Immune editing and surveillance in cancer evolution

Rachel Suzanne Rosenthal

PhD supervisor: Charles Swanton

A dissertation submitted for the degree of

Doctor of Philosophy

University College London

December 2017

Declaration

I, Rachel Rosenthal, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

December 18, 2017

Rachel Rosenthal

Abstract

Cancer is an evolutionary disease, reliant on genetic diversity and sculpted by selective forces from the immune microenvironment. Here, I use genomics data to decipher the tumor's evolutionary trajectory and corresponding shifts in the immune contexture to elucidate the events governing tumor immunogenicity and the immune evasive mechanisms evolved by the tumor.

To better understand the mutational processes contributing to intratumor heterogeneity in individual tumors, a method to quantify the activity of mutational processes in a single tumor sample was developed and applied to temporally dissected mutations.

The clinical relevance of intratumor heterogeneity was examined in the context of immune recognition and modulation. Increased clonal neoantigen burden and minimal neoantigen intratumor heterogeneity were found to associate with improved patient outcome, both in the treatment-naïve and immunotherapy-treated setting. The identification of T-cells recognizing clonal neoantigens further supported the clinical importance of targeting neoantigens present in every cancer cell.

Mechanisms of immune evasion were considered through the development of a method to identify loss-of-heterozygosity at the HLA locus, overcoming the challenges posed by the polymorphic nature of the locus. HLA loss-of-heterozygosity was found to be a frequent subclonal event in NSCLC, under strong selective pressure and associated with increased subclonal neoantigen burden.

Finally, the immune microenvironment was examined through multi-region RNAseq, permitting the quantification of immune infiltration and allowing for the identification of heterogeneously immune infiltrated tumors. Supporting the interplay between genetic events and the immune contexture, a relationship between the genomic features of the tumor and immune infiltration was observed, with HLA loss-of-heterozygosity specifically identified as occurring within a highly active immune microenvironment.

This thesis shows how an improved understanding of the relationship between the tumor and the immune system can illuminate features dictating immune recognition and evasion and how that knowledge may inform the development and implementation of successful immunotherapy.

Impact statement

Cancer affects the lives of millions of people around the world. One of the largest barriers to the successful management of the disease is the development of treatment resistance, often due to the outgrowth of a resilient subclonal population of cells from the heterogeneous tumor. Over the last few decades, the realization that the patient's immune system is capable of targeting and eliminating tumor cells has led to a number of promising advances in cancer treatment. This observation has been exploited through therapies such as immune checkpoint blockade, adoptive T-cell therapy, and cancer vaccines, all of which aim to generate or enhance an immune response specifically against the patient's tumor. My research has focused on understanding the interaction between the tumor and the immune system, with particular emphasis on elucidating the factors that result in tumor immunogenicity and on the mechanisms a tumor may use to evade immune predation. A better understanding of the interplay between the tumor and the immune microenvironment can help identify which patients are more likely to respond to immunotherapy interventions and help select the most potent targets for developing such interventions in a patient-specific manner. Furthermore, learning the various avenues a tumor may take on the path to eventual immune evasion will help to combat resistance mechanisms that may arise.

Acknowledgements

I would like to thank my supervisor, Professor Charles Swanton, for welcoming me into his lab as a displaced graduate student. I could not have imagined a better PhD experience due to his guidance and the wonderfully enthusiastic lab he manages.

I would also like to thank Javier Herrero and the rest of the Bill Lyons Informatics Centre for their constant willingness to answer questions and for affording me so much of their bioinformatics wisdom over the last few years.

The experience of this PhD would not have been nearly as rewarding without the talented and engaging colleagues in the Translational Cancer Therapeutics lab I got to work alongside. Beyond providing invaluable scientific advice, their friendship and laughter made stressful times less so and enjoyable times even more memorable. In particular, I would like to thank Nicholas McGranahan for his always thoughtful direction and discussion, Gerald Goh for helping me to land on my feet in a new environment, and Aśka Przewrocka for her camaraderie as a graduate student and ceaseless excitement over all things, including science.

Finally, none of this work would have been possible without the support of my friends, in London and at home, and my family. I would like to especially thank my parents for their unwavering encouragement. It seems only fitting that a few decades after appearing in my dad's acknowledgements, he should appear in mine, likely playing a larger role in the completion of this PhD than I did in his.

Table of Contents

Table of Tables

Table of Figures

Abbreviations

AI	allelic imbalance
ALL	acute lymphoblastic leukemia
APOBEC	apolipoprotein B mRNA editing enzyme catalytic polypeptide-like
B2M	beta-2-microglobulin
BAF	B-allele frequency
BLCA	bladder adenocarcinoma
BRCA	breast carcinoma
CCF	cancer cell fraction
CLL	chronic lymphocytic leukemia
COAD	colon adenocarcinoma
ESCA	esophageal cancer
FISH	fluorescence *in situ* hybridization
GBM	glioblastoma multiforme
HLA	human leukocyte antigen
HNSC	head and neck squamous cell carcinoma
IMGT	ImMunoGeneTics
ITH	intratumor heterogeneity
logR	log-ratio
LOH	loss-of-heterozygosity
LUAD	lung adenocarcinoma
LUSC	lung squamous cell carcinoma

Indel	small insertion/deletion
MHC	major histocompatibility complex
MSI	microsatellite instability
NGS	next-generation sequencing
NMF	non-negative matrix factorization
NSCLC	non-small cell lung cancer
OS	overall survival
PFS	progression free survival
RFS	relapse free survival
SKCM	skin cutaneous melanoma
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SSE	sum-squared error
TCGA	The Cancer Genome Atlas
TILs	tumor infiltrating lymphocytes
UV	ultraviolet

Chapter 1 Introduction

A developing tumor is subject to ongoing pressure from selective forces from the local microenvironment and immune predation. As cells stochastically acquire new alterations, these selective forces dynamically sculpt the tumor and shape the way in which it evolves, either by pruning away less fit or more immunogenic cells or by enabling cells with beneficial traits to succeed. A record of the tumor's evolutionary trajectory and the contributing mutagenic processes can be unearthed by studying the genetic make-up of the tumor, allowing for the elucidation of events that supported tumor diversification and escape from immune predation. Furthermore by considering the interaction of the tumor with the immune microenvironment, it is possible to understand the factors governing tumor immunogenicity and the evasive mechanisms utilized by the tumor to escape immune recognition.

Historically, cancer studies have focused on mutations in genes thought to drive tumorigenesis, presumably under strong positive selection. Often overlooked is what information the passenger events may contain. It is only via the abundance of passenger mutations that it is feasible to understand which mutational processes are active during tumor evolution, and it is frequently these mutations that are recognized as non-self by the immune system.

Over the last decade next-generation sequencing (NGS) technology has developed to the point where sequencing a tumor's genetic material is no longer reserved for a few select samples. Alongside the improvements in technology, comprehensive and coordinated sequencing studies (for instance, The Cancer Genome Atlas [TCGA]) have allowed for the incorporation of data across multiple –omics platforms to generate an unparalleled understanding of a wide range of cancer types. Together, such studies have allowed for greater insight into the processes involved in generating mutations (Alexandrov et al., 2013a, Helleday et al., 2014, Segovia et al., 2015, Lawrence et al., 2013), the dynamics of tumor clones during the disease course and through treatment (Marusyk and Polyak, 2010, Calbo et al., 2011, Landau et al., 2013, Keats et al., 2012, Murtaza et al., 2013), and have begun to illuminate the tumor immune microenvironment (Rooney et al., 2015, Li et al., 2016, Davoli et al., 2017).

In this thesis, I apply bioinformatics approaches to the wealth of NGS data available to understand the processes that generate mutations contributing to intratumor heterogeneity, the factors determining whether the immune system recognizes such mutations, and the mechanisms through which the tumor may evade detection. Improved understanding of the interplay between the tumor and the immune system may inform the development and implementation of successful immunotherapy.

1.1 Tumor evolution and selection

For decades, tumor progression has been perceived as an evolutionary process reliant on clonal diversity and subsequent selection of subpopulations endowed with a relative fitness advantage (Nowell, 1976, Fidler, 1978, Greaves and Maley, 2012, Gerlinger and Swanton, 2010). At the time of clinical detection, a tumor will have undergone many rounds of cell division, with each generation of cells stochastically acquiring novel somatic alterations (Gerlinger et al., 2014b, Stratton et al., 2009). Selective pressures in the tumor's environment, such as immune activity, therapy, and subclonal interactions actively sculpt the tumor, shaping the way it evolves (Merlo et al., 2006).

The generation of new mutations during each cell cycle continually gives the tumor a chance to adapt to its environment. While the vast majority of the mutations arising have little impact on the overall fitness of the cell (Martincorena et al., 2017, Greenman et al., 2006), a subset of these mutations (known as driver events) will endow a cell with an evolutionary advantage, allowing that cell and its progeny to flourish and outcompete others. From NGS data, it is possible to identify the somatic alterations acquired during the path to tumorigenesis. Analysis of sequencing data provides a historical record of mutational events, including single base substitutions, insertions, and deletions (Koboldt et al., 2012, Cibulskis et al., 2013, Gerlinger et al., 2012b), and copy number alterations, such as amplifications, deletions, and areas where a single parental chromosome has been lost, resulting in loss-of-heterozygosity at the locus (Van Loo et al., 2010, Zack et al., 2013, Favero et al., 2015).

Combined, the evolutionary processes of somatic alteration accumulation and selection acting on the cell may result in the outgrowth of multiple subclones, often with their own distinct driver events, leading to the branched evolutionary phylogeny that has been observed across many cancer types (Gerlinger et al., 2012a, de Bruin et al., 2014, Gundem et al., 2015, Sottoriva et al., 2013, Nik-Zainal et al., 2012b).

By calculating the frequency of somatic events it is also possible to infer when in the life history of the tumor the mutation occurred. This requires incorporating the frequency of the somatic event with overall tumor purity, representing the fraction of cells sequenced that came from the tumor rather than stroma, and ploidy, referring to the number of haploid sets of chromosomes in the cell. (Landau et al., 2015). Clonal events are those found in every cancer cell, indicating they arose early in tumor evolution or after a selective sweep. Subclonal events are only found in a fraction of the cancer cells, suggesting they were later evolutionary events.

Most driver events established to date have been classified as clonal, but increasingly subclonal driver events have been identified across many cancers. An increased power to detect subclonal drivers has been generated through the sequencing of samples at a greater depth, as well as the sequencing of multiple regions per sample and samples at more than one time point. Subclonal drivers likely aid in maintaining the tumor and potentially lead to tumor progression, subclonal expansions and/or the acquisition of drug resistance (McGranahan et al., 2015). Even the most common driver events may occur early in some tumors and late in others (McGranahan et al., 2015, Yates et al., 2015). The delineation between driver and passenger mutations is likely context and tissue dependent, for as selective pressures and the tumor microenvironment change, so do the requirements for tumor survival. For instance, in most cancer types TP53 mutations are almost uniformly clonal (de Bruin et al., 2014, Zhang et al., 2014, Bashashati et al., 2013), but in a minority of cancers, such as chronic lymphocytic leukemia and clear cell renal carcinoma (Landau et al., 2013, Gerlinger et al., 2014a), they are frequently subclonal.

Therapy can also act to alter the dynamics of selection in the tumor. For instance, JAK1/2 loss of function mutations, which are often early drivers in hematological cancers but considerably rarer in melanoma, can undergo clonal selection in melanoma patients upon treatment with immunotherapy, and have been observed in cases with acquired resistance to anti-PD1 therapy (Zaretsky et al., 2016). Similarly, colorectal cancer patients treated with anti-EGFR therapy can develop *de novo* KRAS mutations, which have been shown to drive acquired resistance to the therapy (Misale et al., 2012).

Finally, even the order in which mutations arise can influence the outcome of subsequent selective pressures, restrict evolutionary paths (Papaemmanuil et al.,

2013), and affect the clinical behavior of disease presentation, as well as response to therapy (Ortmann et al., 2015).

The detection of subclonal mutations (and thus subclonal drivers) is also limited by the resolution of bulk tumor sequencing. A mutation may be found in every cancer cell from a single tumor region, but if that region does not accurately reflect the entire tumor, then the mutation may truly be subclonal within the tumor as a whole, thus artificially inflating the number of observed clonal drivers. Examining multiple regions from the same tumor, or even single cells, is one way to improve resolution and more accurately determine the clonal architecture of the tumor (Navin et al., 2011, Tirosh et al., 2016, de Bruin et al., 2014).

In addition to positive selection resulting in the outgrowth of clones with an evolutionary advantage, negative selection also likely impacts the evolution of the tumor by pruning away cells harboring deleterious mutations, resulting in the end of that cell's lineage. As the absence of particular mutations is more challenging to identify, there are few current estimates of how much of an evolutionary force negative selection represents (Martincorena et al., 2017). However, given the increasing recognition of the immune system's role in tumor control, it is conceivable that purifying selection plays a large part in removing clones with immunogenic mutations (Rooney et al., 2015, Rajasagi et al., 2014).

1.2 Intratumor heterogeneity as substrate for selection

1.2.1 Scale and extent of heterogeneity in cancer

Individual cells in a tumor, which may subsequently expand to form subclones, are subject to the activity of mutational processes. Thus the generation of new mutations leads to a heterogeneous cell population. It is this diversity upon which selective pressures may act. Macroscopic ITH has been observed for millennia (Mukherjee, 2011) and microscopic morphological differences between tumor cells were identified as early as the 1800s (Balkwill and Mantovani). As technology has advanced, the scale of heterogeneity observable in tumors has become more finely resolved. Beginning with the observation of differently sized nuclei, improved technology has eventually allowing for characterizations at the karyotypic level (Szollosi et al., 1995) to the gene level and ultimately resulting in the ability to detect single nucleotide differences between single cells (Roth et al., 2016).

Alongside advances in technology, improved bioinformatics methods have also allowed for further dissection of the clonal architecture of the tumor by clustering somatic alterations into tumor subclones (Nik-Zainal et al., 2012b, Shah et al., 2012, Roth et al., 2014, Deshwar et al., 2015). Somatic alterations are grouped by their frequency in the tumor; however, it is possible that distinct events have occurred in separate subclones and present at similar frequencies, potentially hampering clonal reconstruction. Understanding the clonal structure of a tumor is important as it permits the deciphering of its phylogenetic history, providing insight into how it evolved.

Recent studies of single tumor samples, as well as multiple and serial sampling techniques have revealed considerable variability in the extent of diversity both between patients and within individual tumors. Genetic ITH has been identified and characterized across a wide range of cancer types including breast carcinomas (Navin et al., 2011, Nik-Zainal et al., 2012b), clear-cell renal carcinomas (Gerlinger et al., 2012a, Gerlinger et al., 2014a), glioblastomas (Sottoriva et al., 2013), gliomas (Johnson et al., 2014), prostate cancers (Haffner et al., 2013, Gundem et al., 2015), non-small cell lung cancers (de Bruin et al., 2014, Zhang et al., 2014), head and neck squamous cell carcinomas (Mroz et al., 2015), squamous cell melanomas (Ding et al., 2014), high-grade serous ovarian cancer (Schwarz et al., 2015), chronic lymphocytic leukemia (Landau et al., 2013), acute myeloid leukemia (Klco et al., 2014, Ding et al., 2012), and multiple myeloma (Lohr et al., 2014, Bolli et al., 2014).

On the whole, these studies have demonstrated that heterogeneity is observed to varying extents across a wide variety of cancers, with both clonal and subclonal driver mutations identified. However, the majority of studies considering heterogeneity in detail have either been limited to a small number of patients, or have only investigated heterogeneity based on a single sample from each tumor, thereby potentially underestimating the true extent of diversity within tumors.

1.2.2 Mutational processes contributing to ITH

During the evolution of a tumor, mutational processes leave evidence of their activity via the somatic alterations they cause. These somatic alterations can be cataloged and tracked, providing a record of mutational activity and allowing for insights into the routes taken to carcinogenesis (Alexandrov et al., 2013a, Nik-Zainal et al., 2012a). Broadly, mutational processes may be categorized as arising

from an exogenous or an endogenous source. Exogenous mutagens may include tobacco smoke and ultraviolet light, commonly observed in lung and skin cancers, respectively, while endogenous processes encompass defective DNA repair, such as mismatch repair deficiency in colorectal cancers, and the enzymatic modification of DNA, such as APOBEC family activity commonly identified across many cancers (Pfeifer et al., 2002, Pfeifer, 2010, Boland and Goel, 2010, Bhattacharyya et al., 1994, Alexandrov et al., 2013a, McGranahan et al., 2015, Nik-Zainal et al., 2012a). Regardless of the source, these aberrant processes result in characteristic patterns of mutation.

Different periods of a tumor's evolutionary history are associated with the activity of distinct mutational processes (McGranahan et al., 2015, Nik-Zainal et al., 2012b). For instance, smoking mutations in NSCLC and UV-induced mutations in melanoma largely predate tumorigenesis, meaning they are present in the tumor initiating cell and in every subsequent daughter cell. In contrast, therapy-related mutation processes are, by definition, observable only after tumor detection (Johnson et al., 2014).

Thus some mutational processes, which are active later during tumor evolution, disproportionately contribute to an increased level of ITH. This has been observed in NSCLC and bladder cancers, where large numbers of subclonal mutations often reflect activity of the APOBEC family of cytidine deaminases (de Bruin et al., 2014, McGranahan et al., 2015). Similarly, some mutational processes are capable of generating large numbers of mutations both early and late during cancer. In colorectal and prostate cancers, as well as some breast cancers, mismatch repair deficiencies lead to an increased mutational burden, but alterations to the proof-reading machinery can occur early or late in tumor evolution (Kumar et al., 2016a, Uchi et al., 2016, Davies et al., 2017a).

Therapy itself may also directly act as a mutagenic agent, increasing genetic ITH and influencing the evolutionary path of the tumor (Sveen et al., 2016, Murugaesu et al., 2015). The genotoxic effects of chemotherapy are often observable as distinct mutational processes, reflected by changes to the mutational landscape and spectra of the tumor.

In two multi-region exome sequencing analyses of patients with esophageal adenocarcinoma taken before and after treatment with a platinum-containing chemotherapy, an increase in C>A transversions at CpC sites has been identified

among the post-chemotherapy samples of patients with residual disease (Murugaesu et al., 2015, Findlay et al., 2016). Mutations in this particular context have been previously identified in *C. elegans* treated with cisplatin, a platinum based chemotherapeutic (Meier et al., 2014). The majority of the mutations observed in the platinum-associated mutational context were subclonal, consistent with those mutations occurring late in tumor evolution, as would be expected for chemotherapy-induced mutagenesis.

Conceivably, therapy-associated mutational processes may not only leave scars in the genome but also may directly contribute to disease progression. A number of tumors from patients with melanoma and from low-grade glioma that transformed to glioblastomas at recurrence have been found with enormous subclonal mutation burden due to treatment with the alkylating agent temozolomide, selecting for resistant subclones with defective mismatch repair (Johnson et al., 2014, Chan et al., 2015, Alexandrov et al., 2013a, Hunter et al., 2006). Additionally, in the recurrent glioblastomas, novel driver mutations were identified in the RB and Akt-mTOR pathways within the temozolomide associated mutational context, highlighting how chemotherapy-induced mutagenesis is not limited to driving genetic diversification, but can also influence the evolutionary path taken by the tumor (Johnson et al., 2014).

1.2.3 Clinical implications of ITH

Longitudinal analyses of tumor samples have consistently identified shifts in the genomes of samples taken before and after treatment with chemotherapeutics (Johnson et al., 2014, Murugaesu et al., 2015, Mullighan et al., 2008, Landau et al., 2013, Ding et al., 2012, Schuh et al., 2012, Keats et al., 2012, Weston-Bell et al., 2013), indicating that the genomic landscape of a tumor changes in response to cancer therapy.

Even without directly inducing novel mutations as discussed above, cancer therapy results in new selective pressures, which can impact evolutionary trajectories reliant on the genetic variation that existed prior to the start of treatment. Within a heterogeneous tumor, some subclones may be present that originally had no obvious fitness advantage but impart a resistance to therapy and are subsequently selected for. Indeed, there have been numerous reports detailing the outgrowth of resistant subclonal populations in response to therapy across many cancer types including colorectal (Kreso et al., 2013, Diaz et al., 2012), glioblastoma (Cahill et al.,

2007, Yip et al., 2009), melanoma (Wagle et al., 2011, Shi et al., 2014), non-small cell lung cancer (Kosaka et al., 2006, Turke et al., 2010), and CML (Shah et al., 2002). The presence of subclonal drivers has also been associated with poorer outcome and response to therapy (Landau et al., 2013, Landau et al., 2015) in CLL. Thus extensive ITH is likely to limit the impact therapy, with a particularly strong effect on precision medicine approaches and targeted therapeutics (Gerlinger et al., 2014b).

Accordingly, there is some evidence that measures of ITH may be useful as a prognostic biomarker. Patients with low copy number ITH are more likely to respond well to chemotherapy (Marusyk et al., 2014) and, in early stage surgically resected NSCLC, exhibit longer recurrence-free survival (Jamal-Hanjani et al., 2017). Furthermore, increased ITH and clonal expansions have been shown to correlate with disease progression and poor prognosis in multiple cancer types (Mroz and Rocco, 2013, Schwarz et al., 2015, Sveen et al., 2016, Maley et al., 2006).

1.3 Tumor and immune interaction

By recognizing antigenic components of the tumor cell and influencing the local microenvironment, the immune system may exert an evolutionary pressure, shaping the antigenicity of the tumor and its diversity as it evolves. Specifically, immune editing, which describes the interaction between the tumor and immune system wherein the immune system plays the dual role of protecting the host and sculpting the tumor, can impact tumor evolution. Driven by the immune editing process, subclonal populations of tumor cells either lacking immunogenic antigens or able to withstand an immune response may be selected for (Schreiber et al., 2011, Matsushita et al., 2012, DuPage et al., 2012).

1.3.1 History of immune surveillance and immune editing

The immune system has long been thought of being capable of eliminating nascent cancer cells in humans, providing effective protection from tumor development (Schreiber et al., 2011). As early as the 1950s, donor T-cells were observed to recognize antigens on the surface of tumor cells (Barnes et al., 1956). This identification of tumor specific antigens, along with the formal demonstration of immune control and a more thorough understanding of the immune system over the coming decades, eventually led to the formalization of the cancer immune surveillance hypothesis, wherein the role of the adaptive immune response in

preventing tumor development was acknowledged (Burnet, 1957, Thomas, 1982). However, much debate ensued over whether the immune system could effectively curtail tumor growth as critics pointed to experiments showing the growth promoting nature of the pro-inflammatory immune response (Balkwill and Mantovani, 2001, Karin et al., 2002) and argued that tumor cells would be subject to immune tolerance, as they were so closely related to self (Pardoll, 2003).

Throughout the 1990s, mouse studies continued to show that the immune system could mediate tumor rejection and that mice with defective immune responses or lacking functional lymphocytes were more susceptible to tumor formation (Dighe et al., 1994, Kaplan et al., 1998, Street et al., 2001, van den Broek et al., 1996). Furthermore, experiments comparing the immunogenicity of tumors formed in mice without an intact immune system to those formed under immune surveillance led to the realization that the immune system could sculpt tumors as they developed in addition to eliminating them (Shankaran et al., 2001, Dunn et al., 2002), eventually resulting in the immune editing hypothesis.

Subsequently, three phases of tumor/immune interaction have been proposed (Dunn et al., 2002) (Figure 1-1):

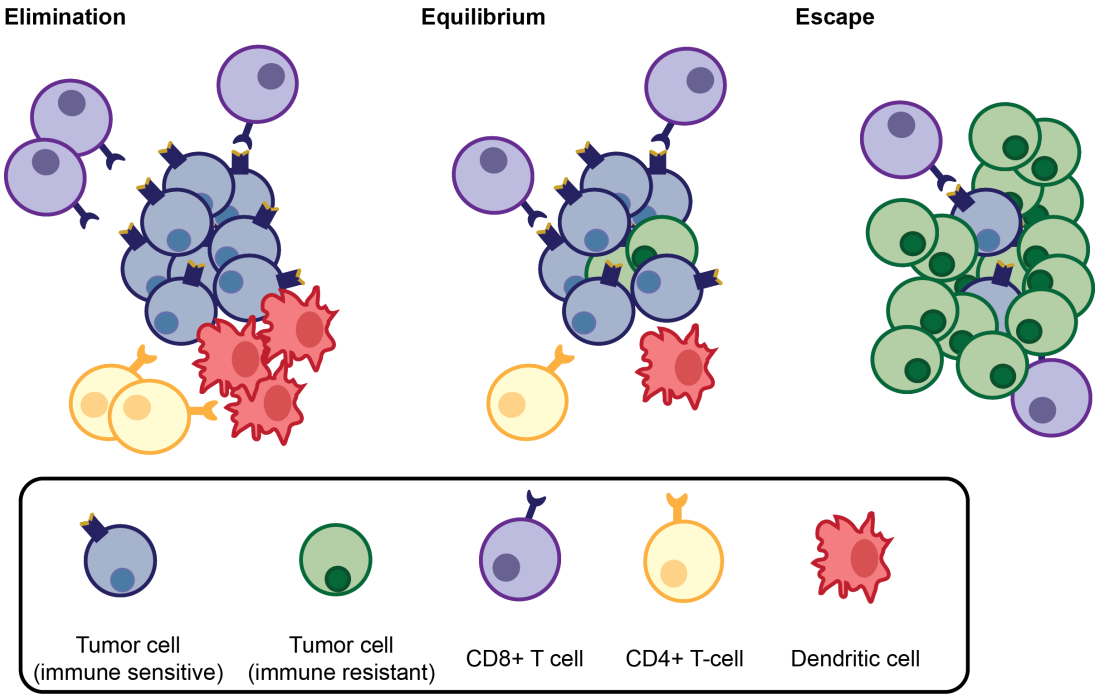

Figure 1-1: Immune elimination, tumor/immune equilibrium, and tumor escape.
Three phases of the tumor/immune interaction are displayed. In the immune elimination phase, the immune system recognizes and destroys tumors cells. Eventually immune pressure can lead to some tumor cells developing ways to avoid recognition (equilibrium). Finally those immune resistant tumor cells can escape immune detection.

1) Elimination: During immune elimination, the immune system is surveilling and destroying transformed cells.

2) Equilibrium: Repeated pressure from the immune system results in the tumor evolving non-antigenic variants to avoid immune detection and/or evasive mechanisms to withstand immune attack. During this phase, genetic instability likely contributes to the ability of the tumor to withstand and respond to immune pressure.

3) Escape: The tumor which has now been sculpted by the immune system eventually escapes and grows unchecked by immune activity.

1.3.2 Tumor antigens

In order for the immune system to actively destroy nascent tumor cells and shape tumor development, it must be capable of recognizing the tumor as distinct from the non-transformed self. Tumor-specific antigens were first inferred to exist after studies in mice showed that T-cells could eliminate cancer cells and again after the realization that mice could not be successfully challenged with the same chemically induced tumor twice, suggesting an immunological memory of the first tumor challenge (Barnes et al., 1956, Old and Boyse, 1964).

In humans, studies have also shown that tumor cells express antigens that are recognizable by T-cells from the same patient. Autologous tumor infiltrating lymphocytes (TILs) from patients with metastatic melanoma have been expanded *ex vivo* and re-administered, resulting in tumor regression (Dudley et al., 2002b), indicating that there are T-cells present capable of recognizing tumor cells and mounting an immune response against them (Dudley et al., 2002b, Rosenberg, 2012).

This response has been exploited therapeutically with adoptive cell therapy (ACT), which allows for the administration of T-cells that have been selected as highly tumor-reactive (Dudley et al., 2002a). If a specific antigen on the tumor cell is known, such as CD19 in lymphoma, CLL, and ALL, then the cells may be genetically engineered to contain specific T-cell receptors (TCRs) or chimeric antigen receptors (CARs). Serving to highlight the potential of T-cell elimination of cancer cells, treatment with ACT can result in long-term remission from disease (Tran et al., 2014, Rosenberg and Restifo, 2015).

A T-cell mediated immune response occurs either from the recognition of self-antigens when T-cell tolerance is incomplete or through recognition of a peptide that is not present in the normal human peptidome (Hacohen et al., 2013, Schumacher and Schreiber, 2015) due to the activity of ongoing mutational processes in a cancer cell (Vogelstein et al.). Self-antigens can arise from genes that are aberrantly expressed in cancer, such as MAGE-1 (van der Bruggen et al., 1991, Chomez et al., 2001, Sahin et al., 1995), which is normally restricted to male germline cells, or from proteins such as Her-2/Neu that are normally expressed in healthy tissue but overexpressed in tumors (Fisk et al., 1995).

While self-antigens were historically easier to identify as they were frequently shared between tumors, targeting them through T-cells engineered to express tumor reactive T-cell receptors can result in significant toxicity since healthy tissue also expresses the same antigen (Johnson et al., 2009, Morgan et al., 2006). Thus in recent years, non-self tumor antigens have been the focus of much study as they may represent a novel class of potent immunotherapy targets.

1.3.3 Neoantigens

Non-self tumor antigens, more commonly referred to as neoantigens, are generated from the multitude of somatic mutations present in a cancer cell (Alexandrov et al., 2013a, Vogelstein et al., 2013, Stratton, 2011). Point mutations that occur in the protein coding regions of the genome (exonic mutations) can either be classified as synonymous, if they do not result in an amino acid change, or non-synonymous, if the mutation results in a new amino acid. In the case of non-synonymous mutations, the immune system may be capable of recognizing the altered peptide sequence as foreign and mounting an immune response against it. Neoantigens represent an interesting potential therapeutic target, as they are tumor specific by definition, yet they are not self-antigens and thus are not limited by central tolerance or treatment associated toxicity.

The first evidence highlighting importance of neoantigenic epitopes mediating a cancer immune response was again found through the use of mouse studies. In 1995 an amino acid substitution from a UV-induced tumor was found to act as a tumor specific antigen capable of being recognized by T-cells, whereas its wildtype counterpart was not (Monach et al., 1995). Shortly thereafter, T-cells from a human patient with metastatic melanoma were also found to recognize a peptide arising from a point mutation (Coulie et al., 1995). These studies, among many others

(Sensi and Anichini, 2006), showed that T-cell responses could be mounted against somatically mutated antigens in both mouse and human. Additionally, cytolytic activity of T-cells against mutated peptides has been observed in patients exhibiting long-term survival (Novellino et al., 2003, Lennerz et al., 2005).

Bioinformatics advances allow the prediction of neoantigens using *in silico* approaches. In a landmark mouse study, potentially immunogenic epitopes arising from a murine melanoma tumor model were predicted from exome sequencing data. When mice were immunized with two of the identified neoantigens, tumor growth was inhibited and a protective immune response was observed (Castle et al., 2012). In a separate study using a mouse tumor cell line, a dominant neoantigen was identified that was recognized by CD8+ T-cells in mice that rejected the tumor when challenged. Additionally, it was shown that tumor cell clones lacking the antigenic mutations were selected for growth in an illustration of tumor immunoediting, where the immune response shapes tumor development towards an immune-resistant direction (Matsushita et al., 2012).

Neoantigens arising from human tumors have also been predicted and confirmed *in vitro*, highlighting their potential benefit in a clinical setting. Segal and colleagues identified potentially immunogenic epitopes in a cohort of breast and colorectal tumor samples (Segal et al., 2008). Putative tumor specific neoantigens have also been identified in CLL (Rajasagi et al., 2014). In a seminal study, van Rooij and colleagues predicted potential neoantigens for a patient with stage IV melanoma that exhibited a response to immunotherapy treatment. They then advanced from the sole use of *in silico* techniques to the screening of TILs obtained from the patient for reactivity against the predicted neoantigens and identified two neoantigens that caused T-cell reactivity *in vitro* (van Rooij et al., 2013). Pan-cancer analyses of large tumor cohorts have identified a link between predicted neoantigenic load, enhanced cytolytic activity, and improved prognosis (Brown et al., 2014, Rooney et al., 2015).

However, due in large part to the heterogeneous composition of many tumors and the polymorphic nature of the HLA locus (discussed below), the number of recurrent neoantigens is astonishingly low. Indeed, an analysis of the predicted neoantigens resulting from a cohort of over 60,000 patients found that each set of neoantigens would be relevant to less than ~0.3% of the population (Hartmaier et al., 2017).

1.3.4 HLA presentation

In order for an antigenic peptide to induce an immune response, it must be presented on the cell surface via major histocompatibility complex (MHC) molecules for potential recognition by T-cells (Neefjes et al., 2011).

There are two primary classes of MHC molecules, known as human leukocyte antigen (HLA) in humans. Class I molecules are expressed by all nucleated cells and present intracellular antigenic peptides that have been degraded by the proteasome to CD8+ T-cells (Figure 1-2A-B). The MHC molecule is a heterodimer, formed by a highly polymorphic heavy and light chain, which is then stabilized by B2M. In humans, the classical class I molecules are encoded by HLA-A, HLA-B, and HLA-C genes on chromosome 6, which show great diversity among the general population. The polymorphisms in the HLA genes result in different peptide binding grooves, such that different HLA alleles bind a diverse set of peptides (The, 1999).

Class II molecules are also encoded by a set of highly polymorphic genes on chromosome 6 (HLA-DP, HLA-DQ, HLA-DR) and present peptides derived from extracellular proteins (The, 1999). Whereas class I MHC molecules are presented by nearly all cells, class II MHC molecule expression is generally restricted to professional antigen presenting cells, such as dendritic cells and B cells (Neefjes et al., 2011) (Figure 1-2C). However, interferon gamma signaling can induce class II expression in many cell types, including tumor cells (Collins et al., 1984, Park et al., 2017). In the MHC class II presentation pathway, a protein is endocytosed and degraded by proteases before being presented to CD4+ T-cells by MHC class II molecules.

One possible way for a neoantigenic peptide to induce a novel immune response is if the mutation affects an amino acid that results in strong peptide binding to the MHC molecule. In such a scenario, if the wildtype peptide was not capable of binding an MHC molecule, T-cell clones recognizing it would not have been subject to removal during the development of central tolerance. If a mutation results in peptide-MHC binding and the mutant peptide is recognized as foreign by T-cells, then it could induce an immune response. Alternatively, if the wildtype peptide is also presented by MHC molecules, then the neoantigenic peptide may be recognized as foreign by T-cells due to a mutation that changes the T-cell receptor exposed area (Fritsch et al., 2014) (Figure 1-2D).

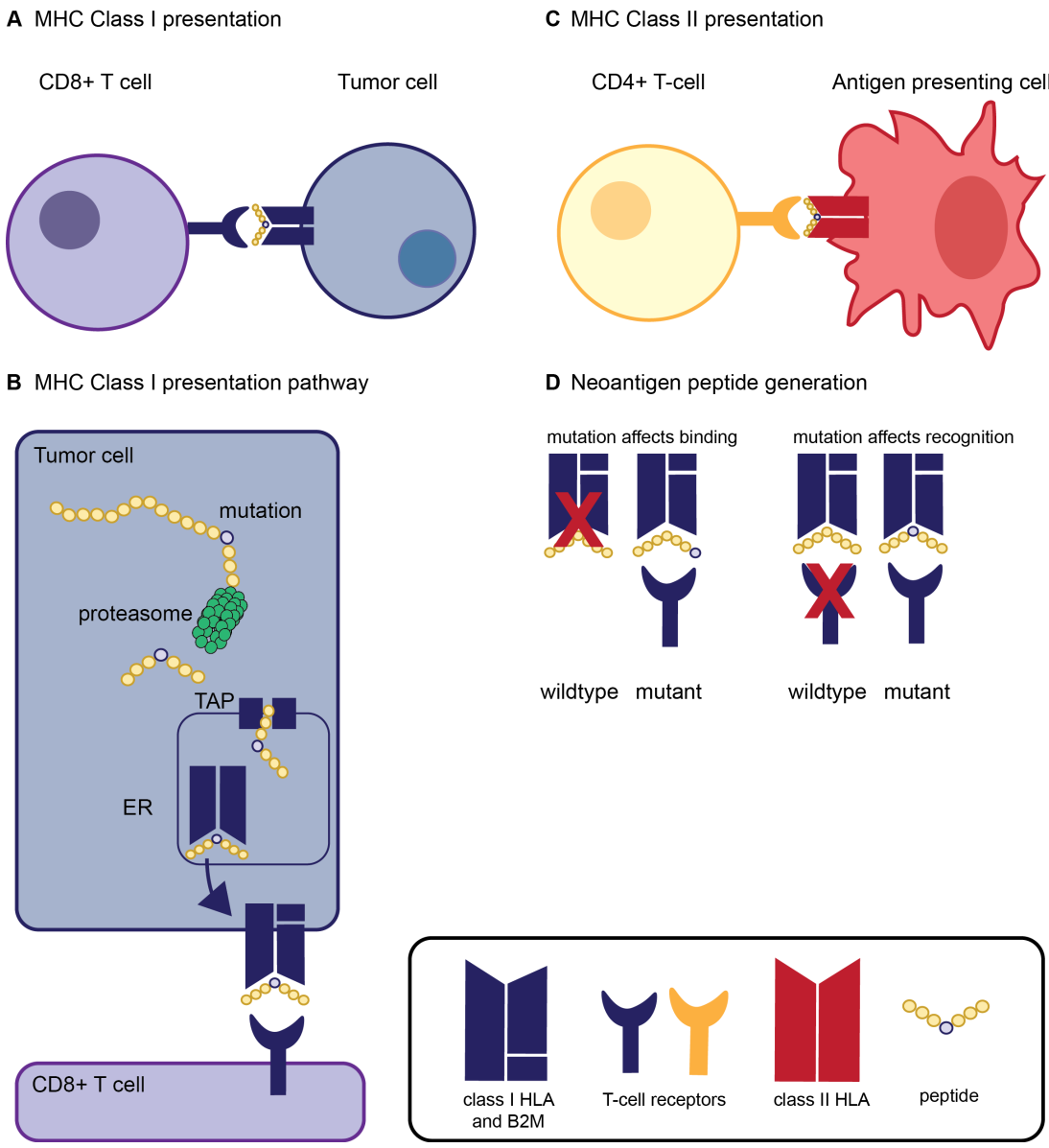

Figure 1-2: Antigen presentation pathways.
A schematic of MHC class I (A) and class II antigen presentation (C). This thesis mainly focuses on MHC class I antigen presentation, so a schematic is also given detailing how mutations present in a tumor cell are processed and eventually presented on the cell surface for potential recognition by CD8+ T-cells (B). Scenarios that could generate an antigenic mutation (D).

1.3.5 Disentangling the immune contexture

Effective tumor recognition requires both the presentation of tumor-specific antigens and a functional immune environment, replete with cells capable of eliciting and sustaining an immune response. The analysis of the different immune cell subpopulations present in and around the tumor, as well as the specific location and proportions of these cells, across a wide range of tumors has allowed for the elucidation of both beneficial and detrimental aspects of immune infiltration (Fridman et al., 2012).

Generally infiltration of cytotoxic T-cells and dendritic cells has been shown to confer favorable prognosis across a variety of cancer types (Clemente et al., 1996, Fridman et al., 2012, Gentles et al., 2015, Roberts et al., 2016). However, in some instances, the association with patient survival has been proven be dependent on the location of cells in the microenvironment, with only the infiltration of CD8+ T-cells within the tumor, rather than at the tumor margin, improving patient survival (Naito et al., 1998).

Other immune cells are associated with poor prognosis, such as myeloid-derived suppressor cells and tumor-associated macrophages (Kumar et al., 2016b, Guo et al., 2016) indicating that understanding the composition of the tumor microenvironment, rather than using an estimate of tumor purity as a proxy for the total level of immune infiltration is key to predicting patient outcome (Racle et al., 2017).

Finally some immune cell subpopulations have a less clear relationship with patient prognosis, owing either to imperfect quantification due to poor cell type markers, unique host microenvironments, tissue type of tumor origin, or factors that remain undetermined. For instance in different cancer types, regulatory T-cells have been found to be associated with worse overall survival (breast, hepatocellular carcinoma) (Bates et al., 2006, Fu et al., 2007), an improved overall survival (colorectal, follicular lymphoma) (Frey et al., 2010, Carreras et al., 2006), or have no relationship at all with survival (breast, brain) (Mahmoud et al., 2011, Jacobs et al., 2010). There have also been inconsistent reports on whether B cell infiltration is associated with improved overall survival (DiLillo et al., 2010, Qin et al., 1998, Schultz et al., 1990).

Estimates of immune infiltration have been developed and applied to large-scale cancer transcriptomics datasets to better understand the relationship between immune infiltration and survival, as well as to identify potential targets for immunotherapy (Gentles et al., 2015, Li et al., 2016, Davoli et al., 2017, Angelova et al., 2015, Danaher et al., 2017). Furthermore, to better stratify tumors based on the activity of the immune cells found in immune microenvironment rather than just quantifying their presence, a number of additional measures have been proposed such as the immune score (Galon et al., 2012), a score of cytolytic activity (Rooney et al., 2015), and the immunophenoscore (Charoentong et al., 2017). Such scores have also been used to predict metastatic potential and response to immunotherapy (Mlecnik et al., 2016, Charoentong et al., 2017).

1.3.6 Promise of immunotherapy and resistance

Recent clinical trials have shown that modulation of the immune system through the blockade of immune checkpoint molecules such as anti-cytotoxic T lymphocyte antigen-4 (CTLA-4), programmed cell death-1 (PD-1), or programmed cell death ligand-1 (PD-L1) results in improved antitumor responses and clinical benefit in a variety of cancers (Hodi et al., 2010, Brahmer et al., 2012, Topalian et al., 2012a, Wolchok et al., 2013). These antibodies function by inhibiting two of the pathways regulating T-cell activation (Figure 1-3). The co-inhibitory receptor CTLA-4 prevents the initial activation of T-cells in lymph nodes by competing with CD28 for binding of shared ligands (Pardoll, 2012). The PD-1 pathway limits T-cell activity in the peripheral tissues as a way to prevent prolonged inflammatory responses and limit autoimmunity (Pardoll, 2012).

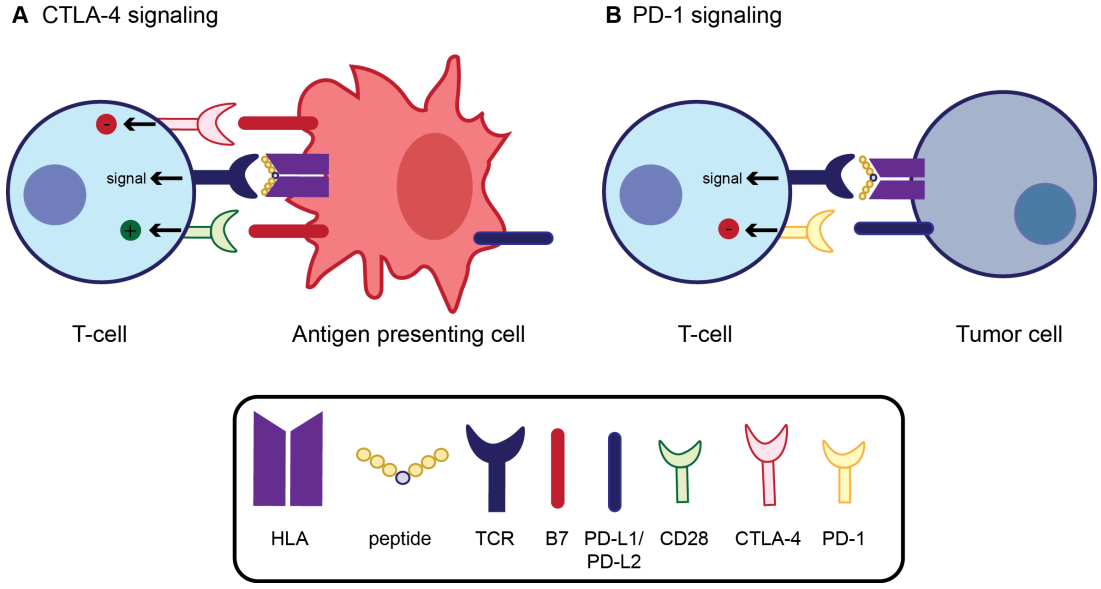

Figure 1-3: CTLA-4 and PD-1 regulatory pathways.
The inhibitory receptors CTLA-4 (A) and PD-1 (B) which act to down-regulate T-cell activation are shown. Antibodies targeting these receptors allow for an increase in T-cell activity by removing the negative signal dampening T-cell response.

Genetic analyses of tumors from patients with malignant melanoma and non-small cell lung cancer (NSCLC) treated with immune checkpoint inhibitors have revealed a relationship between the number of mutations the tumor harbors, its neoantigenic burden, and clinical response, with improved response observed among those patients with a higher mutation/neoantigen burden (Snyder et al., 2014, Rizvi et al., 2015, Van Allen et al., 2015). While cancer immunotherapy has resulted in durable antitumor responses in a subset of patients treated (Hodi et al., 2010, Wolchok et

al., 2013, Topalian et al., 2012b), current genomic and molecular biomarkers predicting response fail to perfectly stratify groups of patients into those that will benefit from immunotherapy and those who will show no clinical benefit, with frequent overlap between responding patients and non-responders for a given biomarker (Hong et al., 2015). Moreover, even among those patients originally improving while on therapy, the mechanisms through which a tumor develops resistance remain incompletely cataloged (Rizvi et al., 2015, Snyder et al., 2014, Roh et al., 2017, Chen et al., 2016).

1.3.6.1 Tumor-intrinsic immune escape mechanisms

Thus far, many tumor-intrinsic mechanisms of immunotherapy resistance have been identified (Figure 1-4). The presence of inhibitory immune checkpoint molecules, such as PD-L1, LAG-3, and TIM-3 down-regulate T-cell activity, usually as a way to maintain immune homeostasis and prevent uncontrolled immune responses. However tumors may overexpress these negative regulatory molecules, providing an avenue for immune escape (Tumeh et al., 2014, Matsuzaki et al., 2010, Sakuishi et al., 2010, Woo et al., 2002, Powles et al., 2014, Herbst et al., 2014). Additionally, oncogenic events themselves, such as increased ß-catenin signaling and PTEN loss have been shown to inhibit T-cell mediated killing (Spranger et al., 2015, Peng et al., 2016). Even in the presence of capable T-cells, lack of tumor antigen recognition due to defects in antigen presentation (Tran et al., 2016, Zaretsky et al., 2016, Zhao et al., 2016) and insensitivity to T-cell effector molecules such as IFN-y signaling (Zaretsky et al., 2016, Minn and Wherry, 2016, Benci et al., 2016, Gao et al., 2016, Sucker et al., 2017) can result in resistance. Additional evidence is emerging that transcriptomic re-programming can lead to immunotherapy resistance as well (Hugo et al., 2016).

1.3.6.2 Tumor-extrinsic immune escape mechanisms

Tumor-extrinsic mechanisms of immunotherapy resistance include immune checkpoints, such as PD-1 and CTLA-4, infiltration of immunosuppressive cell populations, and T-cell exhaustion. Finally, there may be additional characteristics associated with resistance that are not yet well understood, such as the composition of the gut microbiome (Sivan et al., 2015, Vetizou et al., 2015, Routy et al., 2017, Gopalakrishnan et al., 2017).

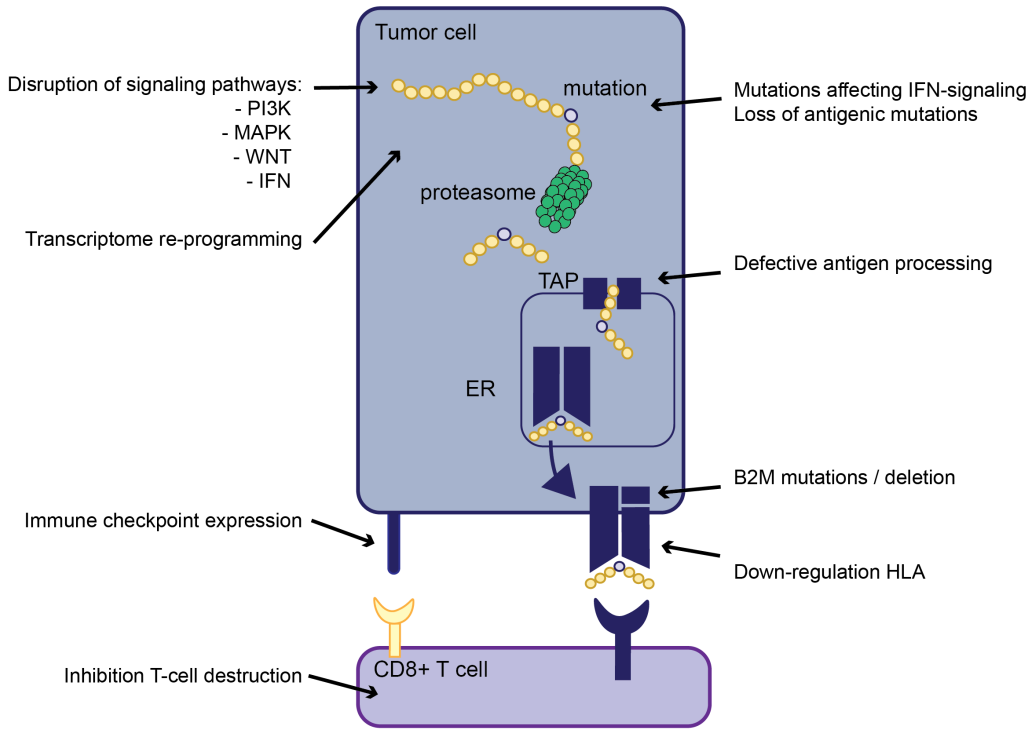

Figure 1-4: Tumor-intrinsic mechanisms of immunotherapy resistance.

1.4 Tools for understanding the tumor-immune interaction

Neoantigen predictions can be generated using the known non-synonymous mutations present in a tumor. Furthermore, the tumor microenvironment can also be investigated to understand the prevalence of infiltrating immune cells and their composition. In combination, these tools allow for a thorough profiling of the components and targets of the immune system, shedding light on the factors that contribute to anti-tumor immunity, response to immunotherapy, and mechanisms of immune evasion.

In order to filter the list of all non-synonymous mutations present in the tumor down to those likely to induce an immune response, the likelihood of a peptide containing the mutation being presented to a T-cell must be considered. There are many steps from mutation generation to class I antigen presentation, broadly classified as proteasomal cleavage, TAP transport, and MHC binding, with tools available to generate predictions at each step (Figure 1-5).

1.4.1 HLA typing

Generally considered the most selective step in antigen presentation, the binding of a mutant peptide to the cell's HLA molecules for presentation is determined by what specific HLA alleles are present in the cell. Thus determining the HLA type of a patient is essential for predicting neoantigenic mutations. This task has proven to be a challenging one, as the HLA locus is highly polymorphic. High-resolution HLA typing was once only achievable at low-throughput via serotyping, but recent algorithms are now capable of using short read DNA or RNA sequence data (Szolek et al., 2014, Shukla et al., 2015, Liu et al., 2013, Boegel et al., 2012, Warren et al., 2012).

Most of these approaches rely on an extensive database of known HLA alleles maintained by international ImMunoGeneTics project (IMGT) (Lefranc et al., 2009). Sequencing reads that may contain HLA sequence information are extracted and mapped to the known alleles from the IMGT database. The allelic combination which maximizes the number of reads explained is then chosen as the individual's likely HLA type (Szolek et al., 2014, Shukla et al., 2015). Alternatively, the sequencing reads may be assembled into contigs and matched to the HLA allele pairs that are nearest to the contig sequence (Warren et al., 2012, Liu et al., 2013).

It is also possible to identify somatic mutations affecting the HLA locus by using the sequence of patient's inferred HLA alleles as the reference and comparing that reference to what was obtained after the re-alignment of the extracted HLA sequencing reads (Shukla et al., 2015). Somatic HLA mutations acquired by the tumor have the potential to disrupt tumor antigen presentation, possibly allowing an avenue for the tumor to escape immune predation (Shukla et al., 2015).

1.4.2 Class I neoantigen predictions

From a given mutation, multiple possible peptides may arise. The exact peptide sequence which may be generated is determined by what sites are cleaved by the proteome, which is responsible for most intra-cellular protein degradation. Large polypeptides are degraded into smaller peptide fragments, usually of length 8-11 amino acids. Neural network based prediction methods exist that allow for the prediction of proteasome cleavage sites to identify what peptides are likely to arise from a given mutation (Nielsen et al., 2005, Kesmir et al., 2002).

Once a mutant peptide has been generated, it may be transported to the endoplasmic reticulum where the association between the peptides and the MHC molecules can occur. Thus prediction methods also exist to determine whether a peptide is likely to bind to the TAP transporter (Bhasin et al., 2007, Diez-Rivero et al., 2010).

After transport of the peptide to the ER, the crucial step of peptide-MHC binding may occur. Arguably, this step is the most important for predicting the immunogenicity of a peptide, (Sette et al., 1994) and correspondingly, the largest number of tools are available to make predictions of this interaction (Hoof et al., 2009, Nielsen and Andreatta, 2016, Andreatta and Nielsen, 2016, O'Donnell et al., 2017, Moutaftsi et al., 2006, Kim et al., 2009). Most tools have been extensively trained on data deposited in the Immune Epitope Database (IEDB) (Vita et al., 2015). For instance, the NetMHC family of tools uses artificial neural networks to predict peptide binding to any input MHC molecule, regardless of the extent of available training data or whether it has even been identified in IMGT (Nielsen and Andreatta, 2016, Andreatta and Nielsen, 2016). Additionally the most recent version of the tools also estimate the likelihood of a peptide naturally occurring as a ligand (Jurtz et al., 2017).

