tumour cell research. *Nat. Rev. Cancer*, 14:623–631, 2014. (Cited on page 63.)

Alix-Panabières, C., Pantel, K. Characterization of single circulating tumor cells. *FEBS Lett*, 591:2241–2250, 2017. (Cited on page 63.)

Altrock, P.M., Liu, L.L., Michor, F. The mathematics of cancer: integrating quantitative models. *Nat Rev Cancer*, 15(12):730–45, 2015a. (Cited on pages 14 and 15.)

Altrock, P.M., Liu, L.L., Michor, F. The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer*, 15(12):730–745, 2015b. (Cited on page 44.)

Alves, J.M., Prieto, T., Posada, D. Multiregional tumor trees are not phylogenies. *Trends Cancer*, 3(8):546–550, 2017. (Cited on page 48.)

American Joint Committee on Cancer. *AJCC Cancer Staging Manual*. Springer, 8th ed., 2018. (Cited on page 9.)

Anderson, A.R.A., Maini, P.K. Mathematical oncology. *Bulletin of Mathematical Biology*, 80(5):945–953, 2018. (Cited on page 44.)

Ang, J.M., Alai, N.N., Ritter, K.R., Machtinger, L.A. Muir-torre syndrome: case report and review of the literature. *Cutis*, 87(3):125–8, 2011. (Cited on page 45.)

Armitage, P., Doll, R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer*, 8:1–12, 1954. (Cited on pages 14 and 44.)

Armitage, P., Doll, R. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *British Journal of Cancer*, 11:161–169, 1957. (Cited on pages 14 and 44.)

Ashworth, T.R. A case of cancer in which cells similar to those in the tumours were seen in the blood after death. *Australian Medical Journal*, 14:146–147, 1869. (Cited on page 63.)

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173:371–385, 2018. (Cited on page 36.)

Barbosa, C., Trebosc, V., Kemmer, C., Rosenstiel, P., Beardmore, R., Schulenburg, H., Jansen, G. Alternative evolutionary paths to bacterial antibiotic resistance cause distinct collateral effects. *Mol Biol Evol*, 34(9):2229–2244, 2017. (Cited on page 7.)

Bell, G., Futuyma, D.J. Evolutionary rescue. *Annu. Rev. Ecol. Evol. Syst.*, 48:605–627, 2017. (Cited on page 7.)

Bozic, I., Gerold, J.M., Nowak, M.A. Quantifying clonal and subclonal passenger mutations in cancer evolution. *PLoS Comput Biol*, 12(2):e1004731, 2016. (Cited on page 90.)

Bozic, I., Paterson, C., Waclaw, B. On measuring selection in cancer from subclonal mutation frequencies. *bioRxiv*, 2019. (Cited on pages 57 and 60.)

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, 2018. (Cited on page 1.)

Brodeur, G.M., Seeger, R.C., Schwab, M., Varmus, H.E., Bishop, J.M. Amplification of n-myc in untreated human neuroblastomas correlates with advanced disease stage. *Science*, 224(4653):1121–4, 1984. (Cited on page 9.)

Brown, J.M., Attardi, L.D. The role of apoptosis in cancer development and treatment response. *Nature reviews cancer*, 5(2):231, 2005. (Cited on page 6.)

Bundred, N., Dixon, J.M. Carcinoma in situ. *BMJ*, 347, 2013. (Cited on page 3.)

Chapman, P.B., Hauschild, A., Robert, C., Haanen, J.B., Ascierto, P., Larkin,

J., Dummer, R., Garbe, C., Testori, A., et al. Improved survival with vemurafenib in melanoma with braf v600e mutation. *N Engl J Med*, 364:2507–2516, 2011. (Cited on page 37.)

Cheng, L., Zhang, S., Wang, L., MacLennan, G.T., Davidson, D.D. Fluorescence in situ hybridization in surgical pathology: principles and applications. *J Pathol Clin Res*, 3(2):73–99, 2017. (Cited on page 9.)

Chkhaidze, K., Heide, T., Werner, B., Williams, M.J., Huang, W., Caravagna, G., Graham, T.A., Sottoriva, A. Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. *bioRxiv*, 2019. (Cited on page 58.)

Coelho, M.C., Pinto, R.M., Murray, A.W. Heterozygous mutations cause genetic instability in a yeast model of cancer evolution. *Nature*, 566(7743):275–+, 2019. (Cited on page 6.)

Cohen, J.D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A.A., Wong, F., et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, 359(6378):926–930, 2018. (Cited on pages 12 and 64.)

Collins, V.P. Observation on growth rates of human tumors. *Am J Roentgenol*, 76:988–1000, 1956. (Cited on pages 14 and 15.)

Cox, N., Eedy, D., Morton, C. Guidelines for management of bowen's disease. *British Journal of Dermatology*, 141:633–641, 1999. (Cited on page 3.)

Croce, C.M. Oncogenes and cancer. *New England journal of medicine*, 358(5):502–511, 2008. (Cited on page 5.)

Cunningham, D., Humblet, Y., Siena, S., Khayat, D., Bleiberg, H., Santoro, A., Bets, D., Mueser, M., Harstrick, A., et al. Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. *New England Journal of Medicine*, 351:337–345, 2004. (Cited on page 19.)

Dagogo-Jack, I., Shaw, A.T. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15:81–94, 2018. (Cited on pages 19 and 63.)

Dang, C.V. Myc on the path to cancer. *Cell*, 149(1):22–35, 2012. (Cited on page 5.)

Davis, C., Naci, H., Gurpinar, E., Poplavska, E., Pinto, A., Aggarwal, A. Availability of evidence of benefits on overall survival and quality of life of cancer drugs approved by european medicines agency: retrospective cohort study of drug approvals 2009-13. *BMJ*, 359:j4530, 2017. (Cited on pages 1 and 89.)

de Bono, J.S., Scher, H.I., Montgomery, R.B., Parker, C., Miller, M.C., Tissing, H., Doyle, G.V., Terstappen, L.W.W.M., Pienta, K.J., et al. Circulating tumor cells predict survival benefit from treatment in metastatic castration-resistant prostate cancer. *Clin Cancer Res*, 14:6302–6309, 2008. (Cited on page 63.)

de Bruin, E.C., McGranahan, N., Mitter, R., Salm, M., Wedge, D.C., Yates, L., Jamal-Hanjani, M., Shafi, S., Murugaesu, N., et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*, 346:251–256, 2014. (Cited on pages 7, 20 and 63.)

Del Monte, U. Does the cell number 10(9) still really fit one gram of tumor tissue? *Cell Cycle*, 8:505–506, 2009. (Cited on page 6.)

Dewar, R., Fadare, O., Gilmore, H., Gown, A.M. Best practices in diagnostic immunohistochemistry: myoepithelial markers in breast pathology. *Arch Pathol Lab Med*, 135(4):422–9, 2011. (Cited on page 9.)

Diehl, F., Schmidt, K., Choti, M.A., Romans, K., Goodman, S., Li, M., Thornton, K., Agrawal, N., Sokoll, L., et al. Circulating mutant dna to assess tumor dynamics. *Nat Med*, 14(9):985–90, 2008. (Cited on pages 63, 65 and 78.)

Drilon, A., Wang, L., Hasanovic, A., Suehara, Y., Lipson, D., Stephens, P., Ross, J., Miller, V., Ginsberg, M., et al. Response to cabozantinib in patients with ret fusion-positive lung adenocarcinomas. *Cancer Discov*, 3:630–635, 2013. (Cited on page 37.)

Druker, B.J., Guilhot, F., O'Brien, S.G., Gathmann, I., Kantarjian, H., Gattermann, N., Deininger, M.W., Silver, R.T., Goldman, J.M., et al. Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *New England Journal of Medicine*, 355:2408–2417, 2006. (Cited on page 19.)

Duffy, M.J. Tumor markers in clinical practice: a review focusing on common solid cancers. *Med Princ Pract*, 22:4–11, 2013. (Cited on page 63.)

Dufresne, A., Brahmi, M., Karanian, M., Blay, J.Y. Using biology to guide the

treatment of sarcomas and aggressive connective-tissue tumours. *Nat Rev Clin Oncol*, 15:443–458, 2018. (Cited on page 37.)

Edge, S.B., Compton, C.C. The american joint committee on cancer: the 7th edition of the ajcc cancer staging manual and the future of tnm. *Ann Surg Oncol*, 17(6):1471–4, 2010. (Cited on page 8.)

Elenbaas, B., Weinberg, R.A. Heterotypic signaling between epithelial tumor cells and fibroblasts in carcinoma formation. *Experimental cell research*, 264(1):169–184, 2001. (Cited on page 5.)

Engell, H.C. Cancer cells in the circulating blood; a clinical study on the occurrence of cancer cells in the peripheral blood and in venous blood draining the tumour area at operation. *Acta chirurgica Scandinavica. Supplementum.*, 201:1–70, 1955. (Cited on page 63.)

Evans, E.K., Gardino, A.K., Kim, J.L., Hodous, B.L., Shutes, A., Davis, A., Zhu, X.J., Schmidt-Kittler, O., Wilson, D., et al. A precision therapy against cancers driven by kit/pdgfra mutations. *Science Translational Medicine*, 9:1–11, 2017. (Cited on page 37.)