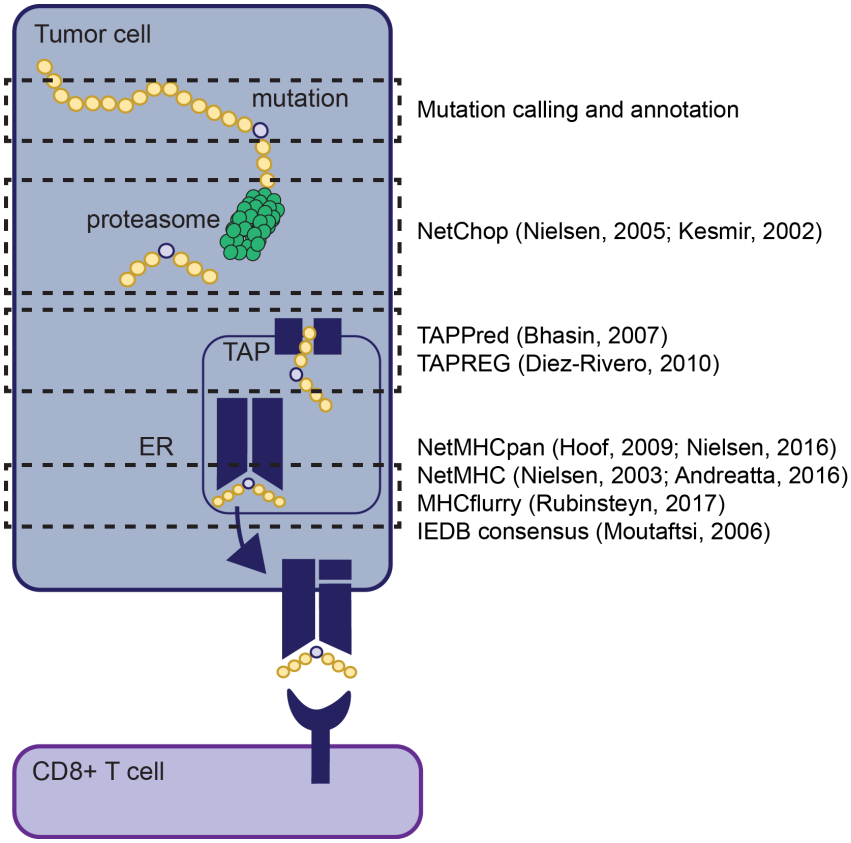

Figure 1-5: Schematic of HLA class I presentation and predictive tools available.
The general steps that occur in order for a mutation to be presented for possible recognition by CD8+ T-cells are depicted. Commonly used bioinformatics methods are provided at each step of the presentation pathway.

In theory, predictions can be made for all the steps comprising class I antigen processing, and prediction software now exists to incorporate all aspects of antigen processing (Larsen et al., 2007). Frequently, however, all possible peptides arising from a given mutation are considered as possible neoantigens and the most emphasis is placed on predicting the peptide-MHC interaction.

1.4.3 Class II neoantigen predictions

Neoantigen predictions can also be made for epitopes presented on professional antigen presenting cells via the class II antigen processing pathway; however, the steps involved in this pathway are more complex, and the epitopes themselves are longer with more flexible MHC binding positions, rendering results less accurate (Roche and Furuta, 2015). Tools are available for predicting the class II MHC/peptide interaction, but fewer benchmarking studies have been performed to determine appropriate thresholds for classification of peptides as weak or strong binders (Nielsen et al., 2008, Nielsen et al., 2010, Lund et al., 2013).

1.4.4 Immune microenvironment

While historically, deciphering the quantity and composition of TILs relied on flow cytometry and immunohistochemcial approaches, it is now possible to use bulk tissue sequencing data to identify different immune cell subpopulations infiltrating the tumor. To achieve this, a number of deconvolution methods have been developed that rely on the expression profiles of different immune cell types to estimate the composition of infiltrating cells resulting in the bulk tumor gene expression values observed (Newman et al., 2015, Li et al., 2016, Becht et al., 2016, Racle et al., 2017). These approaches can use linear regression or quadratic programming and a reference expression matrix generated from sorted immune cells to computationally work backwards to decipher the immune cell subtype fractions (Hackl et al., 2016).

A second class of methods relies on an enrichment based approach to either determine which cell populations are enriched in a subset of patients (GSEA), or which gene sets are coordinately up- or down-regulated in a single tumor sample (ssGSEA) (Hackl et al., 2016, Angelova et al., 2015, Senbabaoglu et al., 2016). These methods provide an estimate of which immune cell subtypes are statistically enriched in the samples considered.

Finally, a number of marker gene enrichment methods have been defined which are based simply on the combination of reference gene expression values (Danaher et al., 2017, Rooney et al., 2015, Davoli et al., 2017, Bindea et al., 2013).

The different methods make various assumptions to determine which genes are best suited to make up the reference immune cell gene sets. For instance, TIMER assumes that the expression of true immune genes must negatively correlate with purity, as purity is the measure of tumor content, and thus this method removes any genes that do not show a strong negative correlation with tumor purity (Li et al., 2016). Alternatively, some immune cell measures assume that gene sets that define the same cell population must strongly correlate with one another (Danaher et al., 2017), and others remove any genes that have been found in the definition of multiple immune cell subtypes or positively select for exclusively expressed genes, leaving only genes unique to a particular immune subtype (Davoli et al., 2017, Rooney et al., 2015). Importantly, creating a reference signature matrix that accurately and specifically identifies the different immune cell subtypes has proven to be challenging, and ultimately, the reference signature matrix can alter output

results drastically. Furthermore, it remains to be determined if tumor specific expression matrices are required, or if the immune cell expression profiles remain consistent across tumor types and cancer in general.

Chapter 2 Data and Methods

2.1 Data

2.1.1 TRACERx multi-region sequencing pilot data

Samples from the TRACERx pilot study (L001, L002, L003, L004, L008, L011 and L012) were obtained from patients diagnosed with non-small cell lung cancer (NSCLC) who underwent definitive surgical resection prior to receiving any form of adjuvant therapy, such as chemotherapy or radiotherapy. Informed consent allowing for genome sequencing had been obtained. All samples were collected from University College London Hospital, London (UCLHRTB 10/H1306/42). Tumor samples were subjected to pathology review to establish the histological subtype: five tumors were classified with CK7+/TTF1+ adenocarcinoma (L001, L003, L008 and L011) or poorly-differentiated CK7+ carcinoma (L004) histology, one tumor (L012) with squamous cell carcinoma histology and one tumor (L002) with adenosquamous histology. Two patients presented with disease in two separate lobes of the lung (L003 and L008). Detailed clinical characteristics are provided in Table 4-2.

Multiple regions (up to five from a single tumor) were taken, separated by 1cm intervals. Peripheral blood was also collected at time of surgery from all patients.

2.1.2 TRACERx multi-region sequencing data

The TRACERx 100 cohort comprises the first 100 patients prospectively analyzed by the lung TRACERx main study (https://clinicaltrials.gov/ct2/show/NCT01888601, approved by an independent Research Ethics Committee, 13/LO/1546) and mirrors the prospective 100 patient cohort described in (Jamal-Hanjani et al., 2017).

All surgically resected tumor samples were macroscopically reviewed by a pathologist. Spatially separated tumor regions, documented by photography, were collected and snap frozen in liquid nitrogen for subsequent DNA extraction. At least two regions from each tumor, separated by at least 3mm, were collected. The samples were taken as to maximize tumor cellularity (areas that were obviously necrotic, fibrotic, or hemorrhagic were avoided) and to reflect the observed macroscopic morphological heterogeneity of the tumor. Peripheral blood was also obtained at time of surgery.

Diagnostic tumor sections from lung adenocarcinoma and lung squamous cell carcinoma cases were subject to histopathological grading.

For analysis using the tool LOHHLA, four patients were excluded due to homozygosity at all three HLA loci or too few mismatch positions between HLA alleles. Lung adenocarcinoma and lung squamous cell carcinoma tumors were considered for downstream analyses. Seven tumors were classified as having a separate histology. Of these, one carcinosarcoma exhibited HLA LOH and three adenosquamous carcinomas, one carcinosarcoma, one large cell carcinoma, and one large cell neuroendocrine tumor did not.

2.1.3 The Cancer Genome Atlas (TCGA) data

Some analysis presented in this thesis was based upon data generated by the TCGA Research Network. Information about TCGA and the investigators and institutions that constitute the TCGA Research Network can be found at http://cancergenome.nih.gov/.

2.1.3.1 TCGA (processed mutational data)

TCGA mutation data for mutational signatures analysis was obtained from Broad Institute MAF dashboard as detailed:

Urothelial bladder carcinoma (BLCA):

- PR_TCGA_BLCA_PAIR_Capture_All_Pairs_QCPASS_v3.aggregated.capture.tcga.uuid.somatic.maf (center: broad.mit.edu, archive version: 0.3.0)

Breast invasive carcinoma (BRCA):

- genome.wustl.edu_BRCA.IlluminaGA_DNASeq.Level_2.1.1.0.curated.somatic.maf (center: genome.wustl.edu, archive version: 1.1.0)

Colon adenocarcinoma (COAD):

- hgsc.bcm.edu_COAD.IlluminaGA_DNASeq.1.somatic.maf (center: hgsc.bcm.edu, archive version: 1.5.0)

Esophageal carcinoma (ESCA):

- An_TCGA_ESCA_External_capture_All_Pairs.aggregated.capture.tcga.uuid.automated.somatic.maf (center: broad.mit.edu, archive version: 1.0.0)

Glioblastoma multiforme (GBM):

- step4_gbm_liftover.aggregated.capture.tcga.uuid.maf2.4.migrated.somatic.maf (center: broad.mit.edu, archive version: 1.4.0)

Head and neck squamous cell carcinoma (HNSC):

- PR_TCGA_HNSC_PAIR_Capture_All_Pairs_QCPASS_v2.aggregated.capture.tcga.uuid.somatic.maf (center: broad.mit.edu, archive version: 0.2.0)

Lung adenocarcinoma (LUAD):

- PR_TCGA_LUAD_PAIR_Capture_All_Pairs_QCPASS_v4.aggregated.capture.tcga.uuid.automated.somatic.maf (center: broad.mit.edu, archive version: 1.5.0)

Lung squamous cell carcinoma (LUSC):

- LUSC_Paper_v8.aggregated.tcga.somatic.maf (center: broad.mit.edu, archive version: 1.5.0)

Skin cutaneous melanoma (SKCM):

- PR_TCGA_SKCM_PAIR_Capture_All_Pairs_QCPASS_v4.aggregated.capture.tcga.uuid.automated.somatic.maf (center: broad.mit.edu, archive version: 1.4.0)

One lung adenocarcinoma patient, TCGA-05-4396, was excluded for having over 7000 low quality mutations called, mostly in a C[C>G]G context. A lung squamous cell carcinoma patient, TCGA-18-3409, was excluded for bearing a strong UV signature, uncharacteristic of a LUSC tumor.

These cancer types were chosen as they had available mutation tables with VAFs calculated and been previously analyzed for mutational timing (McGranahan et al., 2015).

2.1.3.2 TCGA (processed RNAseq data)

Processed RNA-sequencing data was downloaded from the TCGA data portal. For each lung adenocarcinoma and lung squamous cell sample, all available 'Level_3' gene-level data was obtained. This data had been quantified using RSEM. Gene-level data reflecting either raw counts, for use in differential expression analyses, or TPM was considered.

2.1.3.3 TCGA (raw data)

Tumor and matched germline exome sequencing BAM files for both lung adenocarcinoma (n = 397) and lung squamous cell carcinoma (n = 350), were

obtained from TCGA (http://cancergenome.nih.gov/) via https://cghub.ucsc.edu. The BAM files were converted to FASTQ using bedtools bamtofastq (v2.25). The resulting sequence files were processed as described for the TRACERx cohort, with a set of annotated variants being generated as the output.

2.1.3.4 TCGA (clinical data)

Clinical data for TCGA patients was accessed through the TCGA data portal. Patients were first grouped according to quartile of the variable being considered. Survival analyses were then performed in R using the survival package. Complete survival data was available for 139/150 lung adenocarcinoma patients and 122/124 lung squamous cell carcinoma patients.

2.1.4 Pembrolizumab treated NSCLC patient data

A patient cohort of stage IV NSCLC was obtained from (Rizvi et al., 2015). A detailed description of this patient cohort, including tumor processing, can be found in supplementary material of (Rizvi et al., 2015).

2.1.5 Ipilimumab treated melanoma patient data

Samples obtained from (Snyder et al., 2014) reflected a patient cohort of late stage melanoma, and a detailed description of this patient cohort, including tumor processing, can be found in supplementary material of (Snyder et al., 2014).

2.2 Methods

2.2.1 Whole exome sequencing

2.2.1.1 TRACERx multi-region sequencing pilot

Exome capture was performed on 1-2µg of DNA from each tumor and matched germline sample, using the Agilent Human All Exome V4 kit according to manufacturer's protocol. The samples were paired-end multiplex sequenced on the Illumina GAII or HiSeq2500 at the Advanced Sequencing Facility at the London Research Institute (LRI). The desired sequencing depth was 100x.

2.2.1.2 TRACERx multi-region sequencing

Exome capture was performed on 1-2µg of DNA isolated from genomic libraries with median insert size of 190bp for each tumor and matched germline sample. A

customized Agilent Human All Exome V5 kit was used according to the manufacturer's protocol. Samples were 100bp paired-end multiplex sequenced on the Illumina HiSeq 2500 and HiSeq 4000 at the Advanced Sequencing Facility at The Francis Crick Institute. The median sequencing depth was 431 (range 83-986) for tumor regions and 415 (range 107-765) for matched germline. To prevent inter-patient sample swaps, germline SNP profiles were compared, as they should be highly similar between all tumor regions and associated germline sample.

2.2.2 Whole genome sequencing

2.2.2.1 TRACERx multi-region sequencing pilot

Paired-end whole genome sequencing was performed on 1µg of DNA four tumor regions (L002:R1, L002:R3, L008:R1, L008:R3) and matched blood by Illumina Cambridge LTD. The desired sequencing depth was 100x for tumor, 40x for germline.

2.2.3 Multi-region somatic alteration calling

All somatic alteration calling was performed using either the pilot version or the main study version of the TRACERx pipeline designed by Gareth Wilson and Richard Mitter. Full details of the pipelines can be found in (de Bruin et al., 2014) and (Jamal-Hanjani et al., 2017).

2.2.3.1 TRACERx multi-region pilot mutation calling

Raw paired-end sequencing reads were aligned to hg19, including all contigs, obtained from the GATK bundle (v2.8) using bwa mem (bwa-0.5.9). Up to 3-4 mismatches were allowed per read for the GAII or HiSeq, respectively. Files from the same patient region were cleaned, sorted, merged, and duplicate reads removed using Picard tools (v1.8).

SAMtools mpileup (0.1.19) was used to find non-reference positions in tumor and germline samples. Bases with low phred score (<20) or reads with low mapping quality scores (<20) were removed. Somatic variants were identified using VarScan2 somatic (v2.3.6) and extracted using VarScan2 processSomatic (Koboldt et al., 2012). All single nucleotide variant (SNV) calls were filtered for false positives using VarScan2's fpfilter.pl script. Variants were only kept if it was present in >=5%

sequencing reads in one tumor region and present with <=2 reads in germline and >=2 reads in the tumor region.

Small insertions and deletions (indels) were identified using Pindel (v0.2.4) (Ye et al., 2009).

All variants were annotated using ANNOVAR (Wang et al., 2010) and dbNSFP (Liu et al., 2011). All non-silent variants were manually reviewed using Integrated Genomics Viewer (IGV v38).

2.2.3.2 TRACERx multi-region study mutation calling

Raw paired-end sequencing reads were aligned to hg19, including all contigs, obtained from the GATK bundle (v2.8) using bwa mem (bwa-0.7.8). Files from the same patient region were cleaned, sorted, merged, and duplicate reads removed using Picard tools (v1.107).

SAMtools mpileup (0.1.19) was used to find non-reference positions in tumor and germline samples. Bases with low phred score (<20) or reads with low mapping quality scores (<20) were removed. Somatic variants were identified using VarScan2 somatic (v2.3.6) and extracted using VarScan2 processSomatic (Koboldt et al., 2012). All SNV calls were filtered for false positives using VarScan2's fpfilter.pl script. All indel calls in reads classified by VarScan2 as "high confidence" were kept for downstream filtering. Additionally, MuTect (v1.1.4) was used to identify SNVs (Cibulskis et al., 2013). These variants were filtered according to the filter parameter "PASS".

To avoid false positive variant calls, additional filter steps were taken. Variants called by both VarScan2 and MutTect were considered true positives if the variant allele frequency (VAF) was >2%. If the variant was only identified by VarScan2, a VAF of >5% was required. Furthermore, the sequencing depth in each region was required to be >=30 and >=5 sequence reads had to support the call. To ensure the variant was not a germline event, the number of reads containing the variant in the germline data had to be <5 and VAF <=1%.

To utilize the multi-region sequencing aspect of the cohort, individual mutations called across each region from the same tumor were compared. The threshold for detection of a somatic variant in one tumor region was reduced to VAF >=1% if the

same variant had been detected at the VAF>=5% in another tumor region from the same tumor.

Indels were filtered using the same parameters as SNVs, except >=10 reads had to support the variant call and the region had to have a sequencing depth >=50.

All variants were annotated using Annovar (Wang et al., 2010) and COSMIC (v75).

2.2.4 Copy number analysis

Varscan2 copynumber was run to generate copy number data from paired tumor-normal samples, which produced per-region logR values, that were subsequently GC corrected (Koboldt et al., 2012). Homozygous and heterozygous single nucleotide polymorphisms (SNPs) were identified from the germline sample using Platypus (v0.8.1) (Rimmer et al., 2014). The B-allele frequency (BAF) of each SNP was calculated as the proportion of the reads at that position that contained the variant base.

The logR and BAF values were used with ASCAT (v2.3) (Van Loo et al., 2010) in order to generate segmented allele-specific copy number data, purity, and ploidy estimates.

Gene-level amplifications were called if the log2(mean gene copy number/ploidy) was >1. Gene-level deletions were called if the log2(mean gene copy number/ploidy) was <-1.

To compare LOHHLA to additional tools, Sequenza (Favero et al., 2015), and TITAN (Ha et al., 2014) were also implemented. In both cases, default settings were used. For TITAN, the purity estimates from ASCAT were used as input.

2.2.5 Timing of somatic events

2.2.5.1 TRACERx multi-region pilot timing

Clusters of mutations were identified using a Dirichilet process. Ubiquitous clonal mutations, which were identified in every tumor region, were considered events on the trunk of the phylogenetic tree. Heterogeneous mutations were considered events on the branch of the phylogenetic tree.

2.2.5.2 TRACERx multi-region timing

To estimate whether mutations were clonal or subclonal, a modified version of PyClone was used (Roth et al., 2014). For each mutation, an observed CCF (obsCCF) and a phylogenetic CCF (phyloCCF), which took into consideration any subclonal copy number events potentially altering the CCF, was calculated. Mutations were clustered using PyClone Dirichlet process clustering.

2.2.6 Phylogenetic tree construction

Phylogenetic trees for the TRACERx study were constructed using CITUP (v0.1.0) (Malikic et al., 2015), which takes as input mutation clusters and their mean cancer cell fraction. TRACERx tumors with phyloCCF Pyclone output from at least two tumor regions and containing at least two mutation clusters were included. Mutation cluster trees were filtered to include those that contained at least five mutations and adhered with evolutionary principles (Jamal-Hanjani et al., 2017). For six tumors, manual tree construction was required (CRUK0004, CRUK0017, CRUK0032, CRUK0062, CRUK0065, CRUK0069).

2.2.7 Checkpoint blockade clinical efficacy analysis

For each sample cohort obtained from previously published work, clinical efficacy analysis was kept consistent with the original publication.

Rizvi cohort (Rizvi et al., 2015): Objective response to pembrolizumab was assessed by investigator-assessed immune-related response criteria (irRC) by a study radiologist. As outlined in protocol, CT scans were performed every nine weeks. Partial and complete responses were confirmed by repeat imaging occurring a minimum of 4 weeks after the initial identification of response; unconfirmed responses were considered stable or progressive disease dependent on results of the second CT scan. Durable clinical benefit (DCB) was defined as stable disease or partial response lasting longer than 6 months (week 27, the time of third protocol-scheduled response assessment). No durable benefit (NDB) was defined as progression of disease ≤ 6 months following commencement of therapy. For patients with ongoing response to study therapy, progression-free survival was censored at the date of the most recent imaging evaluation. For ‘alive’ patients, overall survival was censored at the date of last known contact.

Snyder cohort (Snyder et al., 2014): Long-term clinical benefit was defined by radiographic evidence of freedom from disease or decreased volume of disease for > 6 months. Conversely, lack of long-term benefit was defined by tumor growth on every computed tomographic scan after the initial treatment (no benefit) or a clinical benefit lasting 6 months or less (minimal benefit).

2.2.8 Comparison of ASCAT and LOHHLA

In order to compare ASCAT and LOHHLA, each tumor region was treated as a separate sample. It was then run it through the LOHHLA pipeline with default settings. All samples were included in this validation analysis, regardless of whether they were subsequently included in downstream LOHHLA analyses.

As there was too little coverage of the HLA alleles to accurately infer the copy number at these loci using ASCAT, the segment overlapping the HLA locus was used as a proxy for the HLA locus. If no segment overlapped the HLA locus, as was the case for 25 tumor regions from seven tumors, the closest genomic segment was used instead.

LOHHLA outputs an allelic imbalance estimate, as well HLA allele specific copy number estimates. Tumor regions were considered to be concordant if ASCAT predicted allelic imbalance across the locus and at least one HLA gene was found to harbor allelic imbalance using LOHHLA. Similarly, for LOH estimates, ASCAT and LOHHLA were considered to be concordant if ASCAT predicted a minor allele of 0 and this was also predicted for at least one HLA gene by LOHHLA.

Conversely, allelic imbalance estimates were classified as discordant if allelic imbalance was predicted in any HLA gene using LOHHA and not with ASCAT or vice versa. Similarly, for LOH was classified as discordant if any HLA gene using LOHHLA was classified as exhibiting a minor allele of 0 and no LOH was identified using ASCAT or vice versa.

2.2.9 HLA Type and HLA Mutations

2.2.9.1 TCGA cohorts

All TCGA patients were HLA typed using Polysolver (POLYmorphic loci reSOLVER) (Shukla et al., 2015), using default settings, as described in (Shukla et al., 2015).

Polysolver uses a normal tissue BAM file as input and employs a Bayesian classifier to determine HLA genotype.

2.2.9.2 TRACERx cohorts

Pilot patients L011 and L012 were serotyped and simultaneously genotyped using Optitype (Szolek et al., 2014) and Polysolver (Shukla et al., 2015), which all produced concordant results. The remaining pilot patients (L001, L002, L004, and L008) were genotyped using Optitype.

Patients from the TRACERx main study were genotyped using both Optitype and Polysolver. When discordant results were given, the Polysolver output was used. Additional details can be found in section 4.2.2. HLA mutations in each tumor region were also assessed using Polysolver.

2.2.10 Predicted neoantigen binders

2.2.10.1 TCGA cohorts, checkpoint blockade treated cohorts, and TRACERx pilot study neoantigen prediction

Novel 9-11mer peptides that could arise from identified non-silent mutations present in the sample (Jamal-Hanjani et al., 2017) were determined. The predicted IC50 binding affinities and rank percentage scores, representing the rank of the predicted affinity compared to a set of 400,000 random natural peptides, were calculated for all peptides binding to each of the patient's HLA alleles using netMHCpan-2.8 (Andreatta and Nielsen, 2016, Nielsen et al., 2003, Hoof et al., 2009). Putative neoantigen binders were those peptides with a predicted binding affinity <500nM.

2.2.10.2 TRACERx main study neoantigen prediction

Neoantigen predictions were made as above, with the exception that netMHC-4.0 was also run for each peptide (Andreatta and Nielsen, 2016, Nielsen et al., 2003). Predicted binders were considered those peptides that had a predicted binding affinity <500nM or rank percentage score <2% by either tool.

When RNAseq data was available, a neoantigen was considered to be expressed if at least five RNAseq reads mapped to the mutation position, and at least three contained the mutated base.

2.2.11 Mapping HLA LOH to phylogenetic trees

LOH events detected in every tumor region tested were considered to be clonal events and mapped to the trunk of the phylogenetic tree. For heterogeneous LOH events, the regional copy number of the HLA allele lost was used in conjunction with the patient tree structure and subclone cancer cell fractions in a quadratic programming approach, using the R package “quadprog”, to determine the best placement of the LOH event.

This was achieved by solving a quadratic programming equation:

min(-d^T b + 1/2 b^T D b)

with the constraints:

A^T b >= bvec

The LOH event was tested at each branch. For each possibility, the phylogenetic tree was broken into two, one containing all clones after the query branch and the other consisting of the remainder of the tree. A 2xn matrix, where n is the number of regions sampled, was constructed containing the regional sum of the cancer cell fractions for each subclone in the subtree and the regional sum of cancer cell fractions from subclones in the remaining tree. The regional cancer cell fraction matrix was multiplied by the transpose of itself to generate a 2x2 matrix for input (*Dmat*) into the quadprog function, solve.QP. The vector to be minimized (*dvec*) was obtained by multiplying the LOHHLA calculated HLA allele copy number for each region by the transpose of the regional cancer cell fraction matrix. Finally, the solve.QP function was called with *Dmat* and *dvec*, using a constraint matrix, *Amat*, such that all results had to be positive and a constraint vector, *bvec*, such that the estimated copy number of HLA allele for the remaining tree was at least 0.5. The errors between observed and predicted copy number values from placing LOH event on each branch were output and the solution providing the least error was selected.

Each mapped event was inspected and events that did not fit the phylogenetic tree or had large error values, either indicating the presence of an additional subclone or multiple independent HLA LOH events, were manually adjusted. Patients CRUK0013, CRUK0061, CRUK0082, and CRUK0084 had HLA LOH events that did not fit the current phylogenetic tree, so additional nodes (indicated in grey) were

included to contain the HLA LOH event. Patients CRUK003, CRUK0032, CRUK0051, and CRUK0062 had multiple independent HLA LOH events which were manually mapped.

2.2.12 Assessing significance of focal and arm-level LOH

In order to assess whether HLA LOH occurred more than expected by chance, LOH events were first described as either being focal or arm-level in nature. To classify LOH as arm-level or focal, the minor allele frequency across the genome was considered. First, any segments (as predicted by ASCAT) with identical minor allele copy numbers were merged. Subsequently, segments that spanned >=75% the length of a given chromosome arm, were classified as 'arm-level', while segments that were <75% were considered focal.

To assess the significance of focal events, for each tumor, the proportion of the genome subject to focal minor allele loss was determined. This value was assumed to reflect the probability for focal minor allele loss in each tumor. Based on this probability, an aberration state (loss or no loss) for each sample was generated and the proportion of samples exhibiting loss was determined. This process was repeated 10,000 times to obtain a background distribution reflecting the likelihood of observing losses given the probability of loss in each sample. A p-value reflecting the likelihood of observing the level of minor allele loss seen at the HLA locus was determined by counting the percentage of simulations showing a higher proportion loss than that observed.

The same procedure was conducted for arm-level events, using the observed frequency of arm-level allele specific loss in each tumor.

2.2.13 Survival analyses

All survival analyses were then performed in R using the survival package. Patients were first grouped according to quartile of the variable (i.e. fraction subclonal neoantigens, number of clonal neoantigens) being considered.

2.2.13.1 Checkpoint blockade treated cohort survival data

Clinical data for the checkpoint blockade treated cohorts was downloaded from the original publications.

2.2.14 RNAseq expression analysis of immune infiltration

Previously defined measures of immune infiltration and activity were used to classify the immune microenvironment of all tumors (and tumor regions) with RNAseq data available (Rooney et al., 2015, Li et al., 2016, Davoli et al., 2017, Racle et al., 2017, Danaher et al., 2017, Newman et al., 2015).

2.2.14.1 Association with HLA LOH

Immune measures were compared between tumors exhibiting HLA LOH at all HLA loci and those without any evidence for HLA LOH. Additionally the expression level of PD-L1, CTLA4, and an IFN score were compared (Tumeh et al., 2014, Herbst et al., 2014, Ribas et al., 2015, Piha-Paul et al., 2016). Significance was determined using a Wilcoxon test and FDR correction. To determine the degree of change between the HLA LOH groups, a ratio of the medians was calculated.

2.2.15 RNAseq differential expression analysis

For differential expression analysis using TCGA data, the raw RNAseq read counts were used as input into the R package DESeq2 for analysis. An FDR cutoff of 0.05 was used to determine genes significantly differentially expressed.

2.2.16 Association of gene expression and copy number

For every gene used in an immune subset definition and the randomly selected genes, the copy number status (deletion, shallow loss, neutral, shallow gain, amplification) and RNA expression value was determined in each tumor region. The correlation between these two variables was determined by the Kendall's test, using copy number status as an ordinal variable. All p-values were FDR adjusted.

2.2.17 Calculation of Shannon entropy

For each tumor region, the Shannon entropy was estimated using the command "entropy.empirical" from the "entropy" R package. This was calculated based on the number and prevalence of different tumor subclones found in that region, such that a tumor region containing only one subclone was assigned a value of 0.

2.2.18 Distance measures

2.2.18.1 Immune distance

The immune distance was determined by taking the Euclidean distance of immune infiltrate estimates (of the Danaher method, as described in Chapter 6) between tumor regions.

2.2.18.2 Genomic distance

The genomic distance was calculated by taking the Euclidean distance of the mutations present between tumor regions. All mutations present in any region from a tumor were turned into a binary matrix, where the rows were mutations and columns tumor regions. This matrix was clustered and the pair-wise distance between any two tumor regions was determined.

2.3 Experimental methods

All experimental and validation results presented in the thesis were performed or provided by: Andrew Furness, Crispin Hiley, Andrew Rowan, and Roberto Salgado.

2.3.1 Isolation of TILs for L011 and L012

Tumors were taken directly from the operating theatre to the department of pathology where the sample was divided into regions. Samples were subsequently minced under sterile conditions followed by enzymatic digestion (RPMI-1640 (Sigma) with Liberase TL research grade (Roche) and DNAse I (Roche)) at 37°C for 30 minutes before mechanical dissociation using gentleMACS (Miltenyi Biotech). Resulting single cell suspensions were filtered and enriched for leukocytes by passage through a Ficoll-paque (GE Healthcare) gradient. Live cells were counted and frozen in human AB serum (Sigma) with 10% dimethyl sulfoxide at -80°C before transfer to liquid nitrogen.

2.3.2 In-vitro expansion of TILs for L011 and L012

TILs were expanded using a rapid expansion protocol (REP) in T25 flasks containing EX-VIVO media (Lonza) supplemented with 10% human AB serum (Sigma), soluble anti-CD3 (OKT3, BioXCell), 6000IU/mL recombinant human (rhIL-2, PeproTech) and $2x10^7$ irradiated PBMCs (30Gy) pooled from 3 allogeneic healthy donors. Fresh media containing rhIL-2 at 6000IU/mL was added every three

days as required. Following 2 weeks of expansion, TILs were counted, phenotyped by flow cytometry and frozen in human AB serum (Sigma) at -80°C before use in relevant assays or long-term storage in liquid nitrogen.

2.3.3 MHC multimer generation and flow cytometry analysis

MHC-multimers holding the predicted neoepitopes were produced in-house (Technical University of Denmark, laboratory of SRH). Synthetic peptides were purchased at Pepscan Presto, NL. HLA molecules matching the HLA-expression of L011 (HLA-A1101, A2402, and B3501) and L012 (HLA-A1101, A2402, and B0702) were refolded with a UV-sensitive peptide, and exchanged to peptides of interest following UV exposure (Toebes et al., 2006, Bakker et al., 2008, Frosig et al., 2015). Briefly, HLA complexes loaded with UV-sensitive peptide were subjected to 366-nm UV light (CAMAG) for one hour at 4°C in the presence of candidate neoantigen peptide in a 384-well plate. Peptide-MHC multimers were generated using a total of 9 different fluorescent streptavidin (SA) conjugates: PE, APC, PE-Cy7, PE-CF594, Brilliant Violet (BV)421, BV510, BV605, BV650, Brilliant Ultraviolet (BUV)395 (BioLegend). MHC-multimers were generated with two different streptavidin-conjugates for each peptide-specificity to allow a combinatorial encoding of each antigen responsive T-cell, enabling analyzes for reactivity against up to 36 different peptides in parallel (Andersen et al., 2012).

2.3.4 Identification of neoantigen-reactive CD8+ T-cells

MHC-multimer analysis was performed on in-vitro expanded CD8+ T lymphocytes isolated from region-specific lung cancer samples and adjacent normal lung tissue. 288 and 354 candidate mutant peptides (with predicted HLA binding affinity <500nM, including multiple potential peptide variations from the same missense mutation) were synthesized and used to screen expanded L011 and L012 TILs respectively. Simultaneously, TIL responses to HLA-matched viral peptides were assessed, demonstrating functionality of the employed MHC-multimer technology. Viral peptides for L011 included A11 EBV-EBNA4 (AVFDRKSDAK), A11 HCMV pp65 (GPISGHVLK), A24 EBV LMP-2 419-427 (TYGPVMCL), A3 EBV EBNA 3A RLR (RLRAEAQVK), A24 HCMV 248-256 (AYAQKIFKIL), B35 Flu Matrix (ASCMGLIY), B35 ENV EBNA 3B (AVLLHEESM), EBV EBNA-3 114-121 (RYSIFFDY) and EBV BZLF1 (APENAYQAY). For L012, these consisted of A11 EBV EBNA4 (AVFDRKSDAK), A11 HCMV pp65 (GPISGHVLK), A24 EBV EBNA-3 114-121 (RYSIFFDY), A24 EBV LMP-2 419-427 (TYGPVFMCL), A24 EBV RTA 28-

37 (DYCNVLNKEF), A24 HCMV 248-256 (AYAQKIFKIL), B7 CMV pp65 RPH-L (RPHERNGFTV), B7 CMV pp65 TPR (TPRVTGGGAM) and B7 EBV EBNA RPP (RPPIFIRLL). Finally, reactivity of healthy donor CD8+ PBMC's against the same peptides was assessed, demonstrating a lack of background/non-specific staining. Response of HD PBMCs was not performed for HLA B35 restricted peptides.

For staining of expanded CD8+ T lymphocytes, samples were thawed, treated with DNAse for 10 minutes, washed and stained with MHC multimer panels for 15 minutes at 37°C. Subsequently, cells were stained with LIVE/DEAD® Fixable Near-IR Dead Cell Stain Kit for 633 or 635 nm excitation (Invitrogen, Life Technologies), CD8-PerCP (Invitrogen, Life Technologies) and FITC coupled antibodies to a panel of CD4, CD14, CD16, CD19 (all from BD Pharmingen) and CD40 (AbD Serotec) for an additional 20 minutes at 4°C. Data acquisition was performed on an LSR II flow cytometer (Becton Dickinson) with FACSDiva 6 software. Cutoff values for the definition of positive responses were ≥0.005% of total CD8+ cells and ≥10 events.

For patient L011, HLA-B3501 MTFR2^{D326Y} -derived multimers were found to bind the mutated sequence FAFQEYDSF (netMHC binding score: 22) but not the wild type sequence FAFQEDDSF (netMHC binding score: 10). No responses were found against overlapping peptides AFQEYDSFEK and KFAFQEYDSF. For patient L012, HLA-A1101 CHTF18^{L769V} -derived multimers bound the mutated sequence LLLDIVAPK (netMHC binding score: 37) but not the wild type sequence: LLLDILAPK (netMHC binding score: 41). No responses were found against overlapping peptides CLLLDIVAPK and IVAPKLRPV. Finally, HLA-B0702 MYADM^{R30W}-derived multimers bound the mutated sequence SPMIVGSPW (netMHC binding score: 15) as well as the wild type sequence SPMIVGSPR (netMHC binding score: 1329). No responses were found against overlapping peptides SPMIVGSPWA, SPMIVGSPWAL, SPWALTQPLGL and SPWALTQPL.

2.3.5 MHC-multimer analysis and phenotyping of non-expanded samples

Tumor samples were thawed, washed and first stained with custom-made MHC-multimers for 10-15 minutes at 37°C in the dark. Cells were thereafter transferred onto wet ice and stained for 30 minutes, in the dark, with a panel of surface antibodies used at the manufacturer's recommended dilution: CD8-V500, SK1 clone (BD Biosciences), PD-1-BV605, EH12.2H7 clone (Biolegend), CD3-BV785, OKT3 clone (Biolegend), LAG-3-PE, 3DS223H clone (eBioscience). A fixable viability dye (eFlour780, eBioscience) was included the surface mastermix. Cells were

permeablized for 20 minutes with use of an intracellular fixation and permeabilization buffer set from eBioscience. An intracellular staining panel was applied for 30 minutes, on ice, in the dark, and consisted of the following antibodies used at the manufacturer's recommended dilution: granzyme B-V450, GB11 clone (BD Biosciences), FoxP3-PerCP-Cy5.5, PCH101 clone (eBioscience), Ki67-FITC, clone B56 (BD Biosciences) and CTLA-4 – APC, L3D10 clone (Biolegend). Data acquisition was performed on a BD FACSAria III flow cytometer (BD Biosciences) and analyzed in Flowjo version 10.0.8 (Tree Star Inc.).

2.3.6 Fragment analysis validation of LOHHLA results

Allelic imbalance was validated using four polymorphic Sequence-Tagged Site (STR) markers located on the short arm of chromosome 6, close to the HLA locus - (D6S2852, D6S2872, D6S248 and D6S1022). 20ng of patient germline and tumor region DNA was amplified using the PCR. The PCR was comprised of 35 cycles of denaturing at 95C for 45 seconds, followed by an annealing temperature of 55C for 45 seconds and then a PCR extension at 720C for 45 seconds. PCR products were separated on the ABI 3730xl DNA analyzer. Fragment length and area under the curve of each allele was determined using the Applied Biosystems software GeneMapper v5. When two separate alleles were identified for a particular marker, the fragments could be analyzed for allelic imbalance using the formula $(A_{tumor}/B_{tumor})/(A_{normal}/B_{normal})$. The output of this formula was defined as the normalized allelic ratio.

2.3.7 PD-L1 immunohistochemistry

Formalin-fixed, paraffin-embedded (FFPE) tissue sections of 4-um thickness were stained for PD-L1 with an anti-human PD-L1 rabbit monoclonal antibody (clone SP142; Ventana, Tucson, AZ) on an automated staining platform (Benchmark; Ventana) with the OptiView DAB IHC Detection Kit and the OptiView Amplification Kit (Ventana Medical Systems Inc.) in a GCP-compliant central laboratory (Targos Molecular Pathology GmbH). PD-L1 expression was evaluated on tumor cells and tumor-infiltrating immune cells. For tumor cells the proportion of PD-L1-positive cells was estimated as the percentage of total tumor cells. For tumor-infiltrating immune cells, the percentage of PD-L1-positive tumor-infiltrating immune cells occupying the tumor was recorded. Scoring was performed by a trained histopathologist (according to previously published scoring criteria (Herbst et al., 2014)).

2.3.8 Pathology TIL estimation

From the pathology slide of a given tumor region, the relative proportion stromal area to tumor area was determined. The percent of TILs identified in the stroma was determined and multiplied that value by the proportion of stromal area. The percent of TILs identified in the tumor was determined and multiplied that value by the proportion of tumor area. Finally to obtain the TILs present on the total slide these two values were summed.

Chapter 3 Identifying mutational processes active during cancer evolution

3.1 Introduction

The generation of novel mutations is key to introducing the genetic diversity upon which selective pressures may act. Mutational processes active in the cell contribute to this ongoing generation of somatic mutational diversity. Some mutational events have known sources, such as exogenous mutagens from tobacco smoke and ultraviolet light, while others are associated with endogenous processes like mismatch repair deficiency (Pfeifer et al., 2002, Pfeifer, 2010, Boland and Goel, 2010, Bhattacharyya et al., 1994, Alexandrov et al., 2013a, McGranahan et al., 2015, Nik-Zainal et al., 2012a). Regardless of the source of mutation generation, these aberrant processes often lead to characteristic patterns of mutation.

By analyzing the composition of mutations and the context they arise in, processes shaping the cancer genome as it evolves can be delineated, allowing for insights into the route the tumor has taken to carcinogenesis and the contributors to subclonal diversity. This could have great impact on understanding the pathophysiology of a cancer type and could help inform patient-specific therapeutic approaches. For instance, the identification of specific defective DNA repair mechanisms may suggest a patient has a favorable chance of positive response to immunotherapy (Le et al., 2015) or PARP inhibition (Farmer et al., 2005, Rottenberg et al., 2008, Alexandrov et al., 2015). Determining the cause of novel signatures identified can shed light on previously overlooked carcinogens, such as aristolochic acid (Poon et al., 2013, Hoang et al., 2013), and can confirm environmental factors implicated by epidemiological studies, such as the contribution of alcohol intake to esophageal squamous cell carcinoma and hepatocellular carcinoma (Chang et al., 2017, Schulze et al., 2015).

Of particular importance is understanding the temporal change in mutational processes, across each stage of tumor evolution, in order to characterize the signatures of clonal mutations, which is further explored in Chapter 4.

3.1.1 Mutational context and signature extraction

Large-scale delineation of mutational signatures was first performed by Alexandrov, Nik-Zainal, and colleagues using a published algorithm (Alexandrov et al., 2013a, Nik-Zainal et al., 2012a). Their Wellcome Trust Sanger Institute (WTSI) mutational signatures framework utilizes non-negative matrix factorization (NMF) followed by model selection to extract the signatures of mutational processes active in a set of cancer genomes (Alexandrov et al., 2013a, Alexandrov et al., 2013b, Nik-Zainal et al., 2012a).

3.1.1.1 Trinucleotide contexts

Specific mutational signatures are defined by considering the substitution type of the mutated base, as well as the context immediately upstream and downstream to the altered base. If strand is ignored the mutated base can be collapsed to one of two possibilities, either C and T (or G and A if the paired base is used as default). This results in six substitution classes (C>A, C>G, C>T, T>A, T>C, and T>G) representing all possible single nucleotide substitutions. Beyond the base pair substitution itself, information about the context in which the mutation occurred can also be included. When the substitution classes are coupled with the information from the bases immediately 5' and 3' to each mutated base, there are 96 possible mutational contexts (4 possible 5' bases * 6 possible substitution classes * 4 possible 3' bases). A mutational signature can then be characterized by the distribution of single nucleotide substitutions at each of 96 trinucleotide contexts. After a signature has been extracted, additional mutation features such as a high frequency of insertions and deletions, dinucleotide mutations, or transcriptional strand bias could also be included in the final definition. Furthermore, the most recent iteration of signature analysis in breast cancer has expanded signature definitions to include structural variation (Nik-Zainal et al., 2016).

3.1.1.2 Signature extraction

The original study performed by Alexandrov and colleagues analyzed over 7,000 cancer genomes and exomes with a total of five million mutations. From this data set, they defined a catalog of twenty-one signatures that contributed to over 30 tumor types (Alexandrov et al., 2013a). About half of these signatures were associated with known mutational processes previously defined in the literature, including tobacco smoke, exposure to ultraviolet light, the APOBEC family of

cytidine deaminases activity, DNA mismatch repair deficiencies, or mutations in *POLE*. Tumor types were first analyzed independently and all mutational signatures extracted were clustered to identify signatures present across different cancer types. Indeed, many signatures were found to be active in a wide variety of tumor types. Since the original report, the current set of mutational signatures has been refined and expanded to include 30 distinct mutational signatures based on the analysis of nearly 11,000 exomes and 1,100 genomes across over 40 cancer types in order to provide an extensive catalog of mutational processes active during cancer development and evolution.

3.1.1.3 Limitations of NMF

The originally published WTSI mutational signatures framework offers an elegant approach to firstly identify the distinct mutational processes active in a set of tumor samples. Once the active signatures have been established, they are then applied to the individual samples in order to quantify the contribution of every mutational process to each sample in the set. However, the extraction step of the NMF-based algorithm was designed to make full use of the abundance of sequencing data currently available in order to identify and define novel mutational signatures. Consequently, to use the tool, it is necessary to begin with a sufficient number of tumor samples for adequate power to accurately deconvolute signatures. In simulations, Alexandrov et al. found that the number of cancer genomes required increased exponentially with the number of active signatures. Indeed, they calculated that in order to accurately identify the signatures of 20 mutational processes, they required at least 200 whole genome samples (Alexandrov et al., 2013b). As exome sequencing only covers ~1% of the human genome, there are fewer mutations available per sample, thus less resolution per sample is gained. After considering these factors, Alexandrov et al. estimated that it would take thousands of exome samples to extract the majority of mutational processes functional over the evolutionary course of a tumor.

Due to the frequency at which many of the identified mutational signatures occur across multiple cancer types as well as the potential for therapeutic intervention based on signature activity, it would be useful to analyze signature prevalence and contributions across additional tumor samples. However, under the current mutational framework, this was not feasible without first acquiring a large enough sample set. In order to address this challenge, I developed a method that uses the

established mutational signatures to circumvent the extraction step and determine the contributions of each mutational process in a single tumor sample.

The work presented in this chapter was published as a first-author paper, (Rosenthal et al., 2016). Additionally, some of the work appears in a review.

3.2 deconstructSigs method

3.2.1 Overview of the tool

The tool for determining the contributions of a set of mutational processes in a single tumor sample was designed as an R package, deconstructSigs, accessible both through the CRAN package webpage (https://cran.r-project.org/) and GitHub site (https://github.com/raerose01/deconstructSigs). Hosting the tool on GitHub in addition to CRAN allows for convenient user feedback and issue reporting, as well as faster access to updated versions.

Briefly, deconstructSigs works by determining the linear combination of a user-provided set of signatures that is capable of most accurately reconstructing the observed mutation profile of a single tumor sample. As the tool begins with a set of pre-defined signatures, it is not necessary to first extract signatures present in the tumor sample(s) being analyzed, thus avoiding the issue of inaccurate deconvolution from a sample set limited by small numbers. The package consists of data processing functions, including those to reshape and normalize the data if required, and the key function to determine relative signature contributions to the sample. This function assigns signature weights by using a multiple linear regression model with the constraint that all coefficients must be greater than 0, as negative contributions of a mutational process do not make biological sense.

3.2.2 Using deconstructSigs

3.2.2.1 Input data

At a minimum in order to run, deconstructSigs requires a data frame containing mutational data for a set of tumors to be analyzed and a reference signatures matrix. As two default options for the reference signatures exist within the package data, the only requisite user input is the list of mutations from the samples to be analyzed.

The mutational data frame must contain:

- Sample identifier
- Chromosome of the mutation
- Base position of the mutation
- Reference base pair
- Alternate base pair

The input mutational data is then converted to an *n*-row and 96-columns data frame where *n* is the number of unique samples present. This step is performed using the command "`mut.to.sigs.input`", as shown below.

```
sigs.input <- mut.to.sigs.input(mut.ref = sample.mut.ref, sample.id
= "Sample", chr = "chr", pos = "pos", ref = "ref", alt = "alt")
```

The output "`sigs.input`" represents the number of mutations found at that particular trinucleotide context (column) in that particular sample (row).

3.2.2.2 Key variables

T : *n*-row x 96-column matrix, where *n* is the number of samples, corresponding to either the number of mutations found in a particular trinucleotide context for a particular tumor sample or the fraction of mutations found in each of the possible 96 trinucleotide contexts for each tumor sample.

S : *k*-row x 96-column matrix, where *k* is the number of supplied signatures, containing the fraction of times a mutation is seen in each of the 96-trinucleotide contexts for a signature *k*.

W : vector of length *k*, where each entry is the weight of one of the supplied signatures.

R : *n*-row x 96-column matrix, representing the reconstructed tumor sample matrix obtained by the matrix multiplication *WS*.

e : error threshold, beyond which deconstructSigs determines it has converged on the best solution.

3.2.2.3 Data normalization

The example output "`sigs.input`" described above can be used directly as input into deconstructSigs. deconstructSigs begins with an input data frame *T*, which contains the per sample counts of each mutation at each trinucleotide context. By default,

the only normalization performed is standardizing *T* such that it contains the fraction of mutations found in each of the possible 96 trinucleotide contexts for each tumor sample (i.e. dividing each row in "`sigs.input`" by the sum of the row).

When *T* contains only the counts of each mutation in each trinucleotide context, as is the case for the output from "`mut.to.sigs.input`", the user may optionally choose to apply an additional normalization step, as some of the published mutational signatures were reported based on the observed trinucleotide frequency of the human exome/genome. To do this, the parameters "`contexts.needed`" and "`tri.counts.method`" are set, which act to normalize *T* by the number of times each trinucleotide context is observed in the portion of the genome sequenced. Trinucleotide counts for exome and genome data are provided in the package for this normalization, but depending on the sequencing design, a user may also create their own counts file. Alternatively, a user may bypass any data cleaning or normalization steps and generate their own *T* data frame to use as input into deconstructSigs, with the stipulation that the input data frame must already be standardized (i.e. the rows must sum to 1).

3.2.2.4 Signature deconvolution

The signatures matrix *S* of *k* rows and 96 columns is either obtained from published data or provided by the user, where *k* is the number of supplied signatures. Each entry in *S* represents the fraction of times a mutation is seen in each of the 96-trinucleotide contexts (columns) for a signature *k* (rows). In theory, if a user has generated their own mutational signatures, they could provide both *T* and *S* with a different number of mutational categories (for instance, 192 if DNA strand was considered, or 6 if trinucleotide context was ignored). However, in the package, only signatures based on a 96 trinucleotide context are provided. Given the two inputs, *T* and *S*, deconstructSigs computes weights W_i (for each signature *i* from 1 to *k*) such that each signature is assigned a weight. Signature weights are determined such that a reconstructed tumor sample matrix *R*, which is computed as *T-(WS)*, minimizes the sum-squared error (SSE).

This step is called with the function "`whichSignatures`" as shown below.

```
output.sigs <- whichSignatures(tumor.ref =
randomly.generated.tumors, signatures.ref = signatures.nature2013,
sample.id = "1")
```

To determine the weights for each mutational signature, *W*, that will best recreate *T*, an iterative approach is taken. First, any signatures containing a single trinucleotide context making up more than 20% of the signature definition which is not present at all in *T* is excluded. This step is performed to account for the fact that some signatures have mutational profiles where only a few specific trinucleotide contexts dominate (for instance, the signature associated with some mutations in *POLE*). Thus, in samples without any mutations found in those contexts, it is unlikely that that signature is active.

From the remaining signatures, an initial mutational signature is chosen that most closely reflects the mutational profile of the given tumor sample. This signature is chosen by minimizing the SSE between the mutational profile of the tumor sample, *T* and the mutational signature S_i. The weights, *W*, are initialized such that the initial signature chosen, S_i is the only signature contributing to the reconstructed tumor mutational profile, thus being assigned a normalized weight of 1.

Next a forward selection process is employed. For each signature, an optimal weight that minimizes the SSE between the given tumor sample and the reconstructed tumor profile is determined. From this set of all possible weights, only the weight corresponding to the signature that results in the overall lowest SSE is next included in *W*. This iterative process repeats until the difference between the SSE before and after the alteration of the weights matrix is less than error threshold, *e* (set to 0.001 by default). The output weights, *W*, are normalized between 0 and 1. A schematic of deconstructSigs is outlined in Figure 3-1.

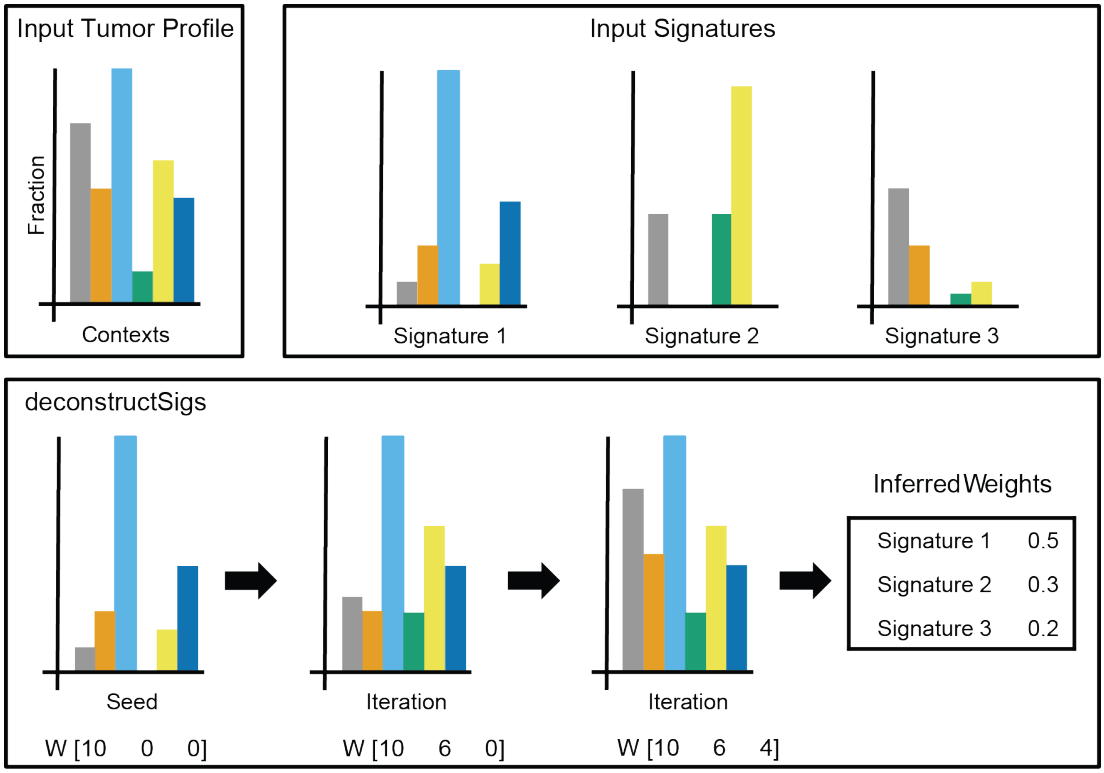

Figure 3-1: Schematic of deconstructSigs workflow.
deconstructSigs requires an input tumor profile and reference input signatures. The tool iteratively infers the contributions of each reference signature by updating the weight of each reference signature until the SSE converges below an empirically chosen error threshold.

3.2.2.5 Filtering signature output

When reporting the output signature weights, any signature with W_i < 6% is excluded (the 6% threshold value is further justified below). A weight threshold was implemented to combat over-fitting of signatures to a tumor sample in an aim to reduce false positives. To determine a suitable threshold, tumors were randomly generated *in silico* using signatures from the published set (Alexandrov et al., 2013a). Five-hundred tumors were simulated with each containing a random combination of up to 10 different mutational processes. An additional perturbation factor was then added, as the original signatures were combined from multiple runs of the WTSI mutational signatures framework algorithm and real-world tumor data will never reflect a perfect combination of mutational signatures. To more accurately reflect the noise in actual data, the simulated tumors were perturbed by randomly altering the assigned value at each trinucleotide context by up to +/- 5%.

These randomly generated tumor samples were analyzed with deconstructSigs and the calculated weights were compared to the known weights used to generate the set of simulated tumors. A false positive was called if deconstructSigs included a

signature that was not used in generating the random tumor sample; a false negative was called if deconstructSigs missed a signature that was used in generating the random tumor sample. Analysis of the false positives revealed that they almost uniformly had weights less than 6% (Figure 3-2A). Additionally, when the 6% cutoff was applied to the randomly generated tumor samples, it only resulted in 38 instances where a signature was incorrectly excluded, leading to a false negative rate of 1.4% (Figure 3-2B).

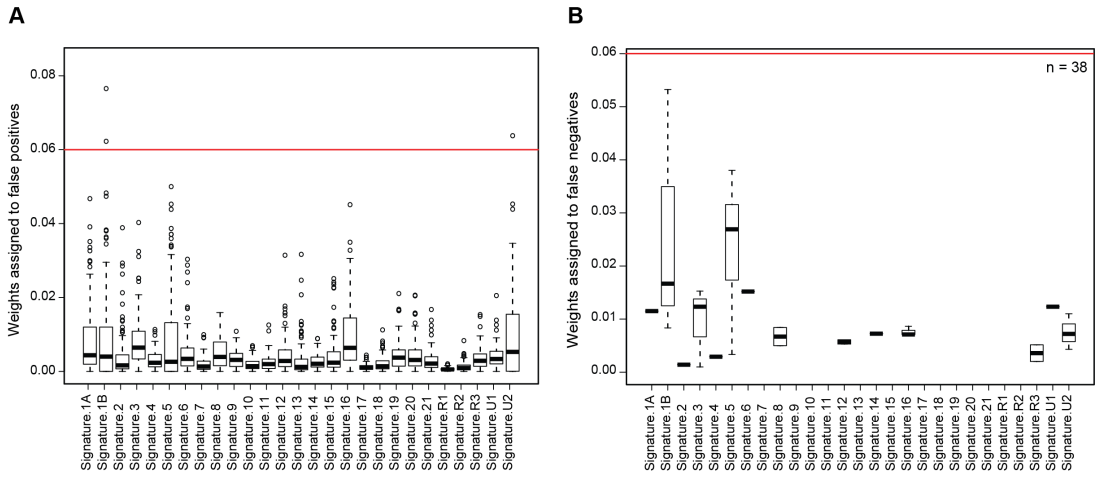

Figure 3-2: False positive and false negative weights in randomly generated tumor cohort.
A randomly generated tumor cohort was constructed consisting of 500 tumors comprised of 2646 signatures of known weights. The frequency of trinucleotide contexts in each tumor was subjected to up to ±5% random perturbations to more accurately reflect "non-ideal" theoretical tumor samples. The false positives, where a signature was erroneously identified as contributing to the samples (A), and false negatives, where a signature was erroneously assigned a weight of 0 (B), were analyzed after inferring signature weights with deconstructSigs. Only three weights assigned to false negatives fell above 0.06, and false negatives using this threshold only occurred at a rate of 1.4% (38/2646).

3.3 Validation of deconstructSigs

3.3.1 Comparison of deconstructSigs to previous analyses

To validate the output generated by deconstructSigs, a comparison was made to the standard mutational signatures algorithm, WTSI mutational signature framework. Publicly available TCGA data was re-analyzed with both signature approaches from bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and skin cutaneous melanoma (SKCM) cancers (https://tcga-data.nci.nih.gov/tcga).

The WTSI mutational signature framework had already been implemented to extract signatures active in each TCGA cancer type, resulting in the extraction of twenty-six mutational signatures as was previously described (McGranahan et al., 2015). As discussed earlier, one limitation of extraction methods is the requisite adequate sample size. Thus some of the TCGA cancer types, which had fewer available samples, lacked the resolution required to extract the full list of signatures originally associated with the cancer type. However, when a signature was successfully extracted from a sample set, it was consistent in profile to those published by Alexandrov et al. (Alexandrov et al., 2013a). Indeed 20/26 of the newly extracted signatures from TCGA could be identified as a previously identified mutational signature (McGranahan et al., 2015). In 2/6 instances where the re-extracted signatures did not match one of the original signatures, they appeared to consist of two or three signatures, again reiterating the importance of adequate sample size for accurate deconvolution of mutational signatures.

To directly compare deconstructSigs to WTSI mutational signature framework, the re-extracted signatures were used as the signatures matrix in deconstructSigs, and the same cohort of TCGA samples was analyzed. Supporting the utility of deconstructSigs to accurately infer mutational signature contribution to a single sample, each signature identified in a given sample showed a highly significant correlation between the contributions of that signature assigned by the two methods (Figure 3-3). These results indicated that by using deconstructSigs it was possible to consistently quantify which signatures were active in an individual tumor sample.

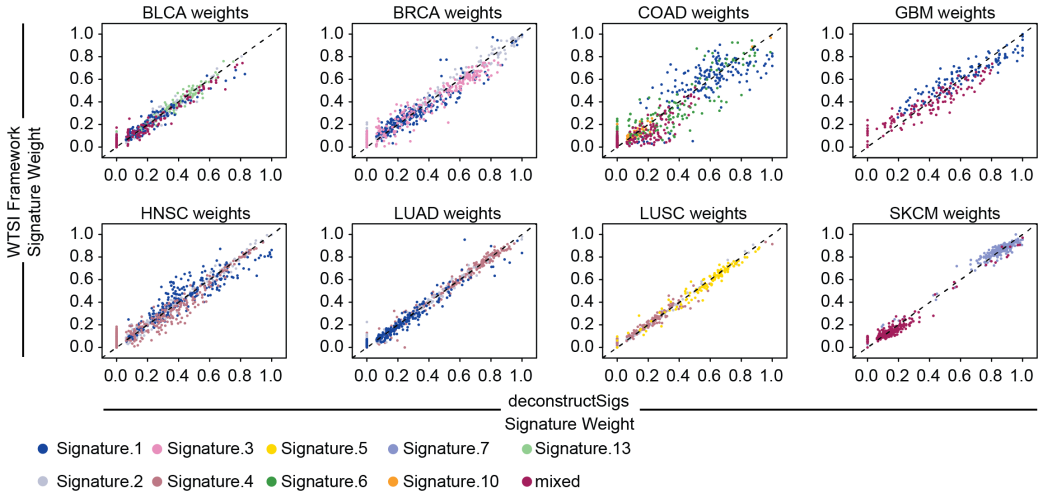

Figure 3-3: Comparison of signature contributions identified between methods.
For each set of tumors in a given TCGA cancer type, the relationship between the weights calculated using the WTSI Mutational Signature Framework and those inferred with deconstructSigs is shown. Each point plotted represents the weights assigned, by both methods, to one signature detected in a patient.

To directly compare the reconstruction error between the two approaches, the SSE was calculated between the reconstructed mutational profile and the observed one by taking the sum-squared difference between the input and reconstructed mutational profiles at each trinucleotide context. The SSE was consistent between the two approaches, and showed a strong correlation, with slope near 1 (Figure 3-4A). The concordance of errors between methods indicates that samples with a poorly reconstructed tumor profile may not be suitable to this type of signature analysis, rather than one method being inherently more capable.

Further supporting the notion that a high SSE may represent a reconstruction issue intrinsic to the tumor sample itself, SSEs calculated from both methods generally decreased as mutation count increased. This trend is indicative of a better fit at higher mutation count (Figure 3-4B). Indeed, the samples with the largest SSE all had very few mutations, highlighting the importance of having a large enough number of mutations to identify and assign signatures when using ones defined by 96-substitution classifications. Attempting to distribute data into more bins than data points will always result in large errors, regardless of the approach that is being used. This is especially critical when the mutational profiles considered are flat, such that each of the 96-trinucleotide contexts is similarly likely to be affected. In these cases the mutational process may be equally likely to affect a greater number of trinucleotide contexts, so the profile would only be observable if there were enough mutations to fully recapitulate the extracted signature. Consequently, deconstructSigs includes a warning for any samples containing fewer than 50 mutations.

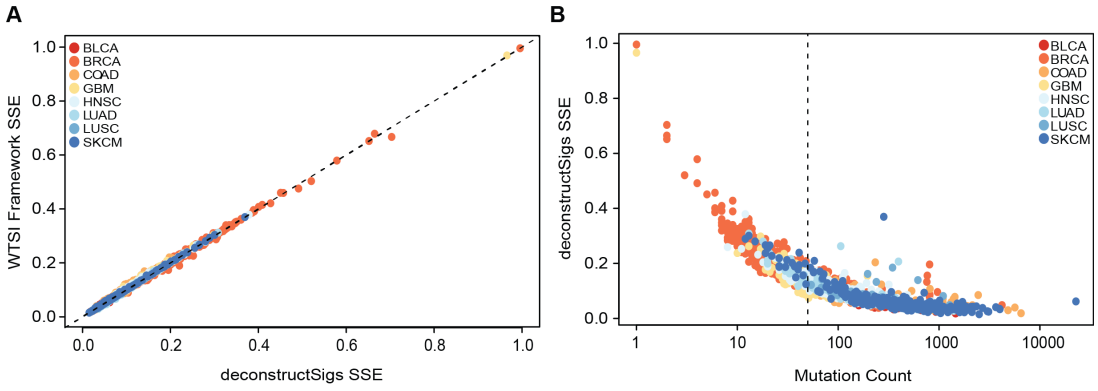

Figure 3-4: Comparison of SSEs using deconstructSigs and WTSI Mutational Signatures Framework. For each tumor, the SSE was calculated to quantify the difference between the input tumor mutational profile and the reconstructed mutational profile. The calculated SSEs from using the WTSI Mutational Signatures Framework were compared with those from using deconstructSigs (A) and the relationship between the SSE and overall mutation count was compared (B). As the mutation count in the tumor increases, the calculated SSE decreases.

3.3.2 Using deconstructSigs with previously defined signatures

As the deconstructSigs package would most likely be used to infer the contributions of mutational processes in a single sample without the need for signature deconvolution, the tool was next tested using an expanded signature search space. In this case, the TCGA tumors were re-analyzed using deconstructSigs, but allowing the tool to search from the set of the originally published signatures (Alexandrov et al., 2013a). To ensure an accurate comparison, cancer types were only included in this analysis if the number of samples was sufficiently large. Thus cohorts that yielded inconclusive composites of multiple published signatures in the McGranahan et al. (McGranahan et al., 2015) study were excluded. In these instances, there would be no signature in the reference set that matched the "mixed" signature obtained, so it would be impossible to accurately judge if deconstructSigs had correctly inferred the signature composition in the samples from those cancer types. The cancer types excluded due to incomplete signature extraction were BLCA, COAD, GBM, and SKCM.

One limitation to this analysis is that re-extracting signatures through a separate iteration of the WTSI mutational signature framework on a different subset of TCGA samples will result in the generation of signatures with slight differences to the originally published ones. This is in part due to the published signatures representing a consensus signature across the multiple cancer types it was identified in. The specific version of the signature extracted from each tissue type likely varies slightly, but for the original publication, the signatures identified across multiple cancer types were combined. While some mutational signatures would likely be identical across all tissue types, that may not be true for all mutational signatures (Nik-Zainal and Morganella, 2017), so combining some signatures may result in ones slightly less robust than their cancer-specific versions. Another reason for observed differences between re-extracted signatures and the originally published ones is that when a subset of samples are used in the analysis, as is the case for the re-analyzed TCGA samples (McGranahan et al., 2015), each signature has a lower level of resolution.

Thus it would have been surprising if the weights assigned by deconstructSigs correlated nearly perfectly with the contributions found using WTSI mutational signature framework (as was found in Figure 3-3). Nevertheless, when the signature contributions assigned by the two methods were compared, a strong positive and statistically significant correlation for nearly all signatures was

observed (Figure 3-5). Two signatures (Signature 3 and Signature 5) which had a weak correlation between the two approaches had flat mutational profiles, exhibiting few distinguishing patterns of trinucleotide context, which render them more susceptible to subtle changes in the signature definition between the originally published signatures and the ones re-extracted from TCGA samples using the same method.

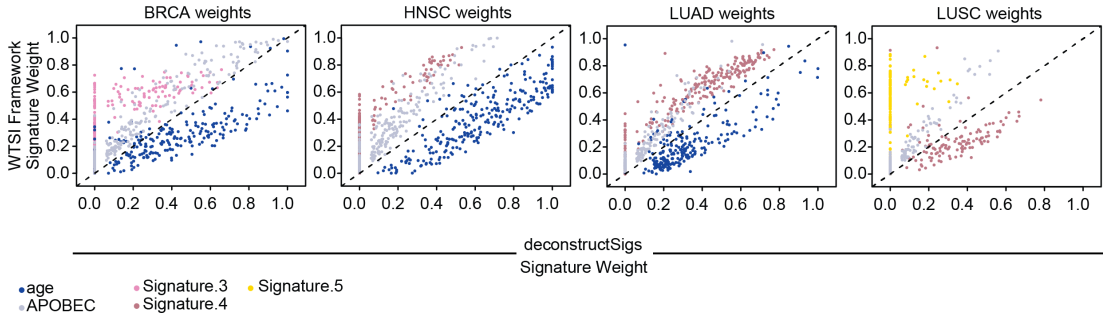

Figure 3-5: Comparison of signature contributions between deconstructSigs and WTSI Mutational Signature Framework using reference signatures as input.
For each TCGA cancer types that contained only unambiguous signatures extracted using WTSI Mutational Signatures Framework, a comparison between the weights assigned to the tumor by deconstructSigs and those calculated by WTSI Mutational Signatures Framework is plotted.

3.3.3 Outlier samples identified with deconstructSigs

Sometimes a signature is found in only a small subset of the total samples analyzed; in extreme cases, a signature may just contribute to a single patient's tumor (Nik-Zainal et al., 2016). With sufficient sample size, these signatures may be extracted successfully; however, often the signal can be diluted to beyond the point of recognition. By using deconstructSigs with an expanded signature set, it was also possible to identify outlier samples, whose tumor profiles had resulted in a distinct set of mutational processes as compared to the rest of the cancer type they were analyzed with.

Because the WTSI mutational signature framework algorithm first extracts signatures from a given set of tumors (broken down by cancer type in the case of the original publication) and then assigned weights using only the extracted set of signatures, contributions would be missed if the signature was not prevalent enough in the sample set to be extracted as a separate entity. For instance, if a signature was only active in a small number of tumors of the cancer type being analyzed, then it may not be extracted, and thus could not later be said to contribute to any tumor sample from that cancer type.

One example of such a false negative was observed in the lung adenocarcinoma sample TCGA-67-6215, which had clear signs of Signature 17 contribution, a signature with a currently unknown etiology (Figure 3-6). However, owing to the fact that Signature 17 was a rare event in the lung adenocarcinoma samples analyzed using the WTSI mutational signature framework, it was not one of the signatures extracted. Thus Signature 17 could not be identified as active in any lung adenocarcinoma sample, resulting in the signature being missed in TCGA-67-6215.

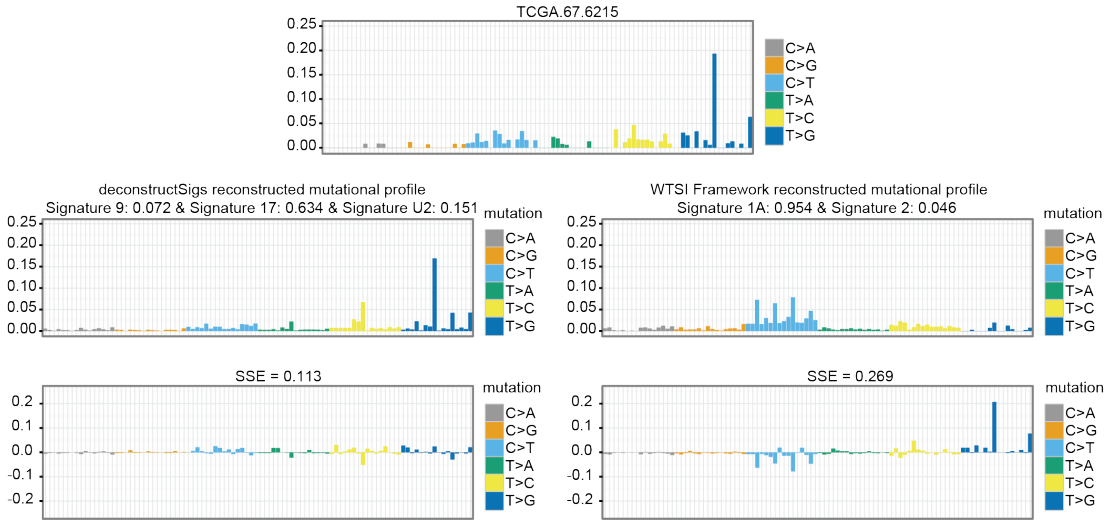

Figure 3-6: Mutational profile exhibiting Signature 17.
The mutational profile of patient TCGA-67-6215 shows activity of Signature 17. The top panel is the tumor mutational profile displaying the fraction of mutations found in each trinucleotide context. The middle panels show the reconstructed mutational profiles generated by multiplying the calculated weights by the input signatures with the contributing signatures annotated. The bottom panels show the error between the tumor mutational profile and the reconstructed mutational profile, with SSE annotated. As Signature 17 was not considered a possible signature extracted in the first step of the WTSI Mutational Signature Framework output, it was only called with deconstructSigs (signature weight = 0.634).

Similarly, a DNA mismatch (MMR) repair deficiency associated signature (Signature 6) was also identified by deconstructSigs in multiple breast cancer samples: TCGA-A8-A08F, TCGA-A8-A09Z, and TCGA-AN-A0AK (Figure 3-7). However, the re-implementation of the WTSI mutational signature framework did not extract Signature 6, nor was this signature originally associated with breast cancer in the first iteration of the pan-cancer signature analysis (Alexandrov et al., 2013a).

Supporting the utility of considering a wider set of signatures that may be active in a sample, two of the breast cancer tumors showing evidence of Signature 6 activity, TCGA-A8-A09Z and TCGA-AN-A0AK, had somatic alterations affecting MMR genes (https://tcga-data.nci.nih.gov/tcga/), which have been previously associated with this defect (Boland and Goel, 2010, Pena-Diaz et al., 2012). TCGA-A8-A09Z

harbored an *MLH1* nonstop mutation, accompanied by separate *MLH1* missense and splice site mutations. In TCGA-AN-A0AK there was an *MSH6* frameshift mutation. Beyond showing the characteristic mutational profile of Signature 6, these two breast cancer tumors also had a higher than average number of mutations and small indels, another characteristic of an microsatellite instability-high (MSI-H) phenotype. While the median number of mutations found in the breast cancer cohort overall was 38, and the median number of indels was 4, both TCGA-A8-A09Z and TCGA-AN-A0AK had substantially higher numbers of SNVs and indels, harboring 1438 mutations with 253 small indels and 1317 mutations with 352 small indels, respectively, providing further evidence for defective mismatch repair.

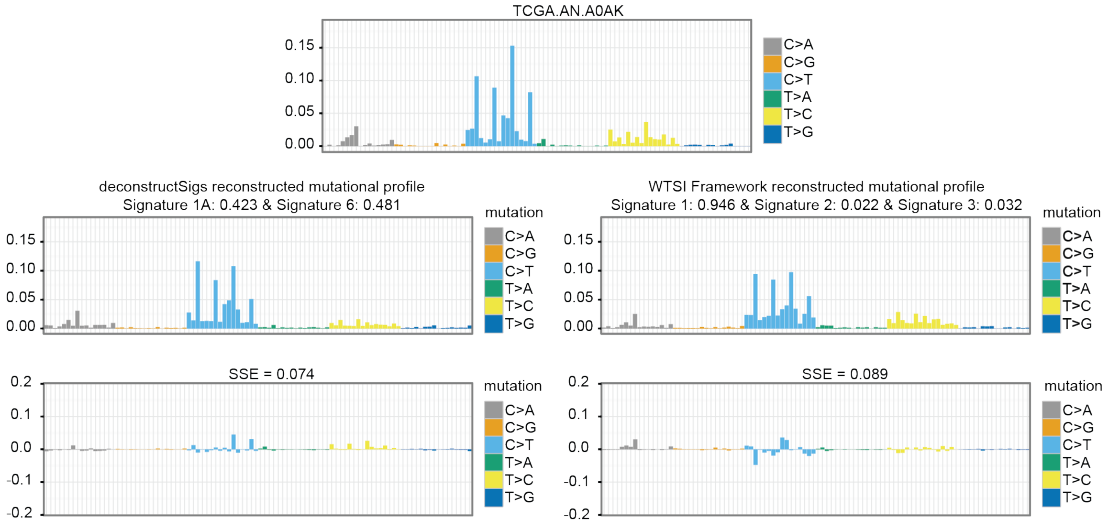

Figure 3-7: Mutational profile exhibiting Signature 6.
The mutational profile and reconstructed mutational profiles are plotted for a TCGA breast cancer patient (TCGA-AN-A0AK). One signature associated with DNA mismatch repair deficiency (Signature 6) identified by deconstructSigs (signature weight = 0.481) in patient, but was not identified by WTSI Mutational Signature Framework. An *MSH6* frameshift mutation was identified in this patient, and the tumor had far more mutations and indels than the cohort median.

Thus by considering tumors on an individual basis, it is possible to accurately detect contributions from mutational processes that are active in only a small number of the samples being considered, allowing for the identification of mutational signatures that may have been overlooked if the samples were only considered as part of a larger group. Indeed a recent analysis of breast cancer whole genomes, where the increased number of samples and greater genome capture provided improved signature extraction resolution, has revealed that the MMR associated signatures can be detected in a small fraction of patients (approximately 2%) (Davies et al., 2017b).

Having validated the deconstructSigs method on single samples, it was also possible to investigate the contribution of mutational processes over the life history of a tumor.

3.4 Using deconstructSigs to refine tumor evolution analyses

3.4.1 Quantifying mutational signatures from multi-region tumor samples

One benefit of analyzing mutational signatures at the single sample level is that it allows for the quantification of differences of signature activity over the evolutionary history of a tumor. Examining how mutational signatures vary over evolutionary time may facilitate the understanding of what processes play a role in tumor initiation (early in evolutionary time) versus tumor propagation and variation (later in evolutionary time). Signatures present late in evolutionary time are also responsible for increasing subclonal genetic diversity of the tumor.

To determine if it was possible to time the activity of specific mutational processes using deconstructSigs, a cohort of 19 tumors with multi-region sequencing available was analyzed (described in the Data and Methods). These samples had been obtained from five patients diagnosed with either lung adenocarcinoma (n=3) or lung squamous cell carcinoma (n=1), with one tumor exhibiting a mixed adenosquamous histological subtype. Multi-region whole-exome and/or genome sequencing had been previously performed in these patients (de Bruin et al., 2014), allowing for temporal dissection (e.g. clonal versus subclonal) of the mutations present. Mutations were annotated as occurring early in the tumor's evolutionary history, along the tumor's trunk, or late in evolutionary time, present only in the tumor's branches.

Given the small number of tumor samples, it had not been feasible to analyze the cohort using the WTSI mutational signature framework. However, the samples could now be considered on an individual basis and were analyzed using deconstructSigs to establish how mutational signatures contributing to the samples mutational catalogs varied with time.

The original analysis of this cohort (de Bruin et al., 2014) had used C>A mutations as a proxy for the smoking signature resulting from tobacco exposure (Pfeifer et al., 2002). However deconstructSigs enabled this analysis to be refined, as specific signature contributions could now be determined, potentially distinguishing C>A mutations in a tobacco exposure context from those arising due to other mutational

processes. The contribution of Signature 4, known to be associated with the number of smoking pack years (Pfeifer, 2010, Alexandrov et al., 2013a) was quantified, rather than relying on the more general C>A mutation class.

Supporting prior analyses and the knowledge that lung cancer may develop after years of cigarette smoking, resulting in a long trunk (Jamal-Hanjani et al., 2017), the smoking associated signature (Signature 4) contributed to a greater extent among the clonal mutations found in the phylogenetic trunk of the tumor and less among the subclonal mutations from the tumor's phylogenetic branches (Figure 3-8A). In these patients, the smoking signature dominated the trunk of the phylogenetic tree, accounting for over 30% of the signature weights in every tumor, with mutations occurring from age-associated signatures (Signature 1, Signature 5) contributing most of the rest. Furthermore, three of the five patients analyzed (L001, L004, and L008) showed no evidence of activity of the smoking associated signature among the branch mutations at all.

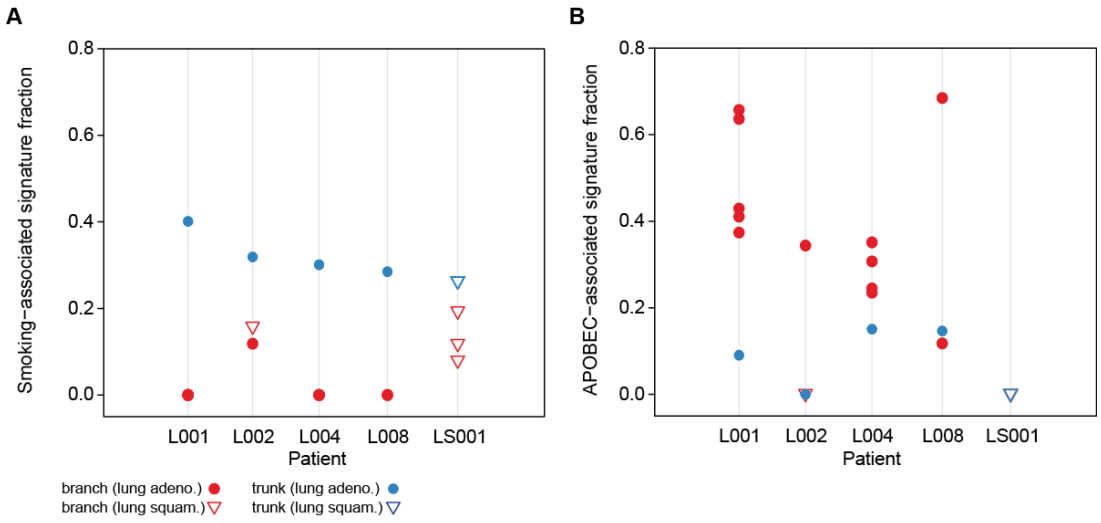

Figure 3-8: Temporal dissection of mutational processes.
The contributions of the smoking associated signature (A) and APOBEC associated signature (B) identified in the individual regions of a multi-region sequencing cohort are displayed. Lung adenocarcinoma regions are shown as circles; lung squamous cell carcinoma regions are shown as triangles. The mutations were temporally dissected into early (trunk, blue) and late (branch, red) mutations prior to signature analysis. The smoking associated signature was seen at higher fractions among early mutations, whereas the signature associated with APOBEC activity was seen to contribute more to late mutations in lung adenocarcinoma as compared to the early mutations.

Another frequent mutational process in NSCLC, expected to contribute to subclonal diversification and expansion, is the activity of the APOBEC family of cytosine deaminases. To determine how APOBEC activity varied over the tumor's life history, the contributions of the APOBEC associated signatures (Signature 2 and Signature 13) were compared in the tumors' trunk and branches. Consistent with

previously published observations, lung adenocarcinoma tumors exhibited far greater APOBEC activity in the branches as compared to the trunk (Figure 3-8B).

Thus, investigating single samples allowed for the tracking of changes in mutational dynamics over time. For instance, in patient L008, the APOBEC-associated signatures increase in prevalence among the late mutations (Figure 3-9). However, it is also clear that different regions from the same patient's tumor evolve differently over time. Two distinct regions show completely different mutational profiles. Region 3 from patient L008 (Figure 3-9C) has a decrease in smoking signature contribution as compared to the early mutations (Figure 3-9A), but still appears similar in profile to the early mutations. Whereas branch mutations present in region 1 from the same patient seem to have been almost entirely driven by APOBEC activity (Figure 3-9B).

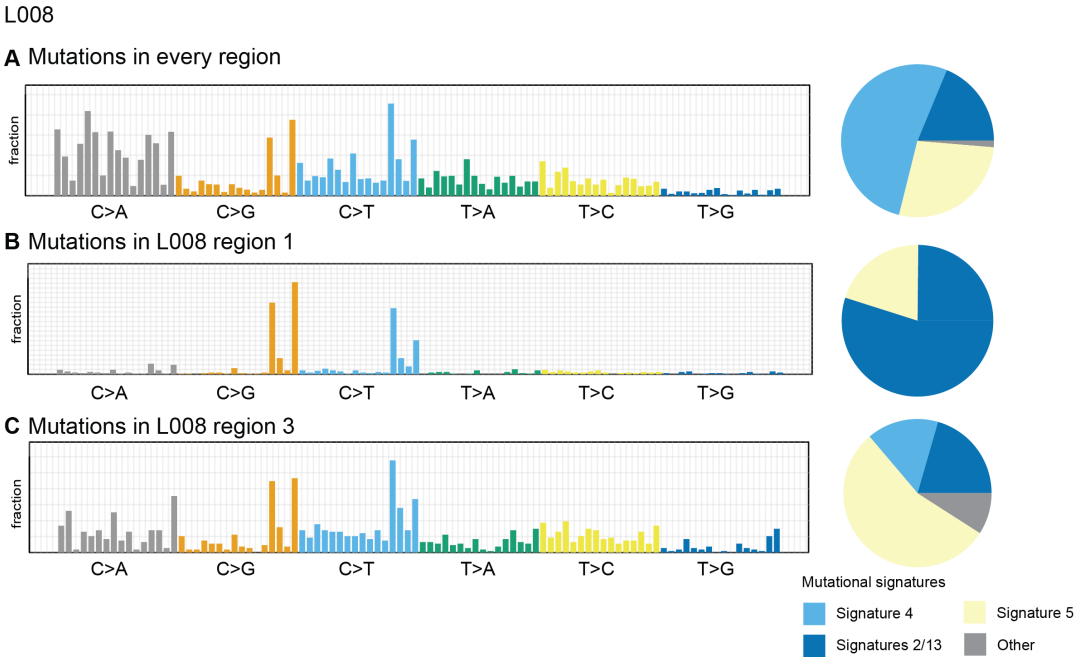

Figure 3-9: Dynamics of mutational signatures in patient L008.
The mutational signatures are shown obtained from considering the early (truncal) mutations (A) or from only considering the mutations specific to a particular region (B, C).

These results demonstrate how it is possible to use deconstructSigs to understand how well-established mutational processes vary over a tumor's lifetime. Importantly, with a more refined analysis of tumor phylogeny, it is also possible to consider what mutational processes are contributing to individual cancer subclones, which will be explored in greater detail below, and in subsequent chapters.

3.4.2 APOBEC associated driver mutations

Multiple groups have identified APOBEC activity as a significant contributor to branched evolution and the acquisition of subclonal mutations (de Bruin et al., 2014, Starrett et al., 2016). A recent analysis of 100 patients with untreated, surgically resected primary NSCLC, enrolled in the TRACERx study has begun to shed light on the extent of ITH present in early-stage disease (Jamal-Hanjani et al., 2017). One key observation is that the overall number of subclonal mutations strongly correlated with the APOBEC-associated signatures, which is consistent with APOBEC being active late in cancer evolution. Furthermore, in 19 patients with an APOBEC signature, subclonal driver events were detected as occurring in the APOBEC mutational context (Figure 3-10).

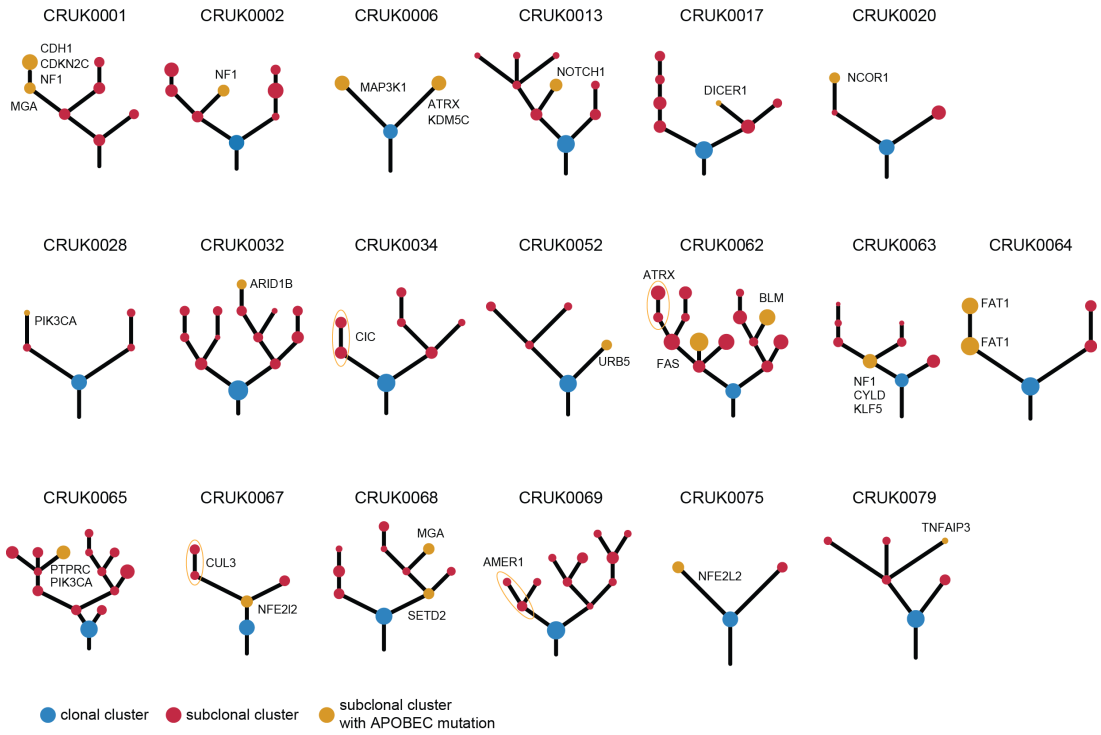

Figure 3-10: TRACERx patients harboring a subclonal driver mutation in an APOBEC preferred motif. The mutation is indicated near the clone in which it occurs. Clonal clusters are shown in blue, subclonal clusters are shown in red, and subclonal clusters containing the driver mutation in a preferred APOBEC motif are shown in yellow. Instances where a mutation could have arisen in multiple possible clones are circled in yellow.

Taken together, this suggests that APOBEC activity is a strong mutagenic force late in cancer evolution. APOBEC activity could drive cancer evolution by permitting the acquisition of late driver mutations. This is further supported by evidence for driver mutations in preferred APOBEC motifs observed across multiple cancer types (Roberts et al., 2012, Henderson et al., 2014, Jamal-Hanjani et al., 2017,

McGranahan and Swanton, 2015, Nik-Zainal et al., 2016). The most striking examples of functional APOBEC driven mutagenesis are two helical domain hot spot mutations in *PIK3CA* commonly observed in human papillomavirus-positive head and neck squamous cell carcinomas (E542K and E545K) (Henderson et al., 2014).

3.5 Conclusions

The identification of mutational signatures active over the life history of a tumor can further the understanding of how distinct mutational processes contribute to a tumor's initiation, subsequent diversification, and progression.

Furthermore, by determining the signatures present in samples with temporally dissected mutations, it is possible to quantify which mutational signatures are active early and late during tumor development, allowing for a deeper understanding of how cancer evolves and how factors such as APOBEC activity, or mutagenic chemotherapeutic agents (Johnson et al., 2014, Murugaesu et al., 2015, Meier et al., 2014, Findlay et al., 2016), may act to alter the evolutionary path of a tumor. A more thorough understanding of cancer etiology has great implications for prevention, as well as for informing the best patient-specific therapeutic choices (Le et al., 2015, Alexandrov et al., 2015).

Different mutational signatures may also affect the local tumor microenvironment to varying extents. Recent reports have shown an association between immune activity, as measured by the activity of cytotoxic T-cells, and mutational load (Rooney et al., 2015). As some mutational signatures, such as those resulting from mismatch repair deficiencies or APOBEC activity, are associated with higher overall mutational loads, they may have a more substantial impact on tumor microenvironment. Indeed, early reports indicate that mismatch repair, double strand break repair, and APOBEC driven mutational landscapes have a significant increase in TIL infiltrate and overall cytolytic activity as compared to tumors where other mutational processes predominate (Connor et al., 2017, Smid et al., 2016).

Alexandrov, Nik-Zainal, and colleagues were first to robustly examine how all mutations, not just the functional drivers of tumorigenesis, could provide insight into what has occurred during the development of a tumor. Their comprehensive analysis of publicly available exomes and genomes resulted in a curated list of mutational signatures recurrent across multiple cancer types (Alexandrov et al.,

2013a, Alexandrov et al., 2013b, Nik-Zainal et al., 2012a). This chapter outlines a computational approach, deconstructSigs, that complements the work already performed to determine what mutational signatures are active in individual tumor specimens from a set of pre-defined input signatures, circumventing the need for large sample sets before signature analysis is viable. Through the use of deconstructSigs, it is possible to consistently identify the same mutational processes active in individual tumor samples as when a similar analysis is performed using the WTSI mutational signature framework (Alexandrov et al., 2013a), providing confidence in the accuracy of the single sample approach.

Finally using this approach, it was possible to illuminate the dynamic nature of mutational processes active in single tumors over time, through the consideration of temporally dissected mutations. Continuation of this sort of analysis will help to elucidate which mutational processes may drive early tumor development, and which may allow for subsequent diversification, subclonal expansion, and potentially immune evasion.

One potential caveat of the reference signatures used in deconstructSigs is that each signature represents the “consensus” signature across the multiple cancer types it was identified in. In practice, there are likely subtle differences in the signatures arising from the same mutational process across different tissue types, which may give rise to inaccurate results depending on the degree of discrepancy between the “consensus” signature and the tissue-type specific one.

However, a key aspect of the tool is that the input signature set is a user-defined parameter, so as additional mutational signatures are identified and the current ones are refined through on-going large-scale genomic analyses, or as cancer-type specific signatures become available, it will be possible to alter the signature set under consideration. For instance, there are currently 30 curated signatures identified by the WTSI (http://cancer.sanger.ac.uk/cosmic/signatures), some of which were only identified in a single tumor sample, and others from tumor types not analyzed in this chapter, such as stomach cancer, kidney clear cell carcinoma, and Hodgkin’s lymphoma. In future studies, and as new signatures are identified, these signatures could be included in the input signatures set by the user.

Furthermore, a constrained multiple linear regression model can be used to deconvolve signatures developed from any number of data types where there is a reference signature available, including recently described copy number variation

signatures (Macintyre et al., 2017, Glodzik et al., 2017). Thus, deconstructSigs can continue to complement other efforts to define and identify mutational processes.

As the sequencing of individual tumors continues to become an aspect of cancer treatment, the ability to focus on single samples will be necessary to understand key characteristics of the tumor from a patient-specific point of view. It may be possible to reveal cancer vulnerabilities that may guide clinical decision-making on a patient-specific basis or even identify potential occult environmental exposures within individual tumors. One such case has been observed in the TRACERx study already (Jamal-Hanjani et al., 2017). A tumor from a patient who had never smoked showed a mutational profile that exhibited undeniable signs of early smoking associated mutations. While the patient himself had never smoked, he had a long history of environmental exposures including arsenic, benzene, bisphenol, polybrominated diphenyl ethers, and coal tar, which may result in mutations similar to those associated with tobacco. Importantly, it was possible to analyze this sample on its own to answer a specific question.

Chapter 4 Determinants of immune recognition

4.1 Introduction

While it is well-established that the immune system is capable of recognizing and eliminating tumor cells, the factors determining tumor cell antigenicity remain incompletely cataloged (Schreiber et al., 2011, Chen and Mellman, 2017). Recent advances in the understanding of the interaction between the tumor and the immune system have resulted in the development of effective cancer immunotherapies, such as immune checkpoint blockade therapy, CAR T-cell therapy, and neoantigen specific vaccines, which all aim to activate the immune system to allow for cancer cell elimination via endogenous T-cell activity (Schumacher and Schreiber, 2015). These immunotherapies act to enhance pre-existing T-cell responses to tumor neoantigens presented on the cancer cell or generate new ones (Liu and Mardis, 2017).

However, the recent clinical success of immunotherapies in a minority of cancer patients has served to highlight both the potential impact of immune modulation as well as our limited understanding of the factors underpinning patient response (Hodi et al., 2010, Brahmer et al., 2012, Topalian et al., 2012a, Wolchok et al., 2013).

As discussed in the previous chapter, ongoing mutational processes active during tumor evolution can result in tumors harboring tens to tens of thousands of somatic alterations (Vogelstein et al., 2013, Stratton, 2011). Many of these mutations lead to amino acid changes that may result in neoantigen generation; thus the non-synonymous mutation/neoantigen landscape of a tumor is capable of directly contributing to its immunogenicity.

Indeed a relationship between the presence of tumor neoantigens, immune activation, and improved prognosis has been documented (Brown et al., 2014, Rooney et al., 2015). Recent analyses of checkpoint blockade treated cohorts designed to elucidate what differentiates the tumors of responding patients from those of non-responders have consistently identified non-synonymous mutation/neoantigen burden as a contributor to response (Snyder et al., 2014, Rizvi et al., 2015, Van Allen et al., 2015, Le et al., 2015). Moreover, T-cell responses elicited towards specific neoantigens have been demonstrated in both pre-clinical and clinical studies, (Castle et al., 2012, Tran et al., 2016, Rizvi et al., 2015, Linnemann et al., 2015). Taken together, these studies indicate that neoantigens

represent an attractive target for cancer therapy and that a greater understanding of the immune system's regulation and response to tumor neoantigens is needed.

Although increased genetic ITH has been shown to correlate with poor prognosis across multiple cancer types (Mroz and Rocco, 2013, Sveen et al., 2016, Schwarz et al., 2015, Jamal-Hanjani et al., 2017), little is known about the impact of neoantigen ITH on the immune response. Thus in this chapter, the effect of neoantigen ITH is considered in relation to both treatment-naïve patients and cohorts of patients treated with immune checkpoint blockade therapy.

The work presented in this chapter was published as a joint-first author paper, (McGranahan et al., 2016). Bioinformatics analysis of the checkpoint blockade treated cohorts was performed in collaboration with Nicholas McGranahan. Experiments to identify and characterize neoantigen reactive T-cells were performed by Andrew Furness and Sine Hadrup's group at the Danish Technical University.

4.2 Neoantigen prediction pipeline

The first step towards identifying putative neoantigens across large tumor cohorts in a streamlined manner, is developing a neoantigen prediction pipeline (Figure 4-1).

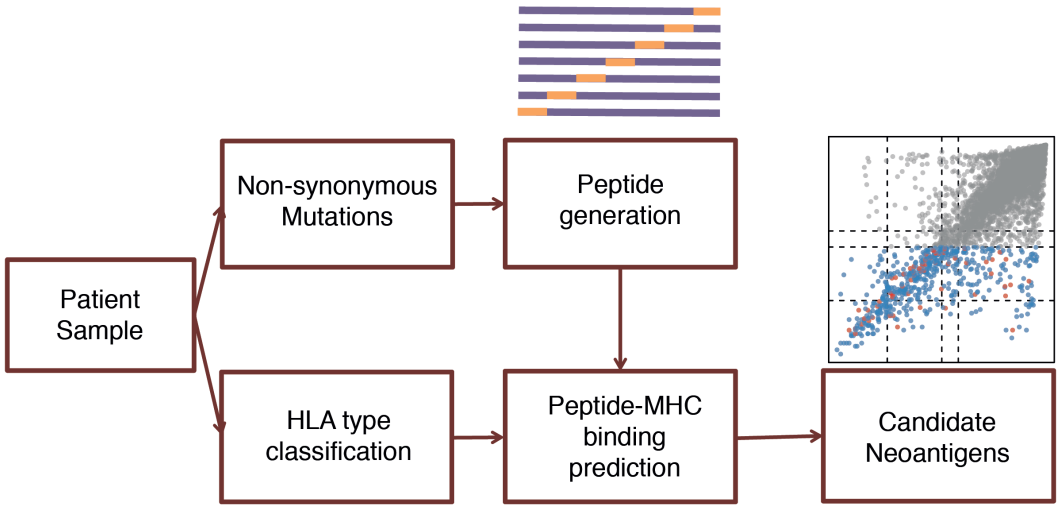

Figure 4-1: Schematic of neoantigen prediction pipeline.
Non-synonymous mutations are used to generate a comprehensive list of peptides 9-11 amino acids in length with the mutated amino acid represented in each possible position. A patient's HLA type is determined using an HLA inference tool. To predict the likelihood of a peptide binding for presentation, every mutant peptide and its corresponding wildtype peptide are used as input to netMHCpan and netMHC. Candidate neoantigens are identified based on binding affinity or rank percentage score, a parameter which can be changed in the pipeline depending on user input.

4.2.1 Peptide prediction

For each non-silent mutation (as determined in the Data and Methods), the pipeline identifies the affected amino acid(s) and the ones surrounding it. Then, using a sliding window approach of fixed length, it generates all possible peptide configurations with the mutation in each position. Because the available tools for proteasomal cleavage do not yield entirely accurate results, potentially eliminating antigenic peptides, every possible mutant configuration is created to be considered as a possible neoantigen. As future steps in the neoantigen prediction pipeline consider HLA class I presentation, a list of 9-11mer peptides generated for each mutation is saved.

4.2.2 HLA typing

For a non-synonymous mutation to result in a neoantigen and induce an immune response, a peptide containing the mutation must be processed and presented on the cell surface via MHC molecules (Neefjes et al., 2011). The antigenic peptide may then be recognized by T-cells, triggering immune activity. While there are many steps involved in antigen presentation ranging from proteasomal cleavage to TAP transport of the peptide to MHC-peptide binding, the most selective step in an immune response is the binding of the antigenic peptide to the MHC molecule for presentation (Nielsen and Andreatta, 2016).

The MHC molecule is highly specific and binds only a small proportion of possible peptides. As a result, the particular alleles encoded at the HLA locus dictate what peptides bind for presentation. Thus, the patient's HLA haplotype must first be determined before it is possible to generate neoantigen predictions. Serotyping is the most accurate way to determine an individual's HLA haplotype; however, on a large scale, serotyping quickly becomes cost and time prohibitive. Accordingly, the last decade has seen the development of computational approaches that rely on NGS data to infer the correct HLA calls.

4.2.2.1 Comparison of HLA inference tools

HLA classification in the neoantigen prediction pipeline described here is performed using multiple approaches which output the HLA allelic combination that maximizes number of sequencing reads explained (Szolek et al., 2014, Shukla et al., 2015). Full details are provided in the Data and Methods. Sequencing reads containing possible HLA information are obtained from BAM files, either by identifying reads

mapped to the HLA locus (on the p-arm of chr6) or by extracting reads partially matching known HLA sequence. The reads are subsequently remapped to a reference file containing most known HLA alleles (6597 possible HLA alleles in total). Finally the set of HLA alleles that best explain the mapped sequencing reads is chosen.

While the published NGS HLA typing methods have been validated using PCR-verified HLA genotypes, an independent comparison was also performed for seven patients from the TRACERx pilot study who had previous serotyping results. These patients had serotype HLA calls for HLA-A and HLA-B alleles, allowing for a comparison with the bioinformatic approaches. The serotyped HLA calls and the calls inferred by both algorithmic approaches, OptiType and Polysolver, for these seven patients showed complete concordance (Table 4-1).

Table 4-1: Comparison of HLA results by serotyping and OptiType/Polysolver

	Serotyping				OptiType/Polysolver			
Sample	HLA-A	HLA-A	HLA-B	HLA-B	HLA-A	HLA-A	HLA-B	HLA-B
L011	A*24:02	A*11:01	B*35:01/42	B*49:01	A*24:02	A*11:01	B*35:01	B*49:01
L012	A*24:02	A*11:01	B*07:02	B*07:02	A*24:02	A*11:01	B*07:02	B*07:02
L013	A*02:01:01	A*32:01:01	B*44:02	B*15:01	A*02:01	A*32:01	B*44:02	B*15:01
L016	A*01:01	A*01:01	B*57:01:01	B*35:01/42	A*01:01	A*01:01	B*57:01	B*35:01
L019	A*02:01:01	A*32:01:01	B*15:01	B*27:05	A*02:01	A*32:01	B*15:01	B*27:05
L021	A*03:01:01	A*30:02:01	B*18:01	B*35:01/42	A*03:01	A*30:02	B*18:01	B*35:01
L022	A*02:01	A*03:01	B*44:03:01	B*57:01:01	A*02:01	A*03:01	B*44:03	B*57:01

Some HLA typing tools offer added functionality, such as HLA mutation calling. This may be an important consideration when determining the likelihood of neoantigen presentation, as alterations affecting the MHC molecule may disrupt antigen binding. In addition to OptiType (Szolek et al., 2014), an HLA inference tool with mutation calling capabilities, Polysolver (Shukla et al., 2015), was employed. A further analysis of 100 TRACERx patients comparing these two sequencing-based HLA inference tools found high agreement between the two sets of results (Figure 4-2).

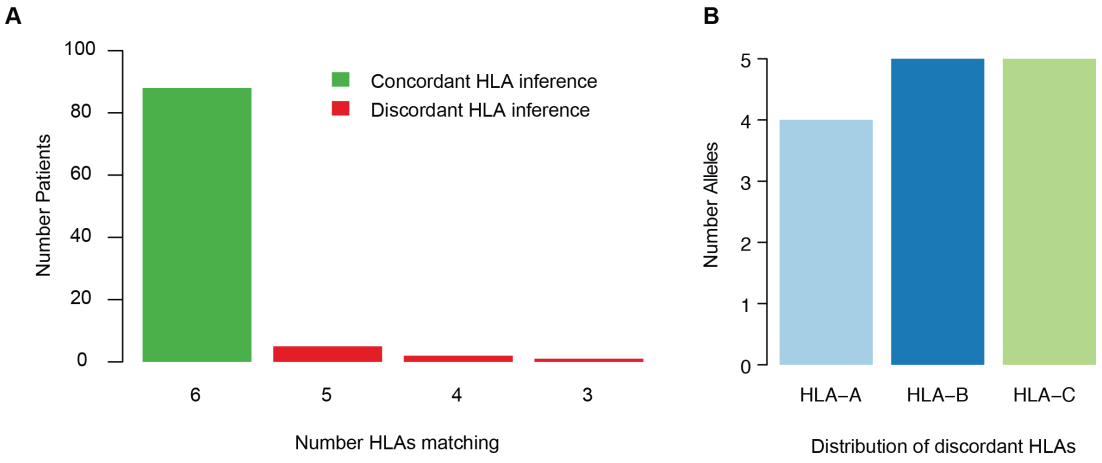

Figure 4-2: Comparison of HLA calls inferred from OptiType and Polysolver.
A) A barplot is displayed showing the number of patients and number of matching HLA calls between OptiType and Polysolver HLA inference tools. If HLA calls at all six possible class I HLA alleles matched, then the patient was considered to have concordant results between the two HLA typing tools (green). If a subset of calls did not match between the tools, the patient was considered to have discordant results (red). B) For the cases of discordant calls, the HLA alleles which did not match are shown.