Evans, M.F. The polymerase chain reaction and pathology practice. *Diagnostic Histopathology*, 15(7):344 – 356, 2009. Mini-Symposium: Cervical Cytology. (Cited on page 9.)

Fanning, D.M., Flood, H. Erythroplasia of queyrat. *Clin Pract*, 2(3):e63, 2012. (Cited on page 3.)

Favero, F., Joshi, T., Marquard, A.M., Birkbak, N.J., Krzystanek, M., Li, Q., Szallasi, Z., Eklund, A.C. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol*, 26(1):64–70, 2015. (Cited on page 50.)

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D., Piñeros, M., Znaor, A., Bray, F. Estimating the global cancer incidence and mortality in 2018: Globocan sources and methods. *International Journal of Cancer*, 144(8):1941–1953, 2019. (Cited on page 1.)

Fujimaki, K., Hattori, Y., Nakajima, H. 10-year complete remission in a philadelphia chromosome-positive acute lymphoblastic leukemia patient using imatinib without high-intensity chemotherapy or allogeneic stem cell

transplantation. *International Journal of Hematology*, 107:709–711, 2018. (Cited on page 19.)

Gale, D., Lawson, A.R.J., Howarth, K., Madi, M., Durham, B., Smalley, S., Calaway, J., Blais, S., Jones, G., et al. Development of a highly sensitive liquid biopsy platform to detect clinically-relevant cancer mutations at low allele fractions in cell-free dna. *PLoS One*, 13(3):e0194630, 2018. (Cited on pages 63 and 65.)

Gatenby, R.A., Maini, P.K. Mathematical oncology: cancer summed up. *Nature*, 421(6921):321, 2003. (Cited on page 44.)

Gatenby, R.A., Silva, A.S., Gillies, R.J., Frieden, B.R. Adaptive therapy. *Cancer Research*, 69(11):4894–4903, 2009. (Cited on page 65.)

Gerlee, P. The model muddle: in search of tumor growth laws. *Cancer Res*, 73(8):2407–11, 2013. (Cited on page 14.)

Gerlee, P., Anderson, A.R.A. An evolutionary hybrid cellular automaton model of solid tumour growth. *J Theor Biol*, 246(4):583–603, 2007. (Cited on page 15.)

Gerlinger, M., Horswell, S., Larkin, J., Rowan, A.J., Salm, M.P., Varela, I., Fisher, R., McGranahan, N., Matthews, N., et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet*, 46:225–233, 2014. (Cited on pages 7, 19, 20, 45 and 63.)

Gerlinger, M., Rowan, A.J., Horswell, S., Math, M., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med*, 366:883–892, 2012. (Cited on pages 7, 19, 20, 45 and 63.)

Giroux, V., Rustgi, A.K. Metaplasia: tissue injury adaptation and a precursor to the dysplasia-cancer sequence. *Nat. Rev. Cancer*, 17:594–604, 2017. (Cited on page 3.)

Gotlib, J. Tyrosine kinase inhibitors in the treatment of eosinophilic neoplasms and systemic mastocytosis. *Hematol Oncol Clin North Am*, 31:643–661, 2017. (Cited on page 37.)

Gozzetti, A., Le Beau, M.M. Fluorescence in situ hybridization: uses and limitations. *Semin Hematol*, 37(4):320–33, 2000. (Cited on page 10.)

Groisberg, R., Hong, D.S., Roszik, J., Janku, F., Tsimberidou, A.M., Javle,

M., Meric-Bernstam, F., Subbiah, V. Clinical next-generation sequencing for precision oncology in rare cancers. *Molecular cancer therapeutics*, 17(7):1595–1601, 2018. (Cited on pages 36 and 37.)

Gruber, M., Bozic, I., Leshchiner, I., Livitz, D., Stevenson, K., Rassenti, L., Rosebrock, D., Taylor-Weiner, A., Olive, O., et al. Growth dynamics in naturally progressing chronic lymphocytic leukaemia. *Nature*, 570(7762):474–479, 2019. (Cited on page 14.)

Hall, J.E. *Guyton and Hall Textbook of Medical Physiology*. Elsevier, 13th ed., 2016. (Cited on pages 3 and 4.)

Hart, D., Shochat, E., Agur, Z. The growth law of primary breast cancer as inferred from mammography screening trials data. *British journal of cancer*, 78(3):382, 1998. (Cited on page 14.)

Hawkes, N. Cancer survival data emphasise importance of early diagnosis. *BMJ*, 364:l408, 2019. (Cited on page 2.)

Heitzer, E., Haque, I.S., Roberts, C.E.S., Speicher, M.R. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nat Rev Genet*, 20:71–88, 2019. (Cited on pages 63, 65 and 91.)

Heitzer, E., Ulz, P., Geigl, J.B. Circulating tumor dna as a liquid biopsy for cancer. *Clin. Chem.*, 61:112–123, 2015. (Cited on page 63.)

Heldin, C.H. Autocrine pdgf stimulation in malignancies. *Ups J Med Sci*, 117(2):83–91, 2012. (Cited on page 5.)

Hench, I.B., Hench, J., Tolnay, M. Liquid biopsy in clinical management of breast, lung, and colorectal cancer. *Frontiers in Medicine*, 5:9, 2018. (Cited on pages 64 and 77.)

Hindson, B.J., Ness, K.D., Masquelier, D.A., Belgrader, P., Heredia, N.J., Makarewicz, A.J., Bright, I.J., Lucero, M.Y., Hiddessen, A.L., et al. High-throughput droplet digital pcr system for absolute quantitation of dna copy number. *Anal Chem*, 83(22):8604–10, 2011. (Cited on page 9.)

Hobbs, G.A., Der, C.J., Rossman, K.L. Ras isoforms and mutations in cancer at a glance. *Journal of Cell Science*, 129(7):1287–1292, 2016. (Cited on page 5.)

Hodi, F.S., O'Day, S.J., McDermott, D.F., Weber, R.W., Sosman, J.A., Haanen, J.B., Gonzalez, R., Robert, C., Schadendorf, D., et al. Improved survival

with ipilimumab in patients with metastatic melanoma. *New England Journal of Medicine*, 363:711–723, 2010. (Cited on page 19.)

Hoffman, J.M., Shah, N.D., Vermeulen, L.C., Doloresco, F., Martin, P.K., Blake, S., Matusiak, L., Hunkler, R.J., Schumock, G.T. Projecting future drug expenditures—2009. *American Journal of Health-System Pharmacy*, 66(3):237–257, 2009. (Cited on page 1.)

Hu, Z., Ding, J., Ma, Z., Sun, R., Seoane, J.A., Scott Shaffer, J., Suarez, C.J., Berghoff, A.S., Cremolini, C., et al. Quantitative evidence for early metastatic seeding in colorectal cancer. *Nat Genet*, 51(7):1113–1122, 2019. (Cited on page 44.)

Imyanitov, E., Sokolenko, A. Molecular diagnostics in clinical oncology. *Frontiers in molecular biosciences*, 5:76, 2018. (Cited on pages 9 and 10.)

Inaba, H., Greaves, M., Mullighan, C.G. Acute lymphoblastic leukaemia. *The Lancet*, 381(9881):1943–1955, 2013. (Cited on page 89.)

Jamal-Hanjani, M., Wilson, G.A., Horswell, S., Mitter, R., Sakarya, O., Constantin, T., Salari, R., Kirkizlar, E., Sigurjonsson, S., et al. Detection of ubiquitous and heterogeneous mutations in cell-free dna from patients with early-stage non-small-cell lung cancer. *Ann Oncol*, 27:862–867, 2016. (Cited on pages 63 and 65.)

Jamal-Hanjani, M., Wilson, G.A., McGranahan, N., Birkbak, N.J., Watkins, T.B.K., Veeriah, S., Shafi, S., Johnson, D.H., Mitter, R., et al. Tracking the evolution of non-small-cell lung cancer. *N Engl J Med*, 376:2109–2121, 2017. (Cited on page 20.)

Kamps, R., Brandão, R.D., Bosch, B.J.v.d., Paulussen, A.D.C., Xanthoulea, S., Blok, M.J., Romano, A. Next-generation sequencing in oncology: Genetic diagnosis, risk prediction and cancer classification. *Int J Mol Sci*, 18(2), 2017. (Cited on page 10.)

Khan, K.H., Cunningham, D., Werner, B., Vlachogiannis, G., Spiteri, I., Heide, T., Mateos, J.F., Vatsiou, A., Lampis, A., et al. Longitudinal liquid biopsy and mathematical modeling of clonal evolution forecast time to treatment failure in the prospect-c phase ii colorectal cancer clinical trial. *Cancer Discov*, 8:1270–1285, 2018. (Cited on pages 64, 65 and 91.)

Kim, S.W., Roh, J., Park, C.S. Immunohistochemistry for pathologists: Pro-

tocols, pitfalls, and tips. *J Pathol Transl Med*, 50(6):411–418, 2016. (Cited on page 9.)

Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903, 1969. (Cited on page 70.)

Kimyai-Asadi, A., Goldberg, L.H., Jih, M.H. Accuracy of serial transverse cross-sections in detecting residual basal cell carcinoma at the surgical margins of an elliptical excision specimen. *J Am Acad Dermatol*, 53(3):469–74, 2005. (Cited on page 10.)

Klein, E.A., Hubbell, E., Maddala, T., Aravanis, A., Beausang, J.F., Filippova, D., Gross, S., Jamshidi, A., Kurtzman, K., et al. Development of a comprehensive cell-free dna (cfdna) assay for early detection of multiple tumor types: The circulating cell-free genome atlas (ccga) study. *Journal of Clinical Oncology*, 36(15):12021–12021, 2018. (Cited on page 64.)

Knudson, A.G. Mutation and cancer: Statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences USA*, 68:820–823, 1971. (Cited on page 6.)

Komarova, N.L., Wodarz, D. Drug resistance in cancer: principles of emergence and prevention. *Proc Natl Acad Sci U S A*, 102:9714–9719, 2005. (Cited on page 19.)

Krimmel, J.D., Schmitt, M.W., Harrell, M.I., Agnew, K.J., Kennedy, S.R., Emond, M.J., Loeb, L.A., Swisher, E.M., Risques, R.A. Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic tp53 mutations in noncancerous tissues. *Proc Natl Acad Sci U S A*, 113:6005–6010, 2016. (Cited on page 37.)

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J.C., et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Research*, 44(11):e108–e108, 2016. (Cited on page 47.)

Le Tourneau, C., Delord, J.P., Gonçalves, A., Gavoille, C., Dubot, C., Isambert, N., Campone, M., Trédan, O., Massiani, M.A., et al. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (shiva): a multicentre, open-label, proof-of-concept, ran-

domised, controlled phase 2 trial. *Lancet Oncol*, 16:1324–1334, 2015. (Cited on pages 19 and 89.)

Lenaerts, T., Pacheco, J.M., Traulsen, A., Dingli, D. Tyrosine kinase inhibitor therapy can cure chronic myeloid leukemia without hitting leukemic stem cells. *Haematologica*, 95:900–907, 2010. (Cited on page 19.)

Leon, S.A., Shapiro, B., Sklaroff, D.M., Yaros, M.J. Free dna in the serum of cancer patients and the effect of therapy. *Cancer Res*, 37(3):646–650, 1977. (Cited on page 63.)

Li, J., Fu, W., Zhang, W., Li, P. High number of circulating tumor cells predicts poor survival of cutaneous melanoma patients in china. *Med Sci Monit*, 24:324–331, 2018. (Cited on page 78.)

Lindblom, A., Liljegren, A. Tumour markers in malignancies. *BMJ*, 320:424–427, 2000. (Cited on page 63.)

Ling, S., Hu, Z., Yang, Z., Yang, F., Li, Y., Lin, P., Chen, K., Dong, L., Cao, L., et al. Extremely high genetic diversity in a single tumor points to prevalence of non-darwinian cell evolution. *Proceedings of the National Academy of Sciences*, 112:E6496–E6505, 2015. (Cited on pages 20 and 45.)

Lloyd, M.C., Cunningham, J.J., Bui, M.M., Gilles, R.J., Brown, J.S., Gatenby, R.A. Darwinian dynamics of intratumoral heterogeneity: Not solely random mutations but also variable environmental selection forces. *Cancer Research*, 76:3136–3144, 2016. (Cited on page 25.)

Lo, Y.M., Zhang, J., Leung, T.N., Lau, T.K., Chang, A.M., Hjelm, N.M. Rapid clearance of fetal dna from maternal plasma. *Am J Hum Genet*, 64(1):218–224, 1999. (Cited on page 65.)

Long, G.V., Weber, J.S., Infante, J.R., Kim, K.B., Daud, A., Gonzalez, R., Sosman, J.A., Hamid, O., Schuchter, L., et al. Overall survival and durable responses in patients with braf v600–mutant metastatic melanoma receiving dabrafenib combined with trametinib. *Journal of Clinical Oncology*, 34:871–878, 2016. (Cited on page 19.)

Lu, T., Yang, X., Huang, Y., Zhao, M., Li, M., Ma, K., Yin, J., Zhan, C., Wang, Q. Trends in the incidence, treatment, and survival of patients with lung cancer in the last four decades. *Cancer Manag Res*, 11:943–953, 2019. (Cited on page 89.)

Luengo-Fernandez, R., Leal, J., Gray, A., Sullivan, R. Economic burden of cancer across the european union: a population-based cost analysis. *Lancet Oncol*, 14(12):1165–74, 2013. (Cited on page 1.)

Lynch, M., Ackerman, M.S., Gout, J.F., Long, H., Sung, W., Thomas, W.K., Foster, P.L. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11):704–714, 2016. (Cited on page 7.)

Mackie, F., Hemming, K., Allen, S., Morris, R., Kilby, M. The accuracy of cell-free fetal dna-based non-invasive prenatal testing in singleton pregnancies: a systematic review and bivariate meta-analysis. *BJOG: An International Journal of Obstetrics & Gynaecology*, 124(1):32–46, 2017. (Cited on page 77.)

Mandel, P., Métais, P. Les acides nucléiques du plasma sanguin chez l'homme. *Comptes Rendus des Seances de la Societe de Biologie et de ses Filiales*, 142:241–243, 1948. (Cited on page 63.)

Manolio, T.A., Rowley, R., Williams, M.S., Roden, D., Ginsburg, G.S., Bult, C., Chisholm, R.L., Deverka, P.A., McLeod, H.L., et al. Opportunities, resources, and techniques for implementing genomics in clinical care. *Lancet*, 394(10197):511–520, 2019. (Cited on page 51.)

Martincorena, I., Campbell, P.J. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015. (Cited on page 45.)

Mazzaferri, E.L. Management of a solitary thyroid nodule. *New England Journal of Medicine*, 328(8):553–559, 1993. PMID: 8426623. (Cited on page 11.)

McCulloch, S.D., Kunkel, T.A. The fidelity of dna synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res*, 18(1):148–61, 2008. (Cited on page 45.)

McGranahan, N., Furness, A.J.S., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S.K., Jamal-Hanjani, M., Wilson, G.A., Birkbak, N.J., et al. Clonal neoantigens elicit t cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*, 351:1463–1469, 2016. (Cited on pages 19 and 45.)

McGranahan, N., Swanton, C. Clonal heterogeneity and tumor evolution: Past, present, and the future. *Cell*, 168:613–628, 2017. (Cited on pages 7, 20 and 63.)

Mehrara, E., Forssell-Aronsson, E., Ahlman, H., Bernhardt, P. Specific growth

rate versus doubling time for quantitative characterization of tumor growth rate. *Cancer Res*, 67(8):3970–5, 2007. (Cited on page 14.)

Meropol, N.J., Schrag, D., Smith, T.J., Mulvey, T.M., Langdon, R.M., Blum, D., Ubel, P.A., Schnipper, L.E. American society of clinical oncology guidance statement: The cost of cancer care. *Journal of Clinical Oncology*, 27(23):3868–3874, 2009. PMID: 19581533. (Cited on page 1.)

Meyerson, M., Gabriel, S., Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, 11:685–696, 2010. (Cited on page 20.)

Mitri, Z., Constantine, T., O'Regan, R. The her2 receptor in breast cancer: Pathophysiology, clinical use, and new advances in therapy. *Chemother Res Pract*, 2012:743193, 2012. (Cited on page 9.)

Moran, P.A.P. Random processes in genetics. *Proceedings of the Cambridge Philosophical Society*, 54:60–71, 1958. (Cited on page 15.)

Neal, R.D., Tharmanathan, P., France, B., Din, N.U., Cotton, S., Fallon-Ferguson, J., Hamilton, W., Hendry, A., Hendry, M., et al. Is increased time to diagnosis and treatment in symptomatic cancer associated with poorer outcomes? systematic review. *Br J Cancer*, 112 Suppl 1:S92–107, 2015. (Cited on page 2.)

Nichol, D., Jeavons, P., Fletcher, A.G., Bonomo, R.A., Maini, P.K., Paul, J.L., Gatenby, R.A., Anderson, A.R.A., Scott, J.G. Steering evolution with sequential therapy to prevent the emergence of bacterial antibiotic resistance. *PLoS Comput Biol*, 11:e1004493, 2015. (Cited on pages 7 and 38.)

Nicholson, A.M., Graham, T.A., Simpson, A., Humphries, A., Burch, N., Rodriguez-Justo, M., Novelli, M., Harrison, R., Wright, N.A., et al. Barrett's metaplasia glands are clonal, contain multiple stem cells and share a common squamous progenitor. *Gut*, 61(10):1380–1389, 2012. (Cited on page 3.)

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149:979–993, 2012. (Cited on page 20.)

Norgauer, J., Idzko, M., Panther, E., Hellstern, O., Herouy, Y. Xeroderma pigmentosum. *Eur. J. Dermatol.*, 13(1):4–9, 2003. (Cited on page 45.)