Ninety-six of the patients were successfully HLA-typed using both tools, with 88 matching at all six possible HLA alleles. Five patients only matched at five HLA alleles, two matched at four HLA alleles, and a single patient had discrepancies between three HLA alleles, for a total agreement between 98% of the HLA alleles (Figure 4-2A). In this cohort, the discordant HLA calls were equally spread over the three HLA class I alleles (Figure 4-2B). Furthermore, unlike previous reports, there were no systematic HLA miscalls identified (Kiyotani et al., 2017).

The four patients that failed one of the HLA-typing tools (OptiType) were found to have extremely low-coverage at the HLA locus. By using both tools in the neoantigen prediction pipeline, it is possible to have increased confidence in inferred alleles and include patients where one tool may fail. However, these

patients should be treated cautiously, as results from a low-coverage sample are more likely to be inaccurate.

4.2.3 Peptide-MHC binding predictions

The list of possible peptides resulting from each non-synonymous mutation (as described in the Data and Methods) and the patient's HLA haplotype are next used as input to determine whether that particular peptide-MHC interaction is likely to occur naturally. To generate quantitative predictions of the affinity for a given peptide-MHC interaction, a prediction tool is used that relies on extensive training data gathered from manually validated peptide-MHC complexes (Nielsen et al., 2007, Hoof et al., 2009). Thus, for each peptide, predictions of how strongly it binds to all the patient's MHC molecules are calculated. For the data presented in this chapter, netMHC-pan v2.8 was used with a neoantigen definition of binding affinity < 500nM. Since publication of this data, netMHC-pan v3.0 has been released, and the tool maintainers have suggested a rank-percentage based neoantigen definition, where the rank of the predicted binding affinity is compared to a set of 400,000 random natural peptides. The rank-percentage measure should not be affected by the availability of training data for that particular HLA allele, which may bias certain HLAs towards higher or lower predicted binding affinities overall.

Candidate neoantigen peptides that are predicted to bind to the patient's MHC molecules can subsequently be filtered further using user-defined parameters, such as peptide length or mutation position, or expression if RNAseq data is also available.

4.3 Neoantigen landscape in multi-region NSCLC

4.3.1 Extent of heterogeneity in neoantigen landscape

While recent research has begun to shed light on the relationship between immune recognition of cancer cells and neoantigens present in the tumor (Matsushita et al., 2012, Rizvi et al., 2015, Castle et al., 2012, Snyder et al., 2014), the impact ITH has on this relationship had not been considered. Thus, to build an accurate picture of the landscape of neoantigen heterogeneity (neoantigen ITH) in early stage NSCLC tumors, the neoantigen prediction pipeline was applied to seven multi-region sequenced tumors, representing a range of histologies, including adenocarcinoma, squamous cell carcinoma, and a single patient with mixed histology (Table 4-2) (de Bruin et al., 2014, Jamal-Hanjani et al., 2016).

Table 4-2: Clinical characteristics of multi-region sequenced NSCLC patients
Abbreviations: RLL, right lower lobe; RUL, right upper lobe; RML, right middle lobe; R, region; LN, lymph node; Undiff, undifferentiated. * A pack-year is defined as the number of packs of cigarettes smoked per day multiplied by the number of years the person has smoked.

Patient ID	Age (years)	Gender	Histology	Lymph node(s)/ location	Stage (I-IV)	Regions sequenced	Smoking status (pack-years*)
L003	84	F	lung ad.	2/Station 4	IIIB	R2 (RLL), R4 (RUL), LN	never-smoker
L008	75	M	lung ad.	2/Hilar	IIIA	R1 (RUL), R3 (RML), LN	ex-smoker (25)
L001	59	F	lung ad.	3/Hilar	IIA	R1-R5, LN	ex-smoker (10)
L004	73	M	Undiff. NSCLC	none	IIB	R1-R4	current smoker (50)
L011	49	F	lung ad.	none	IB	R1-R3	current smoker (45)
L002	78	M	lung ad./ lung squam.	2/Station 5	IIIA	R1-R4	current smoker (>50)
L012	69	F	lung squam.	none	IB	R1-R3	current smoker (40)

From this cohort of seven patients, 2860 putative neoantigens were predicted (median 326; range 80-741) (Figure 4-3). As patients in this dataset had multi-region sequencing performed, the neoantigen ITH could also be accurately considered by determining the proportion of subclonal neoantigens. For this calculation, the same mutation annotations were used as originally published (de Bruin et al., 2014) and a neoantigen was considered subclonal if it was only found in a fraction of tumor regions. Consistent with the observed mutational heterogeneity (de Bruin et al., 2014), the calculated neoantigen ITH varied substantially across the cohort, with an average of 44% of putative neoantigens identified as subclonal per tumor. One of the most heterogeneous tumors, L012, was found to harbor a total of 252 neoantigens of which 74% were heterogeneous. Conversely, for L011, the most homogenous tumor within the cohort, fewer than 10% (39/400) were heterogeneous. The wide range of neoantigen ITH observed in this cohort suggested that it would be possible to study the impact of ITH on neoantigen response among these patients.

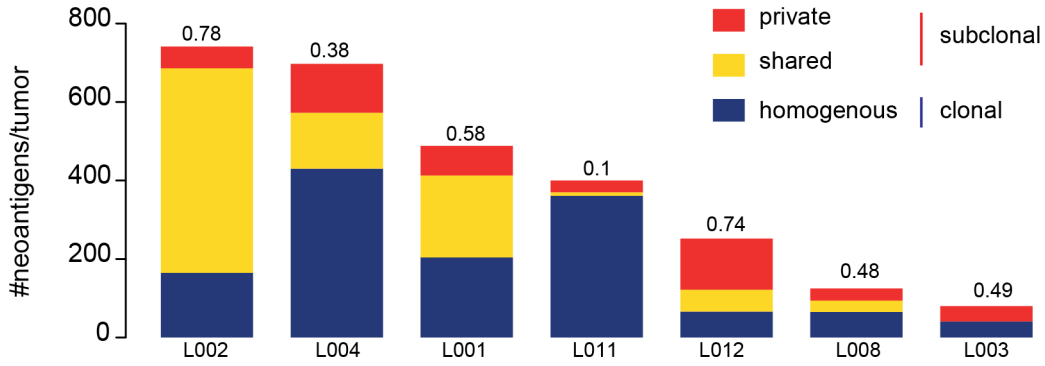

Figure 4-3: Heterogeneity of neoantigen landscape in TRACERx pilot study.
The total predicted neoantigen load is plotted for each of the multi-region NSCLC tumors from the TRACERx pilot study. The number of clonal neoantigens, which could be found in every tumor region are shown in blue. Subclonal neoantigens, which were identified in multiple, but not every tumor region are shown in yellow. Subclonal neoantigens only identified in a single tumor region are show in red. Above each barplot, the fraction of subclonal neoantigens is indicated.

4.3.2 Identification of T-cells reactive to predicted neoantigens

To validate the calls from the neoantigen prediction pipeline and investigate the impact of neoantigen ITH on T-cell recognition, neoantigens from two tumors, L011 and L012, were tested using MHC-multimers (as described in the Data and Methods). Despite having a comparable number of predicted neoantigens and lengthy history of smoking, these two patients were on opposite ends of the neoantigen ITH spectrum (10% vs. 74% heterogeneous predicted neoantigens, Figure 4-4A). HLA-matched multimers loaded with 288 and 354 putative

neoantigens from L011 and L012, respectively, were used to screen CD8+ T-cells for neoantigen reactivities. CD8+ T-cells were collected from tumor regions as well as from adjacent normal tissue and expanded *in vitro*. Finally a high throughput method allowed for testing the expanded CD8+ T-cell populations against many peptides at once (Hadrup and Schumacher, 2010).

Reassuringly, and in validation of some neoantigen predictions generated, CD8+ T-cells reactive to putative neoantigens were identified in both patients. In L011, a mutant peptide arising from MTFR2^{D326Y} (FAFQE***Y***DSF) resulted in a CD8+ T-cell response (Figure 4-4B,D). In L012, two separate CD8+ T-cell responses were identified from peptides arising from a CHTF18^{L769V} mutation (LLDI***V***APK) and a MYADM^{R30W} mutation (SPMIVGSP***W***) (Figure 4-4C,E). As confirmation that these reflected patient-specific CD8+ T-cell responses, four healthy donor CD8+ PBMCs were tested against the same peptides resulting in no observable response (Figure 4-4 D-E).

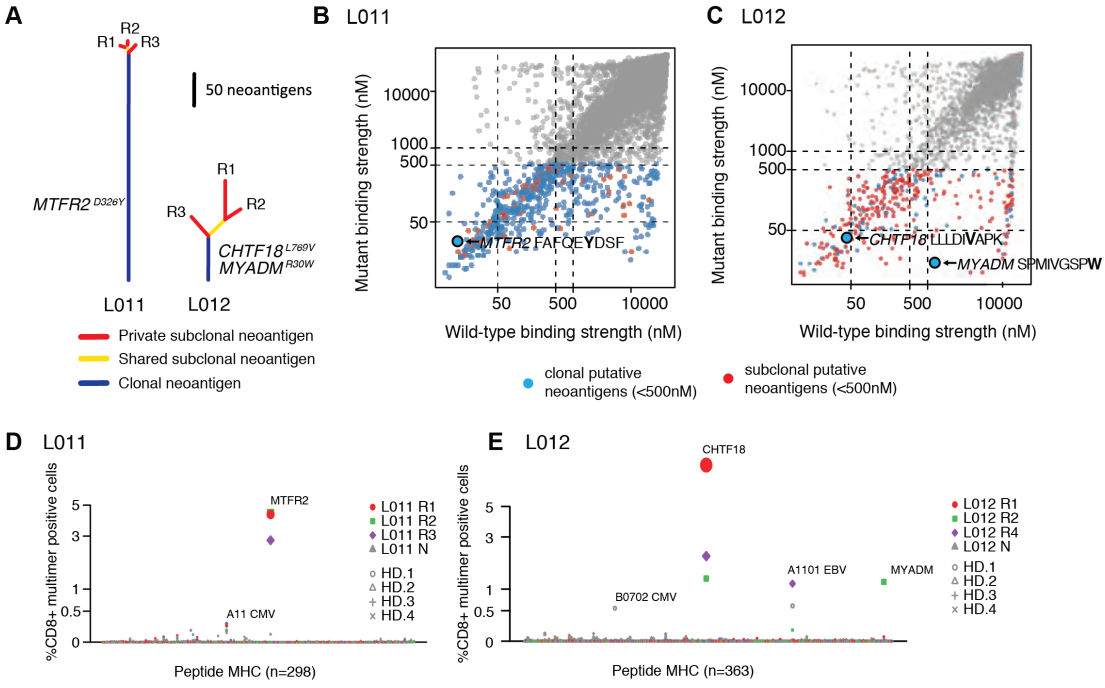

Figure 4-4: Prediction and identification of neoantigen-reactive T-cells.
A) Phylogenetic trees for patients L011 and L012 based on predicted neoantigens are depicted. B) The mutant binding strength and wildtype binding strength are plotted for each predicted neoantigen from all missense mutations in L011. Lower binding affinity indicates a stronger predicted binder. Clonal neoantigens are indicated in blue, subclonal neoantigens are indicated in red. The MTFR2^{D326Y} neoantigen is specifically marked (FAFQE**Y**DSF). C) The mutant and wildtype binding strengths are plotted for each predicted neoantigen from all missense mutations in L012. The CHTF18^{L769V} neoantigen (LLLDI**V**APK) and the MYADM^{R30W} neoantigen (SPMIVGSP**W**) are specifically marked. D) and E) Results from MHC-multimer screening of expanded, region-specific, tumor-infiltrating CD8+ T-cells and healthy donor (HD) CD8+ PBMC controls are shown. Candidate neoantigens (L011, n=288 and L012, n=354) and control HLA-matched viral peptides (L011, n=10 and L012, n=9) were tested. The frequency of CD8+ MHC-multimer positive cells out of total CD3+CD8+ TILs is displayed.

These results proved that neoantigen responses could be identified in patients with both high and low levels of neoantigen ITH, but interestingly, the three mutations that generated identifiable CD8+ T-cell responses all represented clonal neoantigens. In the homogenous patient, L011, this was expected, as the vast majority of predicted neoantigens were clonal; however, the neoantigen landscape of patient L012 was extremely heterogeneous and still the only observable responses arose from clonal mutations. Thus, this suggests that clonal neoantigens can be recognized and initiate immune activity in both homogeneous and heterogeneous tumors.

To ensure the TIL expansion protocol did not bias the observed CD8+ T-cell responses, non-expanded TILs were also considered. Functional CD8+ T-cell responses in L011 from non-expanded TILs, were still identified against the $MTFR2^{D326Y}$ mutation in all tumor regions, with a far greater frequency than the normal region. Similarly, CD8+ T-cells reactive to the $CHTF18^{L769V}$ peptide and $MYADM^{R30W}$ peptide could also be identified in non-expanded samples from all tumor regions in L012, with a lower frequency in the normal lung tissue (Figure 4-5 A-B).

As previously explained in the Introduction, a mutant peptide may generate a tumor-specific immune response either by being a novel binder to a patient's MHC molecule (the mutant peptide has a stronger binding affinity than the wildtype) or by harboring a mutation recognizable by a T-cell (both mutant and wildtype peptides have strong binding affinities). The peptides arising from the $MTFR2^{D326Y}$ mutation in L011 and the $CHTF18^{L769V}$ mutation in L012 belong to the latter category and show very little reactive CD8+ T-cells in normal tissue as compared to tumor tissue (Figure 4-5A and y-axis of Figure 4-5B).

However, $MYADM^{R30W}$-reactive CD8+ T-cells were observed in normal tissue at a higher percent than either of the other two neoantigen generating mutations (x-axis of Figure 4-5B). The mutant peptide arising from $MYADM^{R30W}$ represents an instance of the first category of neoantigen, where the mutant version of the peptide has a stronger binding affinity than the wildtype version. The mutation which generated the $MYADM^{R30W}$ peptide affects the anchor residue, which alters the peptide-MHC binding interaction rather than T-cell recognition. Thus it appears that in patient L012, the increased stability provided within a MHC-multimer system allowed the T-cells to recognize both mutant and wildtype peptides; however *in vivo*, without the artificial stability of the MHC-multimer, the low predicted binding

affinity of the wildtype peptide to the HLA molecule would most likely result in no peptide presentation.

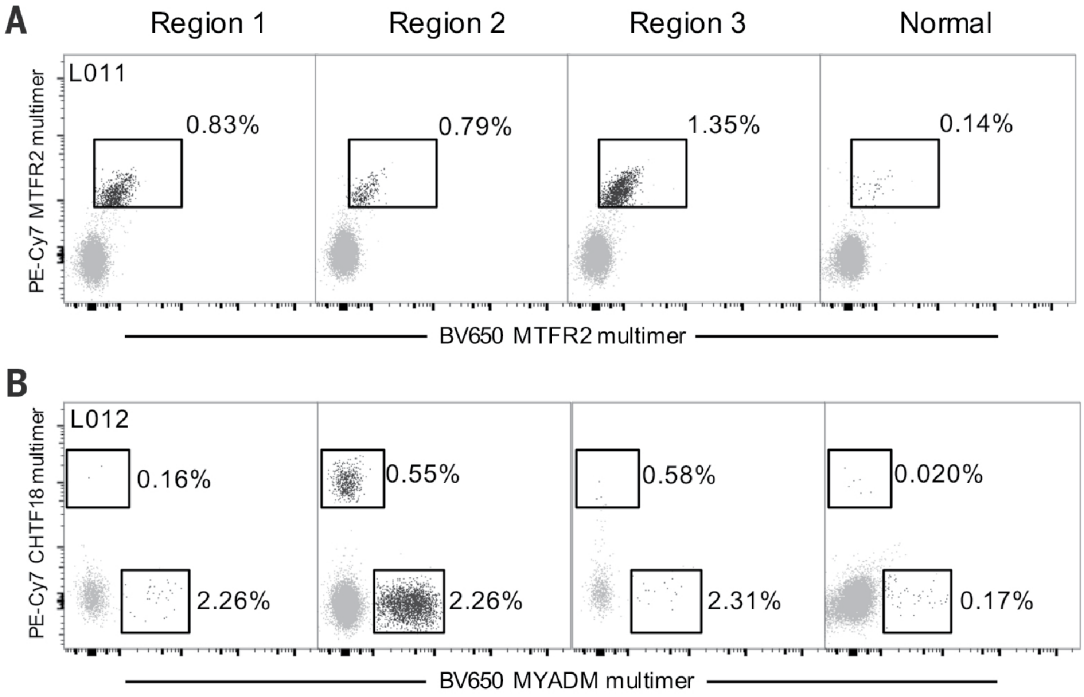

Figure 4-5: Clonal neoantigens identified in multi-region NSCLC.
A) MHC-multimer analysis of non-expanded, tumor-infiltrating CD8+ T-cells isolated from tumor regions and normal lung tissue of L011 showing recognition of mutant MTFR2 peptide. The percentage of CD8+ T-cells recognizing multimers with the fluorescent marker indicated are shown on the x- and y-axes. As there was only one reactivity in L011, both multimers are for the same peptide, so reactive T-cell populations are in the top-right quadrant. B) MHC-multimer analysis of non-expanded, tumor-infiltrating CD8+ T-cells isolated from tumor regions and normal lung tissue of L012 identifies two populations of CD8+ TILs reactive to mutant CHTF18 and MYADM peptides. The percentage of CD8+ T-cells recognizing multimers with the fluorescent marker indicated are shown on the x- and y-axes. There were two reactivities identified in L012, so CD8+ T-cells recognizing the MYADM peptide are to the right on the x-axis, and those recognizing the CHTF18 peptide are towards the top on the y-axis.

4.4 Applying the neoantigen prediction pipeline to TCGA tumors

To further determine the extent of neoantigen ITH in a wider cohort of NSCLC, the neoantigen prediction pipeline was used to analyze data from HLA-typed NSCLC patients available through TCGA, which consisted of 150 lung adenocarcinoma and 124 lung squamous cell cases (Lawrence et al., 2014, Rooney et al., 2015, Shukla et al., 2015). The TCGA cohort represented primarily early-stage disease consisting of 106 stage I/II, 43 stage III/IV lung and 1 unknown stage lung adenocarcinoma and 92 stage I/II and 32 stage III/IV lung squamous cell carcinoma cases. Furthermore, to quantify the neoantigen heterogeneity of each sample, the cancer

cell fraction of each mutation, which describes the proportion of cancer cells that harbors a given mutation, was calculated.

In keeping with previous reports and reflecting the increased mutational burden found in lung squamous cell tumors, in this cohort, lung squamous cell tumors harbored an average of 140 putative neoantigens and lung adenocarcinoma harbored an average of 103 putative neoantigens. (Rajasagi et al., 2014). As has been observed previously (Rooney et al., 2015), the number of predicted neoantigens strongly correlates with the number of missense mutations in the tumor. This relationship held for both clonal (LUAD: cor = 0.96, $p < 2.2e-16$; LUSC: cor = 0.96, $p < 2.2e-16$) and subclonal neoantigens (LUAD: cor = 0.95, $p<2.2e-16$; LUSC: cor = 0.97, $p< 2.2e-16$) (Figure 4-6). Notably, a considerable range in neoantigens per tumor, as well as level of neoantigen heterogeneity was observed (Table 4-3).

Table 4-3: Summary of neoantigens and neoantigen ITH in TCGA cohort.

	Lung adenocarcinoma			Lung squamous cell carcinoma		
	1st Qu.	Median	3rd Qu.	1st Qu.	Median	3rd Qu.
Neoantigen	34	73	128	78	110	150
Clonal Neoantigen	20	57	115	54	84	121
Subclonal Neoantigen	6	12	24	12	21	37
Neoantigen ITH	0.09	0.18	0.31	0.12	0.20	0.34

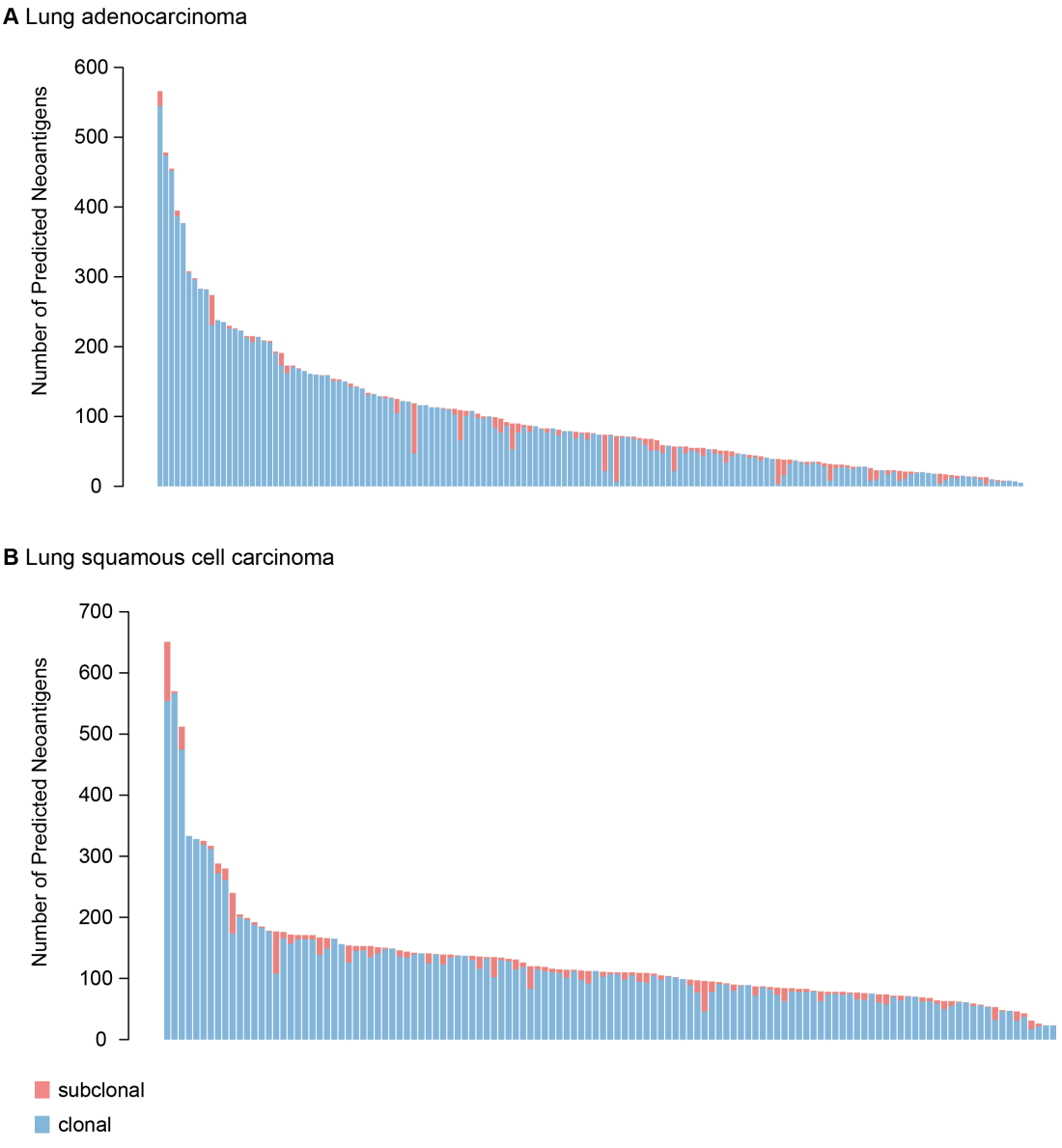

Figure 4-6: Neoantigens predicted in TCGA NSCLC cohort.
Number of predicted neoantigens is shown for lung adenocarcinoma (A) and lung squamous cell carcinoma (B) tumors from TCGA, with neoantigens arising from clonal mutations indicated in blue and neoantigens arising from subclonal mutations indicated in red. Clonality of mutations was determined using the cancer cell fraction.

4.5 Clinical impact of neoantigen heterogeneity

4.5.1 Neoantigen load and heterogeneity associates with survival in the treatment-naïve setting

A large tumor neoantigen burden may increase tumor recognition by T-cells, reducing the potential for immune-evasion (Schreiber et al., 2011). Previous work has provided some evidence of the clinical relevance of tumor neoantigens (Brown et al., 2014). In support of these previous findings, a high neoantigen load (defined as the upper quartile of the number of neoantigens predicted in the cohort) was associated with longer overall survival times in TCGA lung adenocarcinoma samples with matched clinical data (n=139) when compared to the remaining tumors in the cohort. (Figure 4-7A, log-rank p = 0.025). While the lung squamous cell carcinoma tumors had a similar total neoantigen landscape, there appeared to be no relationship between overall survival and total neoantigen load (Figure 4-7B, log-rank p = 0.81).

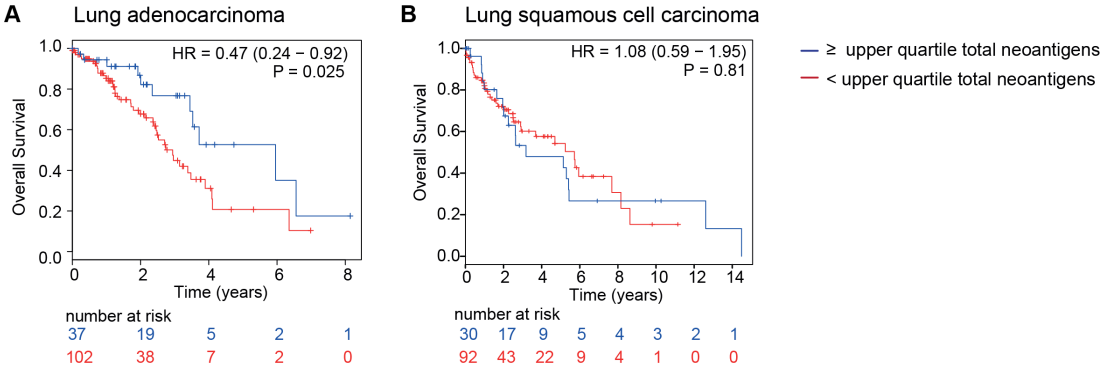

Figure 4-7: Relationship between neoantigen burden and survival in TCGA.
Kaplan-Meier curves are shown for lung adenocarcinoma (A) and lung squamous cell carcinoma (B). The curves are split based on the upper quartile of total neoantigen burden, with high neoantigen burden tumors signified by the blue line and low neoantigen burden tumors signified by the red line. A significant association between overall survival and total neoantigen burden is observed in lung adenocarcinoma, but not lung squamous cell carcinoma.

To determine whether neoantigen clonal status might influence the relationship with survival outcome, potentially allowing for a refined understanding of the impact of neoantigen ITH on immune recognition and response, each putative neoantigen was classified as either arising from a clonal or subclonal mutation. Lung adenocarcinoma patients with homogeneous tumors (neoantigen ITH < 1%) tended to have an increased overall survival time than when compared to patients with heterogeneous tumors (log-rank p = 0.06). Interestingly, a combination of neoantigen ITH and neoantigen burden was more significant than considering either

metric alone and was observed across a range of neoantigen ITH thresholds. In this analysis, a neoantigen ITH=0 represents a tumor where all neoantigens were called as clonal and a neoantigen ITH=0.05 represents a tumor where 5% of the neoantigens were identified as subclonal (Figure 4-8; ITH threshold=0, log-rank p=0.019; ITH threshold=0.01, log-rank p=0.0096; ITH threshold=0.05, log-rank p=0.021). This association remained significant when incorporating tumor stage and patient gender and age in a multivariate analysis (Table 4-4), suggesting that that not just the presence of neoantigens but also their prevalence within the tumor relates to overall survival in treatment-naïve lung adenocarcinoma patients.

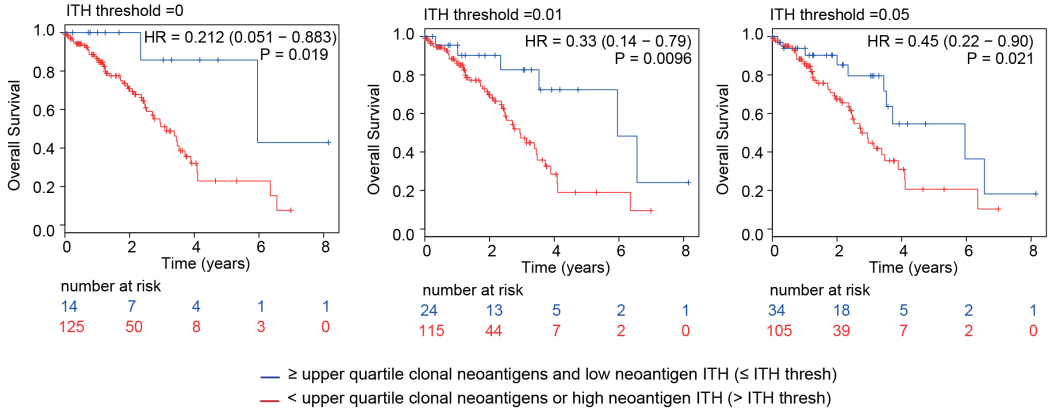

Figure 4-8: Relationship between neoantigen ITH and survival in TCGA lung adenocarcinoma.
Kaplan-Meier curves are shown for TCGA lung adenocarcinoma tumors across multiple levels of neoantigen ITH. Tumors with a high clonal neoantigen burden and neoantigen ITH below the given threshold are shown in blue; tumors with either a low clonal neoantigen burden or neoantigen ITH above the given threshold are show in red.

Table 4-4: Multivariate survival analysis in lung adenocarcinoma.

Neoantigens (without ITH threshold)				
	HR	95% CI lower	95% CI upper	p-value
Number neoantigens	0.996	0.992	1	0.025
Gender	0.675	0.372	1.226	0.2
Early stage (vs. late)	0.259	0.141	0.476	0
Age	1.015	0.985	1.045	0.33
ITH threshold =0				
	HR	95% CI lower	95% CI upper	p-value
High neoantigen and ITH<=0	0.291	0.069	1.237	0.095
Gender	0.66	0.362	1.204	0.18
Early stage (vs. late)	0.313	0.171	0.572	0
Age	1.017	0.989	1.047	0.23
ITH threshold =0.01				
	HR	95% CI lower	95% CI upper	p-value
High neoantigen and ITH<=0.01	0.262	0.103	0.667	0.005
Gender	0.619	0.339	1.132	0.12
Early stage (vs. late)	0.236	0.126	0.442	0
Age	1.017	0.989	1.047	0.23
ITH threshold =0.05				
	HR	95% CI lower	95% CI upper	p-value
High neoantigen and ITH<=0.05	0.366	0.172	0.781	0.009
Gender	0.66	0.363	1.2	0.174
Early stage (vs. late)	0.241	0.13	0.445	0
Age	1.012	0.982	1.042	0.432

An independent cohort of patients from the TRACERx main study was also considered to validate the observed association between clonal neoantigen burden and survival in the treatment-naïve setting. Importantly, the multi-region sequencing nature of this cohort allowed for a more accurate classification of neoantigens as having arisen from a clonal or subclonal mutation. Likely due to the increased sequencing depth, as well as the improved sensitivity to detect subclonal neoantigens gained from multi-region sequencing, both a higher number of clonal neoantigens and subclonal neoantigens were identified in the TRACERx cohort as compared to TCGA (Table 4-5 and Figure 4-9). This observation is consistent with the original publication of the TRACERx study, which reported identifying significantly more mutations from the multi-region TRACERx cohort as compared to only considering a single region from the TRACERx cohort or as compared to using single NSCLC tumor samples from TCGA (Jamal-Hanjani et al., 2017).

Table 4-5: Summary of neoantigens and neoantigen ITH in TRACERx cohort.

	Lung adenocarcinoma			Lung squamous cell carcinoma		
	1st Qu.	Median	3rd Qu.	1st Qu.	Median	3rd Qu.
Neoantigen	68	128	361	178	225	360
Clonal Neoantigen	33	71	171	100	150	213
Subclonal Neoantigen	15	29	87	19	52	88
Neoantigen ITH	0.16	0.28	0.42	0.12	0.22	0.37

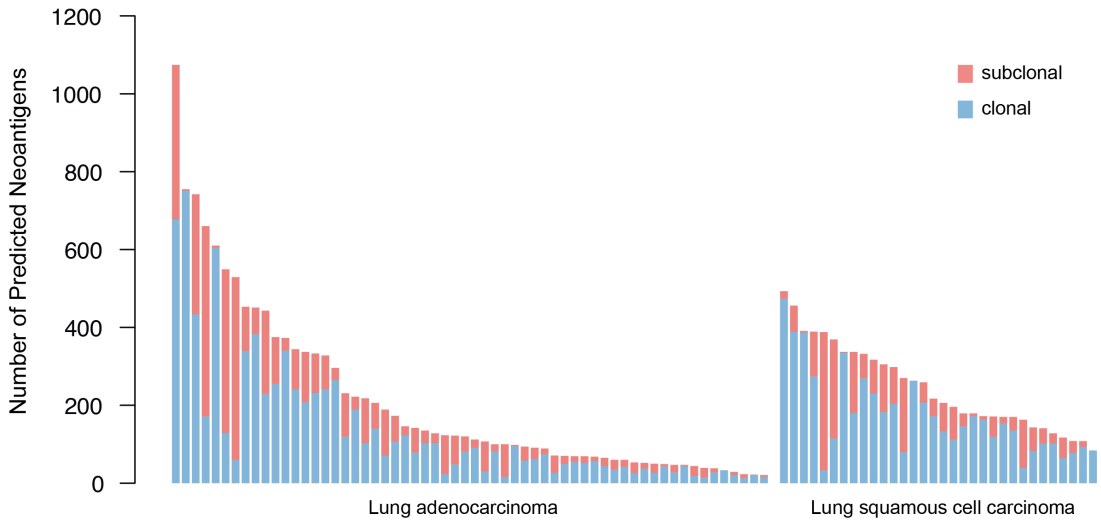

Figure 4-9: Neoantigens predicted in TRACERx cohort.
The number of predicted neoantigens is shown for lung adenocarcinoma and lung squamous cell carcinoma tumors from the TRACERx study, with neoantigens arising from clonal mutations indicated in blue and neoantigens arising from subclonal mutations indicated in red.

Interestingly, whereas in TCGA, neoantigen burden only associated with survival in lung adenocarcinoma, in both TRACERx lung adenocarcinoma and lung squamous cell carcinoma, increased clonal neoantigen burden associated with improved progression free survival (Figure 4-10A-B). This finding supports the potential prognostic relevance of clonal neoantigens and highlights the importance of sensitive subclonal mutation/neoantigen detection, as it is likely that subclonal mutations were either not called or misidentified as clonal mutations in TCGA samples.

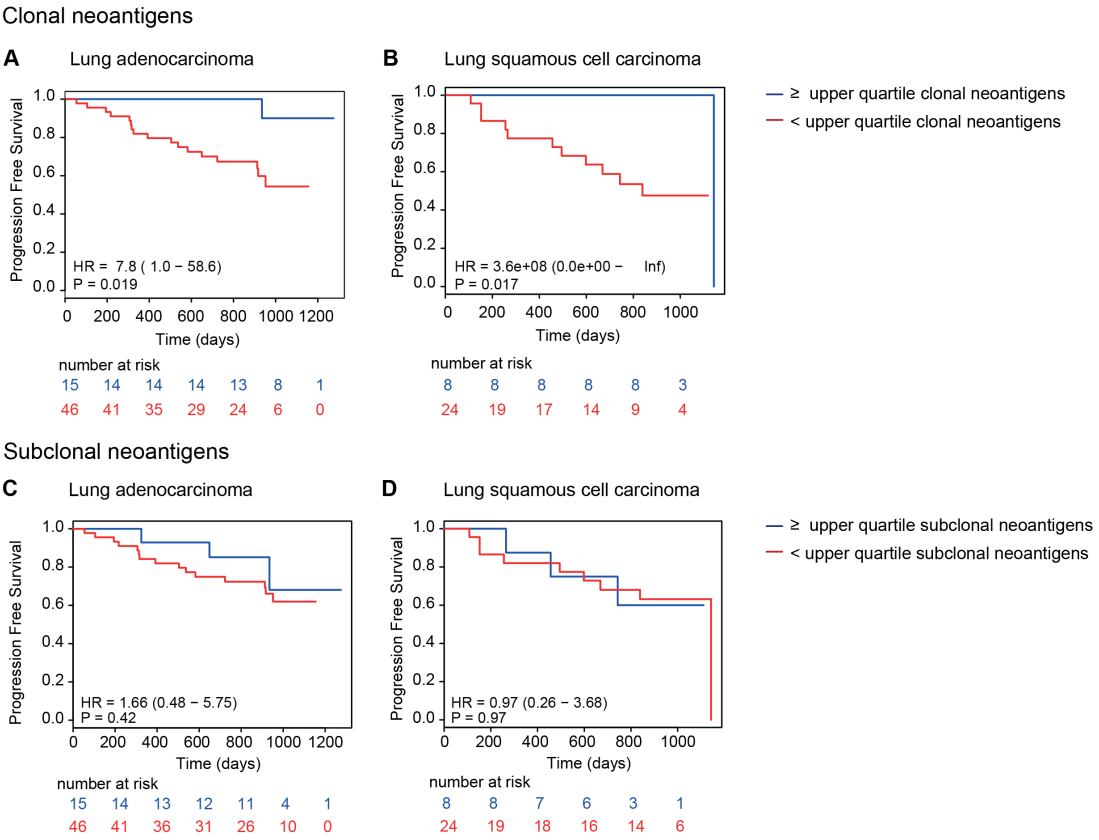

Figure 4-10: Relationship between neoantigen burden and survival in TRACERx main study cohort. Kaplan-Meier curves are shown for lung adenocarcinoma (A,C) and lung squamous cell carcinoma (B,D). The curves are split based on the upper quartile of clonal neoantigen burden (A-B) and on the upper quartile of subclonal neoantigen burden (C-D). Associations between neoantigen burden and progression free survival is only seen when clonal neoantigen burden is considered.

Contrary to clonal antigen burden, no relationship between subclonal neoantigen burden and progression free survival was observed (Figure 4-10C-D). While the numbers in each group were too small to further sub-divide by level of neoantigen ITH, as was done for the analysis of samples from TCGA, the observation that only clonal neoantigen burden and not subclonal neoantigen burden associated with improved progression free survival further supports the hypothesis that clonal

neoantigens, found in every cancer cell, may have a heightened impact on immune response as compared to subclonal neoantigens.

Interestingly, the relationship between survival and clonal neoantigen load in lung squamous cell carcinoma was inconsistent, appearing significant in the TRACERx cohort but not in TCGA. One possibility is that subclonal mutation calling was underpowered in TCGA tumors. However, it is also possible that TCGA lung squamous cell carcinoma utilized additional immune evasive mechanisms, such that high clonal neoantigen burden was not enough to elicit an effective immune response. To investigate if this was a possible reason for the observed differences between TCGA lung squamous cell carcinoma and TCGA lung adenocarcinoma, available RNAseq data, from tumor and adjacent normal samples, was used to explore whether any genes with documented roles in immune regulation were differentially expressed between these cohorts.

Almost all HLA class I genes, including HLA-A (Wald padj = 1.1e-09), HLA-B (Wald padj = 7.9e-06), HLA-C (Wald padj = 3.7e-08), as well as beta-2 microglobulin (β2M) (Wald padj = 2.0e-04), a stabilizing component of the HLA molecule that is required for HLA cell surface expression, were expressed at a significantly lower level in lung squamous cell carcinomas as compared to lung adenocarcinomas. Furthermore, this difference was not dependent on total neoantigen burden, but rather remained significant across all levels of neoantigen burden (Figure 4-11).

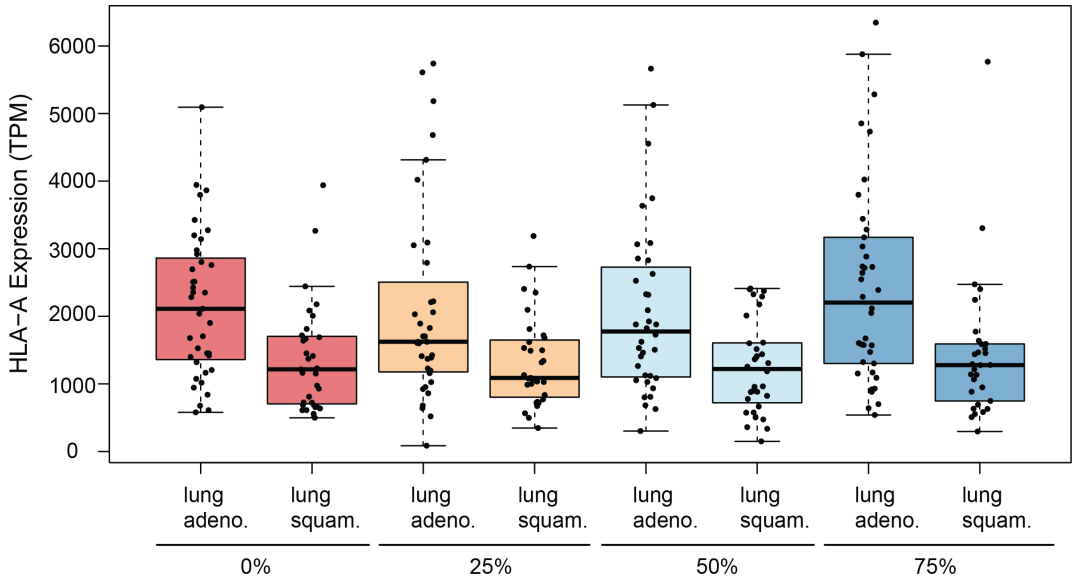

Figure 4-11: HLA-A expression in TCGA lung adenocarcinoma and lung squamous cell carcinoma. The expression of HLA-A is shown across neoantigen burden categories for lung adenocarcinoma and lung squamous cell carcinomas. HLA-A was chosen as a representative of all MHC class I genes.

HLA class I genes were also significantly down-regulated compared to matched normal samples in lung squamous cell carcinomas, but not in lung adenocarcinomas (Figure 4-12). Without HLA class I expression, potential neoantigens that could have been presented on the cell surface may no longer be detectable, thus incapable of instigating an immune response. Indeed, HLA down-regulation has been proposed in many cancer types as one possible immune evasive mechanism (Hicklin et al., 1999, Garrido et al., 2017a, Campoli et al., 2002).

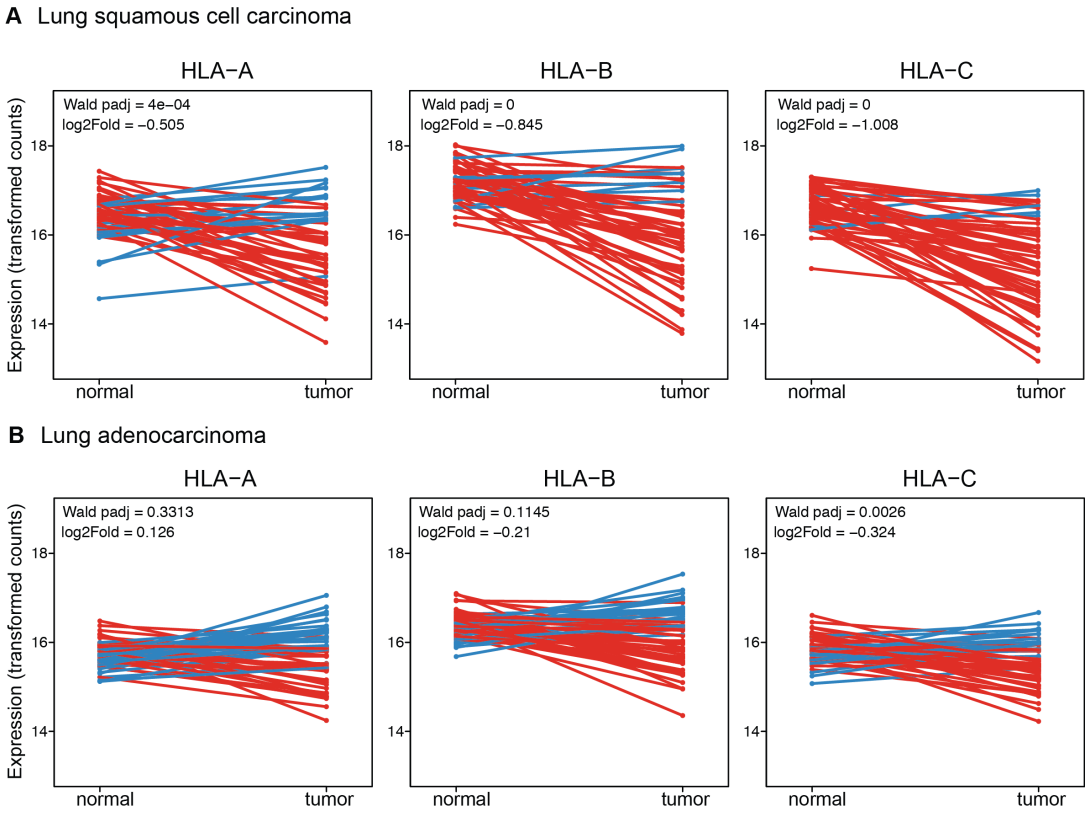

Figure 4-12: Changes in HLA expression between normal and tumor samples.
Expression of HLA class I genes in paired normal and tumor samples from TCGA lung squamous cell carcinoma (A) and lung adenocarcinoma (B) cohorts. Decreased expression from normal to tumor sample is indicated in red, increased expression is indicated in blue. The adjusted Wald p-value and log2 fold change from the differential gene expression analysis is displayed.

Taken together, these analyses support the hypothesis that a high number of clonal neoantigens in homogeneous lung adenocarcinoma may more effectively elicit an immune response, conferring improved patient prognosis. However, in lung squamous cell carcinomas, down-regulation of the HLA alleles may provide one way for the tumor to escape immune detection, a possibility that will be further explored in the following chapter.

4.5.2 Immune microenvironment of high clonal neoantigen tumors

To more completely investigate the impact of clonal neoantigen load on the immune microenvironment, TCGA RNAseq data was further used in a differential gene expression analysis. Homogeneous (≤1% neoantigen ITH) lung adenocarcinoma tumors with a high clonal neoantigen burden (≥ upper quartile) were compared to those lung adenocarcinoma tumors that either had a heterogeneous (>1% neoantigen ITH) neoantigen landscape or low clonal neoantigen burden (< upper quartile). This comparison revealed eight genes that were significantly differentially expressed between the groups. The most significantly differentially expressed genes, programmed cell death ligand-1 (PD-L1) and the pro-inflammatory cytokine, IL-6, were found to be up-regulated among the homogeneous and high clonal neoantigen tumors.

Specifically comparing tumors in the upper quartile of clonal neoantigen burden with tumors in the lower quartile, led to the identification of an additional 25 differentially expressed genes (Figure 4-13). In this analysis, a cluster of genes associated with antigen presentation (TAP-1, TAP-2, STAT-1) and those associated with T-cell infiltration (CD8A, CD8B), T-cell migration (CXCL-10, CXCL-9) , and T-cell function (IFN-γ, granzymes B, H and A) were significantly up-regulated among the high clonal neoantigen tumors.

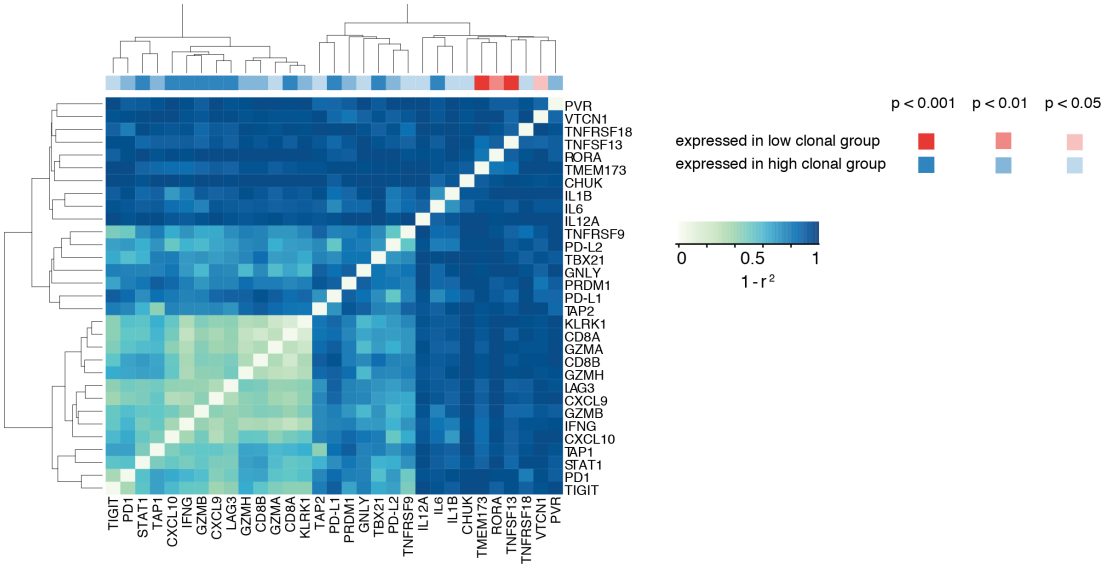

Figure 4-13: Differential expression of immune-related genes.
Significantly differentially expressed genes in high clonal neoantigen tumors (blue) as compared to low clonal neoantigen tumors (red) are shown, clustered by their level of co-expression using a metric $1\text{-}r^2$. The most highly correlated genes are colored more lightly.

Furthermore, negative regulators of T-cell function, PD-1 and lymphocyte activation gene 3 (LAG-3) (Nguyen and Ohashi, 2015), as well as the regulatory ligands PD-L1 and PD-L2 were also found among the set of up-regulated genes. Thus overall, differential gene analysis suggests that, among lung adenocarcinoma, a high clonal neoantigen burden is associated with the presence of activated effector T-cells potentially regulated by the expression of specific immune checkpoint ligands (PD-1, PD-L1/2 and LAG-3).

4.5.3 Characteristics of neoantigen reactive T-cells

In order to further understand the immune microenvironment and confirm results from the differential gene expression analysis, the neoantigen-reactive T-cells which had been identified from the multi-region NSCLC tumors were characterized further using multi-color flow cytometry, allowing for the expression of specific immune checkpoint molecules and effector cytokines to be assessed (Figure 4-14).

Consistent with the findings from TCGA RNAseq, CD8+ T-cells recognizing the $MTFR2^{D326Y}$ mutant peptide highly expressed the negative regulators PD-1 and LAG-3 (Figure 4-14A). Indeed, almost all of the neoantigen-reactive T-cells (97%) expressed high levels of PD-1 compared to only 49% of the tumor-infiltrating CD8+ T-cells which did not recognize the mutant peptide. The subset of CD8+ T-cells that recognized the $MTFR2^{D326Y}$ mutant peptide and expressed PD-1 also expressed high levels of granzyme B (GzmB) (74.8%, Figure 4-14B). Further supporting the immune-signature first identified in TCGA, CD8+ T-cells from patient L012 which recognized $CHTF18^{L769V}$ and $MYADM^{R30W}$ mutant peptides also displayed with high expression of PD-1, observed in 97% and 99.6% of $CHFT18^{L769V}$ and $MYADM^{R30W}$-reactive CD8+ T-cells respectively.

The observed expression of LAG-3 and PD-1 on T-cells reactive to clonal neoantigens supports the immune-signatures identified in TCGA lung adenocarcinoma tumors containing high clonal neoantigen load. These data suggest that immune checkpoint molecules are expressed in response to T-cell activity recognizing clonal neoantigens, providing evidence that targeting such checkpoints in lung adenocarcinoma tumors with high clonal neoantigen burden may be an effective therapy choice.

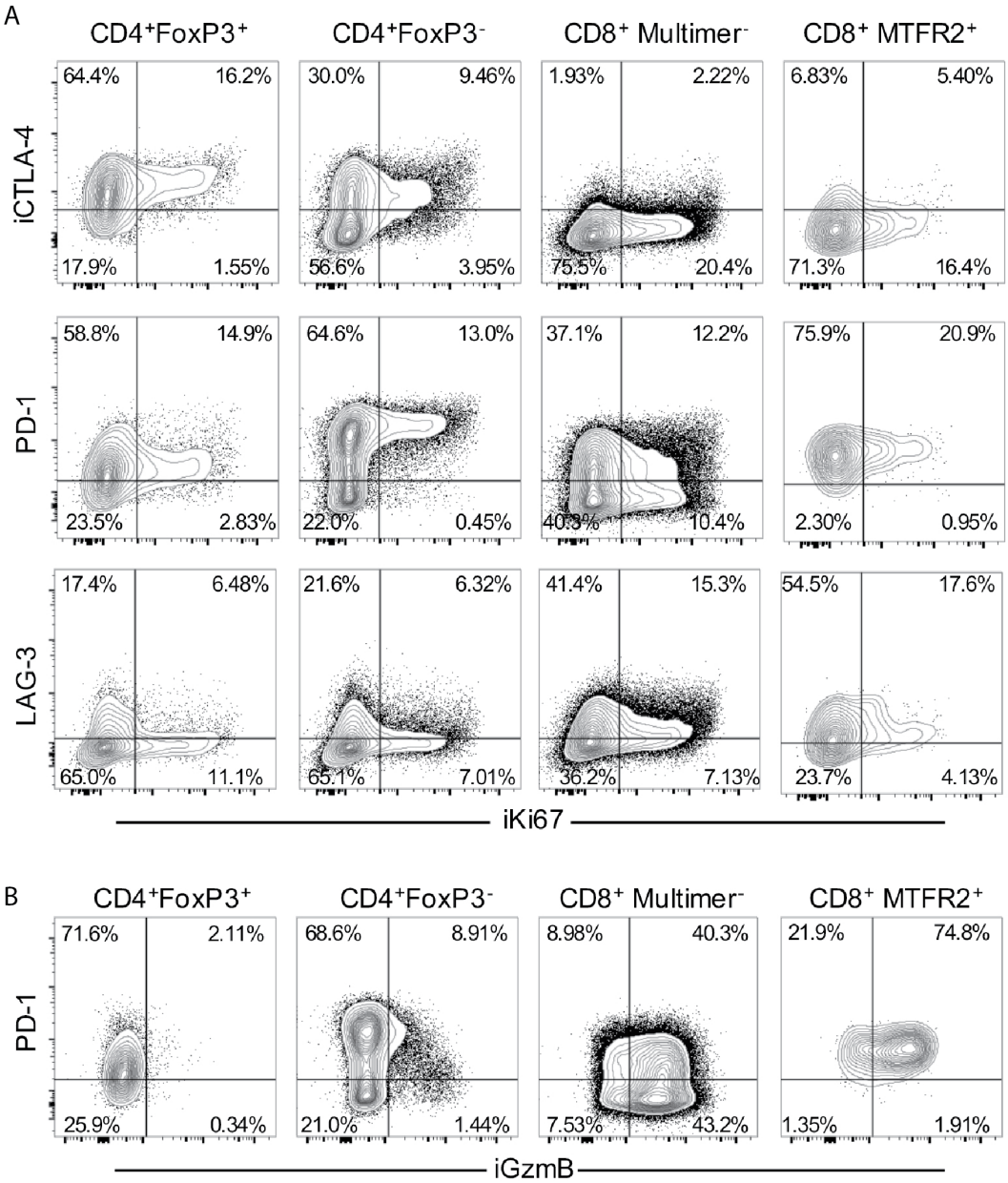

Figure 4-14: Flow cytometry analysis of neoantigen-reactive T-cells.
A) Multi-parametric flow cytometric analysis of TIL subsets isolated from L011 region 3 is displayed. The relative expression of iCTLA-4 (intracellular CTLA-4), surface PD-1, and surface LAG-3 by $CD4^+FoxP3^+$(Tregs), $CD4^+FoxP3^-$ (CD4+ T-cell), CD8+ multimer negative, and CD8+ multimer-reactive (CD8+ MTFR2+) T-cells are displayed. B) The co-expression of PD-1 and iGzmB by tumor-infiltrating T lymphocyte subsets isolated from L011 region 3 is shown.

4.5.4 Neoantigen heterogeneity impacts response to checkpoint blockade

Across a wide variety of cancers, targeting immune checkpoint molecules such as cytotoxic T lymphocyte antigen-4 (CTLA-4), programmed cell death-1 (PD-1), or programmed cell death ligand-1 (PD-L1) through antibody-mediated blockade has

shown great clinical promise (Sharma and Allison, 2015, Rizvi et al., 2015, Topalian et al., 2012a). Furthermore, in both treatment-naïve and checkpoint blockade treated cohorts, the number of tumor neoantigens has been associated with immune activity and overall survival and specific neoantigens have been identified which generate T-cell responses (Brown et al., 2014, Rooney et al., 2015, Castle et al., 2012, Tran et al., 2016, Rizvi et al., 2015, Linnemann et al., 2015).

While the efficacy of anti-PD-1 is linked to predicted neoantigen load of tumors in NSCLC (Rizvi et al., 2015), the impact of ITH upon this relationship is unknown. As earlier results indicated that clonal neoantigens may more effectively lead to neoantigen-reactive T-cell response and that those reactive T-cells expressed elevated levels of PD-1, neoantigen ITH may also be expected to play a role in patient response to anti-PD-1 therapy.

Thus to determine the clinical relevance of neoantigen ITH in the context of immune modulation, a cohort of late-stage NSCLC treated with the antibody targeting PD-1, pembrolizumab, was obtained. This cohort contained exome sequencing data with matched clinical data from a recent study of 34 patients (Rizvi et al., 2015). To investigate the role of neoantigen ITH, the clonal architecture of each tumor was first determined. This was only possible for 31/34 tumors due to sequence quality.

Consistent with the initial publication of this cohort, a high neoantigen burden was related to improved response of the patient while on pembrolizumab (Figure 4-15). Furthermore, clinical response to pembrolizumab also appeared to relate to neoantigen ITH with tumors obtained from patients with non-durable clinical response having significantly higher levels of neoantigen ITH than those tumors obtained from patients exhibiting clinical response ($p = 0.006$), highlighting the potential importance of clonal neoantigens in patient response to immunotherapy (Figure 4-15).

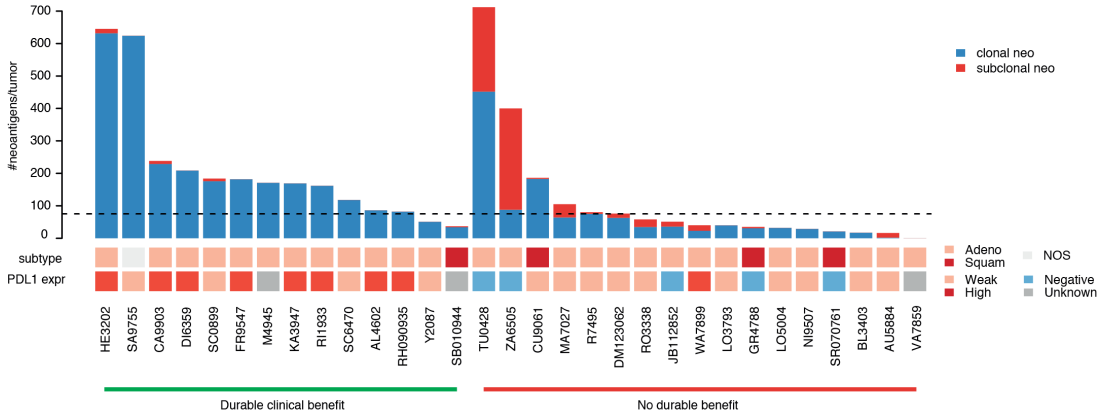

Figure 4-15: Neoantigen load and clinical benefit in anti-PD-1 treated cohort
Neoantigen clonal architecture, clinical response and patient characteristics. Samples are grouped according to response groups, with durable clinical benefit on left and no durable benefit on right (as in (Rizvi et al., 2015)). The barplot depicts clonal neoantigens in blue and subclonal neoantigens in red.

Indeed, nearly every patient (12/13) with a tumor that harbored a high number of clonal neoantigens and few subclonal neoantigens (neoantigen ITH <5%, ≥70, median clonal neoantigens of the cohort) exhibited durable clinical benefit to pembrolizumab. Those patients with tumors that were either highly heterogeneous (>5% neoantigen ITH) or with few clonal neoantigens (<70, median clonal neoantigens of the cohort) frequently demonstrated no durable benefit to pembrolizumab. In fact, there are only two tumors with low neoantigen burden or high neoantigen ITH in the group of responding patients.

Incorporating a measure of neoantigen heterogeneity allowed for a refinement of the initial analysis, which had focused solely on neoantigen burden. Two outlier patients from the original analysis, TU0428 and ZA6505, both had very high neoantigen burden, suggesting that they may respond to pembrolizumab; however, these two patients quickly progressed on the therapy. Analysis of the clonal architecture of these tumors revealed that they were two of the most heterogeneous of the cohort, suggesting a possible explanation for the lack of response on immunotherapy.

Neoantigen ITH also impacted progression free survival in this cohort, with tumors harboring a high clonal neoantigen burden in conjunction with low neoantigen ITH exhibiting an increased progression free survival time. The relationship observed did not depend on choice of ITH threshold, and resulted in a wider separation of the survival curves (lower hazard ratios) than only incorporating total neoantigen burden (Figure 4-16A).

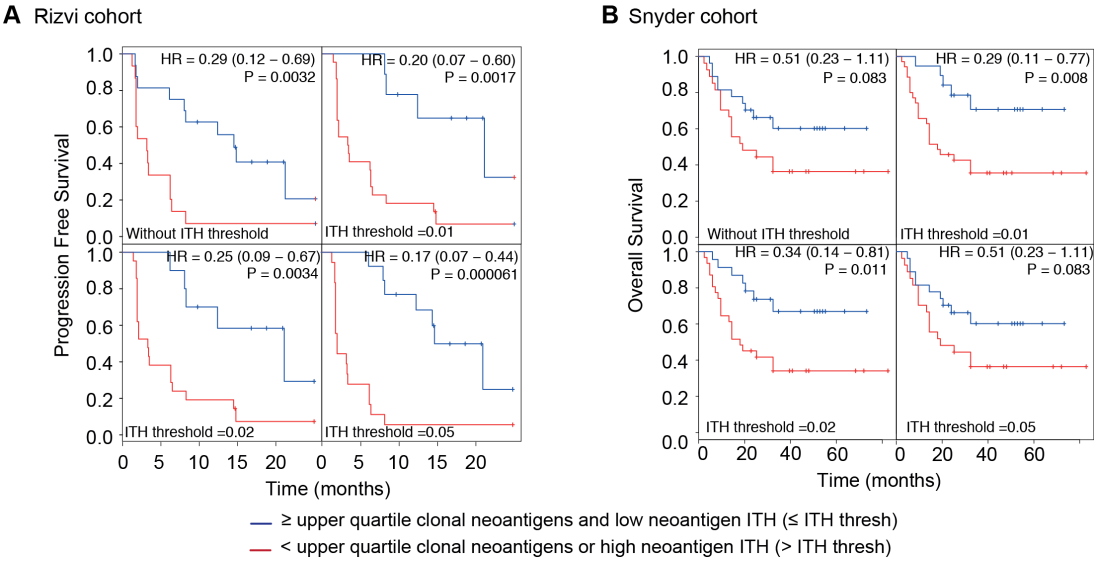

Figure 4-16: Neoantigen clonal architecture and survival following checkpoint blockade therapy. A) Progression free survival in NSCLC (Rizvi et al., 2015) cohort treated with anti-PD-1 either without an ITH threshold, or with an ITH threshold of 0.01, 0.02, or 0.05. B) Overall survival in melanoma (Snyder et al., 2014) cohort treated with anti-CLTA-4 either without an ITH threshold, or with an ITH threshold of 0.01, 0.02, or 0.05.

To validate the impact of neoantigen ITH in a separate checkpoint blockade treated cohort, a cohort of melanoma patients treated with the antibodies targeting CTLA-4, either ipilimumab or tremelimumab, was obtained (Snyder et al., 2014). Of the initial 64 melanoma patients, the clonal architecture of the tumor could be resolved for 57 patients.

Consistent with the results from the anti-PD-1 treated NSCLC cohort, improved overall survival was observed from patients with tumors harboring high number of clonal neoantigens and a low number of subclonal neoantigens, and again, the observed relationship did not depend on a specific ITH threshold, but rather remained robust across multiple thresholds (Figure 4-16B). Intriguingly, in this cohort, considering total neoantigen burden alone did not result in a significant stratification of groups in the survival analysis.

The previously published studies originally investigating these melanoma and NSCLC checkpoint blockade treated cohorts had also sought to identify specific neoantigens engendering an immune response. With the additional information provided by deciphering the tumor's clonal architecture in these cohorts, it was also possible to determine if the potent neoantigens represented clonal or subclonal mutations. Interestingly, and consistent with the results identifying neoantigen reactive T-cells from the TRACERx pilot multi-region sequencing cohort, all

neoantigenic peptides had arisen from clonal non-synonymous mutations (Figure 4-17).

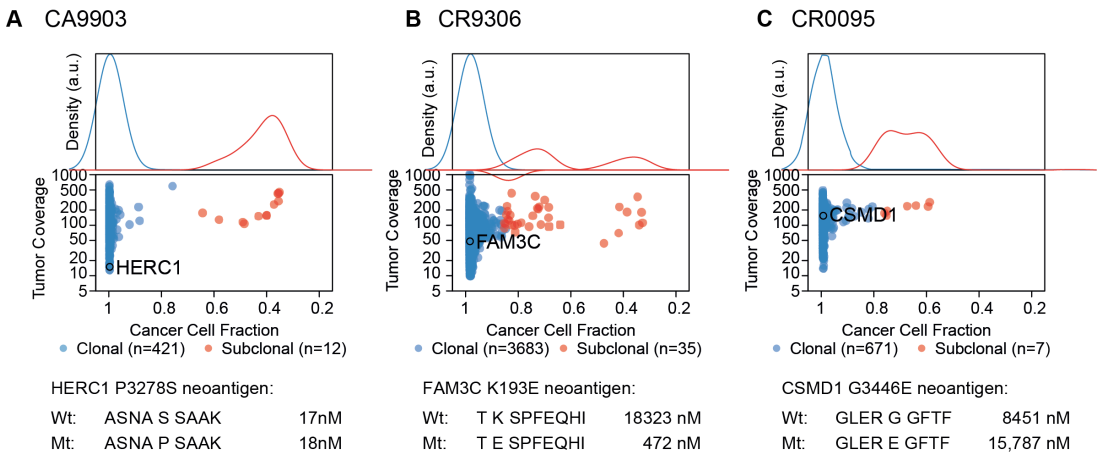

Figure 4-17: Clonality of neoantigens identified in previous publications.
Mutations that had been found to result in peptides which elicited CD8+ T-cell responses in the initial studies (Rizvi et al., 2015, Snyder et al., 2014) were analyzed to determine their clonality. All three mutations had a cancer cell fraction of 1, indicating presence in every cancer cell.

4.6 Conclusions

While previous reports have indicated that overall non-synonymous mutation/neoantigen burden impacts patient response to immune checkpoint blockade (Rizvi et al., 2015, Snyder et al., 2014, Van Allen et al., 2015, Le et al., 2015), this chapter describes the first time the influence of neoantigen ITH was considered. Indeed it appears that clonal neoantigens and subclonal neoantigens do not equally result in immune recognition. Among the checkpoint blockade treated cohorts, all observable T-cell responses were found to react to clonal peptides. While the interpretation of results from these cohorts may be somewhat confounded by the limited ability to identify subclonal mutations from a single tumor sample, results from the multi-region sequencing cohort were consistent. Even though over 250 peptides arising from subclonal mutations were tested, neoantigen reactive T-cells were only found to recognize clonal neoantigens. Further work is needed to establish whether subclonal neoantigens can be identified yielding a T-cell response.

Considering neoantigen clonality also allowed for a more refined analysis of the relationship between patient overall survival and neoantigen burden, significantly associating with longer survival in both treatment-naïve patients and those treated with checkpoint blockade. This suggests that the T-cell responses being generated against clonal neoantigens are also affecting overall tumor immunity. One possible

explanation is that increased heterogeneity of the neoantigen landscape may result in lower antigen dosage. Importantly, high neoantigen ITH could also result in the generation of T-cells which recognize peptides only found in a subset of tumor cells, likely limiting the efficacy of T-cell mediated tumor elimination.

As the mutational processes active over the course of tumor evolution may have distinct timings, early processes may contribute more to the generation of effective neoantigens. In tumor types with an exogenous mutagen, such as melanoma and NSCLC, the abundance of early, clonal non-synonymous mutations may render these diseases vulnerable to vaccination or T-cell therapies targeting multiple clonal neoantigens, to limit potential immune editing, in combination with checkpoint blockade. By targeting clonal mutations, shared by all tumor cells, it may be possible to circumvent the challenges associated with genetic ITH (Yap et al., 2012).

It is also important to consider that mutational processes active late in tumor evolution may be less effective at generating recognized neoantigens. Indeed, a number of melanoma tumors not responding to anti-CTLA4 therapy have been found to harbor huge numbers of subclonal neoantigens in a mutational context associated with prior treatment with alkylating agents, such as DTIC or temozolomide (McGranahan et al., 2016). This suggests that it is important to consider the risk of inducing a multitude of subclonal mutations (Johnson et al., 2014), which may not contribute to a robust anti-tumor immune response against every tumor cell.

While in the treatment-naïve setting, there was an association between neoantigen heterogeneity and overall survival in lung adenocarcinoma, the same relationship was not clearly observed in lung squamous cell carcinoma. Among the lung squamous cell carcinoma cohort, even an abundance of clonal neoantigens did not consistently impart improved overall survival. There are a number of possible explanations for the observed differences between these two subtypes.

Firstly, many possible steps can be taken to improve the neoantigen predictions and clonal neoantigen calling. Currently only the peptide-MHC binding interaction is considered; however as prediction algorithms improve, the likelihood of peptide cleavage and transport may also be included as important factors determining the likelihood of successful antigen presentation. Downstream peptide prediction filtering steps may also be incorporated, such as an expression threshold if RNAseq

data is available. However, as RNAseq only provides a snapshot of gene expression, optimization steps are likely required. Recent work has also suggested that insertion and deletion mutations should be considered as potentially highly immunogenic, as they can result in a greater number of neoantigens, exhibit greater mutant-binding specificity, and are often more dissimilar to self-peptides than the neoantigens arising from SNVs (Turajlic et al., 2017). Finally, some reports suggest trying to capture the likelihood of a given neoantigen being recognized by a TCR by calculating a similarity measure of the peptide to known T-cell antigens from curated immune databases (Luksza et al., 2017).

Another possibility explaining the absence of an association between neoantigen heterogeneity and prognosis is not inferior neoantigen prediction, but rather that presentation of the available clonal neoantigen is dis-functional in many of the lung squamous cell carcinoma tumors. From the available RNAseq data, a marked decrease in HLA expression was observed in lung squamous cell tumors as compared to their paired normal tissue counterparts, independent of clonal neoantigen load. The next chapter will re-visit impaired antigen presentation as means of immune evasion.

As neoantigen predictions improve and more predictions are validated *in vitro* and as the interaction between the immune system and the tumor is more completely understood, it will be possible to more completely understand the complex relationship between neoantigen burden, ITH, and tumor immunogenicity.

Chapter 5 Mechanisms of immune evasion

5.1 Introduction

An evolving tumor and the immune system continuously adapt to each other. As the tumor develops increasing numbers of somatic alterations and dis-regulated genes, it must also find routes to evade detection and elimination by activated immune cells (Hanahan and Weinberg, 2011). This balance between immune detection and evasion is especially clear in the development of resistance to immunotherapy, as the tumor may utilize multiple routes to escape the heightened activity of the immune system. While cancer immunotherapy has resulted in durable antitumor responses in a fraction of patients treated (Hodi et al., 2010, Wolchok et al., 2013, Topalian et al., 2012b), many of the mechanisms through which a tumor develops resistance remain elusive (Rizvi et al., 2015, Snyder et al., 2014, Roh et al., 2017, Chen et al., 2016).

Immunotherapy resistance mechanisms can affect the function of T-cells (for instance via the expression of immune checkpoint molecules or inhibition of T-cell effector activity) (Tumeh et al., 2014, Matsuzaki et al., 2010, Sakuishi et al., 2010, Woo et al., 2002, Powles et al., 2014, Herbst et al., 2014, Spranger et al., 2015, Peng et al., 2016) or they may affect the tumor cell's antigen presentation and response to T-cell activity (Tran et al., 2016, Zaretsky et al., 2016, Zhao et al., 2016, Minn and Wherry, 2016, Benci et al., 2016, Gao et al., 2016, Sucker et al., 2017).

The antigen presentation axis of immunotherapy resistance is of particular interest as it relates to tumor neoantigens, which are frequently the intended targets of immune activation. As described in Chapter 4, one of the most critical steps in the neoantigen-induced generation of an immune response is the binding of the antigenic peptide to the MHC (HLA) molecule, which presents intra-cellular peptides on the cell surface for recognition by T-cell receptors. Thus, disruption of the HLA genes via detrimental mutations (Shukla et al., 2015) or down-regulation of their expression could result in reduced antigen presentation leading to reduced immune recognition. Indeed, many cancer types have been found to down-regulate HLA expression, potentially leading to the evasion of T-cell mediated destruction (Hicklin et al., 1999, Garrido et al., 2017a, Campoli et al.), with reduced HLA expression associating with decreased overall survival and tumor progression (Mehta et al., 2008).

While the down-regulation of HLA expression may be transient and therapeutically counteracted by the release of specific cytokines (Garrido et al., 2017b), there are other mechanisms of HLA loss that cannot be reversed, such as mutation of a β-2 microglobulin (B2M) allele followed by deletion of the other allele (Benitez et al., 1998, D'Urso et al., 1991, Drake et al., 2006).

Another irreversible means of HLA disruption is via loss-of-heterozygosity (LOH) at the HLA locus (Koopman et al., 2000). Much of the diversity in antigen presentation is due to three HLA genes (*HLA-A*, *HLA-B*, and *HLA-C*), located on the homologous paternal and maternal chromosome 6. An individual heterozygous at each of these loci would have six different alleles available for antigen presentation. However, loss of either the maternal or paternal HLA haplotype, resulting in LOH at the locus, would likely reduce the diversity of peptides that are presented and impair the immune system's ability to recognize tumor antigens.

Indeed, Tran and colleagues reported LOH at the HLA locus in a resistant lesion from a patient with metastatic colorectal cancer that had spread to the lung (Tran et al., 2016). The patient was treated with TILs composed of T-cell clones reactive to a KRAS G12D mutation found in the metastatic lesions. After infusion, the metastatic lesions initially regressed, but within a year, a resistant lesion had lost expression of HLA-C*08:02 via LOH at the locus and stopped responding to treatment. The HLA-C*08:02 allele was responsible for presentation of the KRAS G12D neoantigen, allowing for tumor recognition by the T-cells. Thus the loss of this specific HLA allele appeared to enable immune evasion.

Beyond the case study presented by Tran et al., the extent of LOH at the HLA locus in human tumor samples and the impact it may have on the relationship between the tumor and the immune system has not been fully explored. This is due in large part to the polymorphic nature of the HLA locus. The same HLA diversity that allows for a wide range of antigens to be presented to the immune system also hinders the alignment of sequencing reads, as an individual's HLA genes differ too greatly from the human reference genome. Thus copy number at the locus cannot be inferred using standard approaches.

In order to further understand the impact HLA LOH may have on tumor immunity, neoantigen presentation, and subsequent tumor evolution, this chapter will describe a tool developed to estimate haplotype specific copy number at the HLA locus, LOHHLA (Loss Of Heterozygosity in Human Leukocyte Antigen).

The work presented in this chapter was published as a joint-first author paper, (McGranahan et al., 2017). Tool development and bioinformatics analyses was performed in conjunction with Nicholas McGranahan. PCR-based validation experiments to validate the tool described were performed by Andrew Rowan.

5.2 HLA Mutations in TRACERx

HLA mutations have been described in many cancer types (Shukla et al., 2015). Due to the role the HLA molecules play in binding potential neoantigens and presenting them to the immune system, HLA mutations have the potential to disrupt neoantigen recognition and have been associated with immune escape and increased cytolytic activity (Lawrence et al., 2014, Shukla et al., 2015). However, HLA mutations are not common events in tumor evolution, only found to occur at a maximum frequency of ~10% in the head and neck squamous cell carcinoma cancer type (Shukla et al., 2015).

In a cohort of 100 NSCLC TRACERx patients, only three non-synonymous HLA mutations were detected using the state-of-the-art tool, Polysolver (Shukla et al., 2015). One additional lung adenocarcinoma tumor also harbored a mutation in the MHC class I stabilizing molecule, B2M; however recent studies have shown tumors with immunotherapy resistant metastatic lesions containing B2M mutations also have LOH at the locus, suggesting that one function B2M allele may still allow for antigen presentation (Li et al., 2016). No other mutations that were predicted to disrupt antigen presentation were identified in the TRACERx cohort. Similarly, a TCGA pan-cancer study that included 174 lung squamous cell and 223 lung adenocarcinoma patients only identified HLA mutations in 8% and 5% of tumors, respectively (Shukla et al., 2015). These findings indicate that while HLA mutations have the capacity to disrupt neoantigen presentation, they are infrequent events, particularly in a majority early-stage NSCLC cohort.

5.3 LOH at the HLA locus

Frequently during tumor evolution, one copy of a gene or the surrounding chromosomal region is subject to loss, resulting in only one parental copy remaining at the locus. The first step to infer allele specific copy number, allowing for the identification of regions where one allele is present at a copy number of 0, is to identify single nucleotide polymorphisms (SNPs) in the tumor and matched normal and determine their relative coverage and variant allele frequency (Van Loo et al.,

2010, Shen and Seshan, 2016, Carter et al., 2012). However, in regions of the genome where there are no identifiable heterozygous SNPs, determining if a parental copy has been lost is not possible.

5.3.1 Difficulty of the HLA locus

Very few sequencing reads successfully align to the HLA region of the genome due to its polymorphic nature, rendering copy number inference at this locus is unfeasible, as the coverage is too low. With low coverage of the locus, SNPs that allow for the identification of the maternal and paternal allele cannot be identified.

Indeed, while the TRACERx pipeline uses a highly sensitive SNP caller (Rimmer et al., 2014), fewer than 1 heterozygous SNP from the HLA locus was identified on average in the first cohort of 100 TRACERx patients (Jamal-Hanjani et al., 2017). This suggests that in order to accurately infer allele specific copy number at the HLA locus, it is impractical to rely on established copy number calling approaches.

5.3.2 Using patient-specific HLA information

Instead of relying on the human reference genome, LOHHLA uses information gathered from patient-specific HLA typing, which can either be performed using computational approaches (Szolek et al., 2014, Shukla et al., 2015, Warren et al., 2012, Liu et al., 2013, Bai et al., 2014, Xie et al., 2017) or via serotyping. HLA typing is performed as part of the neoantigen prediction pipeline outlined in Chapter 4 using multiple algorithmic approaches. Sequencing reads are then aligned to the inferred HLA types, rather than the human reference HLA, resulting in coverage of the locus whereas previously reads had failed to align (Figure 5-1).

As the sequence of the inferred HLA alleles is known, it is possible to perform sequence alignment of the alleles found on homologous chromosomes to determine where they differ, effectively identifying the heterozygous SNP positions. By mapping sequencing reads to the individuals' germline alleles directly and using the known mismatch positions between the two HLA alleles as heterozygous SNPs, LOHHLA circumvents the issues generated due to poor coverage at the HLA locus, allowing for the determination of HLA haplotype specific copy number (Figure 5-2).

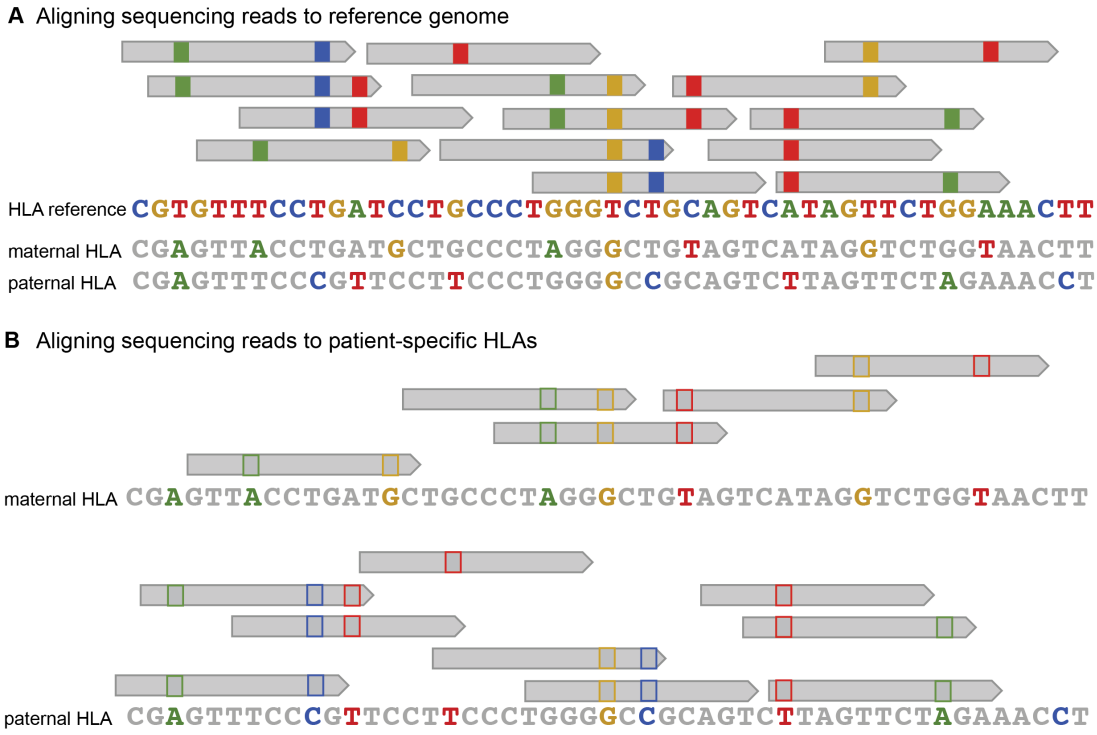

Figure 5-1: Schematic of sequencing read alignment to HLA locus.
A) Aligning sequencing reads to the human reference HLA results in individual reads containing many mismatches, indicated in color along the grey read. The HLA locus is highly polymorphic, so the patient's HLA alleles do not match the reference. B) Instead, using the patient-specific HLA alleles as reference allows sequencing reads to map without mismatches. Positions that would have originally been called a mismatch appear as a box along the sequencing read.

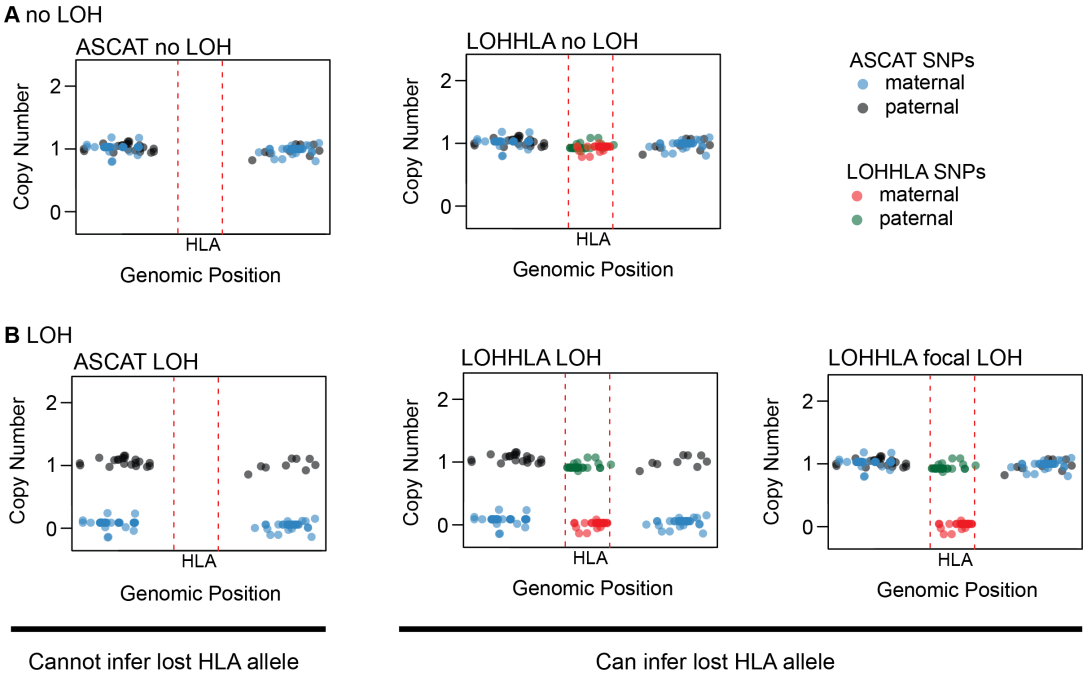

Figure 5-2: Information gained by using known HLA alleles as reference.
(legend on following page)

Schematic of SNPs that are identified using the standard allele-specific copy number tool, ASCAT, and LOHHLA at the HLA loci and surrounding, less polymorphic, regions. A) In instances without any LOH, maternal and paternal SNPs are positioned at the same copy number, with no separation between them. ASCAT does not have any identifiable SNPs at the HLA locus, whereas LOHHLA has used the mismatch positions between the two known patient HLA types B) In cases with evidence for LOH, there is separation between the maternal and paternal SNPs reflecting the allele that was lost. The lack of identifiable heterozygous SNPs at the HLA locus makes it impossible for ASCAT to infer LOH at the HLA locus, whereas LOHHLA can infer LOH at the HLA locus and determine which HLA allele was subjected to loss. LOHHLA can also identify focal HLA LOH, where surrounding regions of the genome do not exhibit any evidence of LOH.

5.3.3 LOHHLA method

To infer HLA-allele specific copy number, LOHHLA performs five steps (Figure 5-3):

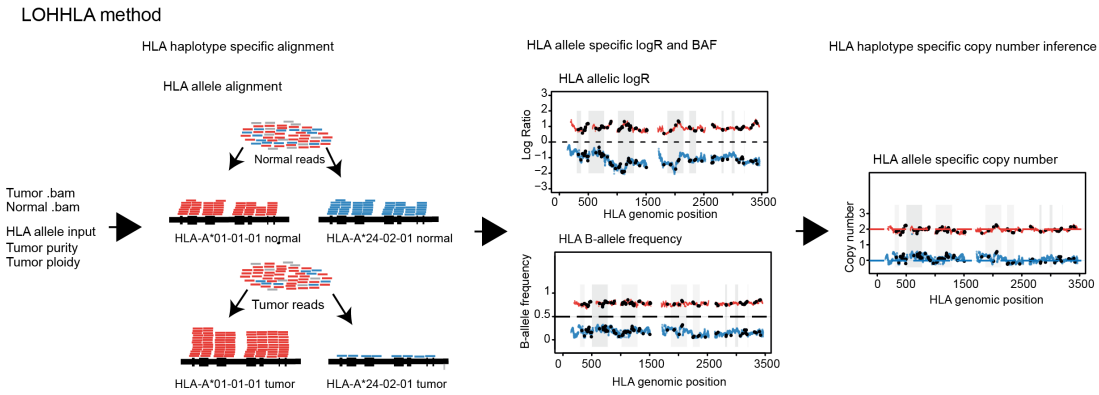

Figure 5-3: Overview of the LOHHLA method.
LOHHLA uses normal and tumor sequencing BAM files, along with patient HLA type information and tumor purity and ploidy to identify HLA LOH in tumor samples. After mapping tumor and normal sequencing reads to the patient-specific HLA alleles, LOHHLA uses the relative coverage at mismatch SNP positions to calculate the log-ratio and B-allele frequency (BAF). Finally LOHHLA incorporates tumor purity and ploidy to infer the HLA allele specific copy number.

5.3.3.1 Step 1: Extract HLA reads

Tumor and germline reads that map to the HLA region of the genome as well as chromosome 6 contigs (chr6_cox_hap2, chr6_dbb_hap3, chr6_mann_hap4, chr6_mcf_hap5, chr6_qbl_hap6, chr6_ssto_hap7) are extracted using samtools view. An additional (optional) step extracts any read that perfectly matches a list of unique k-mer (default: 38mer) sequences generated from the full reference HLA FASTA file. Any unpaired mates from this step are removed and the output is converted to FASTQ format.

5.3.3.2 Step 2: Create HLA allele-specific BAM files

The entries corresponding to each of the patient's heterozygous HLA alleles are extracted from the input HLA FASTA file to generate a patient-specific reference FASTA. The FASTQ files generated in the previous step are used to generate HLA

specific BAM files, using mapping parameters that allow for reads to map to multiple HLA alleles, using similar mapping parameters to those previously published (Shukla et al., 2015). Post-alignment filtering is performed such that reads whose mates mapped to a different allele were discarded, as well as any reads that contained more than one insertion, deletion, or mismatch event compared to the reference HLA allele. For each filtered tumor/germline HLA allele-specific BAM file, coverage is calculated using samtools mpileup.

5.3.3.3 Step 3: Determine coverage at mismatch positions between homologous HLA alleles

For each HLA gene under consideration, a local pairwise alignment is performed between the two homologous HLA alleles using the R biostrings package to determine the polymorphic sites. The HLA-specific coverage calculated in Step 2 is used to determine differences in relative (tumor/normal) coverage at each of the mismatch positions. Because some mismatch positions fall within a sequencing read-length from one another, to avoid over-counting reads that spanned more than one mismatch position, an additional coverage file containing the coverage at every mismatch position, counting each read only once, is also generated.

5.3.3.4 Step 4: Obtain HLA specific logR and BAF

Tumor coverage relative to germline (logR) and b-allele frequencies (BAF) are inferred at each HLA locus, making use of identified polymorphic sites. LogR's across each HLA gene are obtained by binning the coverage across both homologous alleles at 150 base pair intervals, for both tumor and normal. For each bin, the tumor/normal coverage ratio was multiplied by the multiplication factor, M, corresponding to number of unique mapped reads in the germline, divided by the number of unique mapped reads in the tumor region. The BAF is calculated at each polymorphic site, and simply reflects the coverage of HLA allele 1 divided by the coverage of HLA allele 1 + coverage of HLA allele 2.

5.3.3.5 Step 5: Determine HLA haplotype specific copy number

Finally, HLA allele specific copy number is determined for each HLA gene, accounting for tumor purity and ploidy (obtained from a copy number caller e.g. ASCAT (Van Loo et al., 2010) or FACETs (Shen and Seshan, 2016)). At each polymorphic site, an estimate of the major and minor allele copy number is obtained using the following equations, with the logR value from the corresponding bin in

which the polymorphic site was found to reside utilized and the BAF of the polymorphic site.

$$Allele\ 1 = \frac{\rho - 1 + (BAF \times 2) \times logR \times (2(1-\rho) + \rho \times \varphi)}{\rho}$$

$$Allele\ 2 = \frac{\rho - 1 - 2(BAF - 1) \times logR \times (2(1-\rho) + \rho \times \varphi)}{\rho}$$

where ρ = tumor purity and ψ= tumor ploidy, which are input at the start. The logR value from the corresponding bin in which the polymorphic site was found to reside is used as well as the BAF of the polymorphic site.

For each bin, the median Allele 1 and Allele 2 copy number was then determined. To estimate copy number of Allele 1, the median value across bins was calculated. Likewise, to estimate the copy number of Allele 2, the median value across bins was calculated.

A copy number <0.5, was classified as subject to loss. In addition, to avoid over-calling LOH, a p-value was calculated relating to allelic imbalance (AI) for each HLA gene. This p-value corresponded to the difference in logR values at mismatch sites between the two HLA homologues, adjusted to count each sequencing read if it spanned more than one mismatch site. AI was determined if $p < 0.01$ using the paired Student's t-Test between the two distributions.

5.3.4 Validation of LOHHLA

To ensure confidence in the LOHHLA output, a number of validation steps were performed. Firstly, LOHHLA calls for each sample were compared to the corresponding calls from the surrounding region as generated by a standard allele specific copy number tool, ASCAT. This did not provide a way to test if LOHHLA was calling the correct HLA allele as subject to LOH, and there is a possibility of LOH only affecting the HLA locus; however, excluding highly focal events, ASCAT would likely identify a nearby chromosome segment exhibiting LOH. Secondly, LOHHLA results were compared to those generated from an independent approach, not reliant on exome sequencing data. Finally, LOHHLA output was explored using incorrect HLA calls to quantify potential sources of false positives. Each of these validation steps is explained in greater detail below.

5.3.5 Comparison to ASCAT

As there is currently no other method to infer allele specific copy number at the HLA locus, LOHHLA was first tested against a standard allele specific copy number tool, ASCAT (Van Loo et al., 2010). In order to perform these comparisons, the assumption was made that the segments adjacent to the HLA locus would show the same copy number profile as the HLA locus itself. This assumption is valid as long as the resulting LOH was not due to a highly focal event (

Figure 5-2). Such a comparison is not able to perfectly determine whether the copy number estimation determined by LOHHLA is correct, nor is it capable of validating LOHHLA's inference as to which HLA haplotype is subject to loss, as ASCAT is not designed to infer HLA specific events. However, in conjunction with other validation methods, it serves as a good assessment against the current state-of-the-art copy number calling tools.

Thus ASCAT was independently used to estimate the frequency of AI and LOH in the genomic regions surrounding the HLA locus in 288 TRACERx NSCLC exomes from 96 patients (Jamal-Hanjani et al., 2017). These were compared to LOHHLA copy number estimation in order to determine whether ASCAT and LOHHLA exhibited concordant copy number profiles.

The minor copy number obtained from LOHHLA and ASCAT exhibited a highly significant relationship (Figure 5-4A, $p < 0.001$, rho = 0.70), providing confidence in the copy number estimates calculated by LOHHLA. Furthermore, on an individual tumor region basis, both AI estimates and LOH estimates were largely concordant (Figure 5-4B-C). Out of 288 tumor regions considered, AI estimates agreed between LOHHLA and ASCAT in 246 cases. Using LOHHLA, evidence of AI in thirty-four additional tumor regions was uncovered, while only 8 tumor regions exhibited evidence of AI using ASCAT and not LOHHLA. Similarly, there were concordant LOH calls in 258/288 cases, where twenty-one tumor regions with LOH identified using LOHHLA alone, and 9 tumors regions which only showed signs of LOH using ASCAT. In many cases, LOHHLA has provided finer resolution for identifying true AI or LOH events not spanning the entire HLA locus. For instance, as ASCAT cannot directly infer haplotype specific copy number at the HLA locus, a highly focal event would be missed by ASCAT by relying on the adjacent 5' or 3' segment.

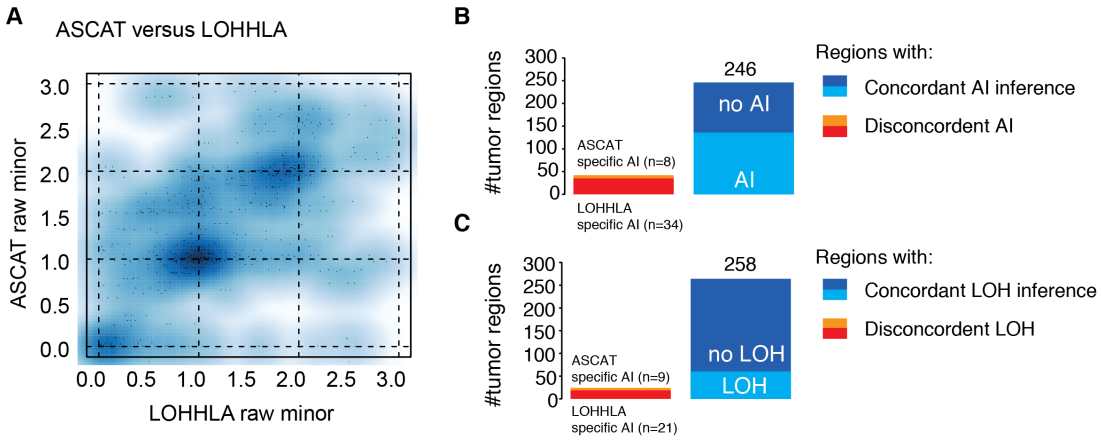

Figure 5-4: Comparison of LOHHLA and ASCAT copy number profiles.
A comparison between the HLA copy number profile generated by LOHHLA and the copy number profile from the regions surrounding HLA by ASCAT was made. A) The minor copy number estimated by both LOHHLA and ASCAT were highly concordant ($p<0.001$, rho=0.70). (B-C) Summary of concordant and discordant tumor regions in terms of allelic imbalance (B) and LOH (C).

5.3.6 PCR-based fragment analysis

To validate LOHHLA without using a method also reliant on exome sequencing data, PCR based fragment analysis was performed (Figure 5-5A-B). Highly polymorphic markers located in close proximity to the HLA locus were analyzed in 82 tumors regions from 27 tumors. The fragment length and area under the curve of each allele could be determined, and when two separate alleles were identified for a particular marker, the fragments could be analyzed for AI using the formula $(A_{tumor}/B_{tumor})/(A_{normal}/B_{normal})$. The output of this formula was defined as the normalized allelic ratio.

For comparison to LOHHLA, tumor regions analyzed were either predicted to have all loci (HLA-A, HLA-B and HLA-C) subject to LOH, or no loci affected. Reassuringly, there were highly significant differences in the normalized allelic ratio between tumors classified as exhibiting either LOH, AI without LOH, or no AI or LOH ($p = 1.07e-19$ [LOH versus no imbalance], $p= 4.57e-05$ [LOH versus AI],). These results provide independent confirmation that LOHHLA is able to accurately classify LOH and AI. Furthermore, the separation in normalized allelic ratios between tumor regions only exhibiting signs of HLA AI and those also harboring HLA LOH events was clearer using LOHHLA than the standard copy number tool ASCAT (Van Loo et al., 2010) (Figure 5-5C-D).

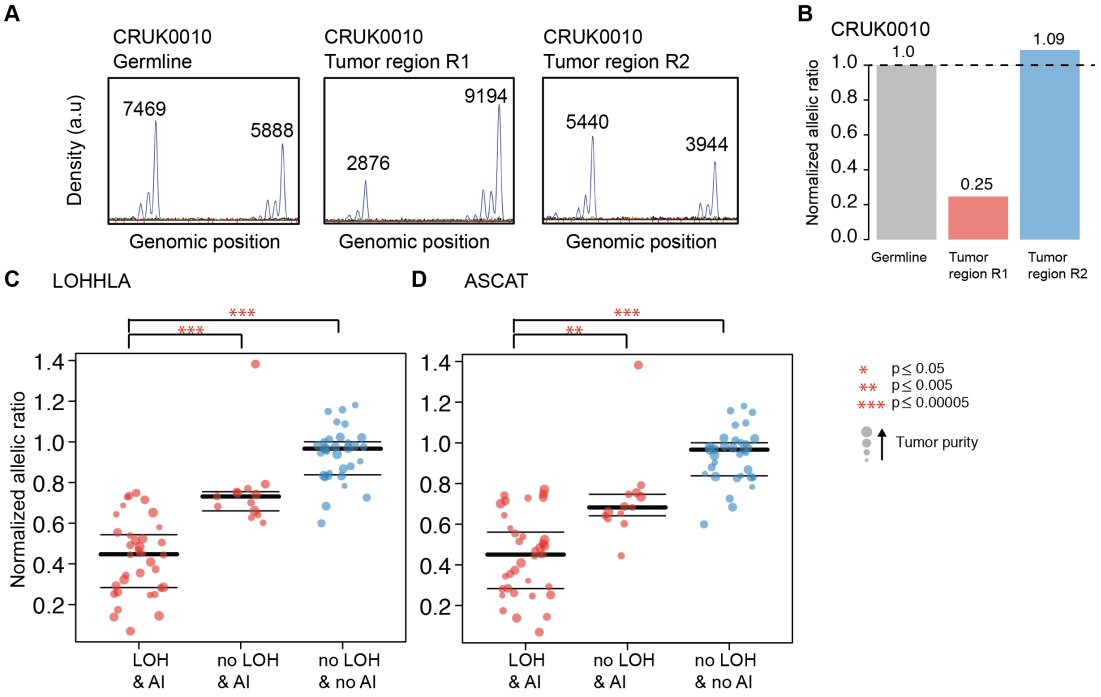

Figure 5-5: PCR-based fragment analysis validation of LOHHLA.
A) Area under the curve of each allele (as determined through the use of polymorphic markers near the HLA locus) for germline and tumor regions R1 and R2 in CRUK0010, given in arbitrary units (a.u). While the germline alleles had approximately equal expression, tumor region R1 showed evidence of decreased expression for one HLA allele. B) Normalized allelic ratio determined using the formula $(A_{tumor}/B_{tumor})/(A_{normal}/B_{normal})$. Region R1 shows clear evidence of allelic imbalance and likely LOH, while region R2 appears similar to germline. (C-D) Normalized allelic ratio for tumor regions showing either LOH and allelic imbalance; no LOH but allelic imbalance; or no LOH or allelic imbalance classified by LOHHLA (C) and ASCAT (D).

5.3.7 Incorrect HLA alleles

Finally, the impact of using the incorrect HLA allele as input to LOHHLA was determined to ensure that false positive results were not generated due to inaccurate starting data, as there are many competing tools designed to infer HLA types from sequencing data, and they do not always give concordant results.

When the input to LOHHLA was altered such that an HLA allele was replaced by the one least similar to it, as determined by percent identity following multiple sequence alignment, there were no likely false positives generated (Figure 5-6). Instead the majority of changes occurred because LOHHLA could no longer identify LOH at the locus where it once had, usually due to a decrease in coverage, as sequencing reads do not map well to the incorrect HLA allele. Indeed, in 55% the cases (140/256) LOHHLA was unable to perform allele specific copy number estimation due to insufficient coverage and returned no result. Of cases with sufficient coverage, 97% of the time LOHHLA was unable to identify HLA LOH, and

in 3% of cases (4/256), results were concordant to those inferred using the correct allele. Importantly, similar results were seen, when either both HLA alleles were incorrectly used as input, and when an incorrect allele with higher similarity to the correct allele was used. In these cases, a maximum of 1 false positive was observed, representing 0.4% of cases, providing confidence that LOHHLA will not erroneously call HLA LOH due to incorrect input. Additionally, the tool includes warnings if the coverage of the locus is too low or differs drastically between alleles, so that the user may check the validity of the HLA input being used.

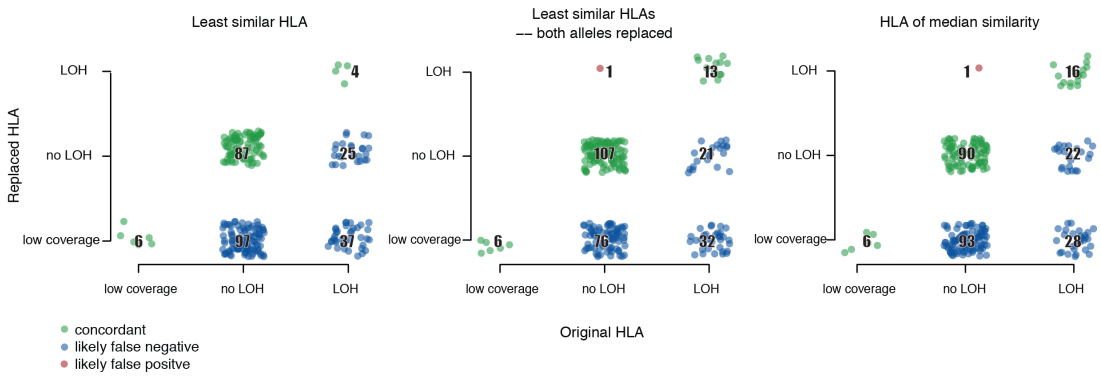

Figure 5-6: Impact of incorrect HLA type input to LOHHLA.
Multiple instances of using the incorrect HLA alleles as input to LOHHLA were considered. Either one HLA allele was replaced by the one least similar to it, both HLA alleles were replaced by the ones least similar to them, or one HLA allele was replaced by the allele of median similarity. Concordant calls are plotted in green, likely false negatives, where the LOH is now missed due to low coverage, are plotted in blue, and likely false positive calls are plotted in red.

Together the validation steps show that is possible to use LOHHLA to correctly infer both AI and LOH in tumor samples. Furthermore, LOHHLA will not give misleading results if provided with poor input data. Additionally, LOHHLA provides greater resolution to detect highly focal HLA LOH events, exhibiting an increased sensitivity and specificity.

5.4 Prevalence and timing of HLA LOH

5.4.1 HLA LOH is a common event in NSCLC

In contrast to the low frequency of HLA mutations observed in the TRACERx cohort, a far higher percentage of patients were found to exhibit HLA AI and HLA LOH (Figure 5-7A). Forty-percent (36/90) of the NSCLC patients had tumors harboring HLA LOH (six patients had tumors with a histology other than lung adenocarcinoma or lung squamous cell carcinoma, and thus not considered further for analysis), where either one maternal or paternal HLA allele was lost, resulting in

HLA homozygosity. Over 60% of NSCLC tumors showed signs of HLA AI, where the maternal and paternal HLA alleles were present at different, non-zero copy numbers.

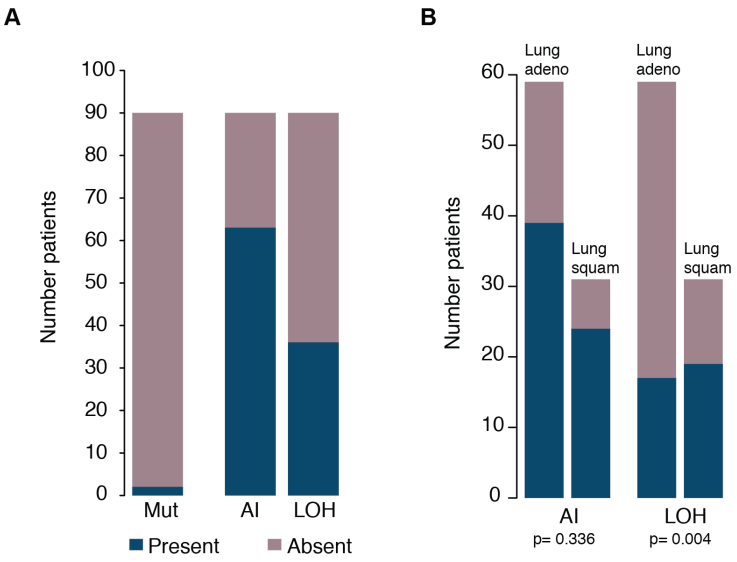

Figure 5-7: Observations of HLA LOH in TRACERx.
A) The number of patients whose tumors harbored an HLA mutation, HLA AI, or HLA LOH are shown. B) The number of lung adenocarcinoma patients exhibiting HLA LOH or HLA AI as compared to lung squamous cell carcinoma patients exhibiting HLA LOH or HLA AI.