Nowell, P. The clonal evolution of tumour cell populations. *Science*, 194:23–28, 1976. (Cited on page 45.)

O'Bryan, J.P. Pharmacological targeting of ras: Recent success with direct inhibitors. *Pharmacol Res*, 139:503–511, 2019. (Cited on page 5.)

Opasic, L., Zhou, D., Werner, B., Dingli, D., Traulsen, A. How many samples are needed to infer truly clonal mutations from heterogenous tumours? *BMC Cancer*, 19:403, 2019. (Cited on pages 17, 47, 51 and 70.)

Pál, C., Papp, B., Lázár, V. Collateral sensitivity of antibiotic-resistant microbes. *Trends Microbiol*, 23(7):401–7, 2015. (Cited on page 7.)

Papanicolaou, G.N., Traut, H.F. The diagnostic value of vaginal smears in carcinoma of the uterus. 1941. *Archives of pathology & laboratory medicine*, 121(3):211, 1997. (Cited on page 11.)

Pastorino, U., Silva, M., Sestini, S., Sabia, F., Boeri, M., Cantarutti, A., Sverzellati, N., Sozzi, G., Corrao, G., et al. Prolonged lung cancer screening reduced 10-year mortality in the mild trial: new confirmation of lung cancer screening efficacy. *Ann Oncol*, 30(7):1162–1169, 2019. (Cited on page 2.)

Paterson, C., Nowak, M.A., Waclaw, B. An exactly solvable, spatial model of mutation accumulation in cancer. *Sci Rep*, 6:39511, 2016. (Cited on page 93.)

Pellettieri, J., Sánchez Alvarado, A. Cell turnover and adult tissue homeostasis: from humans to planarians. *Annu Rev Genet*, 41:83–105, 2007. (Cited on page 3.)

Piva de Freitas, P., Senna, C.G., Tabai, M., Chone, C.T., Altemani, A. Metastatic basal cell carcinoma: A rare manifestation of a common disease. *Case Rep Med*, 2017:8929745, 2017. (Cited on page 4.)

Popic, V., Salari, R., Hajirasouliha, I., Kashef-Haghighi, D., West, R.B., Batzoglou, S. Fast and scalable inference of multi-sample cancer lineages. *Genome Biol*, 16:91, 2015. (Cited on page 47.)

Rack, B., Schindlbeck, C., Jückstock, J., Andergassen, U., Hepp, P., Zwingers, T., Friedl, T.W.P., Lorenz, R., Tesch, H., et al. Circulating tumor cells predict survival in early average-to-high risk breast cancer patients. *J Natl Cancer Inst*, 106, 2014. (Cited on page 63.)

Rasche, L., Chavan, S.S., Stephens, O.W., Patel, P.H., Tytarenko, R., Ashby, C., Bauer, M., Stein, C., Deshpande, S., et al. Spatial genomic heterogeneity in multiple myeloma revealed by multi-region sequencing. *Nat Commun*, 8:268, 2017. (Cited on pages 7, 20 and 63.)

Reiter, J.G., Baretti, M., Gerold, J.M., Makohon-Moore, A.P., Daud, A., Iacobuzio-Donahue, C.A., Azad, N.S., Kinzler, K.W., Nowak, M.A., et al. An analysis of genetic heterogeneity in untreated cancers. *Nat Rev Cancer*, 2019. (Cited on pages 7, 8, 48, 54 and 91.)

Riethdorf, S., Fritsche, H., Mueller, V., Rau, T., Schindibeck, C., Rack, B., Janni, W., Coith, C., Beck, K., et al. Detection of circulating tumor cells in peripheral blood of patients with metastatic breast cancer: A validation study of the cellsearch system. *Clin. Cancer Res.*, 13:920–928, 2007. (Cited on page 63.)

Rockne, R.C., Hawkins-Daarud, A., Swanson, K.R., Sluka, J.P., Glazier, J.A., Macklin, P., Hormuth, D.A., Jarrett, A.M., Lima, E.A.B.F., et al. The 2019 mathematical oncology roadmap. *Phys Biol*, 16(4):041005, 2019. (Cited on page 13.)

Rockne, R.C., Scott, J.G. Introduction to mathematical oncology. *JCO Clinical Cancer Informatics*, (3):1–4, 2019. PMID: 31026176. (Cited on page 13.)

Rodriguez-Brenes, I.A., Komarova, N.L., Wodarz, D. Tumor growth dynamics: insights into evolutionary processes. *Trends in ecology & evolution*, 28(10):597–604, 2013. (Cited on page 14.)

Saarenheimo, J., Eigeliene, N., Andersen, H., Tiirola, M., Jekunen, A. The value of liquid biopsies for guiding therapy decisions in non-small cell lung cancer. *Frontiers in Oncology*, 9:129, 2019. (Cited on page 77.)

Sacher, A.G., Paweletz, C., Dahlberg, S.E., Alden, R.S., O'Connell, A., Feeney, N., Mach, S.L., Jänne, P.A., Oxnard, G.R. Prospective validation of rapid plasma genotyping for the detection of egfr and kras mutations in advanced lung cancer. *JAMA Oncol*, 2(8):1014–22, 2016. (Cited on page 64.)

Salgia, R., Kulkarni, P. The genetic/non-genetic duality of drug 'resistance' in cancer. *Trends Cancer*, 4:110–118, 2018. (Cited on page 19.)

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W.K., Luna, A., La, K.C., Dimitriadoy, S., Liu, D.L., Kantheti, H.S., et al. Oncogenic signaling path-

ways in the cancer genome atlas. *Cell*, 173:321–337, 2018. (Cited on pages 5 and 36.)

Sante, G.D., Pagé, J., Jiao, X., Nawab, O., Cristofanilli, M., Skordalakes, E., Pestell, R.G. Recent advances with cyclin-dependent kinase inhibitors: therapeutic agents for breast cancer and their role in immuno-oncology. *Expert Review of Anticancer Therapy*, 19(7):569–587, 2019. PMID: 31219365. (Cited on page 5.)

Savage, D.G., Antman, K.H. Imatinib mesylate — a new oral targeted therapy. *New England Journal of Medicine*, 346:683–693, 2002. (Cited on pages 10 and 19.)

Schirrmacher, V. From chemotherapy to biological therapy: A review of novel concepts to reduce the side effects of systemic cancer treatment (review). *Int J Oncol*, 54(2):407–419, 2019. (Cited on page 1.)

Schütz, E., Fischer, A., Beck, J., Harden, M., Koch, M., Wuensch, T., Stockmann, M., Nashan, B., Kollmar, O., et al. Graft-derived cell-free dna, a noninvasive early rejection and graft damage marker in liver transplantation: A prospective, observational, multicenter cohort study. *PLOS Medicine*, 14(4):1–19, 2017. (Cited on page 77.)

Schwartz, R., Schaeffer, A.A. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18:213–229, 2017. (Cited on page 20.)

Schwarze, K., Buchanan, J., Taylor, J.C., Wordsworth, S. Are whole-exome and whole-genome sequencing approaches cost-effective? a systematic review of the literature. *Genet Med*, 20(10):1122–1130, 2018. (Cited on page 51.)

Scott, J.G., Maini, P.K., Anderson, A.R., Fletcher, A.G. Inferring tumour proliferative organisation from phylogenetic tree measures in a computational model. *Systematic biology*, 1–41, 2019. (Cited on page 63.)

Sehgal, R., Sheahan, K., O'Connell, P.R., Hanly, A.M., Martin, S.T., Winter, D.C. Lynch syndrome: an updated review. *Genes (Basel)*, 5(3):497–507, 2014. (Cited on page 45.)

Servedio, M.R., Brandvain, Y., Dhole, S., Fitzpatrick, C.L., Goldberg, E.E., Stern, C.A., Van Cleve, J., Yeh, D.J. Not just a theory—the utility of mathematical models in evolutionary biology. *PLoS biology*, 12(12):e1002017, 2014. (Cited on pages 13 and 15.)

Shackney, S.E., McCormack, G.W., Cuchural, G.J. Growth rate patterns of solid tumors and their relation to responsiveness to therapy: an analytical review. *Annals of internal medicine*, 89(1):107–121, 1978. (Cited on page 14.)

Sharma, P., Hu-Lieskovan, S., Wargo, J.A., Ribas, A. Primary, adaptive, and acquired resistance to cancer immunotherapy. *Cell*, 168:707–723, 2017. (Cited on page 19.)

Shaw, A.T., Engelman, J.A. Alk in lung cancer: past, present, and future. *J Clin Oncol*, 31(8):1105–11, 2013. (Cited on page 5.)

Siegmund, K., Shibata, D. At least two well-spaced samples are needed to genotype a solid tumor. *BMC Cancer*, 16:250, 2016. (Cited on pages 20, 51 and 90.)

Singer, S. *Psychosocial Impact of Cancer*, 1–11. Springer International Publishing, Cham, 2018. (Cited on page 1.)

Siravegna, G., Marsoni, S., Siena, S., Bardelli, A. Integrating liquid biopsies into the management of cancer. *Nat Rev Clin Oncol*, 14:531–548, 2017. (Cited on page 63.)