5.4.2 Enrichment for HLA LOH among lung squamous cell carcinomas

Just as HLA mutations have found to be more common in lung squamous cell carcinomas (Shukla et al., 2015), HLA LOH was also enriched among lung squamous cell carcinomas as compared to lung adenocarcinomas ($p = 0.004$) (Figure 5-7B). Interestingly, there was no enrichment for AI in lung squamous cell carcinoma as compared to lung adenocarcinoma ($p = 0.336$). The cause of the difference appears to be due to the fact that fewer lung adenocarcinoma with AI also exhibit LOH. As many lung adenocarcinoma tumors also are genome doubled, this suggests that the genome doubling event has occurred before the HLA locus is altered. Thus timing HLA LOH in tumor evolution could help understand what role it plays.

5.4.3 HLA LOH is a late event in tumor evolution

As with other genomic alterations in cancer, it is possible to characterize HLA LOH as an early event in the tumor's evolution, present clonally in every cancer cell, or as a subclonal event, occurring only in a subset of cancer cells. Subclonal HLA LOH indicates that the locus was altered later in tumor evolution and potentially as

a response to a change in the balance between immune recognition and evasion. However, unlike mutations, where the cancer cell fraction can be calculated, allowing for clustering approaches to estimate clonality, copy number alterations are more difficult to time.

Instead the multi-region nature of the TRACERx dataset was utilized to understand when HLA LOH occurred during tumor evolution. Events at each of the HLA A/B/C loci were considered clonal if they were found in every patient region and subclonal if they were found in only a subset of tumor regions. A tumor sample was considered to exhibit clonal HLA LOH if all of the loci subject to loss in the tumor were classified as clonal events.

HLA LOH was a frequently subclonal event in both histological subtypes, with 13/17 lung adenocarcinoma and 9/17 lung squamous cell carcinomas exhibiting loss of an HLA allele in a subset of cancer cells (Figure 5-8). Two lung squamous cell carcinomas exhibiting HLA LOH, but with only a single region available for copy number analysis could not be considered.

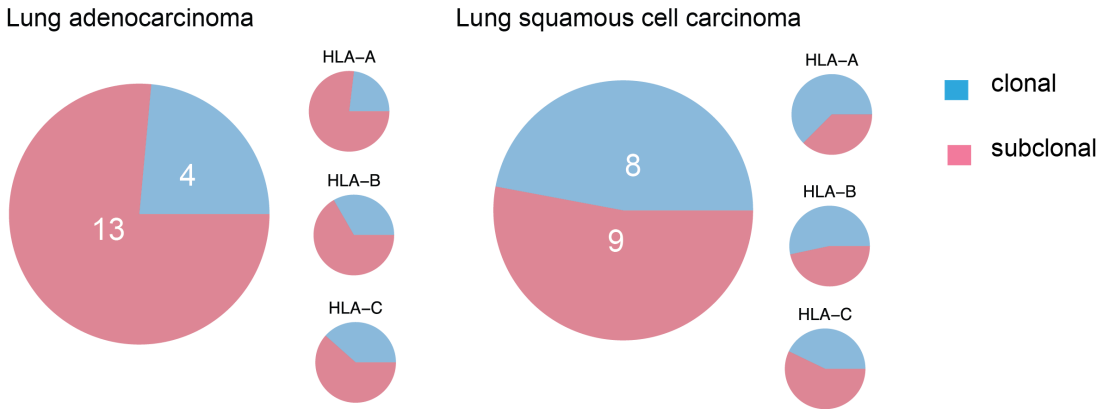

Figure 5-8: Timing of HLA LOH events in NSCLC.
Clonal HLA LOH events are shown in blue and subclonal HLA LOH events are shown in red for lung adenocarcinomas and lung squamous cell carcinomas.

Because this cohort of multi-region NSCLC patients had been previously analyzed (Jamal-Hanjani et al., 2017), phylogenetic trees built from the integrated mutation and copy number information were available. Thus the individual HLA LOH events were mapped to probable subclones from the tumor's evolutionary tree, allowing for a more refined analysis of their timing (Figure 5-9).

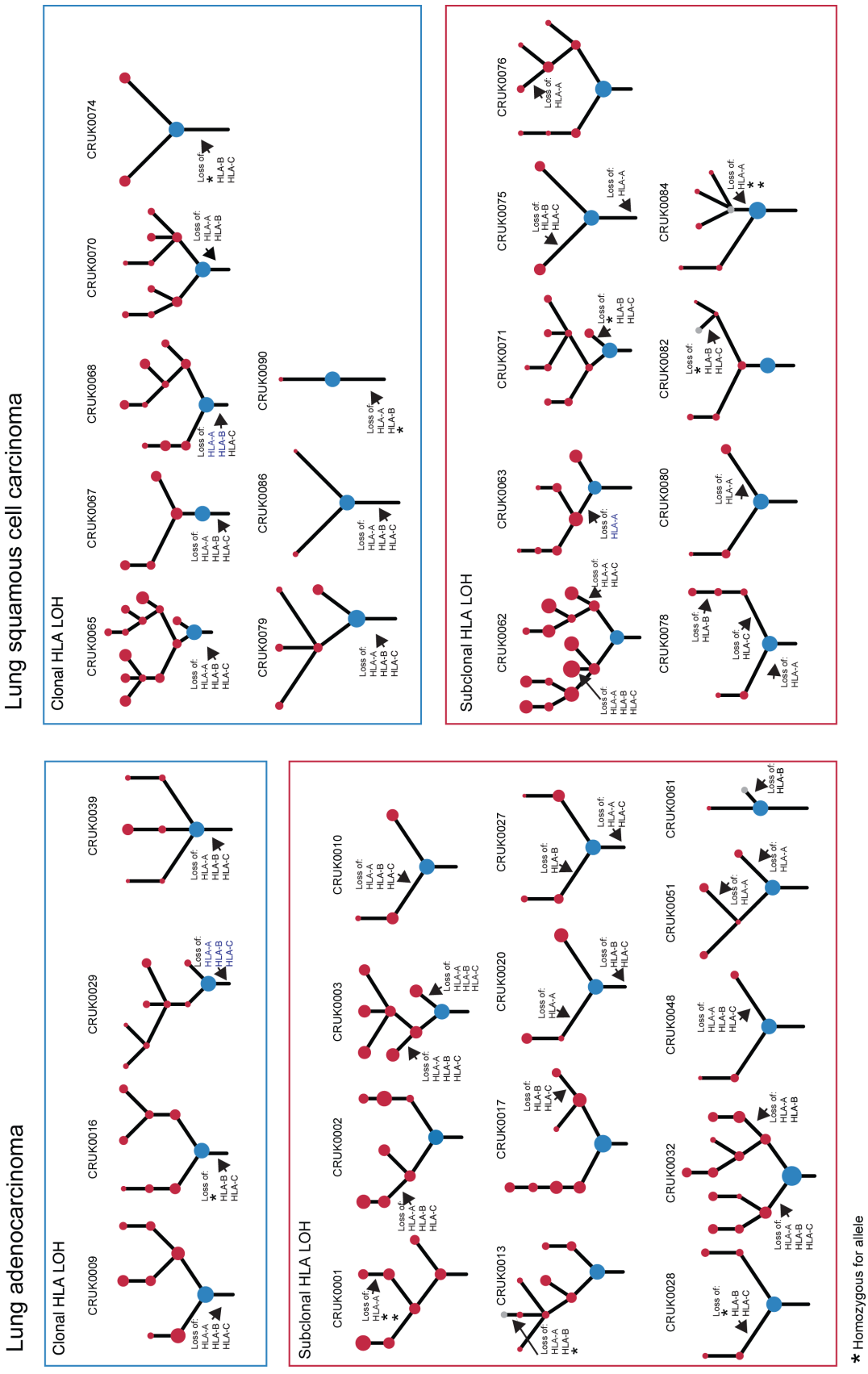

Figure 5-9: Phylogenetic mapping of HLA LOH events.
(legend on following page)

Phylogenetic trees for each lung adenocarcinoma and lung squamous cell carcinoma tumor showing evidence of HLA LOH have been annotated with the most likely timing of the HLA LOH event. Homozygous HLA alleles, where HLA LOH is not possible, are indicated by an asterisk. Clones on the phylogenetic tree (nodes) are indicated as clonal (blue) or subclonal (red). In cases where the HLA LOH event did not map to a possible clone on the phylogenetic tree, an additional grey subclone was included.

To determine placement on the tumor's phylogenetic tree, an algorithm was developed. LOH events that were classified as clonal events (found in every tumor region considered) were simply mapped to the trunk of the phylogenetic tree. To determine the placement of heterogeneous LOH events, the regional copy number of the HLA allele lost was calculated. These values were incorporated with the patient tree structure and subclone cancer cell fractions in a quadratic programming approach to determine where on the phylogenetic tree the LOH event most likely occurred (Data and Methods). All of the subclonal events that were mapped were further inspected. Events sometimes did not fit the phylogenetic tree generated from mutation data or some had large error values. This indicated the presence of an additional subclone or multiple independent HLA LOH events. As such events failed the algorithmic approach, they were manually adjusted. Patients with HLA LOH events that did not fit the current phylogenetic tree had additional nodes included to contain the HLA LOH event.

5.4.4 Enrichment of HLA LOH in metastatic samples

The TRACERx cohort examined consisted of mostly early stage, primary tumors. To gain greater understanding on the timing of HLA LOH in NSCLC tumor evolution, a second cohort was considered, consisting of 37 NSCLC patients who had primary tumors with matched brain metastases (Brastianos et al., 2015). LOH at the HLA locus was identified in 17/37 (46%) of the patients' tumors, which was similar to the prevalence observed in the early stage disease cohort (Figure 5-10A).

To time the LOH event, a similar classification was used as previously described. Patients with HLA LOH identified across the same HLA loci in both the primary tumor and every brain metastasis were classified as having clonal HLA LOH. Patients with either different HLA loci subject to LOH or those with HLA LOH identified in only a subset of the samples available were classified as having subclonal HLA LOH. Again, the LOH event was found to occur predominantly later in tumor evolution, occurring subclonally in 11/17 (65%) cases (Figure 5-10A).

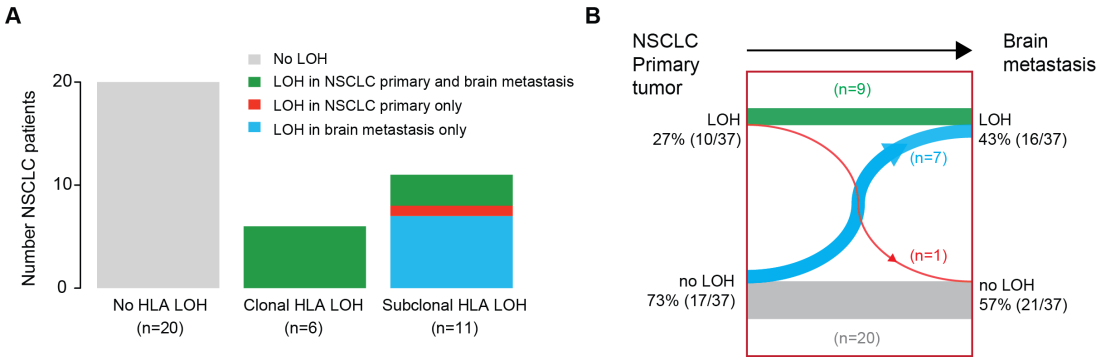

Figure 5-10: HLA LOH occurrence in metastatic samples.
(A) Number of NSCLC patients from Brastianos et al (Brastianos et al., 2015) exhibiting no HLA LOH (grey), HLA LOH in both the primary tumor and brain metastasis (green), HLA LOH only in the primary tumor (red), or HLA LOH only in the brain metastasis (blue). (B) The number of events that were found in the primary and/or brain metastasis is shown. Clonal HLA LOH events occur in both the primary tumor sample and the brain metastases (green), whereas subclonal HLA LOH events either arise in the brain metastases (blue) or have occurred in a subclone of the primary tumor that does not seed the brain metastasis (red).

Furthermore, the availability of paired primary and metastatic samples from the same patient allowed for a comparison of event frequencies over the course of disease development (Figure 5-10B). Overall, an increase in HLA LOH was observed in the brain metastatic samples as compared to the primary tumor (27% to 43%) and a corresponding decrease was observed in brain metastatic samples exhibiting no HLA LOH (73% to 57%). While seven patients harbored HLA LOH in the metastatic sample alone, there was only one patient with HLA LOH in the primary tumor alone. Thus there was a trend towards enrichment of HLA LOH in brain metastases compared to the matched primary tumor (p=0.08). These results provide support that HLA LOH occurs later in cancer evolution, and indicate that over the course of late stage disease, there may be additional selection for immune evasive mechanisms.

5.5 Positive selection for HLA LOH

5.5.1 Recurrent HLA LOH events

Four patients from the TRACERx cohort had tumors where the HLA LOH event was equally likely to map to multiple branches of the phylogenetic tree (Figure 5-9). The only possibility for reconciling the observed phylogenetic and LOH data was if multiple instances of HLA LOH had occurred during the evolution of the tumor. In all four cases, the same alleles from a patient were subject to loss on separate branches.

Indeed, in these four cases, loss of the same HLA haplotype occurred as separate events on different branches of tumor's phylogenetic trees. Multiple losses could be an indication of parallel evolution with convergence upon the loss of a particular HLA haplotype (Figure 5-11). The fact that the same alleles were subjected to loss multiple times during tumor evolution indicates that loss of these alleles specifically may endowed the subclones harboring the loss events with an evolutionary advantage.

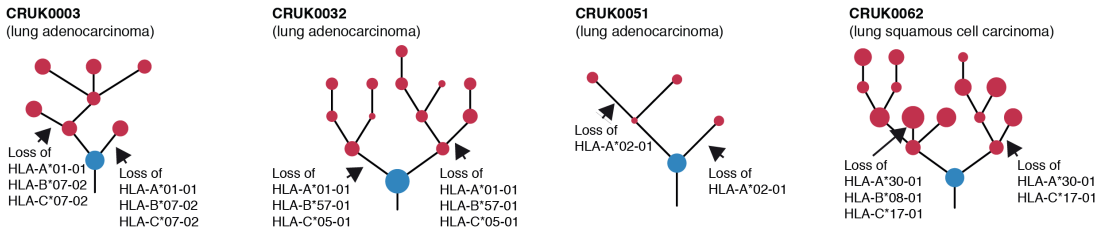

Figure 5-11: Recurrence of HLA LOH events in tumor evolution.
Parallel evolution of HLA LOH, with allele specific HLA loss shown on phylogenetic trees.

5.5.2 Focal LOH more frequent than expected by chance

The previous results show that HLA LOH is a common event in NSCLC, frequently occurring late in cancer evolution. To formally test whether HLA LOH is selected for in tumor evolution, the expected frequencies of both focal (<75% chromosome arm) and arm-level (>= 75% chromosome arm) events were simulated, taking into account the baseline frequency of LOH in every tumor considered.

For tumors harboring focal events, first the proportion of the entire genome with evidence of focal LOH was determined. This value was used as the background level of focal minor allele loss in each tumor, so that tumors were simulated with an LOH rate that reflected what was observed in each specific sample. Using the background value as probability of LOH, each tumor sample was randomly assigned an aberration state (loss or no loss), and the proportion of samples assigned loss was determined. This entire process was repeated 10,000 times to obtain a background distribution, which reflected the likelihood of observing a loss given the respective probabilities of LOH in each sample. Then a p-value was calculated by counting the percentage of simulations showing a higher proportion of loss at the HLA locus than observed.

Supporting the putative role of HLA LOH in tumor evolution, focal LOH events affecting the HLA locus were observed in the TRACERx cohort significantly more

frequently than expected by chance (Figure 5-12). There was a clear peak in focal LOH centered around the HLA locus for both histological subtypes, strongly suggesting the HLA locus is subject to selective pressure during NSCLC evolution.

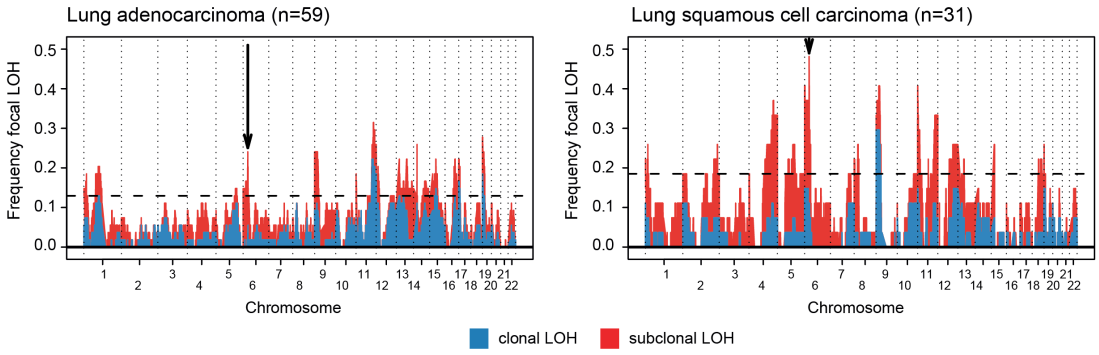

Figure 5-12: Selection for focal LOH in NSCLC.
The frequency of focal LOH events in lung adenocarcinoma and lung squamous cell carcinoma is shown. Focal LOH was defined as <75% of a chromosome arm. An arrow indicates location of HLA locus. A horizontal dashed line depicts significant focal LOH at p=0.05, using simulations. Clonal LOH is shown in blue, with subclonal LOH shown in red.

A similar simulation process was used to investigate whether arm-level events occurred more frequently than expected by chance. However, arm-level LOH events affecting the HLA locus were not any more common than the background simulation would suggest (Figure 5-13). The increased frequency of focal LOH events and not arm-level LOH events further suggests that alteration of the HLA locus specifically is selected for, rather than as the result of large-scale events that happen to affect the LOH locus.

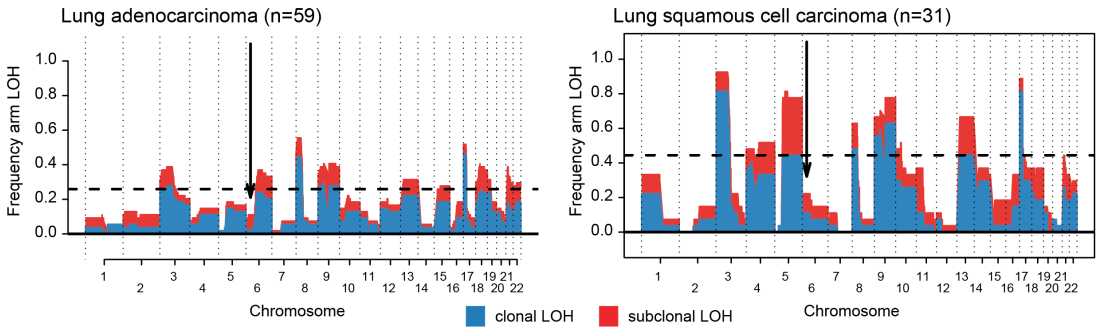

Figure 5-13: Selection for arm-level LOH in NSCLC.
The frequency of arm-level LOH in lung adenocarcinoma and lung squamous cell carcinoma is shown. Arm-level LOH was defined as >=75% of a chromosome arm. An arrow indicates location of HLA locus. A horizontal dashed line depicts significant arm-level LOH at p=0.05, using simulations. Clonal LOH is shown in blue, with subclonal LOH shown in red.

The results from the simulation, in conjunction with the observation of parallel evolution of HLA LOH, indicate that HLA LOH is under strong selection late in cancer evolution.

5.6 Impact of HLA LOH on tumor evolution

As each HLA allele present on a cell is capable of presenting a different subset of antigenic peptides to the immune system, the loss of one of the HLA haplotypes could result in fewer putative neoantigens being presented to the T-cells for recognition. Thus after the HLA LOH event, there may be less ongoing immune surveillance, potentially allowing for subclonal expansions that would have been otherwise subjected to immune restriction. If this hypothesis is true, then the expected result would be an increased mutation/neoantigen burden among tumors harboring an HLA LOH event.

5.6.1 Increased mutation burden in tumors with HLA LOH

To test this hypothesis, first the total number of non-synonymous mutations were compared between tumors samples with and without an LOH event at the HLA locus, without considering the timing of the HLA LOH event, such that all events were considered together. Overall there was a significant increase in the number of non-synonymous neoantigens in tumor samples exhibiting HLA LOH, but this did not remain significant when the subtypes were considered separately (Figure 5-14A) (NSCLC p=0.016; lung adenocarcinoma p=0.07; lung squamous cell carcinoma p = 0.82). Interestingly, when the tumors were divided into mutational burden categories (low defined by the lowest quartile of NSCLC mutation burden), there were only 3/36 NSCLC tumors harboring an HLA LOH event with a low mutation burden. Among the 54 tumors without an HLA LOH event, far more tumors harbored low mutational burden (21/54).

Because the HLA LOH event tends to occur later in tumor evolution, the increase in mutational burden may be limited to subclonal mutations, so next the clonal nature of the mutations was considered. There was a significant increase in the number of subclonal, but not clonal, non-synonymous mutations. The increase in subclonal mutational burden was only observed among the lung adenocarcinoma subtype (Figure 5-14B-C) (NSCLC p=0.008; lung adenocarcinoma p=0.01; lung squamous cell carcinoma p=0.6). This observation indicates that HLA LOH may allow for the accumulation of potentially antigenic subclonal mutations.

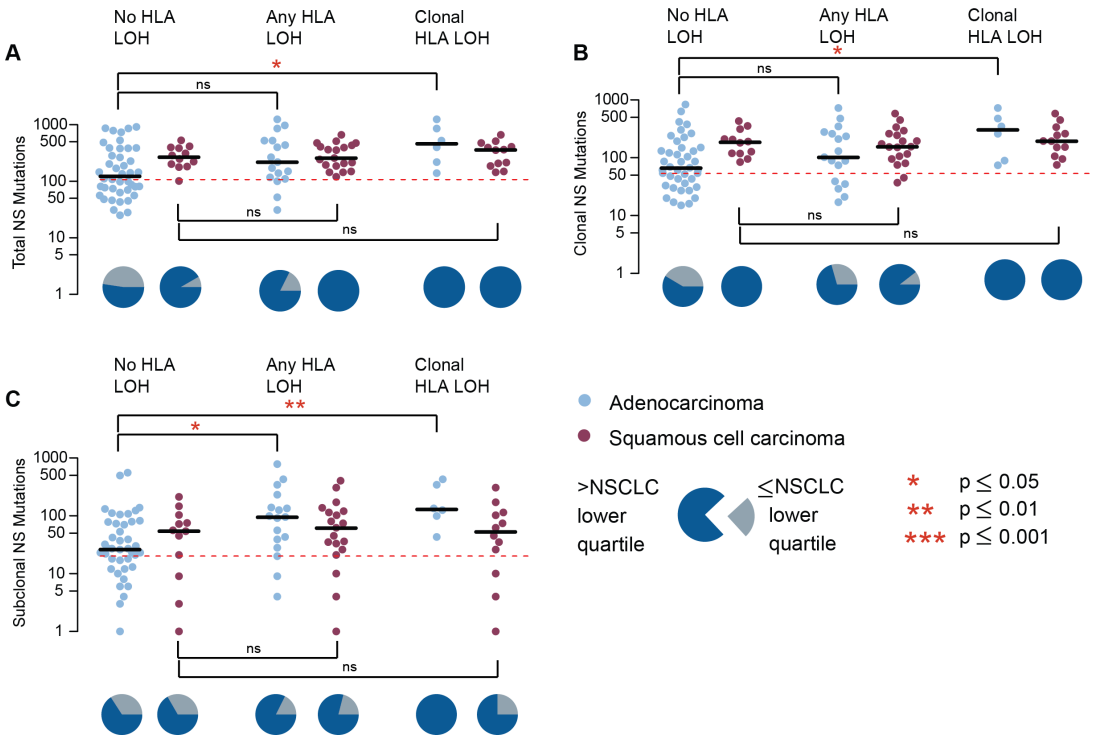

Figure 5-14: Non-synonymous mutation burden in tumors with HLA LOH.
(A) The total number of non-synonymous mutations is plotted across different categories of HLA LOH for lung adenocarcinoma (light blue) and lung squamous cell carcinomas (magenta). Tumors were classified as having: no HLA LOH; any HLA LOH event, without taking into account the timing of the event; or clonal HLA LOH. The lowest quartile of total non-synonymous mutation is indicated by the dashed red line and the proportion of tumors with a total non-synonymous mutational burden greater or less than the lowest quintile is indicated by the pie charts for each HLA LOH classification group. (B) The number of clonal non-synonymous mutations is plotted across different categories of HLA LOH. (C) The number of subclonal non-synonymous mutations is plotted across different categories of HLA LOH. All p-values are calculated using an unpaired wilcoxon test.

When the timing of the HLA LOH event itself was included in the analysis, there was a significant association between early HLA LOH events, events mapped to the trunk of the phylogenetic tree, and an elevated clonal (NSCLC p=0.002; lung adenocarcinoma p=0.01; lung squamous cell carcinoma p=0.29) and subclonal (NSCLC p=0.03; lung adenocarcinoma p=0.004; lung squamous cell carcinoma p=0.89) non-synonymous mutational burden (Figure 5-14B-C). These results show that when the HLA LOH event occurs early in tumor evolution, there is an increase among both clonal and subclonal non-synonymous mutations in lung adenocarcinoma.

5.6.2 Increased mutation burden in tumor regions with HLA LOH

The observed intratumor heterogeneity of the HLA LOH events meant that some patients had tumors where only a subset of regions harbored LOH at the HLA locus. To determine if the increased non-synonymous mutational burden was confined to

tumor regions that harbored the LOH event, HLA LOH events were next considered at the region-level. Consistent with the previous results, tumor regions exhibiting HLA loss had a significant increase in subclonal non-synonymous mutations as compared to tumor regions from patients without HLA LOH (NSCLC p=1.9e-05; lung adenocarcinoma p=0.009; lung squamous cell carcinoma p=0.07) (Figure 5-15).

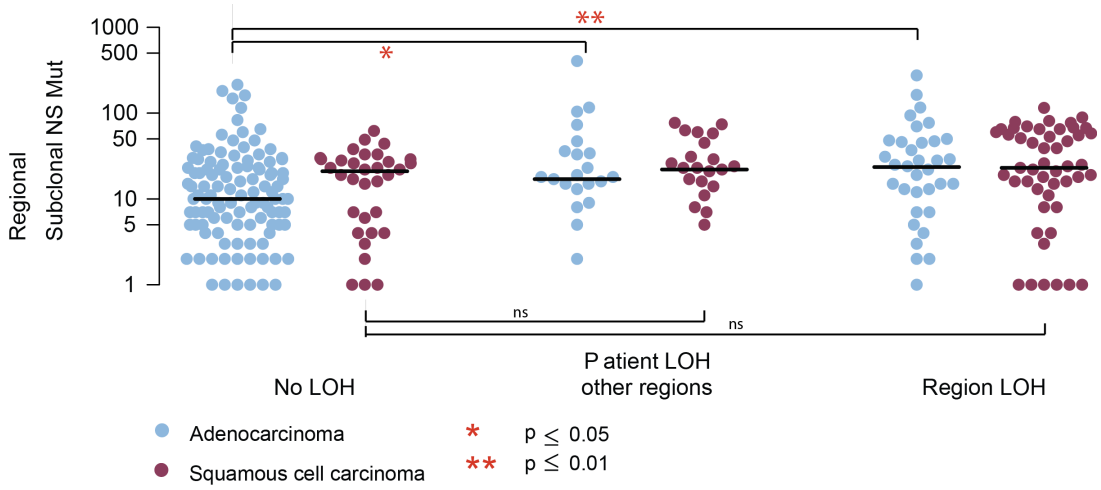

Figure 5-15: Subclonal non-synonymous mutational burden at the region-level.
The number of subclonal non-synonymous mutations is plotted for tumor regions from tumors without any indication of HLA LOH, for tumor regions without HLA LOH from a tumor where other regions exhibit HLA LOH, and for tumor regions containing an HLA LOH event. All p-values are calculated using an unpaired wilcoxon test. Lung adenocarcinoma is shown in light blue and lung squamous cell carcinomas is shown in magenta.

Interestingly, even among tumor regions that had no HLA loss event, but HLA LOH was observed in other regions from the same tumor (i.e. the non-affected regions from patients with subclonal HLA LOH), there was a significantly higher subclonal non-synonymous mutational burden as compared to tumor regions where the entire tumor had both HLA haplotypes (Figure 5-15). This result suggests that not only may HLA LOH allow for subclonal expansions in affected tumor regions, but that the tumors harboring an elevated mutational burden may be under increased evolutionary pressure for losing an HLA haplotype.

5.6.3 Increased mutation burden in clones with HLA LOH

Because phylogenetic analysis had allowed for each HLA LOH event to be mapped to specific clones present during tumor evolution, it was possible to consider the impact HLA LOH had on specific cancer subclones. In tumors with subclonal HLA LOH the number of mutations present in the cancer subclone harboring HLA loss

could be directly compared with the mutational load of its sister subclone without HLA loss (i.e. clones that shared an ancestor clone but had diverged) (Figure 5-16). As only sister clones were being considered, and they had only differentially evolved after the acquisition of an HLA LOH event, it was possible to assess the immediate impact of HLA LOH on non-synonymous mutation acquisition.

Of the 36 tumors exhibiting any HLA LOH, there were 19 instances where the event was subclonal and not on a terminal node for which a comparison between sister subclones could be made. Events that had occurred immediately prior to a terminal node had to be excluded, as there was no sister clone to compare against.

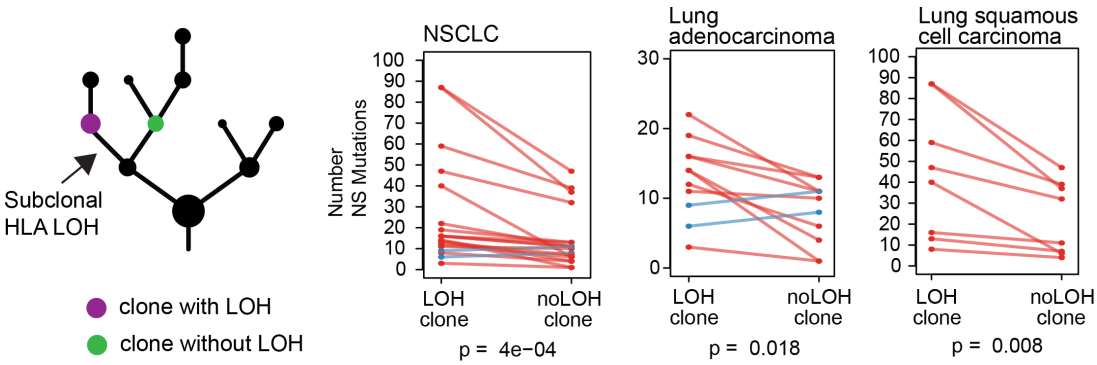

Figure 5-16: Non-synonymous mutational burden at the tumor subclone level.
Schematic of the clones considered for the comparison performed. Here, the cancer subclone harboring HLA loss (purple) is shown with its sister subclone, descended from the same ancestral cancer cell, but without HLA loss (green). The number of non-synonymous mutations found in the clone with HLA LOH is compared to the number of non-synonymous mutations found in the sister clone without HLA LOH. If the HLA LOH containing clone has a higher non-synonymous mutation burden than its sister clone, the line is shown in red; if it has a lower non-synonymous mutation burden, the line is shown in blue. All p-values are calculated using a paired wilcoxon test.

Further confirming the hypothesis that HLA LOH was permissive for subclonal expansions, the subclones harboring an loss of an HLA allele consistently showed an increased non-synonymous mutational burden as compared to their sister clones without HLA LOH (NSCLC p=4e-04; lung adenocarcinoma p=0.018; lung squamous cell carcinoma p=0.008) (Figure 5-16). Unlike the observations at the tumor-level and tumor region-level, subclones with HLA LOH had an increased mutational burden regardless of histological subtype. Overall, there were only 2/19 instances (blue lines in Figure 5-16) of the subclone with HLA LOH having fewer non-synonymous mutations than its counterpart subclone without loss of an HLA allele.

Taken together, these results, at multiple levels of resolution, suggest that HLA LOH directly contributes to the observed increase in subclonal non-synonymous mutations among tumors harboring HLA LOH. While lung squamous cell carcinoma

subclones harboring an HLA LOH event had higher non-synonymous mutational burden as compared to their wildtype counterparts, the overall trend for HLA LOH allowing for higher non-synonymous mutational burden was more strongly observed for lung adenocarcinomas. Yet there were very few lung squamous cell carcinomas harboring an HLA LOH event that were categorized as having low non-synonymous mutational burden. This suggests that while HLA LOH may allow for acquisition of subclonal mutations in lung squamous cell carcinomas, there are also additional mechanisms beyond HLA LOH and HLA down-regulation contributing to the observed high subclonal mutational burden.

5.6.4 Validation in TCGA

To further validate the findings from the TRACERx cohorts in larger cohort, 383 lung adenocarcinomas and 309 lung squamous-cell carcinomas samples from TCGA were analyzed (Campbell et al., 2016). Similar to what was observed in the TRACERx cohort, HLA LOH frequently occurred in lung squamous-cell carcinomas (133/309) and lung adenocarcinomas (118/383) tumors, again being a significantly more common event in lung squamous cell carcinomas (p=0.001) (Figure 5-17A).

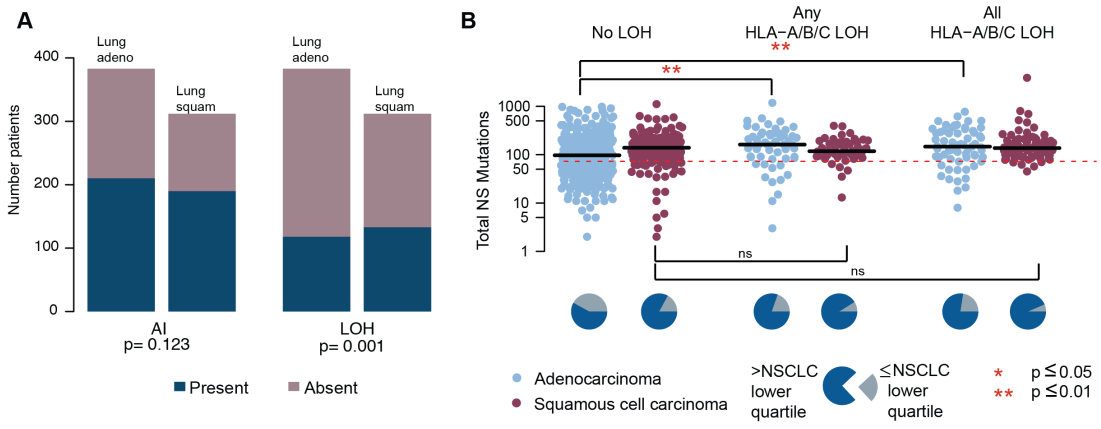

Figure 5-17: Prevalence and impact of HLA LOH in TCGA NSCLC.
(A) The total number of TCGA patients exhibiting AI or LOH at the HLA locus is shown. (B) The total number of non-synonymous mutations is plotted across different categories of HLA LOH for lung adenocarcinoma (light blue) and lung squamous cell carcinomas (magenta). Tumors were classified as having: no HLA LOH; any HLA LOH event; or HLA LOH at all three HLA loci. The lowest total non-synonymous mutation quartile is indicated by the dashed red line and the proportion of tumors with a total non-synonymous mutational burden greater or less than that is indicated by the pie charts for each HLA LOH classification group.

As the TCGA dataset contained more patients, samples could also be categorized as either having HLA LOH at a single locus (56 lung squamous cell carcinoma, 56 lung adenocarcinoma) or HLA LOH affecting all three HLA loci (77 lung squamous

cell carcinoma, 62 lung adenocarcinoma). In agreement with the TRACERx samples, a significantly higher non-synonymous mutation burden was observed in lung adenocarcinomas tumors exhibiting HLA LOH (p=0.0001), regardless of whether a single HLA locus was affected (p=0.002) or all three HLA loci were (p=0.003) (Figure 5-17B).

5.6.5 Mutational signatures in tumors with HLA LOH

A known contributor to subclonal mutations often active late in tumor evolution is the APOBEC family of enzymes. To determine if APOBEC activity was contributing to the elevated subclonal mutational load observed in tumors harboring an HLA LOH event, the mutational signatures active in each tumor sample were calculated (Figure 5-18) (Rosenthal et al., 2016, Alexandrov et al., 2013a).

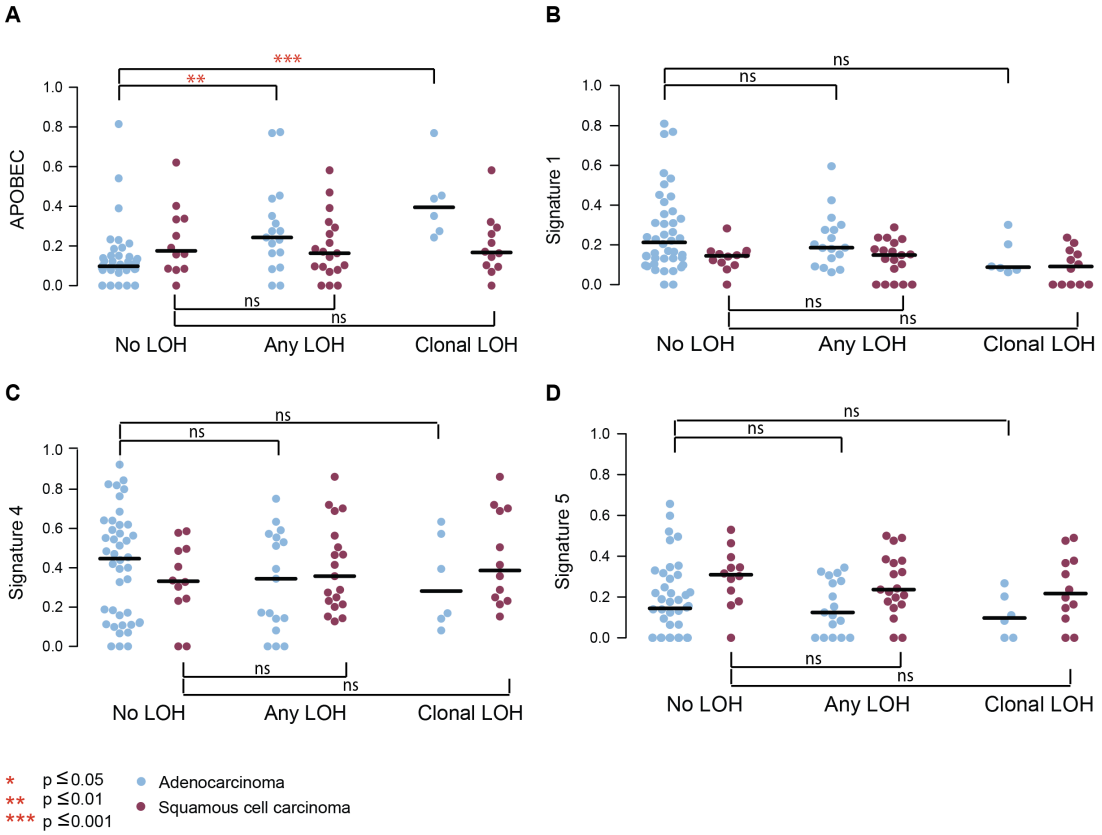

Figure 5-18: Weights of mutational signatures in tumors by HLA LOH status.
For each lung adenocarcinoma (blue) and lung squamous cell carcinoma (purple) tumor, the relative contributions of APOBEC mutational signatures (A), Signature 1 (B), Signature 4 (C), and Signature 5 (D) are shown. p-values are calculated using an unpaired wilcoxon test.

Among lung adenocarcinoma tumors that exhibited HLA LOH, there was a significant increase in the APOBEC signatures (Signature 2 and Signature 13) (NSCLC p=0.03; lung adenocarcinoma p=0.003, lung squamous cell carcinoma

p=0.63); however, no other signatures found in this cohort (Signatures 1, 4, and 5) were found to be differentially contributing between groups.

5.6.6 Enrichment for neoantigens bound to lost HLA allele

After loss of one of the HLA haplotypes, only one set of HLA alleles will be still available to present neoantigens to the immune system. If the decrease in neoantigen presentation has allowed for subclonal expansions, as suggested by the data described above, then there may be an enrichment for subclonal neoantigens that bind to the allele that was lost, as it is those neoantigens that are no longer affected by immune surveillance and elimination. To test this hypothesis, subclonal neoantigens binding to lost and kept HLA alleles were compared from tumors with 6 distinct HLA alleles (i.e. not homozygous at HLA-A,B, or C) and loss of one HLA haplotype in at least one tumor region (n=20; 9 lung adenocarcinomas and 11 lung squamous cell carcinoma).

Consistent with the hypothesis that loss of an HLA allele facilitates the accumulation of subclonal neoantigens, there was an enrichment for subclonal neoantigens predicted to bind to the lost HLA alleles as compared to the kept alleles (Figure 5-19A) (NSCLC p=0.0083; lung adenocarcinoma p=0.29; lung squamous cell carcinoma p=0.02, paired wilcoxon test).

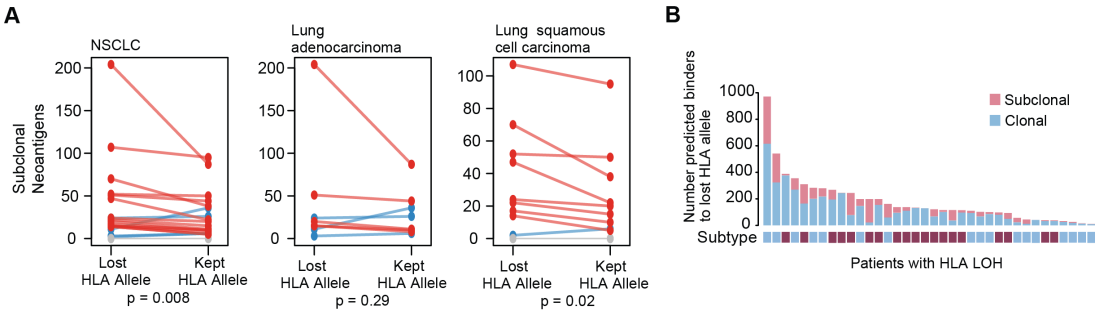

Figure 5-19: Neoantigens predicted to bind to the lost HLA allele.
(A) The number of subclonal neoantigens predicted to bind to either the lost HLA allele or the kept HLA allele is indicated for tumors exhibiting HLA LOH. All NSCLC tumors are first considered, and then lung adenocarcinomas and lung squamous cell carcinomas are considered separately. The p-value is calculated using a paired wilcoxon test. (B) The total number of mutations predicted to result in a neoantigen binding to the lost allele is shown for all patients with at least one HLA LOH event. The mutation clonality is also indicated as either clonal (light blue) or subclonal (light red).

To determine the potential impact that HLA LOH could have on which neoantigen repertoire of a tumor, the neoantigens binding to the lost alleles from the entire TRACERx cohort harboring an HLA LOH event were determined (37 patients total) (Figure 5-19B). All patients had mutations that had previously been classified as

neoantigens but were predicted to bind to a now lost HLA allele. This finding highlights the potential clinical impact HLA LOH could have on the targeting of putative neoantigens (Ott et al., 2017, Sahin et al., 2017).

5.7 Conclusions

In order to avoid immune predation, evolving tumors must acquire mechanisms to escape immune detection or withstand its activity. One such mechanism could be via loss of antigen presentation, as then somatic mutations present in the tumor cell as a result of ongoing mutational processes do not have the chance to be recognized as foreign by T-cells.

While HLA down-regulation has been explored previously as a means of disrupting the antigen presentation pathway, the irreversible loss of an HLA allele had mostly been determined by immunohistochemistry staining. For large-scale analyses of the impact of HLA loss, a computational tool to identify HLA LOH from sequencing data was required. As standard copy number tools fail to resolve the HLA locus, due to its highly polymorphic nature, this chapter has described the computational tool, LOHHLA. LOHHLA is capable of not only determining the major and minor copy numbers at the HLA locus, but also which specific HLA haplotype is subject to copy number loss.

Supporting the notion that loss of antigen presentation may play an important role in immune evasion during tumor evolution, HLA LOH was identified in 40% of the NSCLC samples analyzed, frequently as a late event in tumor evolution. HLA LOH appeared to be an event under strong selection, occurring multiple times over the course of a single tumor's evolutionary history, with consistent HLA alleles subject to loss during each event, suggesting preferential selection for loss of one set of alleles over the other. A formal analysis of focal HLA LOH found that the LOH event occurred more frequently than expected based on random simulations, supporting the hypothesis that HLA LOH is under strong selective pressure.

The subclonal nature of HLA LOH, as well as the tendency for brain metastases to harbor an HLA LOH event more frequently than their matched primaries, supports the theory that immune evasion is a key factor in tumor evolution. Furthermore, mapping the HLA LOH events to the tumors' phylogenetic tree allowed for a direct comparison of the non-synonymous mutational burden between sister clones with and without an LOH event and revealed a significantly elevated non-synonymous

mutation burden among clones exhibiting HLA LOH. This increase in non-synonymous mutation burden was accompanied by an enrichment of neoantigens predicted to bind to the lost HLA alleles. Together, these results suggest that decreased antigen recognition following loss of HLA alleles may be permissive for subclonal expansions and could allow mutations that may have once instigated an immune response to go undetected by the immune system.

Importantly, the characterization of HLA LOH in this chapter was performed in strictly treatment-naïve cohorts, but given the prevalence of LOH events detected, it may be important to consider HLA LOH when designing patient-specific immunotherapy approaches, such as TIL based therapies and neoantigen vaccines. Targeting neoantigens predicted to bind to HLA alleles already lost in the tumor may not effectively elicit a T-cell response. Furthermore, as HLA allele specific loss has already once been observed in an immunotherapy-resistant lesion, it will be intriguing to investigate how frequently HLA LOH results in acquired immunotherapy resistance. As more cohorts become available containing data from both pre- and post-therapy samples, it will be possible to address such questions.

The previous two chapters have explored the determinants of tumor immunogenicity and the mechanisms facilitating immune escape. In the following chapter I'll consider how these factors can be combined with information about the tumor microenvironment to generate a more complete understanding of the tumor/immune interaction in cancer.

Chapter 6 Interaction between tumor and immune microenvironment

6.1 Introduction

Much of the previous work reported in this thesis, as well as in the field of cancer immunology on the whole, has focused on understanding the genomic correlates of immune evasion and patient response to immunotherapy. While immune recognition of a tumor does require functional presentation of tumor-specific antigens, the surrounding microenvironment must also be capable of generating an immune response. Thus, the degree of immune infiltration and composition of the infiltrating cells is an equally important consideration in the understanding of tumor immunogenicity and has been shown to have prognostic relevance (Galon et al., 2006, Charoentong et al., 2017).

Indeed, there have been many recent efforts made to develop bioinformatic methods capable of quantifying immune infiltration from RNAseq data in large tumor cohorts. These approaches can help to elucidate the complex relationship between the immune microenvironment, genomic features of the tumor, and overall patient outcome or response to checkpoint blockade (Gentles et al., 2015, Li et al., 2016, Davoli et al., 2017, Angelova et al., 2015, Danaher et al., 2017, Charoentong et al., 2017). However, recent publications have only considered a single tumor region as reflective of the entire tumor's microenvironment. As genomic ITH has been widely observed across nearly all cancer types, it is likely that if immune infiltration is influenced by somatic tumor alterations then the level and composition of immune infiltrate will also be heterogeneous.

In order to determine the degree of immune infiltration in NSCLC and how it varies on the backdrop of a genomically heterogeneous tumor, this chapter uses multi-region RNAseq data to estimate the abundance of various immune cell populations and determine what characteristics of the tumor are associated with changes in the immune microenvironment.

The pathology determination of region-specific TIL scores used as a ground-truth measure was performed by Roberto Salgado. The PD-L1 staining described in this chapter was performed by Crispin Hiley and Roche.

6.2 Improving immune signatures of infiltration

Traditionally, the immune landscape of tumors has been quantified using flow cytometry and immunohistochemical staining. However, these methods are labor intensive and not well suited for the analysis of large expression datasets. Therefore, to fully leverage the data available, bioinformatic methods to estimate the quantity and composition of infiltrating immune cells have been developed. These scores, which all make different assumptions for their analysis, are generally benchmarked against datasets of known immune cells, but have rarely been compared with each other or across different cancer types. Thus, the first step in estimating the immune infiltration in the multi-region RNAseq cohort was to determine which measure can best describe the data available using direct pathological TIL quantification methods as a benchmark.

6.2.1 Summary of methods

6.2.1.1 CIBERSORT (Newman et al., 2015)

CIBERSORT requires a reference gene expression input matrix containing expression data for each immune cell subtype to quantify. The gene matrix developed by the authors contains 547 genes that are capable of distinguishing 22 different cell types; however, it was designed based on microarray data and not RNAseq. The approach then uses linear support vector regression to de-convolve bulk expression data from tumor samples into cell subsets. The CIBERSORT authors claim their approach can handle noise in the data and collinearity effects due to closely related cell types, though the latter claim is questioned by the authors of TIMER (Li et al., 2016).

6.2.1.2 TIMER (Li et al., 2016)

TIMER first selects genes from the tumor sample set of interest which negatively correlate with tumor purity, under the assumption that these genes are more likely to be representative of infiltrating immune cells. They then use an external reference set of purified immune cells and filter it to only include contributions from the genes they identified to negatively correlate with purity and remove those genes that were in the top 1% most highly expressed for a particular cell subtype to avoid outliers driven by the large variance of highly expressed genes. Finally, they use constrained least squares fitting to estimate the abundance of immune cell types. The final estimates represent the relative abundance of the immune cell subtypes,

but they are not comparable between cancer types or between different immune cells within the same cancer type. TIMER is capable of resolving six different immune cell populations.

6.2.1.3 EPIC (Racle et al., 2017)

Like the previous two deconvolution methods, EPIC also uses a reference gene expression profile, provided with the tool, to model the bulk expression data as a linear combination of the different cell types. The reference genes are only lowly expressed in non-immune cells. Also included among the different populations is an uncharacterized group of cells in order to simultaneously estimate the fraction of cancer cells in addition to the different immune cells.

6.2.1.4 Davoli (Davoli et al., 2017)

Davoli et al. estimate immune cell composition by identifying gene sets from the ImmGen (https://www.immgen.org/) gene expression database which uniquely define a specific cell type (i.e. they are only expressed in a single immune cell type and not others). They considered the average gene expression level of genes comprising each cell subtype.

6.2.1.5 Danaher (Danaher et al., 2017)

Danaher et al. also begin with previously identified immune cell subtype markers identified from studies of individual immune cell populations performed by the Galon group (Bindea et al., 2013). To fill in candidate gene markers for immune cell subtypes that were not considered by Bindea et al., they included genes which were highly enriched in a particular immune cell subtype as reported by Newman et al. (Newman et al., 2015). Finally they included well-characterized markers for exhausted CD8+ T-cells and regulatory T-cells. Danaher et al. next attempted to remove poor marker genes in order to only use reference genes that were stably expressed in a single cell type and at roughly the same level within that cell type. To accomplish this, the expression levels for all reference genes for a given immune cell subtype were compared pair-wise and only those that correlated well, with a slope of ~1, were retained. For each cell subtype, a score was calculated from the mean of the log-transformed expression values of the reference genes. This score does not reflect absolute quantification, but rather allows for a comparison of the same immune cell subtype between different tumor samples.

6.2.1.6 Rooney (Rooney et al., 2015)

Rooney et al. defined cytolytic activity as the geometric mean of the genes *GZMA* and *PRF1*. Specific cell type enrichment was calculated using ssGSEA using marker genes that had been defined as having at least 2-fold greater expression in the immune cell subtype of interest as compared to any other immune cell subtype.

Importantly, all of the immune signature measures described rely heavily on the fidelity of the reference gene set used to either define marker genes delineating the immune cell subtype of interest or in the reference matrix used for deconvolution. Furthermore, they often ignore the immune component of the normal tissue. Marker gene approaches do not depend so strongly on the data sets used to define an immune subset signature, but they do require that the gene expression be confined to only a single immune cell subtype. Finally, it's important to consider that tumor cells may also express genes that are typically associated with a particular immune subset. Deconvolution methods which have not taken this into account, for instance, if they have only used expression profiles generated from purified immune gene populations, may not correctly capture the infiltrating immune cell population.