Sonnenberg, A., Marciniak, J.Y., Rassenti, L., Ghia, E.M., Skowronski, E.A., Manouchehri, S., McCanna, J., Widhopf, 2nd, G.F., Kipps, T.J., et al. Rapid electrokinetic isolation of cancer-related circulating cell-free dna directly from blood. *Clin Chem*, 60:500–509, 2014. (Cited on page 63.)

Sottoriva, A., Kang, H., Ma, Z., Graham, T.A., Salomon, M.P., Zhao, J., Marjoram, P., Siegmund, K., Press, M.F., et al. A big bang model of human colorectal tumor growth. *Nature Genetics*, 47:209–216, 2015. (Cited on pages 19, 20 and 44.)

Sottoriva, A., Spiteri, I., Piccirillo, S.G.M., Touloumis, A., Collins, V.P., Marioni, J.C., Curtis, C., Watts, C., Tavaré, S. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci U S A*, 110:4009–4014, 2013. (Cited on page 20.)

Stanková, K., Brown, J.S., Dalton, W.S., Gatenby, R.A. Optimizing cancer treatment using game theory: A review. *JAMA Oncol*, 5(1):96–103, 2019. (Cited on page 44.)

Stroun, M., Anker, P., Maurice, P., Lyautey, J., Lederrey, C., Beljanski, M.

Neoplastic characteristics of the dna found in the plasma of cancer patients. *Oncology*, 46:318–322, 1989. (Cited on page 63.)

Sugita, S., Hasegawa, T. Practical use and utility of fluorescence in situ hybridization in the pathological diagnosis of soft tissue and bone tumors. *Journal of Orthopaedic Science*, 22(4):601 – 612, 2017. (Cited on page 9.)

Sun, R., Hu, Z., Sottoriva, A., Graham, T.A., Harpak, A., Ma, Z., Fischer, J.M., Shibata, D., Curtis, C. Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nature Genetics*, 49:1015–1024, 2017. (Cited on pages 20, 25, 45, 55 and 57.)

Sun, Z., Aubry, M.C., Deschamps, C., Marks, R.S., Okuno, S.H., Williams, B.A., Sugimura, H., Pankratz, V.S., Yang, P. Histologic grade is an independent prognostic factor for survival in non–small cell lung cancer: An analysis of 5018 hospital- and 712 population-based cases. *The Journal of Thoracic and Cardiovascular Surgery*, 131(5):1014 – 1020, 2006. (Cited on page 8.)

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, 156(6):1324–1335, 2014. (Cited on page 58.)

Syn, N.L., Teng, M.W.L., Mok, T.S.K., Soo, R.A. De-novo and acquired resistance to immune checkpoint targeting. *Lancet Oncol*, 18:e731–e741, 2017. (Cited on page 19.)

Tainio, K., Athanasiou, A., Tikkinen, K.A.O., Aaltonen, R., Cárdenas, J., Hernándes, Glazer-Livson, S., Jakobsson, M., Joronen, K., et al. Clinical course of untreated cervical intraepithelial neoplasia grade 2 under active surveillance: systematic review and meta-analysis. *BMJ*, 360, 2018. (Cited on page 3.)

Talevich, E., Shain, A.H., Botton, T., Bastian, B.C. Cnvkit: Genome-wide copy number detection and visualization from targeted dna sequencing. *PLoS Comput Biol*, 12(4):e1004873, 2016. (Cited on page 49.)

Thress, K.S., Paweletz, C.P., Felip, E., Cho, B.C., Stetson, D., Dougherty, B., Lai, Z., Markovets, A., Vivancos, A., et al. Acquired egfr c797s mutation mediates resistance to azd9291 in non-small cell lung cancer harboring egfr t790m. *Nat Med*, 21(6):560–2, 2015. (Cited on page 78.)

Tie, J., Wang, Y., Tomasetti, C., Li, L., Springer, S., Kinde, I., Silliman, N.,

Tacey, M., Wong, H.L., et al. Circulating tumor dna analysis detects minimal residual disease and predicts recurrence in patients with stage ii colon cancer. *Sci Transl Med*, 8:346ra92, 2016. (Cited on page 65.)

Turajlic, S., Sottoriva, A., Graham, T., Swanton, C. Resolving genetic heterogeneity in cancer. *Nature Reviews Genetics*, 20(7):404–416, 2019. (Cited on pages 7, 15 and 56.)

Uecker, H. Evolutionary rescue in randomly mating, selfing, and clonal populations. *Evolution*, 71(4):845–858, 2017. (Cited on page 7.)

Vanneman, M., Dranoff, G. Combining immunotherapy and targeted therapies in cancer treatment. *Nature Reviews Cancer*, 12:237–251, 2012. (Cited on page 19.)

Venkatesan, S., Swanton, C., Taylor, B.S., Costello, J.F. Treatment-induced mutagenesis and selective pressures sculpt cancer evolution. *Cold Spring Harb Perspect Med*, 7(8), 2017. (Cited on page 56.)

Vinay Kumar, Abul K. Abbas, J.C.A. *Robbins basic pathology*. Robbins Pathology. Elsevier, 10th ed., 2017. (Cited on pages 3, 4, 8, 10 and 11.)

Waclaw, B., Bozic, I., Pittman, M.E., Hruban, R.H., Vogelstein, B., Nowak, M.A. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature*, 525:261–264, 2015. (Cited on pages 15, 24, 33, 35 and 93.)

Wall, D.P., Tonellato, P.J. The future of genomics in pathology. *F1000 medicine reports*, 4, 2012. (Cited on page 8.)

Wang, K., Li, M., Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164, 2010. (Cited on page 47.)

Wang, Y., Schmid-Bindert, G., Zhou, C. Erlotinib in the treatment of advanced non-small cell lung cancer: an update for clinicians. *Ther Adv Med Oncol*, 4(1):19–29, 2012. (Cited on page 51.)

Warnakulasuriya, S., Reibel, J., Bouquot, J., Dabelsteen, E. Oral epithelial dysplasia classification systems: predictive value, utility, weaknesses and scope for improvement. *Journal of Oral Pathology & Medicine*, 37(3):127–133, 2008. (Cited on page 3.)

Weinberg, R.A. The retinoblastoma protein and cell cycle control. *Cell*, 81(3):323–330, 1995. (Cited on page 6.)

Werner, B., Case, J., Williams, M.J., Chkhaidze, K., Temko, D., Fernandez-Mateos, J., Cresswell, G.D., Nichol, D., Cross, W., et al. Measuring single cell divisions in human cancers from multi-region sequencing data. *bioRxiv*, 2019. (Cited on page 6.)

Werner, B., Traulsen, A., Sottoriva, A., Dingli, D. Detecting truly clonal alterations from multi-region profiling of tumours. *Scientific Reports*, 7:44991, 2017. (Cited on pages 20, 21, 38, 47, 51, 72 and 90.)

West, J.B., Dinh, M.N., Brown, J.S., Zhang, J., Anderson, A.R., Gatenby, R.A. Multidrug cancer therapy in metastatic castrate-resistant prostate cancer: An evolution-based strategy. *Clinical Cancer Research*, 2019. (Cited on page 78.)

Williams, M.J., Sottoriva, A., Graham, T.A. Measuring clonal evolution in cancer with genomics. *Annu Rev Genomics Hum Genet*, 2019a. (Cited on pages 57 and 58.)

Williams, M.J., Werner, B., Barnes, C.P., Graham, T.A., Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nature Genetics*, 48:238–244, 2016. (Cited on pages 7, 20, 25, 45, 49, 50, 52, 54, 55, 56, 57, 69, 70 and 76.)

Williams, M.J., Werner, B., Heide, T., Curtis, C., Barnes, C.P., Sottoriva, A., Graham, T.A. Quantification of subclonal selection in cancer from bulk sequencing data. *Nat Genet*, 50(6):895–903, 2018. (Cited on pages 45, 55, 56 and 70.)

Williams, M.J., Zapata, L., Werner, B., Barnes, C., Sottoriva, A., Graham, T.A. Measuring the distribution of fitness effects in somatic evolution by combining clonal dynamics with dn/ds ratios. *bioRxiv*, 2019b. (Cited on pages 49 and 58.)

Wodarz, D., Komarova, N.L. *Dynamics of cancer: mathematical foundations of oncology*. World Scientific, 2014. (Cited on pages 14 and 15.)

Wood, H.M., Daly, C., Chalkley, R., Senguven, B., Ross, L., Egan, P., Chengot, P., Graham, J., Sethi, N., et al. The genomic road to invasion-examining the

similarities and differences in the genomes of associated oral pre-cancer and cancer samples. *Genome Med.*, 9(53), 2017. (Cited on pages 3, 56 and 57.)

Xu, L.L., Yang, Y., Wang, Z., Wang, X.J., Tong, Z.H., Shi, H.Z. Malignant pleural mesothelioma: diagnostic value of medical thoracoscopy and long-term prognostic analysis. *BMC Pulm Med*, 18(1):56, 2018. (Cited on page 11.)