6.2.2 Consistency of immune signatures

To investigate the variability of different methods for immune subset quantification, immune cell types were compared across the large number of TCGA lung adenocarcinoma and lung squamous cell carcinoma tumors available using the different approaches. Even if an approach characterizes relative immune contribution for a given cell type, a correlation between predictions generated by the different methods should be observed.

Two cell types, CD8+ T-cells and B cells, were chosen for analysis as they both could be estimated by multiple immune signature approaches. Surprisingly CD8+ T-cells showed poor correlation between the different estimation methods, particularly when comparing tools that relied on deconvolution (EPIC and TIMER) to those that used marker genes (Davoli and Danaher) (Figure 6-1A).

Importantly, such differences could confound conclusions associating particular immune cell infiltrate levels with patient prognosis. The correlations observed among the B cell estimates were much stronger across methods (Figure 6-1B), suggesting that some immune cell types are more challenging to accurately estimate than others.

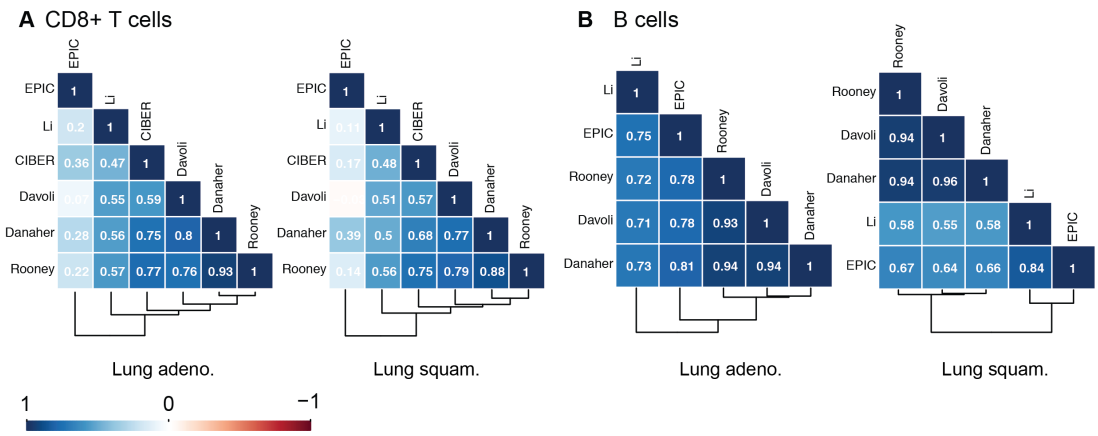

Figure 6-1: Correlations of immune cell type estimations.
Immune cell type estimations for TCGA lung adenocarcinoma and lung squamous cell carcinoma are compared between various approaches for CD8+ T-cells (A) and B cells (B). CIBERSORT, which had originally been designed for microarray data, was only trusted to quantify CD8+ T-cells. The steps taken to determine what cell types CIBERSORT could accurately identify from RNAseq data are detailed in the Methods.

The main explanation for the differences observed in immune cell estimation is that the genes used in the reference signatures, or the assumptions used to filter the gene lists, differ between approaches. To explore this, the genes used in the various immune cell type definitions were compared. Interestingly, for both CD8+ T-cells and B cells, there was not a single gene that was shared between all of the immune signatures considered (Figure 6-2).

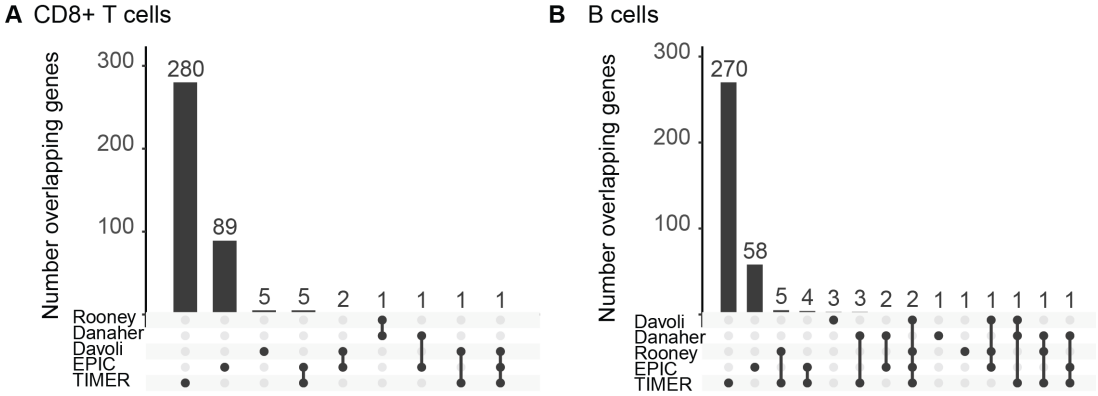

Figure 6-2: Overlapping gene between immune signatures.
Number of genes shared between different signature definitions are shown for CD8+ T-cells (A) and B cells (B). Dots represent the number of genes in each category, and connected dots represent the number of genes shared between those categories. No gene is shared between all of the immune signatures considered for either immune subset.

6.2.3 Choosing an immune signature approach

Given the number of possible immune infiltration tools available and their observed variability, choosing the most reliable estimate to describe the TRACERx dataset is

imperative, as inaccurate estimates may generate misleading conclusions. As the genes used to define immune subtypes will be expressed on infiltrating immune cells, a negative correlation should exist between the expression of the immune genes and the purity of the tumor. Indeed, this assumption is incorporated into the first step of the TIMER approach. To test how the expression of the genes used in other immune signatures related to tumor purity, correlations were determined for each gene used to define an immune subset (Figure 6-3).

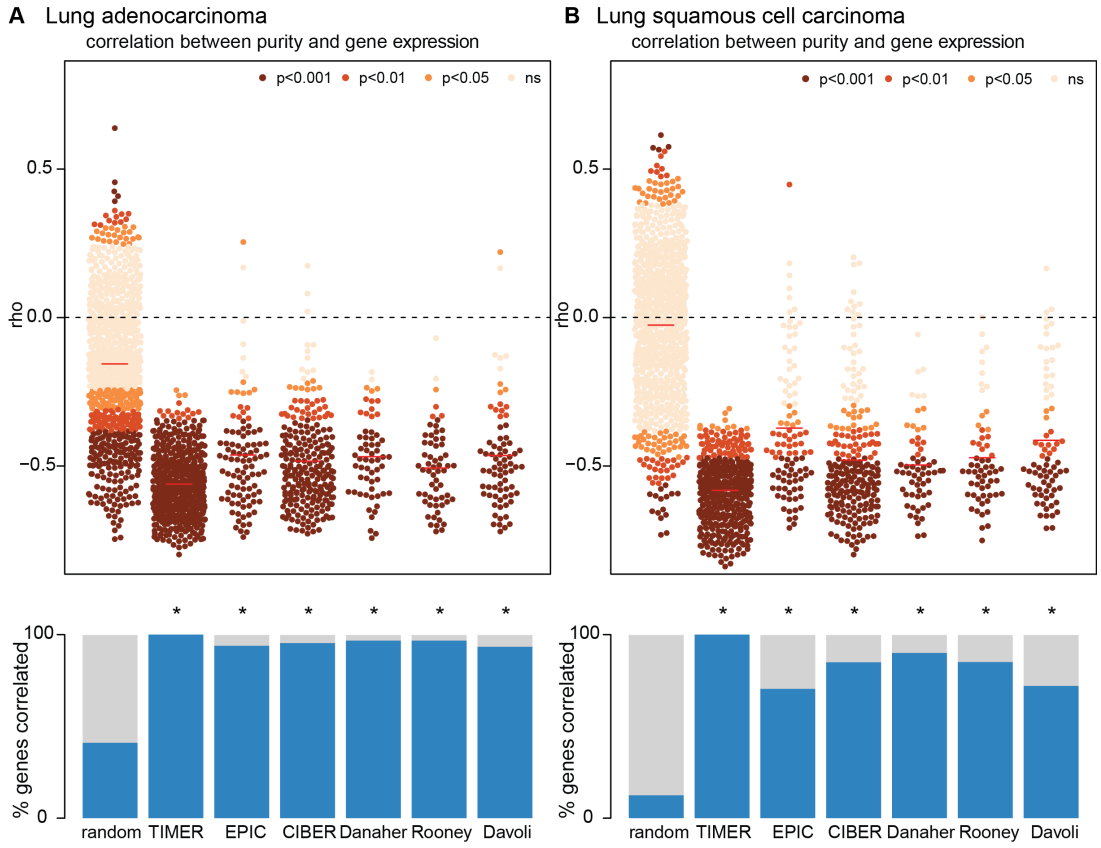

Figure 6-3: Immune signature genes correlated with tumor copy number.
The expression of the genes used in each immune signature definition is correlated against tumor purity. The rho value of the correlation is plotted for the immune genes comprising each signature definition, colored by the p-value of the association. The comparisons are performed separately for lung adenocarcinoma (A) and lung squamous cell carcinoma (B). The median rho value for each immune signature set is indicated by the red line. The fraction of genes whose expression value is significantly correlated with purity is shown and compared to a set of random genes. For every immune signature, there was significant enrichment of genes whose expression negatively correlated with tumor purity as compared to the random selection of genes.

Reassuringly, for both lung adenocarcinoma and lung squamous cell carcinoma tumors, most of the immune genes used in cell subset definitions were significantly negatively correlated with tumor purity. Furthermore, there was significant enrichment of immune genes that negatively correlated with tumor purity as compared to a random subset of genes. As expected, all of the genes used in

TIMER were negatively correlated with purity. The immune signature which had the next highest proportion of immune genes anti-correlated with tumor purity for both lung cancer histologies was the Danaher approach. Interestingly, the Danaher approach did not specifically require this characteristic of their gene lists.

6.2.4 Using copy number to refine immune signatures

Building on the assumption made by TIMER that immune gene expression should negatively correlate with tumor purity if the expression of these genes is due to infiltrating immune cells, it holds that there should also be no relationship between immune gene expression and tumor copy number at the gene locus. A positive correlation may indicate that the gene is expressed by the tumor cell in addition to tumor infiltrating immune cells, thereby confounding any immune estimates made using a reference gene list obtained from purified immune cell populations. In any cancer type characterized by extensive copy number aberrations, such as NSCLC, an immune gene also expressed by the tumor may have a profound impact on the immune infiltrate estimations.

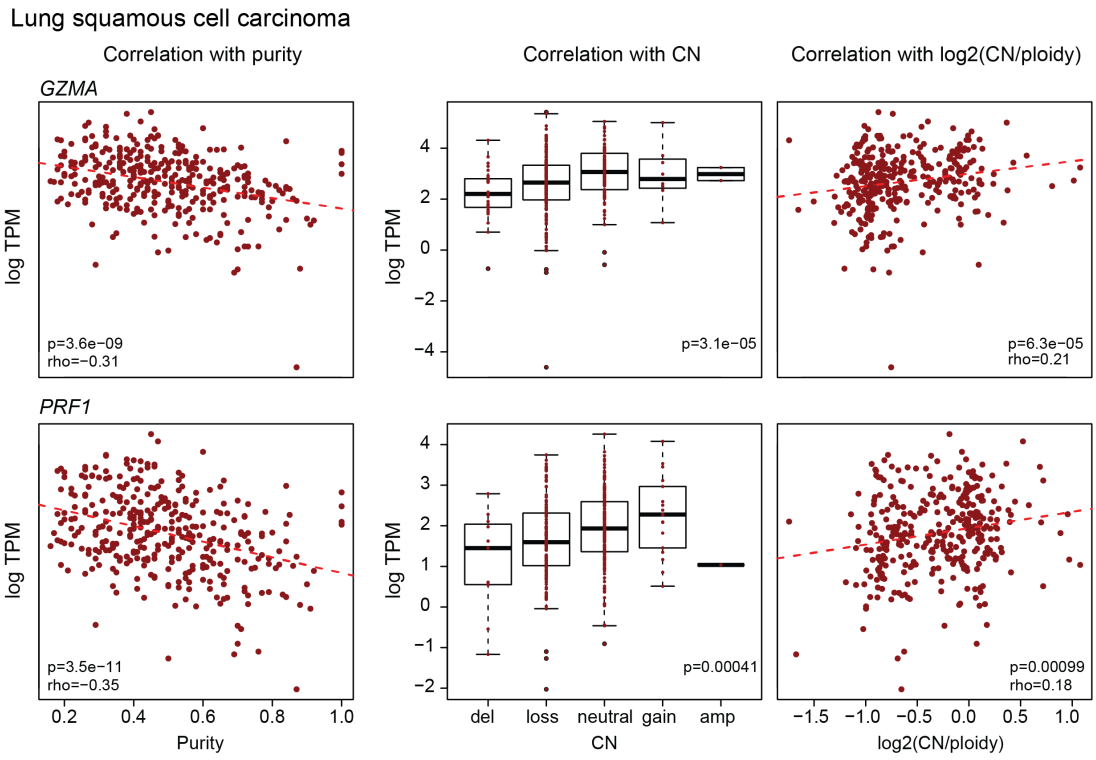

Figure 6-4: Association between CYT genes expression, copy number, and purity in lung squamous cell carcinoma.
The expression levels of the two genes comprising the CYT score (*GZMA* and *PRF1*) was correlated against the copy number status of the genes in the tumor as well as the purity of the tumor for lung squamous cell carcinoma.

For instance, the two genes comprising the CYT score (*GZMA* and *PRF1*) are both significantly negatively correlated with tumor purity, as would be expected by genes expressed on infiltrating immune cells. However, in lung squamous cell carcinoma, expression of both these genes is also positively correlated with tumor copy number at the *GZMA* and *PRF1* loci, suggesting that these genes are not solely indicative of immune cytolytic activity in this cancer type (Figure 6-4).

No correlation was observed between tumor copy number and gene expression for the CYT score components in lung adenocarcinoma (Figure 6-5). Interestingly, in the paper where the CYT score was first defined, the authors observe a correlation between mutational load and cytolytic activity in most cancer types including lung adenocarcinoma, but there was no significant association in lung squamous cell carcinoma (Rooney et al., 2015), further suggesting that in this cancer type CYT score may be an inaccurate measure of immune activity.

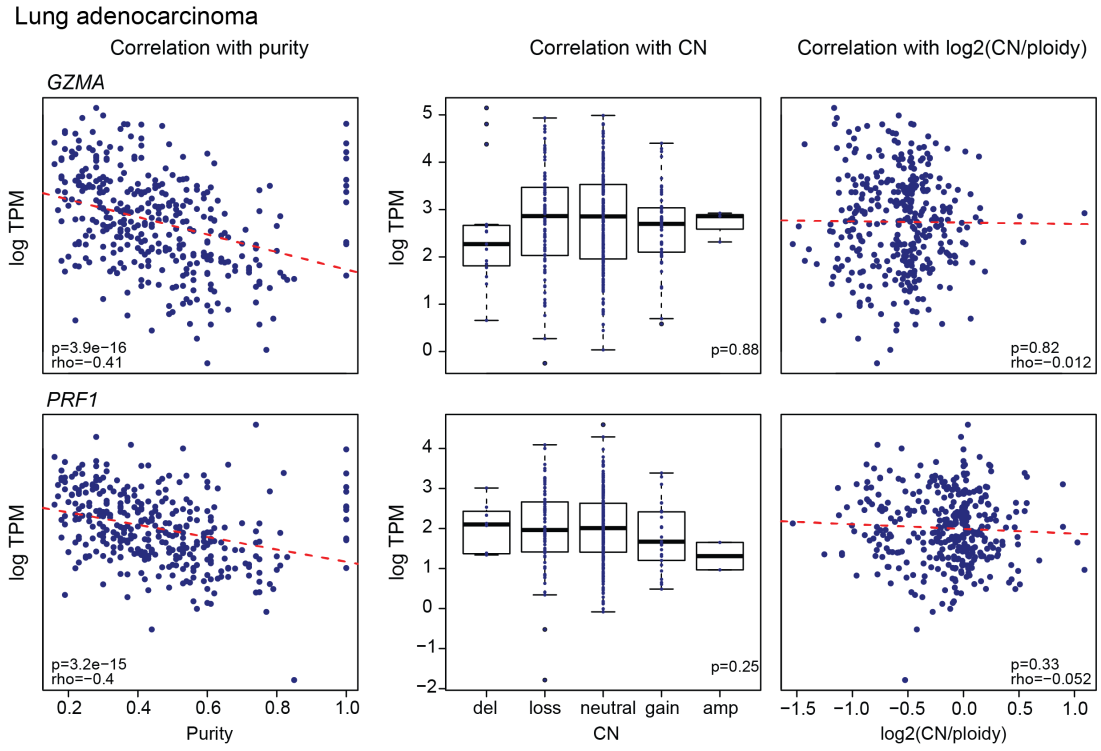

Figure 6-5: Association between CYT genes expression, copy number, and purity in lung adenocarcinoma.
The expression levels of the two genes comprising the CYT score (*GZMA* and *PRF1*) was correlated against the copy number status of the genes in the tumor as well as the purity of the tumor for lung squamous cell carcinoma.

Thus, to refine immune signatures, the genes used by each approach were tested for a correlation with tumor copy number. Reassuringly, a significantly lower proportion of the immune signature genes exhibited expression that was positively

correlated with copy number as compared to a random selection of genes. However, a large number of genes used in immune signature definitions had expression values that were significantly positively correlated with tumor copy number, even among the TIMER-defined genes, where these selected genes must also have shown a negative correlation with tumor purity (Figure 6-6).

These results suggest that for each immune signature, there are genes expressed by the tumor that are possibly contributing to inaccurate estimates of immune infiltration. The immune signature with the fewest genes showing a positive correlation between expression value and copy number was the Danaher set. For this signature, only two genes (*MS4A4A* and *TPSAB1*, used in the definitions of macrophages and mast cells, respectively) had a positive relationship. This suggests the Danaher immune signature approach may be most well-suited for use in this data set.

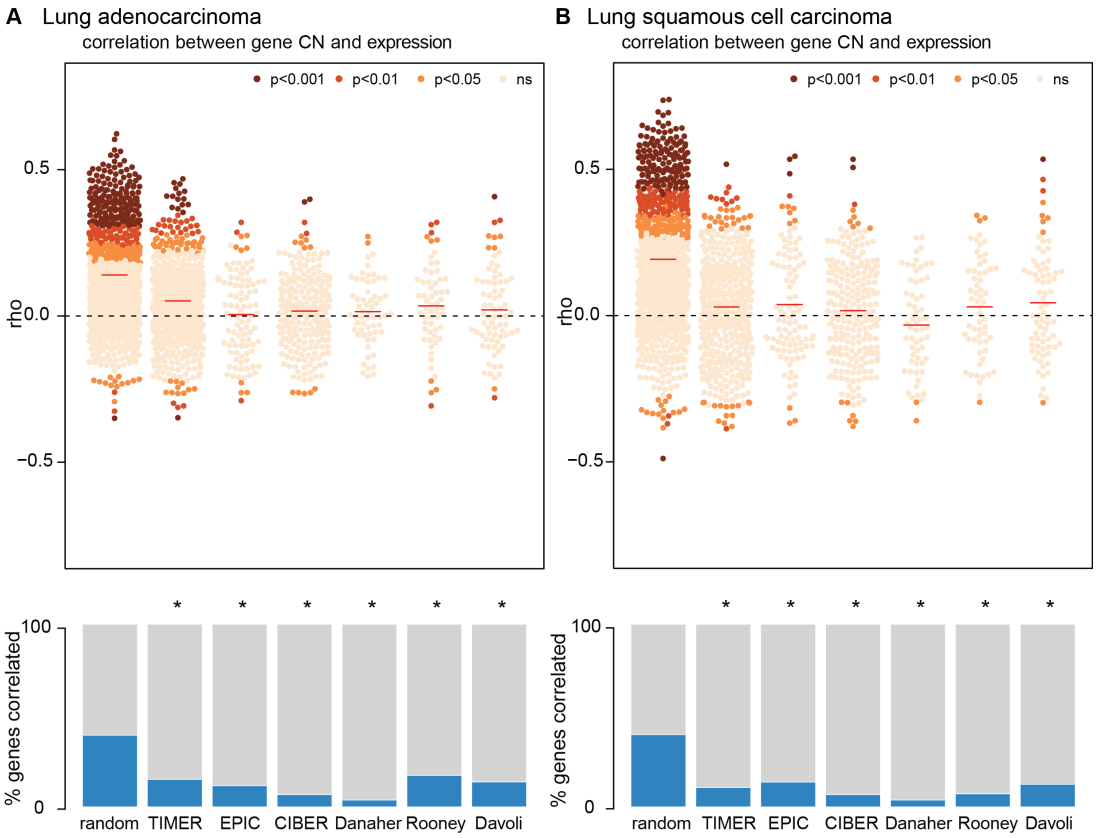

Figure 6-6: Immune signature genes correlated with tumor copy number.
The expression of the genes used in each immune signature definition are compared against tumor copy number. The rho value of the correlation is plotted for the immune genes comprising each signature definition, colored by the p-value of the association. The comparisons are performed separately for lung adenocarcinoma (A) and lung squamous cell carcinoma (B). The median rho value for each immune signature set is indicated by the red line. The fraction of genes whose expression value is positively correlated with is shown and compared to a set of random genes. For every immune signature, there was significant depletion of genes whose expression positively correlated with tumor copy number as compared to the random selection of genes.

6.2.5 Copy number associations depend on cancer type

Further complicating the de-convolution of immune cell subsets, the immune genes expressed by the tumor cell may not be consistent between cancer types. Thus, if different cancer types express different immune genes, adjusting immune signatures for tumor expressed genes will not be as simple as further filtering the gene lists. The genes that are used in immune subset definitions, in large part, do not depend on tumor type, with the only exception being the genes used by TIMER, as those are identified on a cohort-by-cohort basis (Figure 6-7A). However, the genes that were found to positively correlate with tumor copy number were almost entirely unique to the cancer type being studied (Figure 6-7B). Only eight genes exhibiting a correlation with tumor copy number were shared between lung adenocarcinoma and lung squamous cell carcinoma.

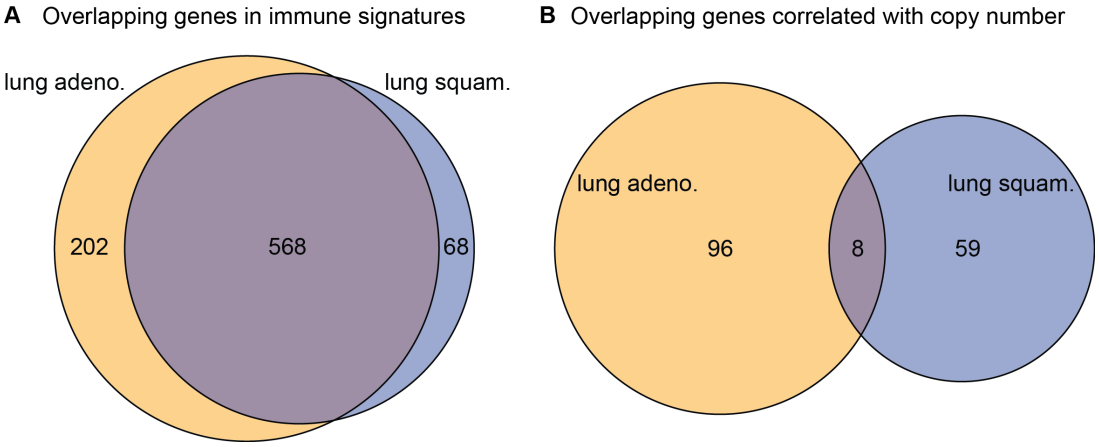

Figure 6-7: Overlap in immune signature genes between cancer types.
The vast majority of genes used in the definitions of the different immune signatures overlap between lung adenocarcinoma and lung squamous cell carcinoma (A). However, the genes whose expression is significantly correlated with copy number rarely overlap between the two histologies (B).

6.2.6 Comparison to pathology determined TIL scores

To determine if the observed correlations between immune gene expression and tumor copy number had any bearing on the accuracy of immune infiltrate estimates, the immune infiltrate estimates were compared to pathology determined TIL scores found in that tumor region (Figure 6-8). Many of the measures of immune infiltration showed a significant positive correlation with the TIL scores obtained, with only a few measures exhibiting a negative correlation (macrophages as estimated by TIMER, CD4+ T-cells in lung squamous cell carcinoma as estimated by EPIC, endothelial cells and CAFs in lung adenocarcinoma as estimated by EPIC, mast cells in lung adenocarcinoma as estimated by Danaher et al., and neutrophils in

lung squamous cell carcinoma as estimated by CIBERSORT). However, none of these negative correlations were significant.

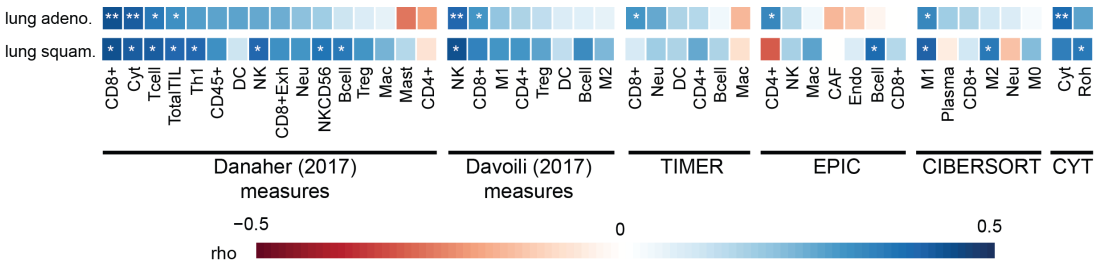

Figure 6-8: Correlations between TIL scores and immune infiltrate estimates.
Correlations between the estimates of different immune cell subtypes calculated by different methods and the TIL scores identified in that tumor region. Significance indicators are FDR corrected, q<0.01 (**), q<0.05 (*).

The immune score showing the strongest association with TIL scores was the CD8+ T-cell score estimated using the Danaher approach, supporting the notion that taking into account copy number variation may be critical to accurately estimating immune infiltrate. Indeed, many of the Danaher measures significantly correlated with the TIL score for lung adenocarcinoma and/or lung squamous cell carcinoma tumors, with the exception of CD4+ T-cells. This is likely because Danaher et al. found no suitable genes to describe the CD4+ T-cell population, and instead, estimated this population by using the total T-cell score minus CD8+ T-cells.

Overall the Danaher immune signatures consistently showed a superior correlation with TIL scores, even when comparing the lung adenocarcinoma and lung squamous cell carcinoma subtypes separately, indicating that this immune estimate was valid in both histological subtypes. For instance, the Danaher immune signatures were the only ones to show a significant relationship between CD8+ T-cells and TIL score for both histology subtypes (Figure 6-9).

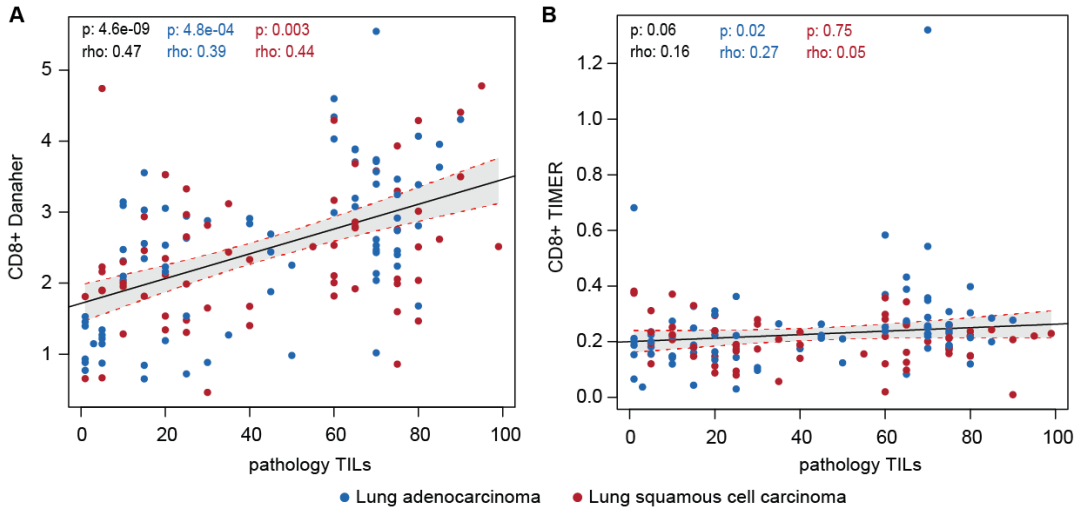

Figure 6-9: Relationship between CD8+ T-cells and TIL scores.
Scatterplots show the correlation between TIL scores and CD8+ T-cells as measured by the Danaher (A) and TIMER (B) approaches. Blue dots indicate regions from a lung adenocarcinoma tumor, red dots indicate regions from a lung squamous cell carcinoma tumor. Rho values and p-values are given for all tumor regions (black), lung adenocarcinoma tumor regions (blue), and lung squamous cell carcinoma tumor regions (red).

Additionally, there was no correlation of tumor purity with pathology determined TIL scores (lung adeno: p=0.73, rho=-0.04; lung squam: p=0.24, rho=-0.17), suggesting that the immune measures provide further information than would be gained from considering stromal content alone (Figure 6-10).

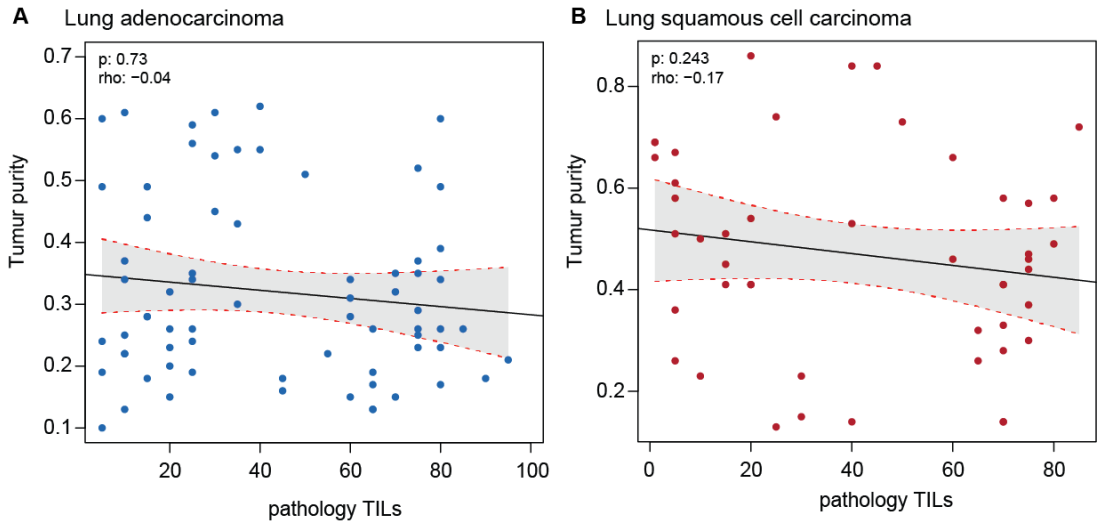

Figure 6-10: Relationship between tumor purity and TIL scores.
Scatterplots show the correlation between TIL scores tumor purity for lung adenocarcinoma (A) and lung squamous cell carcinoma (B). Rho values and p-values are displayed.

Thus, for analysis of the TRACERx data, the Danaher immune signatures were used for all immune subtypes except CD4+ T-cells. For CD4+ T-cells, the Davoli immune signature was used, as it also relied on a marker gene approach.

6.3 Classifying immune activity in NSCLC

The Danaher immune signatures were used to estimate of immune cell populations for each TRACERx region using available RNAseq data. This cohort was taken from a subset of the original TRACERx first 100 cohort (Jamal-Hanjani et al., 2017) and was comprised of 172 tumor regions from 64 patients (41 lung adenocarcinoma, 16 lung squamous cell carcinoma, 7 other histology). The majority of patients had RNAseq data from multiple regions available, with 19 only having a single region sequenced (Figure 6-11).

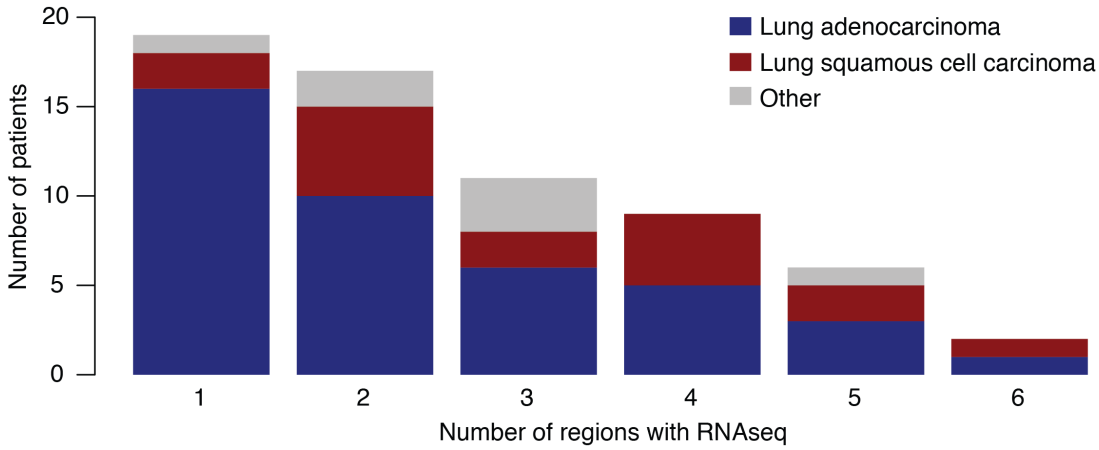

Figure 6-11: TRACERx multi-region RNA-sequencing
The number of regions from each patient with available RNAseq data is shown, colored by histological subtype.

There was a wide range of immune infiltration observed in this cohort, both between tumor samples and between separate regions from the same tumor. However, individual tumor regions from both lung squamous cell carcinomas and lung adenocarcinomas could be stratified as either having high immune infiltrate for the majority of the measures considered or nearly uniform low levels of immune infiltrate (Figure 6-12 & Figure 6-13).

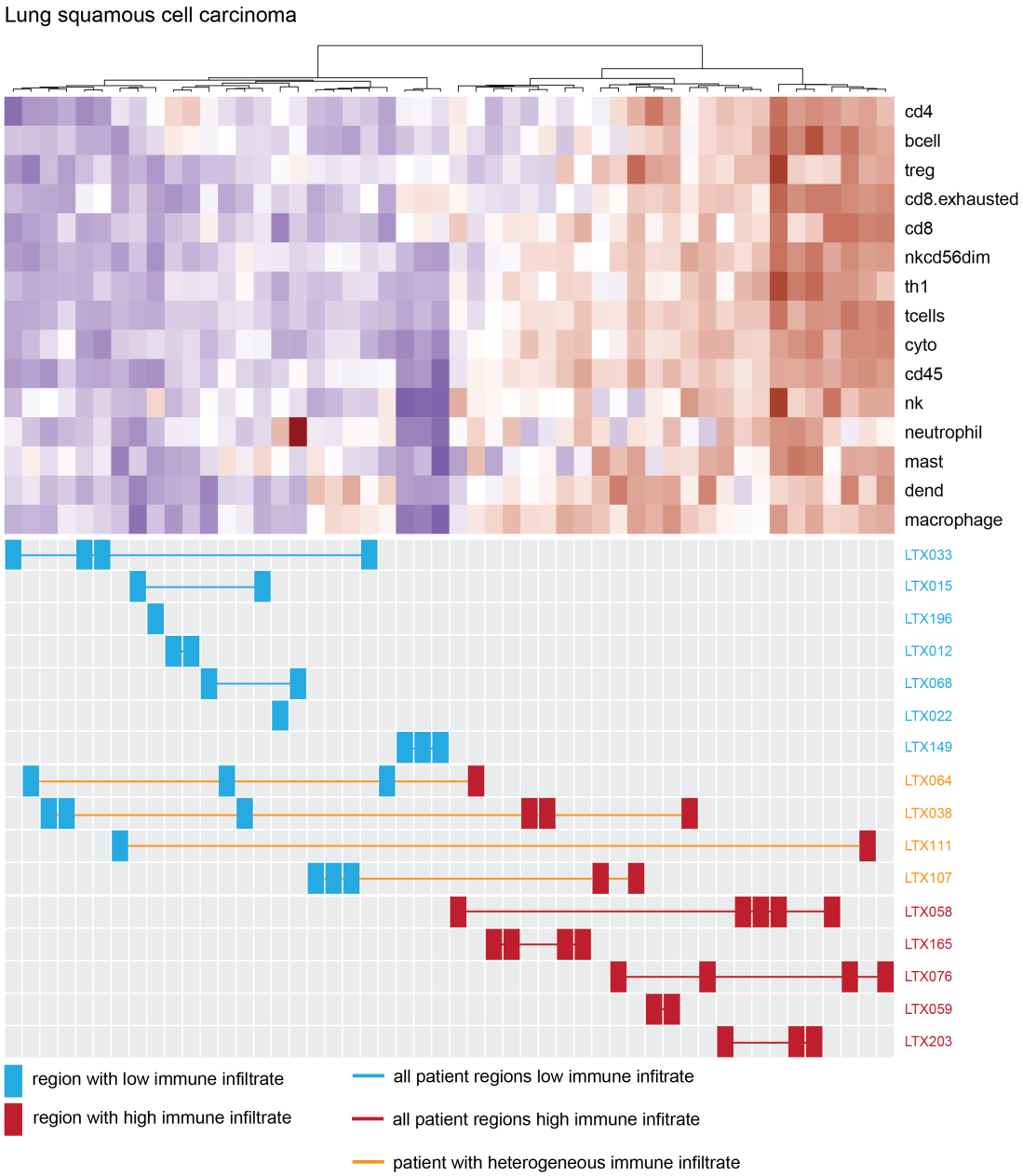

Figure 6-12: Heatmap of immune infiltrates for lung squamous cell carcinoma.
Regions from lung squamous cell carcinoma TRACERx patients are shown, clustered by the level of estimated immune infiltrate. Each column represents a tumor region from the patient indicated on the barplot below. Regions classified as having low levels of immune infiltration are shown in blue, whereas regions classified as having high levels of immune infiltration are shown in red. If all regions from a patient's tumor are classified as having low immune infiltrate, that patient is indicated in blue. If all regions from a patient's tumor are classified as having high immune infiltrate, that patient is indicated in red. Patients with tumors containing heterogeneous levels of immune infiltration are indicated in orange.

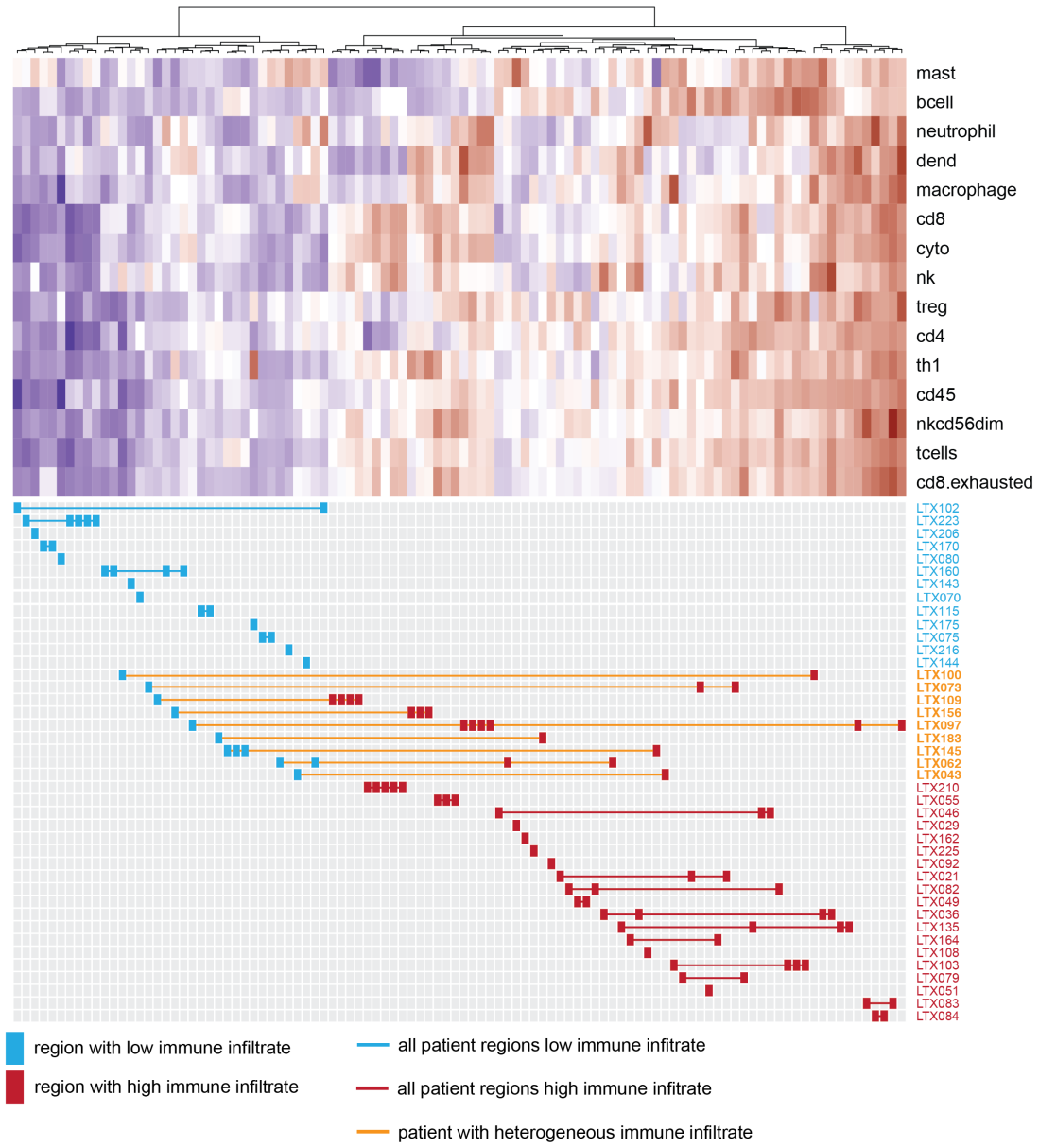

Figure 6-13: Heatmap of immune infiltrates for lung adenocarcinoma.
Regions from lung adenocarcinoma TRACERx patients are shown, clustered by the level of estimated immune infiltrate. Each column represents a tumor region from the patient indicated on the barplot below. Regions classified as having low levels of immune infiltration are shown in blue, whereas regions classified as having high levels of immune infiltration are shown in red. If all regions from a patient's tumor are classified as having low immune infiltrate, that patient is indicated in blue. If all regions from a patient's tumor are classified as having high immune infiltrate, that patient is indicated in red. Patients with tumors containing heterogeneous levels of immune infiltration are indicated in orange.

6.3.1 Heterogeneity of immune infiltration

The multi-region nature of the RNAseq data also allowed the investigation of the heterogeneity of immune infiltration across different tumor regions from the same patient. While some patients had tumors with consistently low or high levels of

immune infiltration, many patients had tumors with regions of disparate levels of immune infiltration. For instance, for LTX111, a lung squamous cell carcinoma, (Figure 6-12) and LTX097, a lung adenocarcinoma, (Figure 6-13), tumor regions were found on completely opposite ends of the immune infiltration spectrum. In total, 4/14 lung squamous cell carcinomas and 9/25 lung adenocarcinomas with more than a single region had a heterogeneous immune landscape.

Thus, individually, tumor regions could be classified as having high or low levels of immune infiltration, and tumors on the whole could also be classified as having consistent (low or high) levels of immune infiltration or heterogeneous levels of immune infiltration. The classification at the tumor-level is of particular importance when trying to determine which patients may respond from immunotherapy using only a single region. If a tumor is heterogeneously infiltrated, then a single biopsy will provide an inaccurate picture of the tumor as a whole.

6.3.2 Composition of immune clusters

Consistent with previous reports (Tamborero et al., 2017), tumor regions with high levels of immune infiltration also had a significantly increased ratio of immune effector cells (CD8+ T-cells, NK cells) to immune suppressor cells (regulatory T-cells, macrophages, neutrophils) (Figure 6-14). This suggests that the composition of immune cells, as well as their quantity, differs between high and low immune infiltrate tumor regions.

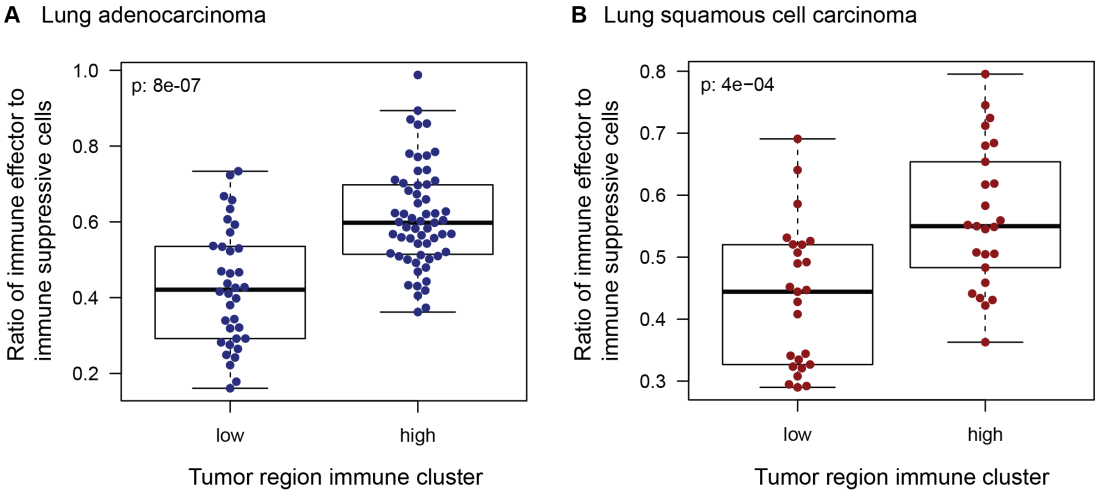

Figure 6-14: Ratio of immune effector to immune suppressive cells.
For each tumor region, the ratio of immune effector to immune suppressive cells is plotted by the immune cluster indicating whether that tumor region had high or low levels of immune infiltration). Plots are shown for lung adenocarcinoma (A) and lung squamous cell carcinoma (B). The p-value tests the effector to suppressive ratio between the immune high regions and immune low regions.

Interestingly, while both lung adenocarcinoma and lung squamous cell carcinoma tumors showed an increase in immune effector cells relative to immune suppressive cells in high infiltrate tumor regions, the ratios calculated for tumors with heterogeneous immune infiltrate differed. Regardless of whether the region itself had high or low levels of immune infiltration, regions from lung adenocarcinoma tumors with heterogeneous immune infiltrate tended to resemble regions from tumors with consistently high immune infiltrate (Figure 6-15A). On the other hand, regions from heterogeneous lung squamous cell carcinoma tumors tended to resemble regions from tumors with consistently low levels of immune infiltration (Figure 6-15B).

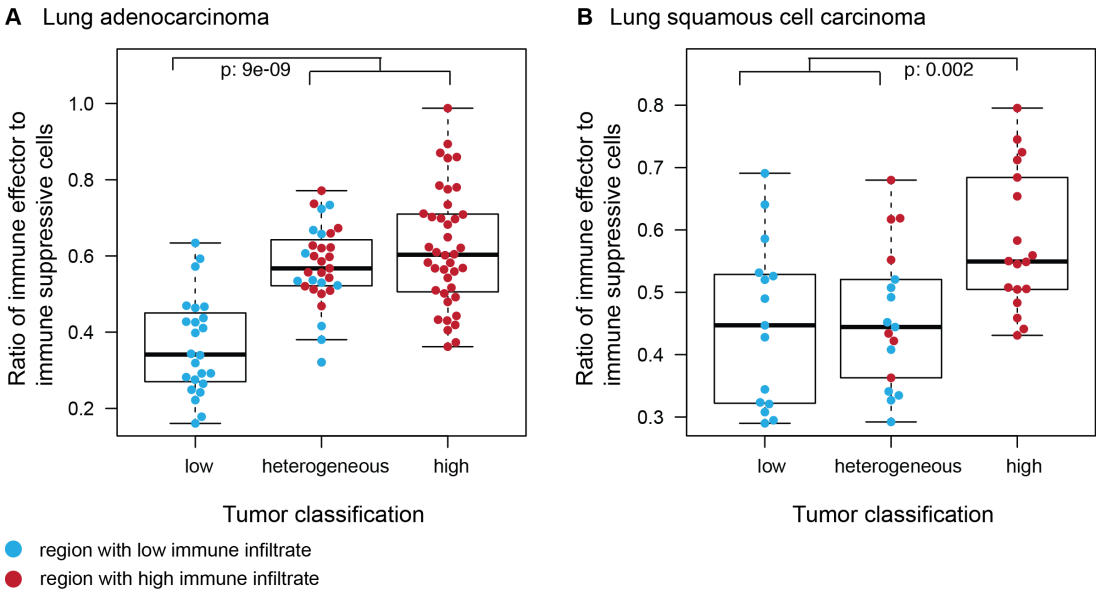

Figure 6-15: Immune effector to suppressive cell ratio by tumor classification.
The ratio of immune effector cells to immune suppressive cells is shown for each region, broken down by whether that region came from a tumor with consistently low immune infiltration, consistently high immune infiltration, or a heterogeneous level of immune infiltration. Plots for lung adenocarcinoma (A) and lung squamous cell carcinoma (B) tumors are displayed. Regions with low levels of immune infiltrate are shown in blue, those with high levels of immune infiltrate are shown in red.

When the regions from heterogeneous lung adenocarcinomas were split into those containing high or low immune infiltration, all regions had a high ratio of immune effector to immune suppressive cells (Figure 6-16A). Indeed, there was no significant difference between the effector to suppressor ratio between low immune infiltrate regions and high immune infiltrate regions from heterogeneously infiltrated tumors. This indicated that lung adenocarcinoma tumors may exhibit a strong “tumor phenotype”, wherein all the regions from the tumor, regardless of whether they have low or high levels of immune infiltration, share specific characteristics.

Lung squamous cell carcinoma tumors tended to show a stronger “region phenotype”. In this case, the low immune infiltrate regions from heterogeneously infiltrated tumors tended to resemble the regions from tumors with consistently low levels of immune infiltration. Similarly, the high immune infiltrate regions from heterogeneously infiltrated tumors resembled the regions from tumors with consistently high levels of immune infiltration (Figure 6-16B).

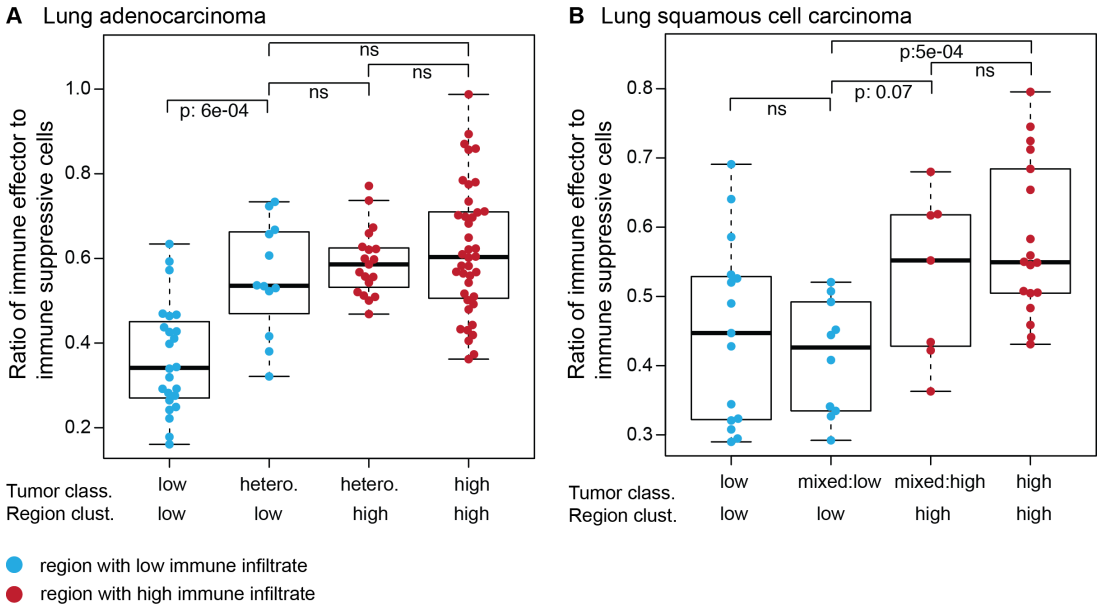

Figure 6-16: Immune effector to suppressive cell ratio by tumor classification and region cluster.
The ratio of immune effector cells to immune suppressive cells is shown for each region, broken down by whether that region came from a tumor with consistently low immune infiltration, consistently high immune infiltration, or a heterogeneous level of immune infiltration. Regions from heterogeneously infiltrated tumors are further divided based on whether the region itself had a low or high level of immune infiltration. Plots for lung adenocarcinoma (A) and lung squamous cell carcinoma (B) tumors are displayed. Regions with low levels of immune infiltrate are shown in blue, those with high levels of immune infiltrate are shown in red.