Xu, X.L., Singh, H.P., Wang, L., Qi, D.L., Poulos, B.K., Abramson, D.H., Jhanwar, S.C., Cobrinik, D. Rb suppresses human cone-precursor-derived retinoblastoma tumours. *Nature*, 514(7522):385–388, 2014. (Cited on page 6.)

Zhang, B.O., Xu, C.W., Shao, Y., Wang, H.T., Wu, Y.F., Song, Y.Y., Li, X.B., Zhang, Z., Wang, W.J., et al. Comparison of droplet digital pcr and conventional quantitative pcr for measuring egfr gene mutation. *Exp Ther Med*, 9(4):1383–1388, 2015. (Cited on page 9.)

Zhang, J., Cunningham, J.J., Brown, J.S., Gatenby, R.A. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. *Nat Commun*, 8(1):1816, 2017. (Cited on pages 44 and 78.)

Zhang, L., Riethdorf, S., Wu, G., Wang, T., Yang, K., Peng, G., Liu, J., Pantel, K. Meta-analysis of the prognostic value of circulating tumor cells in breast cancer. *Clin. Cancer Res.*, 18:5701–5710, 2012. (Cited on page 63.)

Zhao, E.Y., Jones, M., Jones, S.J. Whole-genome sequencing in cancer. *Cold Spring Harbor perspectives in medicine*, 9(3):a034579, 2019. (Cited on page 51.)

Supplementary Figures

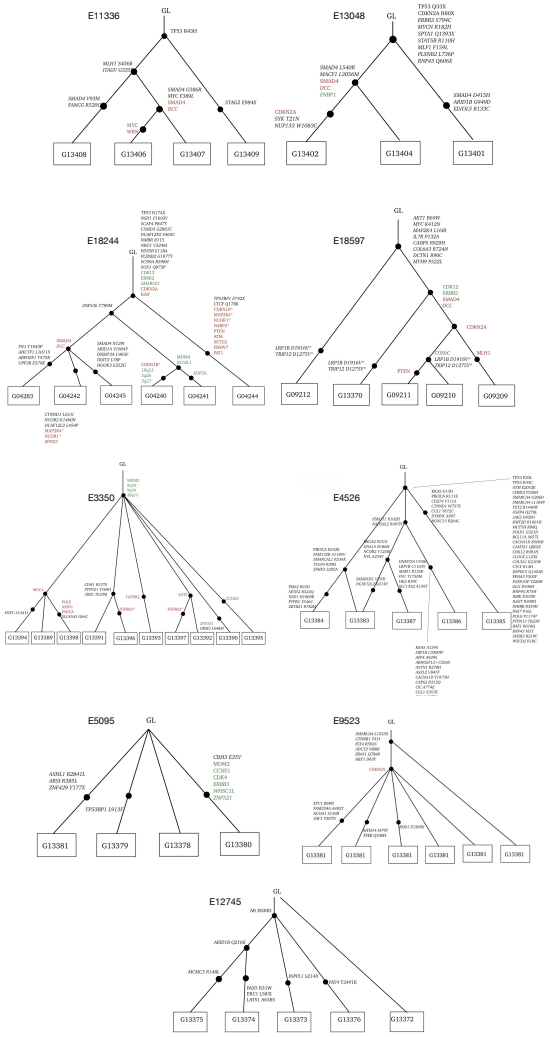

Figure 6.1: Phylogenetic trees reconstructed by our collaborators based on the single nucleotides and copy number variations using the LiCHE software. Branch defining alterations are used for our clonality analysis.

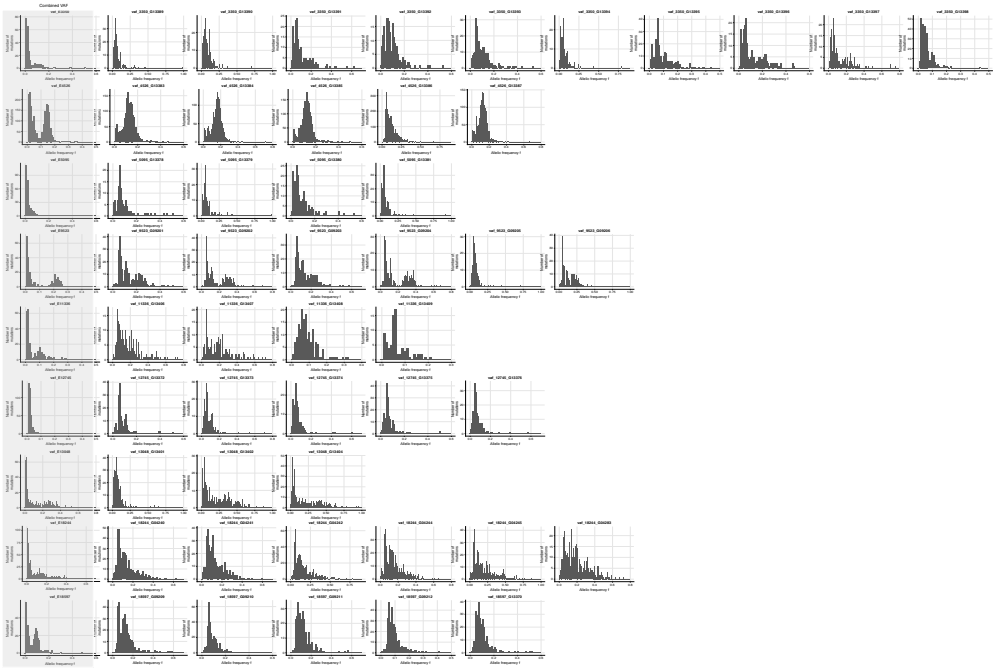

Figure 6.2: Output of the sequencing data presented as a histogram of variant allelic frequencies (VAF histogram) for each sample within patients tumour. The first column (marked grey) shows combined allelic frequencies of all samples combined.

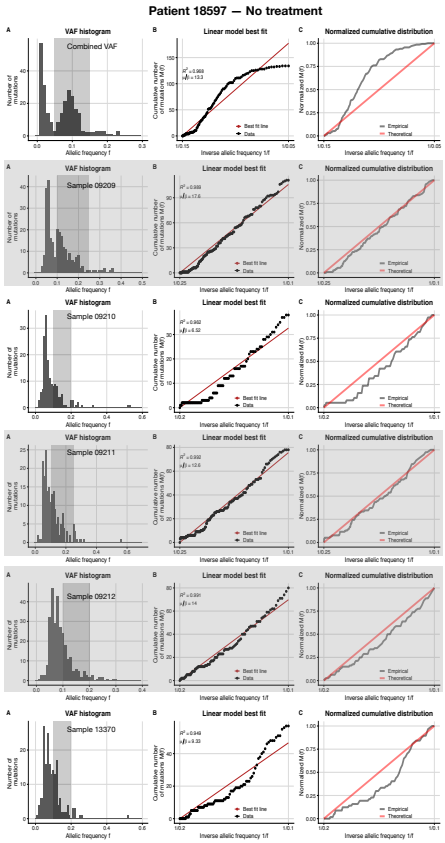

Figure 6.3: Neutrality analysis of samples from one patient. Panel A show variant allele frequency histogram. Dark grey shade marks interval used for comparison with the neutral model. Panel B show increment in cumulative number of mutation with inverse allelic frequency $1/f$ (black dots) and linear model best fit (red line). Light grey marks samples that are in agreement with the neutral model $R^2 \geq 0.98$. Panel C Shows normalized cumulative distribution of mutations and theoretical model. Distance between distributions was quantified using Kolmogorov-Smirnov test.

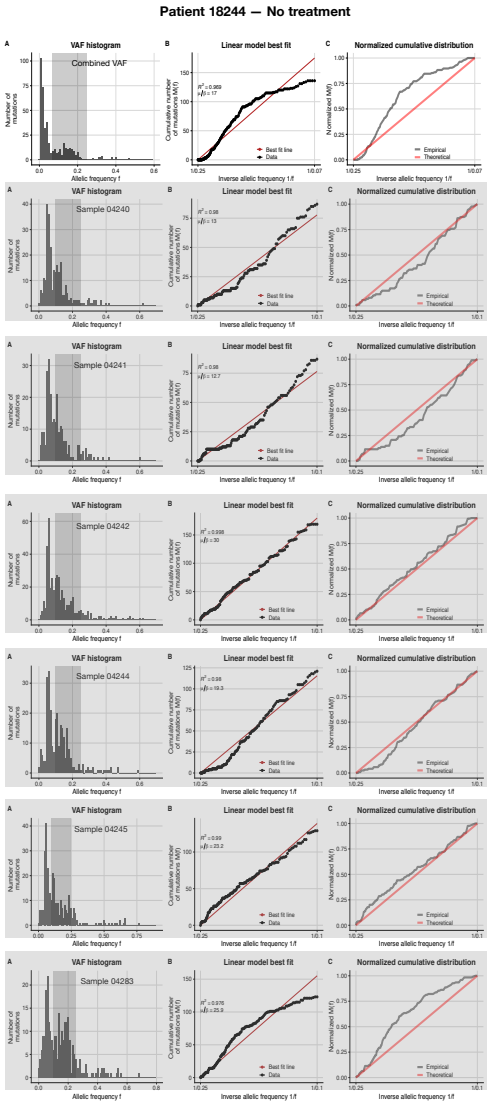

Figure 6.4: Neutrality analysis of samples from a patient.

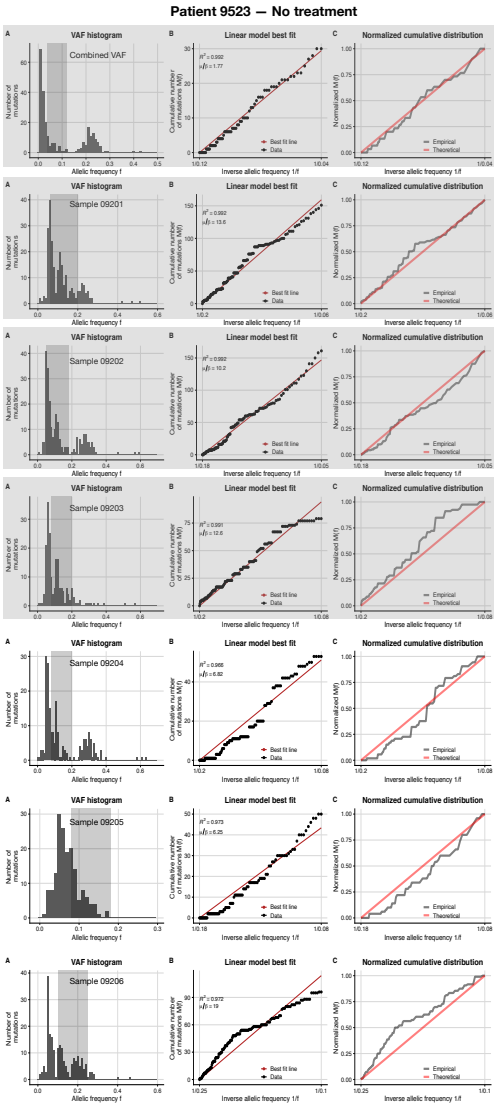

Figure 6.5: Neutrality analysis of samples from a patient.

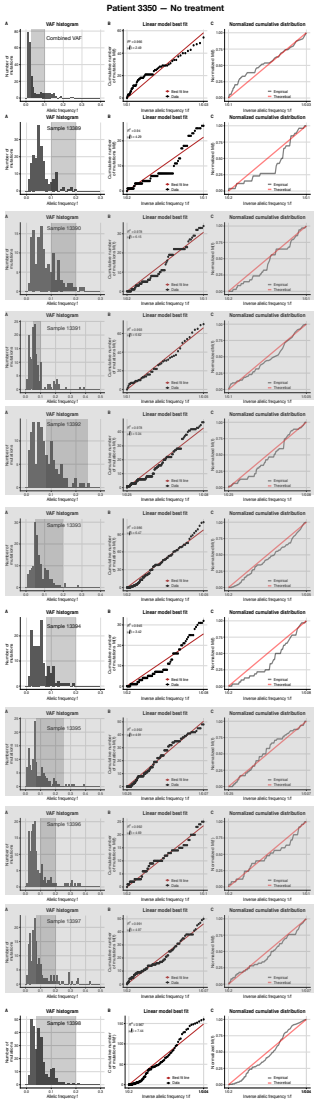

Figure 6.6: Neutrality analysis of samples from a patient.

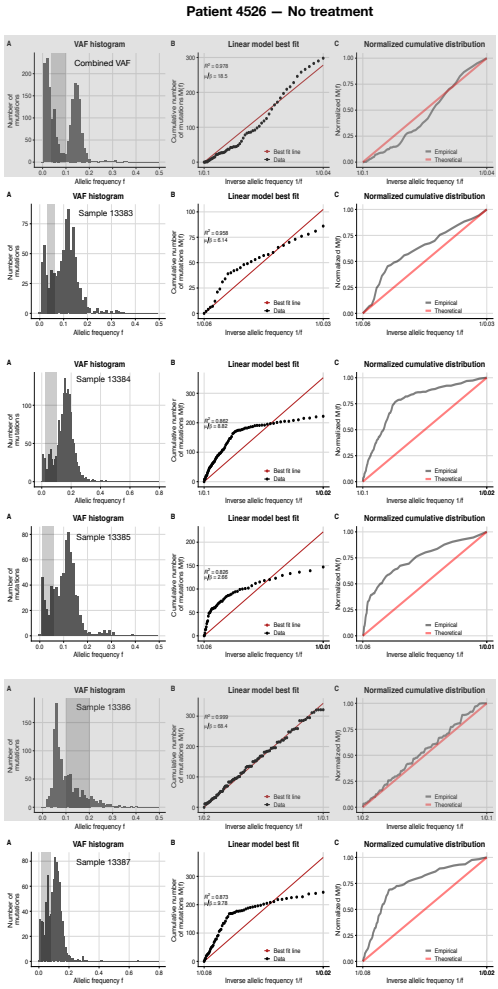

Figure 6.7: Neutrality analysis of samples from a patient. This patient has a microsatellite instability which leads to accumulation of a large number of mutations. Large clonal peak obfuscates the distribution of sub-clonal mutations and makes a fit to neutral model problematic.

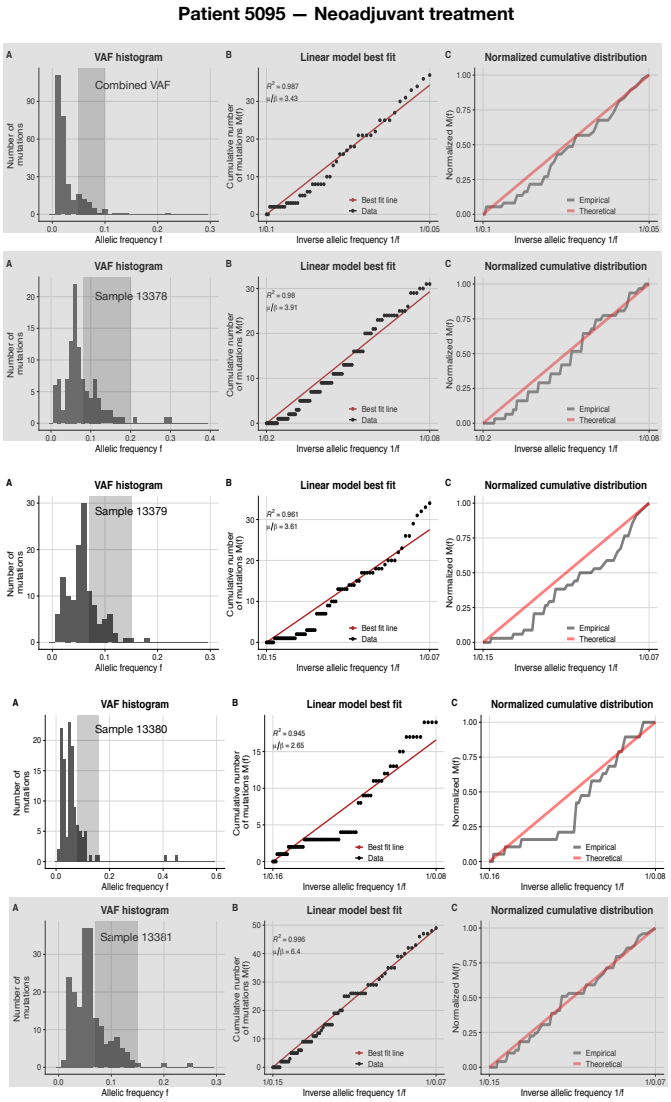

Figure 6.8: Neutrality analysis of samples from a patient.

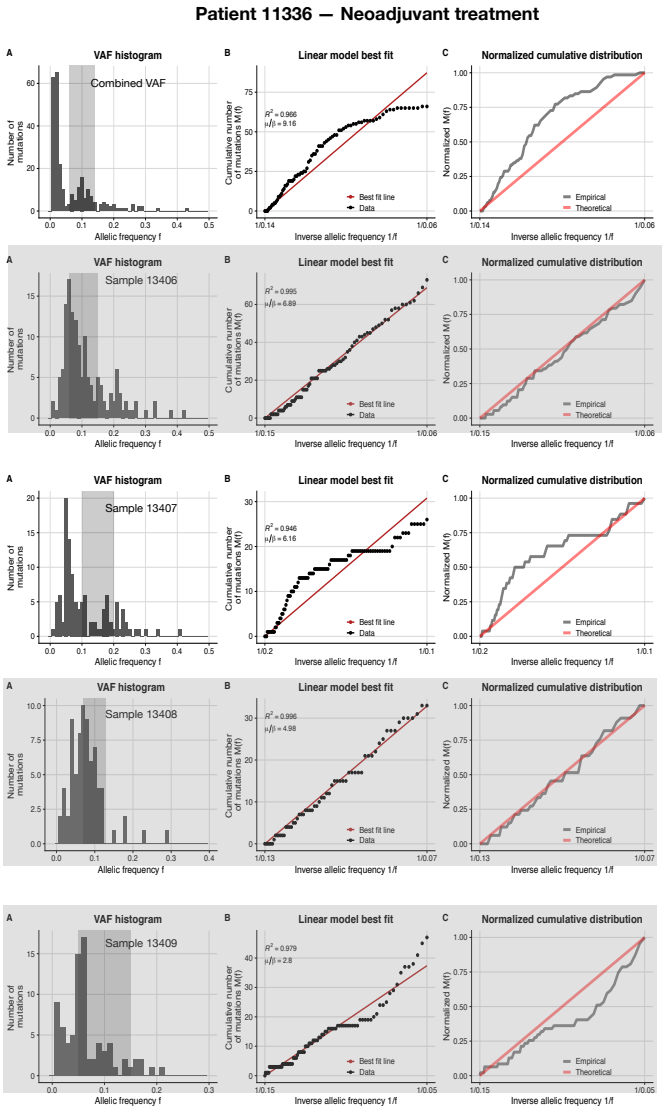

Figure 6.9: Neutrality analysis of samples from a patient.

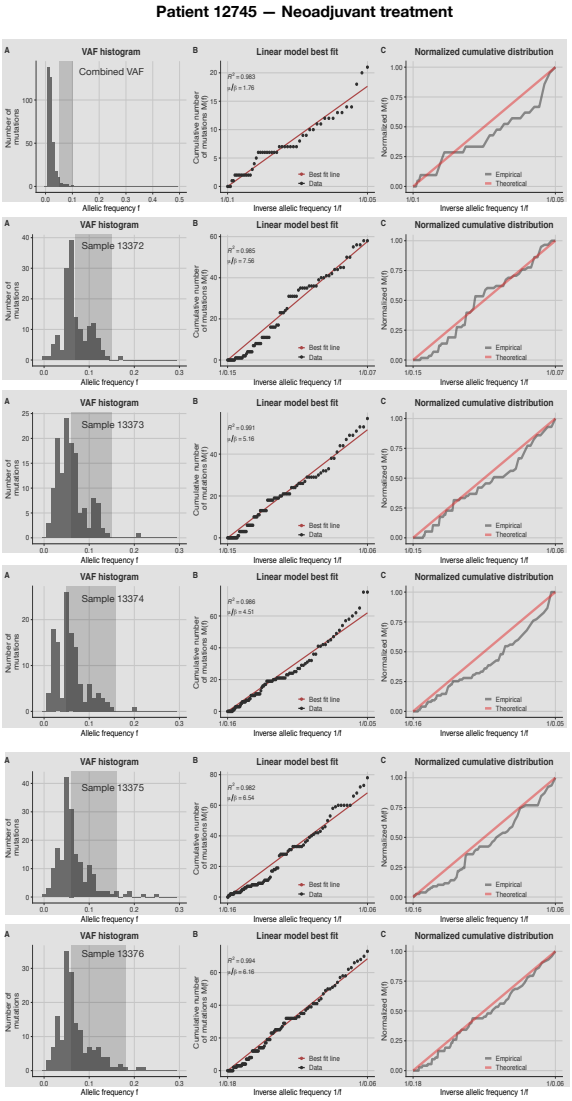

Figure 6.10: Neutrality analysis of samples from a patient.

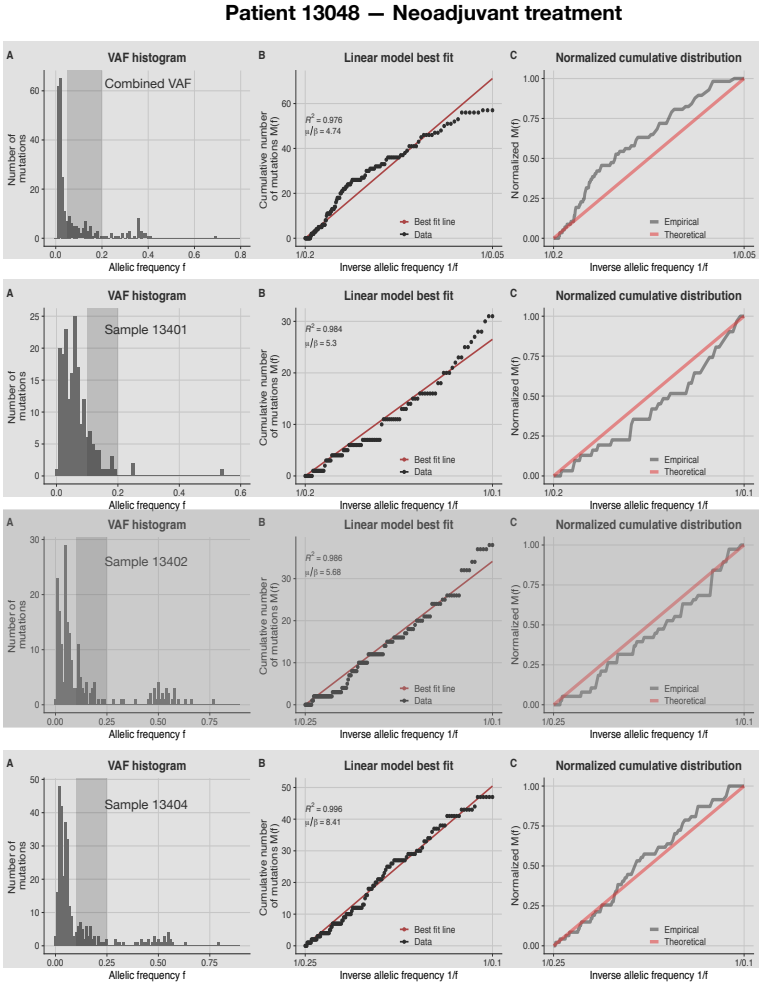

Figure 6.11: Neutrality analysis of samples from a patient.

Author's contributions

Chapter 2: All authors developed the concept. Luka Opasic implemented simulations and analysed the model with input from all authors. Luka Opasic and Arne Traulsen wrote the manuscript with input from David Dingli, Benjamin Werner and Da Zhou.

Chapter 3: Results presented in this chapter will be included in the manuscript of a study led by Christoph Roecken. Christoph Roecken provided material and obtained ethical committee aproval. Philip Rosenstiel performed the sequencing of DNA samples. Tobias Meissner and Anu Amallraja performed the bioinformatic analyis (variant calling and construction of phylogenetic trees). Luka Opasic analyzed the data. Luka Opasic and Arne Traulsen interpreted the data.

Chapter 4: Luka Opasic conceptualized the idea for the project. Da Zhou contributed with the probability theory. Luka Opasic performed the simulations and analysed the results. Luka Opasic wrote the manuscript with input from Arne Traulsen, Da Zhou and Jacob Scott.

Chapter 5: Luka Opasic conceptualized and created the model with input from Jacob Scott and Arne Traulsen. Carsten Fortmann-Grote adapted the code for git release. Luka Opasic and Carsten Fortmann-Grote wrote the manuscript.

Affidavit

I declare:

- That the content and design of this dissertation is product of my own work, except where reference is made to the work of others or otherwise noted, and apart from my supervisors guidance;
- Has not been submitted elsewhere partially or wholly as a part of a doctoral degree and no other materials are published or submitted for publication than indicates in the thesis;
- The work and thesis has been performed and prepared following the Rules of Good Scientific Practice of the German Research Foundation.

Acknowledgements

First of all, I would like to thank my supervisor Arne Traulsen for the opportunity, support and patience while guiding me on this rough journey. I am very grateful to be working with a group of amazing collaborators Benjamin Werner, David Dingli, Christoph Roecken and Carsten Fortman-Grote. It was great working with you on exciting joint projects. Definitely the most memorable moment of my doctoral studies was my visit to the USA and Cleveland Clinic. Working with Jacob Scott, his team of wonderful people and living with Andrew Dhawan was a really enriching experience I will remember and cherish. Next, I would like to thank my collegues Michael Ratz, Andrew Farr, Neva Skrabar and Christin Nyhoegen, who proofread parts of my thesis and provided me with constructive feedback.

Gloomy winter days of Schleswig Holstein were much nicer with great officemates Yuriy Pichugin and Marvin Boettcher. Outside of the office, I had a great time with my dear friends Jordi Arranz and Loukas Theodosiou. Thank you guys for making my time here fun. Immense gratitude goes to Da Zhou. Your passion, knowledge and kindness inspired me deeply and showed me that I am able to accomplish more. Our hours-long scientific discussions were so much fun and reminded me why I love science. Finally, the biggest thanks goes to my Petra who followed me to Ploen and supported me on this endevour, and my two sons Karlo and Leon who gave meaning to all of my work.

Curriculum Vitæ

2016 – 2020	Ph.D. candidate Max Planck Institute for Evolutionary Biology, Plön, Germany
2015 –2016	Intern Physician Psychiatric Hospital Sveti Ivan, Zagreb, Croatia
2009 – 2015	Doctor of Medicine, MD School of Medicine, University of Zagreb, Croatia
16.5.1990	Born in Rijeka, Croatia