6.4 Characteristics of immune clusters

Previous chapters in this thesis have shown that clonal neoantigens can be recognized by T-cells in the tumor and that high clonal neoantigen burden is related to good clinical outcome, potentially due to enhanced T-cell activation by clonal neoantigens. In contrast, high neoantigen heterogeneity is associated with poor clinical outcome and response to immunotherapy. To determine if the level of immune infiltration was related to other factors associated with tumor immunity, such as neoantigen load and genomic ITH, the information gathered from the exome analysis of these tumors was also considered. The goal of these analyses was to determine any tumor intrinsic characteristics that could explain the level or

heterogeneity of immune infiltration, potentially shedding light on the relationship between the neoantigen repertoire and immune activation.

6.4.1 Increase of clonal neoantigens in high immune infiltrate tumors

Consistent with previous reports (Brown et al., 2014, Giannakis et al., 2016) and the work reported in Chapter 4, both total and clonal neoantigen load in lung adenocarcinomas were positively correlated with the average CD8+ T-cell infiltrate estimate from that tumor (Figure 6-17). A patient-level average measure was considered here, as the clonal neoantigen load will be consistent across regions from the same tumor.

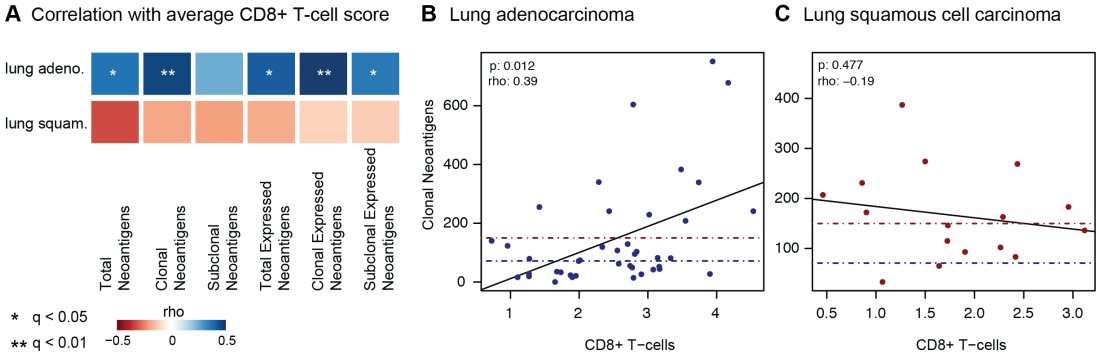

Figure 6-17: Correlation of immune infiltrate and neoantigen load.
(A) The correlation between average CD8+ T-cell score per patient and neoantigen burden in the tumor is shown for lung adenocarcinomas and lung squamous cell carcinomas, which positive correlations indicated in blue and negative in red. An FDR corrected p-value is shown where significant. (B-C) The correlation between CD8+ T-cells and clonal neoantigens is shown for lung adenocarcinoma (B) and lung squamous cell carcinoma (C). The red dashed lines on each plot represent the median clonal neoantigen burden for lung squamous cell carcinomas, and the blue dashed lines represent the median clonal neoantigen burden for lung adenocarcinomas.

Interestingly, and in agreement with the findings from Chapter 4, clonal neoantigen load was more significantly correlated with immune infiltration than total neoantigen load, suggesting that clonal neoantigens can more potently induce an immune response. Furthermore, this association was strengthened when the number of clonal neoantigens was filtered by whether the underlying mutation was expressed or not (Figure 6-17A). A neoantigen was considered to be expressed if at least five RNAseq reads mapped to the mutation position, and at least three contained the mutated base to ensure that the gene was expressed as well as the mutant copy of the gene.

As reported in Chapter 5, overall lung squamous cell carcinoma tumors had an increased clonal neoantigen burden as compared to lung adenocarcinomas (lung

squam: median clonal neo = 150, lung adeno: median clonal neo = 71) (Figure 6-17B-C). It may be that there is no observed correlation in lung squamous cell carcinoma between neoantigen load and CD8+ T-cell infiltrate because all the tumors considered of this cancer type had high clonal neoantigen burden. Indeed when only the subset of lung adenocarcinoma tumors with >150 clonal neoantigens was considered, there was no longer an association between neoantigen load and CD8+ T-cell estimate, suggesting that CD8+ T-cell infiltrate may not correlate with neoantigen burden in high burden tumors ($p = 0.6$, $rho = 0.18$).

6.4.2 High immune infiltration associates with low genomic ITH

The previous results have shown that clonal neoantigen load is an important factor in generating an immune response and influencing patient response to checkpoint blockade, but under the immune editing hypothesis, the immune system may also shape the evolution of the tumor. Thus, to determine if there was evidence of immune activity altering the genomic landscape of the tumor, the level of CD8+ T-cells was compared to the observed heterogeneity of each tumor region, as measured by the Shannon diversity index (calculation described in the Data and Methods).

Consistent with the hypotheses that T-cell activity can prune away tumor subclones or that T-cell infiltrate is driven by high clonal neoantigen burden, a significant negative correlation was observed between the level of immune infiltration and ITH in that tumor region, with regions having high T-cell activity generally having lower ITH. This association was only observed among tumor regions from lung adenocarcinomas (lung adeno: $p=0.035$, $rho=-0.22$; lung squam: $p=0.91$, $rho=-0.02$) (Figure 6-18).

The level of heterogeneity observed in a tumor region also differed by immune classification of the tumor. Lung adenocarcinoma tumors whose regions had consistently low levels of immune infiltration had higher levels of ITH, and those with high levels of immune infiltrate had low levels of ITH; however, heterogeneously infiltrated tumors tended to also have low levels of genomic heterogeneity (Figure 6-19). In these tumors, it is possible that earlier T-cell infiltration which had subsequently been diminished due to an acquired mutation already had an effect in narrowing the genomic landscape. Alternatively, the T-cell activation from heavily infiltrated neighboring regions may still affect clonal expansion in immune sparse tumor regions.

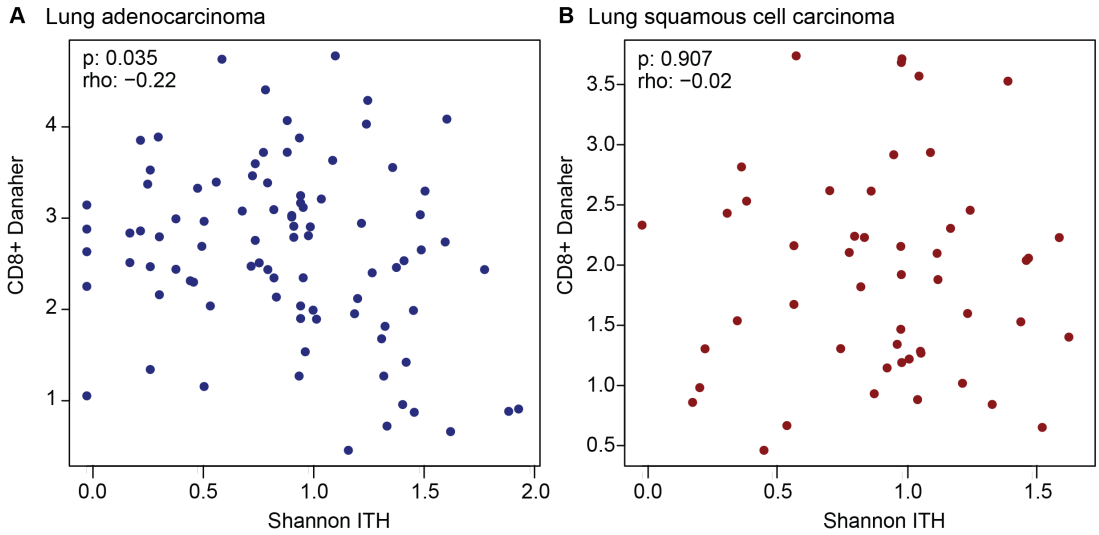

Figure 6-18: Relationship between immune infiltration and heterogeneity.
A) For each region from the lung adenocarcinoma tumors, the CD8+ T-cell score is plotted against the Shannon ITH score. Lung adenocarcinomas (A) and lung squamous cell carcinomas (B) are shown.

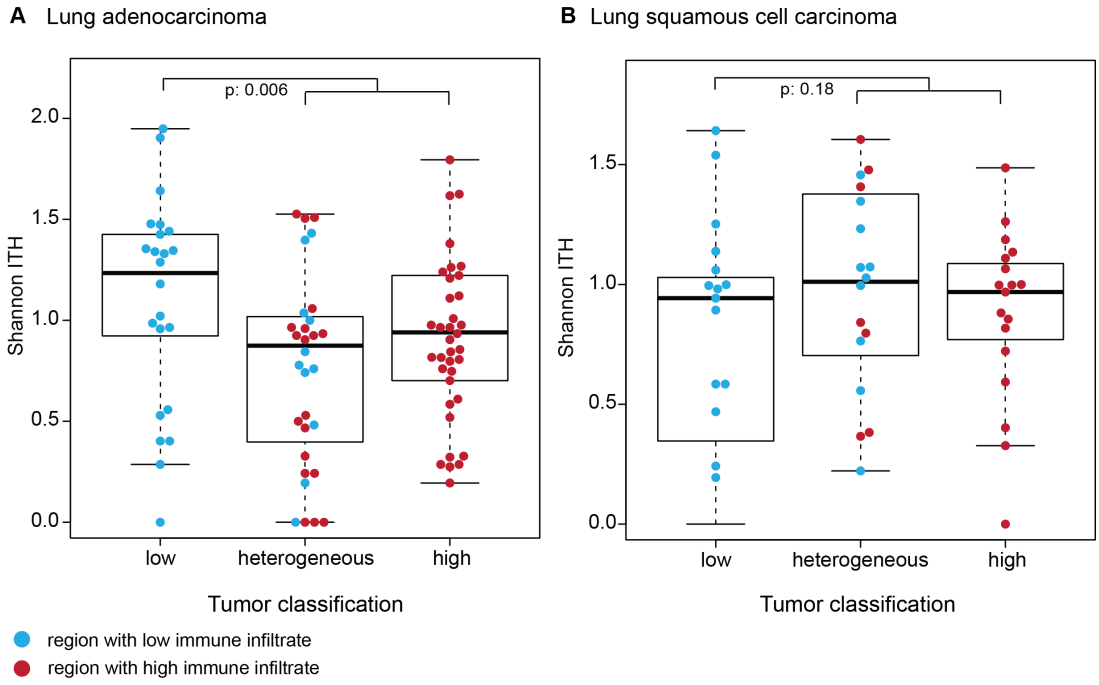

Figure 6-19: Relationship between heterogeneity and immune classification.
The Shannon ITH score for each region is shown grouped by whether the patient had consistently low levels of immune infiltration, heterogeneous levels of immune infiltration, or high levels of immune infiltration. Lung adenocarcinomas (A) and lung squamous cell carcinomas (B) are shown.

6.4.3 Immune distance mirrors with genomic distance

To better understand the associations between genomic features of a tumor and the immune microenvironment, the pairwise genomic and immune distances were calculated between every two tumor regions from the same patient. The immune

distance was determined by taking the Euclidean distance of immune infiltrate estimates between tumor regions, whereas the genomic distance was calculated by taking the Euclidean distance of the mutations present between tumor regions. Further details are contained in the Data and Methods.

Interestingly, there was a significant correlation between the two distance measures observed among both lung adenocarcinomas and lung squamous cell carcinomas (Figure 6-20), indicating that tumor regions that are more closely related to each other in genomic space also have more closely linked immune microenvironments. This is further evidence of an interaction between immune infiltrate and the tumor and suggests that there may be specific genomic events either influencing or selected in response to the immune landscape of the tumor.

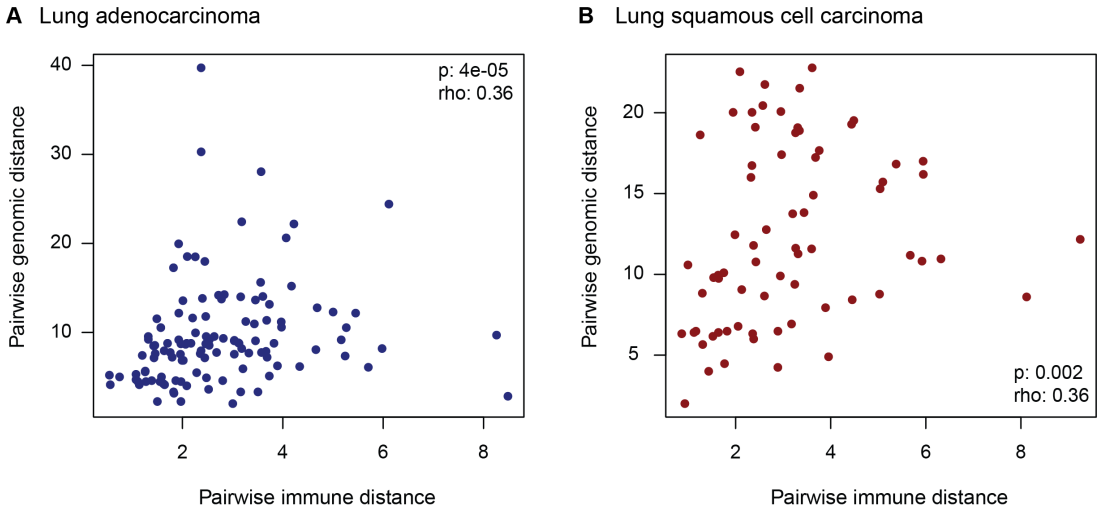

Figure 6-20: Comparison of pairwise genomic and immune distances.
The pairwise genomic and immune distances between every two tumor regions from the same patient are compared for lung adenocarcinoma (A) and lung squamous cell carcinoma (B) patients. There is a significant association between the two measures.

6.5 Genomic basis of immune infiltration

6.5.1 Elevated PD-L1 staining in HLA LOH tumors

Evidence from the previous chapter suggested that there is heavy selective pressure in tumors that develop LOH at the HLA locus, as it was frequently a subclonal event, occurring more often than expected by chance, sometimes at multiple points during the evolution of a single tumor. Furthermore, loss of the HLA alleles resulted in subsequent subclonal expansions, suggesting that upon loss of antigen presentation, the tumor had fewer restrictions on growth.

To investigate whether evidence from the tumor microenvironment supported high immune activity, immunohistochemistry analysis was performed to determine the expression of PD-L1 on both tumor and immune cells. As the PD-L1 ligand binds to the inhibitory receptor PD1, the expression of PD-L1 may reflect a response to an active immune system.

Consistent with loss of the HLA alleles occurring in immune hot microenvironments, tumors in which the HLA LOH event occurred early in evolution had significantly elevated PD-L1 staining of immune cells as compared to tumors without any HLA LOH (p=0.029) (Figure 6-21).

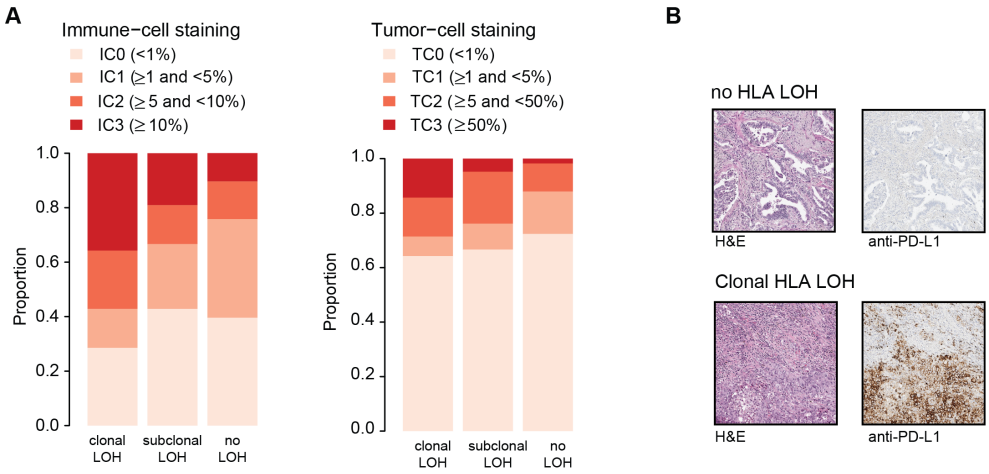

Figure 6-21: PD-L1 immunohistochemistry staining of tumors with HLA LOH.
(A) anti-PD-L1 staining on FFPE diagnostic blocks from tumors with clonal HLA LOH, subclonal HLA LOH and no observed HLA LOH. Immune-cell based staining and tumor-cell staining is depicted. (B) Staining from two representative tumors, one without HLA LOH and one with clonal HLA LOH is shown.

Non-significant trends were also observed for elevated PD-L1 staining on tumor cells. Additionally, the proportion of high PD-L1 staining tumors increased from tumors without any evidence for HLA LOH to those with subclonal LOH to those with clonal LOH.

These data are consistent with the hypothesis that HLA LOH may facilitate immune escape in response to an active immune microenvironment. The observed trends suggest that investigating the expression of checkpoint molecules and immune infiltrate in a larger cohort of tumors may help resolve whether HLA loss is associated with an immune replete microenvironment.

6.5.2 High immune infiltrate in HLA LOH tumors

By analyzing the full set of immune signatures used for immune estimation in the multi-region RNAseq data set, it was possible to investigate whether HLA loss was associated with a more active tumor microenvironment. There was RNAseq data available for 13 lung adenocarcinoma patients (29 regions) and 12 lung squamous cell carcinoma patients (36 regions) exhibiting HLA LOH and 27 lung adenocarcinoma patients (62 regions) and 3 lung squamous cell carcinoma patients (10 regions) without HLA LOH.

Supporting the results found from the PD-L1 IHC staining in the TRACERx cohort, *PDL1* was again found to be significantly up-regulated in the lung adenocarcinomas harboring loss of HLA (Figure 6-22). Furthermore, the estimated abundance of nearly every subpopulation of infiltrating immune cells was significantly elevated among lung adenocarcinoma tumor regions harboring an HLA LOH event, as was a score designed to capture the extent of immune activity (IFN score), all indicating a highly active immune microenvironment.

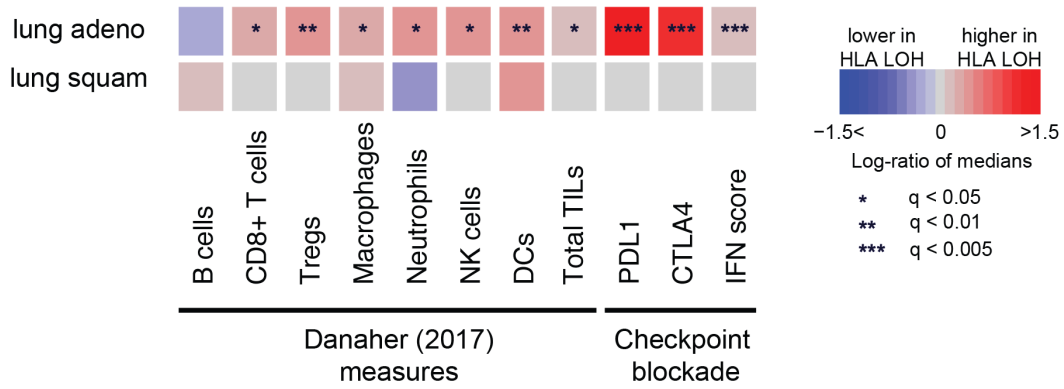

Figure 6-22: Immune signatures in TRACERx by HLA LOH status.
The log-ratio of medians between tumors containing an HLA LOH event at all loci and those without any HLA LOH event is shown for published immune microenvironment measures and signatures. Increase of an immune measure among tumors with HLA LOH is shown in red, and a decrease is shown in blue. FDR (q) values comparing the distribution of immune measures between the HLA LOH groups are indicated by asterisks (*).

To validate these findings, a larger set of TCGA NSCLC samples was used, where HLA LOH had already been identified as reported in the previous chapter. For this analysis, tumors with an HLA LOH event that affected all three loci were compared to those without any evidence of HLA LOH. For lung adenocarcinoma, 62 tumors had ubiquitous loss of the HLA alleles, compared to 265 tumors without any HLA LOH. For lung squamous cell carcinomas 77 tumors with ubiquitous HLA LOH were

compared to 179 tumors without any HLA LOH. Consistent with results observed in the TRACERx cohort, in lung adenocarcinoma with loss of the HLA locus, there was significantly elevated CD8+ T-cells and NK cells (Figure 6-23).

Together, these data suggest that tumors harboring an HLA loss event have a more active immune microenvironment, with higher immune cell infiltration observed in lung adenocarcinoma.

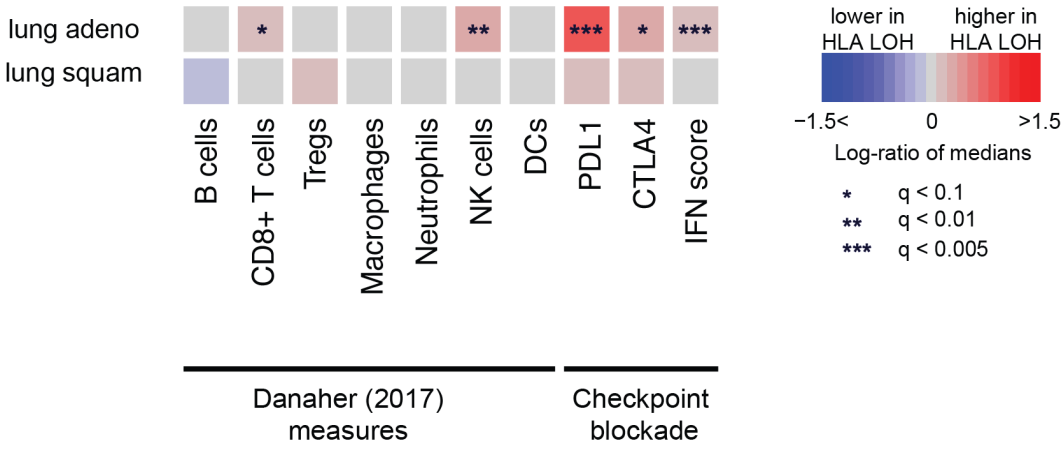

Figure 6-23: Immune signatures in TCGA by HLA LOH status.
The log-ratio of medians between tumors containing an HLA LOH event at all loci and those without any HLA LOH event is shown for published immune microenvironment measures and signatures. Increase of an immune measure among tumors with HLA LOH is shown in red, and a decrease is shown in blue. FDR (q) values comparing the distribution of immune measures between the HLA LOH groups are indicated by asterisks (*).

6.6 Conclusions

The interaction between the tumor and immune system can only be fully understood by combining genomic features of the tumor with knowledge of the immune microenvironment. To decipher the degree and composition of infiltrating immune cells using transcriptomic data, many bioinformatic methods have been published. However, these published methods often produce inconsistent estimates. Here a number of methods were compared to determine which was best suited to describe the TRACERx multi-region RNAseq data. Building on previous approaches, three criteria were considered.

Firstly, the reference genes used in immune subset definitions should exhibit a negative correlation with tumor purity in order to reflect their expression on infiltrating immune cells. Secondly, these genes should also not correlate with tumor copy number at the gene locus, as expression of the reference gene on both the immune cells and the tumor will confound any estimates of immune infiltration.

Finally, the immune cell predictions should correlate well with a ground truth measure. After analyzing potential immune signature methods with the TRACERx data, the immune subset definitions that best matched the above criteria was selected and used to determine the quantity and composition of infiltrating immune cells in each tumor region with RNAseq data. The fact that purity did not correlate with the ground truth measure used in validating the immune signatures suggested that the immune measures provide further information than would be gained from considering stromal content alone.

While a wide range of immune infiltration was observed in both the lung adenocarcinoma and lung squamous cell carcinoma subtypes, tumor regions could be classed as having either high or low immune infiltrate. Surprisingly, individual regions from the same tumor did not always closely resemble each other, often exhibiting disparate levels of immune infiltration. Indeed tumors were identified with heterogeneous levels of immune infiltration, a finding that may have implications for the robustness of predictive or prognostic immune-based biomarkers, as different conclusions may be reached on the basis of the particular sample selected.

Tumor regions containing high levels of immune infiltration appeared to differ from the sparsely infiltrated tumor regions not just in extent of infiltrate but also in composition, with highly infiltrated tumor regions also being shifted towards a higher ratio of immune effectors cells compared to immune suppressive cells. Furthermore, increased levels of effector CD8+ T-cell infiltration in lung adenocarcinoma correlated with a reduction in tumor region heterogeneity and increase in clonal neoantigen burden.

Interestingly, a difference between tumors with a heterogeneous level of immune was observed between histological subtypes. All the regions from heterogeneously infiltrated lung adenocarcinoma tumors tended to resemble tumors with high levels of immune infiltration, regardless of whether the region itself was highly immune infiltrated. This suggested that lung adenocarcinoma tumors had more of a tumor intrinsic phenotype, such that certain characteristics of the tumor were shared among all regions. On the contrary, the behavior of lung squamous cell carcinomas was driven at the regional level, with regions from heterogeneously infiltrated tumors not closely resembling each other, but rather driven by the level of immune infiltration in the sample. While there were fewer lung squamous cell carcinomas considered in this chapter, if a regional phenotype is more common in this subtype, it could be one reason why associating properties at the tumor level, such as

number of neoantigens and overall survival, have been inconclusive. As such, this is a key area for further study.

Finally a relationship between genomic features of the tumor and immune infiltration was observed. In both lung adenocarcinoma and lung squamous cell carcinoma, as the genomic distance between two tumor regions grew, so did the immune distance between those tumor regions. This suggested that specific genomic events may either influence the immune infiltration of the tumor or that they may be selected for in response to the immune landscape of the tumor.

One such event, hypothesized to be an immune evasive mechanism developed in response to a highly active immune microenvironment was LOH at the HLA locus. Indeed, supporting this hypothesis, elevated PD-L1 staining was observed among tumors exhibiting clonal HLA LOH and a significant increase in infiltrating immune cells, such as CD8+ T-cells, NK cells, and DCs, was observed in lung adenocarcinoma exhibiting HLA LOH.

While only one specific genomic event related to level of immune infiltration was identified, the observed relationship between pairwise genomic distance and pairwise immune distance indicates that there may be others. Thus, important next steps would include trying to identify any known driver alterations common in regions exhibiting high or low levels of immune infiltration. Furthermore, there may be specific immune pathway genes that are either disrupted or amplified, such as *B2M*, *PDL1,* or genes in the JAK-STAT pathway. However, given the ever-growing number of identified genes that modulate the immune system, identifying a signal in a limited number of samples and integrating that signal with poorly understood tumor extrinsic factors, such as host genetics and the microbiome, will present a challenge.

Chapter 7 Discussion

The immune system is capable of recognizing antigens present on the surface of tumor cells and eliminating them, effectively providing some protection from tumor development. While early attempts to exploit this were largely considered failures, the recent success of immunotherapies, ranging from checkpoint-blockade to neoantigen vaccines to cellular therapies, has shown that harnessing and re-directing the power of the immune system can have a significant impact of patient outcome. However, by recognizing antigenic components of the tumor cell, the immune system imposes a selective pressure on the growing cancer, shaping its antigenicity and diversity as the tumor evolves.

This thesis has investigated the mutational processes active that may generate potentially antigenic mutations and endeavored to understand what factors are important in immune recognition of the tumor, including the clonality of neoantigens present, mechanisms through which the tumor may evade detection, and the immune microenvironment.

7.1 Mutational processes and immune recognition

7.1.1 Identification of mutational signatures is possible in single tumor samples

As sequencing of patient tumors continues to become more affordable, the clinical utility of single sample analysis to understand key features of the tumor will be increasingly relevant. A more thorough understanding of mutational processes active over the course of tumor development has great implications for understanding how the tumor evolved and for informing the best patient-specific therapeutic choices (Le et al., 2015, Alexandrov et al., 2015).

As part of this thesis, a tool (deconstructSigs) was developed that allows for the determination of mutational signatures present in individual tumor samples (Chapter 3). Through the use of deconstructSigs on temporally dissected mutations, it was possible to more fully understand the dynamic nature of mutational processes active in single tumors over time. Understanding the timing of different mutational processes will help to elucidate which may most contribute to early tumorigenesis, and which may aid subsequent diversification, subclonal expansion, and potentially immune evasion.

7.1.2 Relationship between mutation generation and immune recognition

Increased mutation burden of a tumor has been previously shown to have both prognostic capabilities, associating with improved overall survival, and can be used as a biomarker predicting patient response to immunotherapy (Rizvi et al., 2015, Snyder et al., 2014, Brown et al., 2014). This, coupled with the fact that some mutational processes contribute high numbers of mutations, suggests that some mutational processes may generate more antigenic tumors than others. However, mutational processes active during tumor evolution often have distinct timings (McGranahan et al., 2015, de Bruin et al., 2014). Those that are active late during tumor evolution disproportionately contribute to intratumor heterogeneity and can result in the expansion of clones that have acquired new driver mutations. On the other hand, clonal mutations are often the result of mutational processes active early in tumor evolution. As described below, Chapter 4 of this thesis found that clonal mutations that are recognized by the immune system can result in effective neoantigens and may represent the most potent targets of immune activation. Thus by understanding the contribution and dynamics of mutational processes, it may be possible to select specific patients likely to respond to immunotherapeutic interventions.

7.1.3 Clonal neoantigens elicit T-cell responses and influence patient survival

In this thesis (Chapter 4), *in silico* neoantigen prediction made using a bioinformatic pipeline were tested and validated *in vitro* using T-cells collected from individual patient tumors. Interestingly, among both checkpoint-blockade treated cohorts and treatment-naïve patients, all observable T-cell responses were identified against clonal peptides. Furthermore, patients with a large number of clonal neoantigens and few subclonal neoantigens were found to fare better both without treatment and in response to checkpoint blockade therapy. Together these findings suggested that clonal neoantigens and subclonal neoantigens differentially elicited an immune response and highlighted the potential gain in immunotherapy efficiency that could be made by considering the clonality of tumor neoantigens prior to treatment.

While the cohort investigated in this thesis was small, no neoantigen reactive T-cells were identified that recognized peptides arising from subclonal mutations. As this may have been confounded by the challenge of making accurate subclonal mutation calls, further study in a larger cohort is warranted to determine whether

subclonal mutations generate peptides capable of being recognized by T-cells. However, even if an immune response is elicited, the presentation of a subclonal neoantigen on the surface of a subset of tumor cells will likely inherently limit the ability of the T-cells to target the entire tumor. A recent publication suggests that inactivating DNA repair processes in the tumor to increase the burden of neoantigens may be exploited therapeutically, without regard for the clonality of the mutations generated. Thus determining if subclonal neoantigens are indeed capable of driving effective anti-tumor immunity is of paramount importance for patient care (Germano et al., 2017).

As more neoantigen reactive T-cells are identified, the properties determining what leads to effective antigen presentation can be further elucidated. This will allow for the prioritization of peptides likely to be antigenic based on the most discriminating attributes, such as expression, predicted binding affinity, how the mutant binding affinity compares to the wildtype binding affinity, similarity to self, similarity to known antigens, and likelihood of *in vivo* processing. Currently, in an effort to circumvent the difficulties predicting *in vivo* antigen processing, large-scale mass spectrometry approaches are underway to determine which neoantigens are being recognized from the tumor (Abelin et al., 2017, Bassani-Sternberg et al., 2017, Muller et al., 2017). Importantly, by being prediction-agnostic, mass spectrometry approaches could also allow for the identification of peptides presented by MHC class II molecules (Muller et al., 2017, Gfeller et al., 2016), which have proven even harder to predict than their MHC class I counterparts.

Additionally fitness models, similar to ones used to model immune interactions in viral disease, have been recently developed. These models aim to determine the similarity between presented peptides and known T-cell antigens in order to better predict whether a given peptide is likely to be recognized once it is successfully processed and presented on the tumor cell (Luksza et al., 2017).

These developments will all serve to improve neoantigen predictions and can subsequently be used to refine neoantigen predictions to more completely understand the complex relationship between neoantigen burden, ITH, and tumor immunogenicity.

7.2 HLA LOH as an immune evasive mechanism in NSCLC

Under the immune editing hypothesis, the selective pressure of the immune system can result in a tumor acquiring mechanisms to evade immune detection in order to avoid immune predation. As neoantigens are displayed on the cell surface for potential immune recognition, disruption of antigen presentation is one possible avenue to escape immune detection. Extending the observations of reversible HLA down-regulation found in a number of cancer types (Hicklin et al., 1999, Garrido et al., 2017a, Campoli et al.), the LOHHLA method was developed as a part of this thesis (Chapter 5) to detect irreversible LOH at the HLA locus. LOHHLA exploited patient specific HLA information in order to overcome the challenges posed by the highly polymorphic nature of the HLA locus and its resulting poor mapability.

7.2.1 HLA LOH occurs under heavy selection late in tumor evolution

Keeping with the notion that defective antigen presentation may play a key role during tumor evolution, LOHHLA identified HLA LOH in 40% of the NSCLC samples analyzed. This was in contrast to the low frequency (~3%) of HLA mutations cataloged in the cohort analyzed, suggesting that HLA LOH may represent a far more prevalent means of antigen presentation disruption in NSCLC.

Mapping LOH events to the phylogenetic tree allowed for the characterization of the timing of HLA LOH. HLA LOH was often a late occurrence in tumor evolution, occurring subclonally in only a portion of tumor cells. There also appeared to be strong selection for HLA LOH, as it occurred multiple times over the course of a single tumor's evolutionary history, with consistent HLA alleles subject to loss during each event. This implied preferential selection for the loss of one set of alleles over the other. Formal testing using simulated focal LOH events found that LOH at the HLA locus occurred more frequently than expected by chance. Together, these observations suggested HLA LOH is strongly selected for in NSCLC evolution.

While lung adenocarcinoma and lung squamous cell carcinoma were the only tumor types investigated in this chapter, aneuploidy is a hallmark of cancer, and other LOH events are commonly observed in tumor development (Ryland et al., 2015, Merajver et al., 1995, Knudson, 1971). As such, further work is warranted to determine the frequency of HLA LOH across other tumor types and its association with immune evasion.

7.2.2 HLA LOH is permissive for subclonal expansion

Further analysis showed that tumor subclones harboring an HLA LOH event had a significantly higher non-synonymous mutation burden as compared to sister subclones descended from the same ancestral cancer cell but without HLA LOH. Tumors exhibiting HLA LOH also showed an enrichment for neoantigens predicted to bind to the lost allele. Together, these results suggest that decreased antigen recognition following loss of HLA alleles may be permissive for subclonal expansions and could allow mutations that may have once instigated an immune response to go undetected by the immune system.

One hypothesis generated from this work is that HLA LOH facilitates immune escape in response to an active immune microenvironment. Indeed, a single case study following a patient treated with tumor-infiltrating lymphocytes composed of T-cell clones reactive to a particular neoantigen observed loss of the HLA allele responsible for that neoantigen's presentation (Tran et al., 2016). Thus while the cohorts analyzed with LOHHLA were treatment-naïve, it would be expected that HLA LOH may also be a mechanism of acquired resistance to checkpoint blockade therapy.

A recent publication has since investigated the HLA locus of patients treated with checkpoint blockade therapy and found that patients harboring an HLA LOH event had poorer survival (Chowell et al., 2017). While this study only considered pre-treatment samples, as more data is published from patients who have relapsed on immunotherapy, the prevalence of HLA LOH as a mechanism of immune evasion can be investigated further.

7.3 Immune microenvironment is heterogeneous in NSCLC

As the picture of tumor immunity is incomplete without also considering the immune microenvironment, Chapter 6 of this thesis used multi-region RNAseq data to quantify immune infiltration and to explore the extent to which the immune contexture varied across different regions from the same tumor. A re-analysis of published immune signatures led to the identification of a set of signatures that were not informed by immune genes also expressed by the tumor and could be validated using *in vitro* data. Thus the degree of immune infiltration and its cell composition could be determined on a region-by-region basis.

Paralleling the genetic ITH reported in NSCLC, Chapter 6 found that regions from the same tumor exhibited vastly different immune infiltration estimates. Not only did the abundance of infiltration vary between tumor regions, but they also exhibited different immune cell composition, with highly infiltrated tumor regions harboring a significantly higher ratio of immune effector cells to immune suppressive cells. These findings also highlighted a potential limitation of using immune infiltration estimates generated from a single tumor sample to guide therapeutic decisions, as the information gathered from a single tumor region may not be accurately reflect the immune landscape of the entire tumor.

7.3.1 Genomic events reflect shifts in immune contexture

With increasing levels of effector CD8+ T-cell infiltration in lung adenocarcinoma, a reduction in tumor region heterogeneity and increase in clonal neoantigen burden was also observed, providing evidence for a relationship between the immune cells infiltrating the tumor and the genomic landscape. In further support of a tumor/immune interaction, a relationship was observed between the pair-wise genomic distance and pair-wise immune distances calculated for every two regions of the same tumor.

By investigating the immune microenvironment it was possible to identify whether specific genomic events either influence or are selected in response to the immune landscape of the tumor.

One such genomic event suspected to arise due to the strong selective pressure of the immune system was HLA LOH. Indeed, consistent with loss of the HLA alleles occurring in immune hot microenvironments and potentially as an immune evasive mechanism, lung adenocarcinoma tumors exhibiting LOH at the HLA locus had a significant increase in immune infiltrate, including CD8+ T-cells and NK cells. These results were recapitulated using a larger cohort of lung adenocarcinomas from TCGA. Furthermore, PD-L1 staining of immune cells was performed and found to be elevated among tumors in which the HLA LOH event occurred early in evolution.

Using bulk transcriptomic data presents inherent challenges in deciphering immune cell subpopulations, particularly as all current deconvolution or marker gene approaches rely on reference genes expressed identified from purified immune cells. As shown in this thesis, tumor cells, as well as the infiltrating immune cells, may express these reference immune genes. Furthermore, the expression profiles

of tumor infiltrating immune cells may not be well represented by immune cells purified from the peripheral blood. To overcome such challenges, the analysis of single cells may be required using methods such as CyTOF (Newell et al., 2012, Chevrier et al., 2017, Lavin et al., 2017) or single-cell RNAseq (Tirosh et al., 2016, Singer et al., 2017).

Finally, as more aspects contributing to an effective immune response are understood and as more data is generated, it may be possible to design a comprehensive model to predict the tumor immune interaction and patient response to immunotherapy. Such a model would have to consider tumor intrinsic properties, such as genomic events altering the immune landscape (i.e. disrupted antigen presentation or altered/unresponsive immune pathways), tumor cytokine secretion, and epigenetic events (i.e. demethylation of the *PD-L1* promoter (Gettinger et al., 2015)). Incorporating tumor genomic features with TCR/BCR sequencing and monitoring the immune repertoire may also help to shed light on how a patient is responding (Liu and Mardis, 2017). In addition, tumor extrinsic properties are likely to play a role in dictating the immune response. Host genetics, such as the particular HLA haplotype of an individual, can influence response to immunotherapy (Chowell et al., 2017), and the impact of recent/current infection and the gut microbiome remains under-studied (Routy et al., 2017, Gopalakrishnan et al., 2017).

As small variations in any one of these components could tip the immune balance in favor of immunity or tolerance (Chen and Mellman, 2017), it will take a nuanced understanding to decipher how they connect with and impact each other, likely requiring much more data.

References

Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., Clauser, K. R., Hacohen, N., Rooney, M. S., Carr, S. A. & Wu, C. J. 2017. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity,* 46**,** 315-326.

Alexandrov, L. B., Nik-Zainal, S., Siu, H. C., Leung, S. Y. & Stratton, M. R. 2015. A mutational signature in gastric cancer suggests therapeutic strategies. *Nat Commun,* 6**,** 8683.

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

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. 2013b. Deciphering signatures of mutational processes operative in human cancer. *Cell reports,* 3**,** 246-59.

Andersen, R. S., Kvistborg, P., Frosig, T. M., Pedersen, N. W., Lyngaa, R., Bakker, A. H., Shu, C. J., Straten, P., Schumacher, T. N. & Hadrup, S. R. 2012. Parallel detection of antigen-specific T cell responses by combinatorial encoding of MHC multimers. *Nat Protoc,* 7**,** 891-902.

Andreatta, M. & Nielsen, M. 2016. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics,* 32**,** 511-7.

Angelova, M., Charoentong, P., Hackl, H., Fischer, M. L., Snajder, R., Krogsdam, A. M., Waldner, M. J., Bindea, G., Mlecnik, B., Galon, J. & Trajanoski, Z. 2015. Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol,* 16**,** 64.

Bai, Y., Ni, M., Cooper, B., Wei, Y. & Fury, W. 2014. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genomics,* 15**,** 325.

Bakker, A. H., Hoppes, R., Linnemann, C., Toebes, M., Rodenko, B., Berkers, C. R., Hadrup, S. R., van Esch, W. J., Heemskerk, M. H., Ovaa, H. & Schumacher, T. N. 2008. Conditional MHC class I ligands

and peptide exchange technology for the human MHC gene products HLA-A1, -A3, -A11, and -B7. *Proc Natl Acad Sci U S A,* 105**,** 3825-30.
Balkwill, F. & Mantovani, A. 2001. Inflammation and cancer: back to Virchow? *Lancet,* 357**,** 539-45.
Barnes, D. W., Corp, M. J., Loutit, J. F. & Neal, F. E. 1956. Treatment of murine leukaemia with X rays and homologous bone marrow; preliminary communication. *Br Med J,* 2**,** 626-7.
Bashashati, A., Ha, G., Tone, A., Ding, J., Prentice, L. M., Roth, A., Rosner, J., Shumansky, K., Kalloger, S., Senz, J., Yang, W., McConechy, M., Melnyk, N., Anglesio, M., Luk, M. T., Tse, K., Zeng, T., Moore, R., Zhao, Y., Marra, M. A., Gilks, B., Yip, S., Huntsman, D. G., McAlpine, J. N. & Shah, S. P. 2013. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J Pathol,* 231**,** 21-34.
Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P. O., Kandalaft, L. E., Coukos, G. & Gfeller, D. 2017. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput Biol,* 13**,** e1005725.
Bates, G. J., Fox, S. B., Han, C., Leek, R. D., Garcia, J. F., Harris, A. L. & Banham, A. H. 2006. Quantification of regulatory T cells enables the identification of high-risk breast cancer patients and those at risk of late relapse. *J Clin Oncol,* 24**,** 5373-80.
Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautes-Fridman, C., Fridman, W. H. & de Reynies, A. 2016. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol,* 17**,** 218.
Benci, J. L., Xu, B., Qiu, Y., Wu, T. J., Dada, H., Twyman-Saint Victor, C., Cucolo, L., Lee, D. S. M., Pauken, K. E., Huang, A. C., Gangadhar, T. C., Amaravadi, R. K., Schuchter, L. M., Feldman, M. D., Ishwaran, H., Vonderheide, R. H., Maity, A., Wherry, E. J. & Minn, A. J. 2016. Tumor Interferon Signaling Regulates a Multigenic Resistance Program to Immune Checkpoint Blockade. *Cell,* 167**,** 1540-1554 e12.
Benitez, R., Godelaine, D., Lopez-Nevot, M. A., Brasseur, F., Jimenez, P., Marchand, M., Oliva, M. R., van Baren, N., Cabrera, T., Andry, G., Landry, C., Ruiz-Cabello, F., Boon, T. & Garrido, F. 1998. Mutations of the beta2-microglobulin gene result in a lack of HLA class I molecules on melanoma cells of two patients immunized with MAGE peptides. *Tissue Antigens,* 52**,** 520-9.
Bhasin, M., Lata, S. & Raghava, G. P. 2007. TAPPred prediction of TAP-binding peptides in antigens. *Methods Mol Biol,* 409**,** 381-6.
Bhattacharyya, N. P., Skandalis, A., Ganesh, A., Groden, J. & Meuth, M. 1994. Mutator phenotypes in human colorectal carcinoma cell lines. *Proc Natl Acad Sci U S A,* 91**,** 6319-23.
Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., Angell, H., Fredriksen, T., Lafontaine, L., Berger, A., Bruneval, P., Fridman, W. H., Becker, C., Pages, F., Speicher, M. R., Trajanoski, Z. & Galon, J. 2013. Spatiotemporal dynamics of intratumoral immune

cells reveal the immune landscape in human cancer. *Immunity,* 39**,** 782-95.

Boegel, S., Lower, M., Schafer, M., Bukur, T., de Graaf, J., Boisguerin, V., Tureci, O., Diken, M., Castle, J. C. & Sahin, U. 2012. HLA typing from RNA-Seq sequence reads. *Genome Med,* 4**,** 102.

Boland, C. R. & Goel, A. 2010. Microsatellite instability in colorectal cancer. *Gastroenterology,* 138**,** 2073-2087 e3.

Bolli, N., Avet-Loiseau, H., Wedge, D. C., Van Loo, P., Alexandrov, L. B., Martincorena, I., Dawson, K. J., Iorio, F., Nik-Zainal, S., Bignell, G. R., Hinton, J. W., Li, Y., Tubio, J. M., McLaren, S., S, O. M., Butler, A. P., Teague, J. W., Mudie, L., Anderson, E., Rashid, N., Tai, Y. T., Shammas, M. A., Sperling, A. S., Fulciniti, M., Richardson, P. G., Parmigiani, G., Magrangeas, F., Minvielle, S., Moreau, P., Attal, M., Facon, T., Futreal, P. A., Anderson, K. C., Campbell, P. J. & Munshi, N. C. 2014. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun,* 5**,** 2997.

Brahmer, J. R., Tykodi, S. S., Chow, L. Q., Hwu, W. J., Topalian, S. L., Hwu, P., Drake, C. G., Camacho, L. H., Kauh, J., Odunsi, K., Pitot, H. C., Hamid, O., Bhatia, S., Martins, R., Eaton, K., Chen, S., Salay, T. M., Alaparthy, S., Grosso, J. F., Korman, A. J., Parker, S. M., Agrawal, S., Goldberg, S. M., Pardoll, D. M., Gupta, A. & Wigginton, J. M. 2012. Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. *The New England journal of medicine,* 366**,** 2455-65.

Brastianos, P. K., Carter, S. L., Santagata, S., Cahill, D. P., Taylor-Weiner, A., Jones, R. T., Van Allen, E. M., Lawrence, M. S., Horowitz, P. M., Cibulskis, K., Ligon, K. L., Tabernero, J., Seoane, J., Martinez-Saez, E., Curry, W. T., Dunn, I. F., Paek, S. H., Park, S. H., McKenna, A., Chevalier, A., Rosenberg, M., Barker, F. G., 2nd, Gill, C. M., Van Hummelen, P., Thorner, A. R., Johnson, B. E., Hoang, M. P., Choueiri, T. K., Signoretti, S., Sougnez, C., Rabin, M. S., Lin, N. U., Winer, E. P., Stemmer-Rachamimov, A., Meyerson, M., Garraway, L., Gabriel, S., Lander, E. S., Beroukhim, R., Batchelor, T. T., Baselga, J., Louis, D. N., Getz, G. & Hahn, W. C. 2015. Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets. *Cancer Discov,* 5**,** 1164-1177.

Brown, S. D., Warren, R. L., Gibb, E. A., Martin, S. D., Spinelli, J. J., Nelson, B. H. & Holt, R. A. 2014. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res,* 24**,** 743-50.

Burnet, M. 1957. Cancer; a biological approach. I. The processes of control. *Br Med J,* 1**,** 779-86.

Cahill, D. P., Levine, K. K., Betensky, R. A., Codd, P. J., Romany, C. A., Reavie, L. B., Batchelor, T. T., Futreal, P. A., Stratton, M. R., Curry, W. T., Iafrate, A. J. & Louis, D. N. 2007. Loss of the mismatch repair protein MSH6 in human glioblastomas is associated with tumor progression during temozolomide treatment. *Clin Cancer Res,* 13**,** 2038-45.

Calbo, J., van Montfort, E., Proost, N., van Drunen, E., Beverloo, H. B., Meuwissen, R. & Berns, A. 2011. A functional role for tumor cell

heterogeneity in a mouse model of small cell lung cancer. *Cancer Cell,* 19**,** 244-56.

Campbell, J. D., Alexandrov, A., Kim, J., Wala, J., Berger, A. H., Pedamallu, C. S., Shukla, S. A., Guo, G., Brooks, A. N., Murray, B. A., Imielinski, M., Hu, X., Ling, S., Akbani, R., Rosenberg, M., Cibulskis, C., Ramachandran, A., Collisson, E. A., Kwiatkowski, D. J., Lawrence, M. S., Weinstein, J. N., Verhaak, R. G., Wu, C. J., Hammerman, P. S., Cherniack, A. D., Getz, G., Cancer Genome Atlas Research, N., Artyomov, M. N., Schreiber, R., Govindan, R. & Meyerson, M. 2016. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat Genet,* 48**,** 607-16.

Campoli, M., Chang, C. C. & Ferrone, S. 2002. HLA class I antigen loss, tumor immune escape and immune selection. *Vaccine,* 20 Suppl 4**,** A40-5.

Carreras, J., Lopez-Guillermo, A., Fox, B. C., Colomo, L., Martinez, A., Roncador, G., Montserrat, E., Campo, E. & Banham, A. H. 2006. High numbers of tumor-infiltrating FOXP3-positive regulatory T cells are associated with improved overall survival in follicular lymphoma. *Blood,* 108**,** 2957-64.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M. & Getz, G. 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol,* 30**,** 413-21.

Castle, J. C., Kreiter, S., Diekmann, J., Lower, M., van de Roemer, N., de Graaf, J., Selmi, A., Diken, M., Boegel, S., Paret, C., Koslowski, M., Kuhn, A. N., Britten, C. M., Huber, C., Tureci, O. & Sahin, U. 2012. Exploiting the mutanome for tumor vaccination. *Cancer Res,* 72**,** 1081-91.

Chan, K., Roberts, S. A., Klimczak, L. J., Sterling, J. F., Saini, N., Malc, E. P., Kim, J., Kwiatkowski, D. J., Fargo, D. C., Mieczkowski, P. A., Getz, G. & Gordenin, D. A. 2015. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet*.

Chang, J., Tan, W., Ling, Z., Xi, R., Shao, M., Chen, M., Luo, Y., Zhao, Y., Liu, Y., Huang, X., Xia, Y., Hu, J., Parker, J. S., Marron, D., Cui, Q., Peng, L., Chu, J., Li, H., Du, Z., Han, Y., Tan, W., Liu, Z., Zhan, Q., Li, Y., Mao, W., Wu, C. & Lin, D. 2017. Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. *Nat Commun,* 8**,** 15290.

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., Hackl, H. & Trajanoski, Z. 2017. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep,* 18**,** 248-262.

Chen, D. S. & Mellman, I. 2017. Elements of cancer immunity and the cancer-immune set point. *Nature,* 541**,** 321-330.

Chen, P. L., Roh, W., Reuben, A., Cooper, Z. A., Spencer, C. N., Prieto, P. A., Miller, J. P., Bassett, R. L., Gopalakrishnan, V., Wani, K., De

Macedo, M. P., Austin-Breneman, J. L., Jiang, H., Chang, Q., Reddy, S. M., Chen, W. S., Tetzlaff, M. T., Broaddus, R. J., Davies, M. A., Gershenwald, J. E., Haydu, L., Lazar, A. J., Patel, S. P., Hwu, P., Hwu, W. J., Diab, A., Glitza, I. C., Woodman, S. E., Vence, L. M., Wistuba, II, Amaria, R. N., Kwong, L. N., Prieto, V., Davis, R. E., Ma, W., Overwijk, W. W., Sharpe, A. H., Hu, J., Futreal, P. A., Blando, J., Sharma, P., Allison, J. P., Chin, L. & Wargo, J. A. 2016. Analysis of Immune Signatures in Longitudinal Tumor Samples Yields Insight into Biomarkers of Response and Mechanisms of Resistance to Immune Checkpoint Blockade. *Cancer Discov,* 6**,** 827-37.

Chevrier, S., Levine, J. H., Zanotelli, V. R. T., Silina, K., Schulz, D., Bacac, M., Ries, C. H., Ailles, L., Jewett, M. A. S., Moch, H., van den Broek, M., Beisel, C., Stadler, M. B., Gedye, C., Reis, B., Pe'er, D. & Bodenmiller, B. 2017. An Immune Atlas of Clear Cell Renal Cell Carcinoma. *Cell,* 169**,** 736-749 e18.

Chomez, P., De Backer, O., Bertrand, M., De Plaen, E., Boon, T. & Lucas, S. 2001. An overview of the MAGE gene family with the identification of all human members of the family. *Cancer Res,* 61**,** 5544-51.

Chowell, D., Morris, L. G. T., Grigg, C. M., Weber, J. K., Samstein, R. M., Makarov, V., Kuo, F., Kendall, S. M., Requena, D., Riaz, N., Greenbaum, B., Carroll, J., Garon, E., Hyman, D. M., Zehir, A., Solit, D., Berger, M., Zhou, R., Rizvi, N. A. & Chan, T. A. 2017. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science*.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S. & Getz, G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol,* 31**,** 213-9.

Clemente, C. G., Mihm, M. C., Jr., Bufalino, R., Zurrida, S., Collini, P. & Cascinelli, N. 1996. Prognostic value of tumor infiltrating lymphocytes in the vertical growth phase of primary cutaneous melanoma. *Cancer,* 77**,** 1303-10.

Collins, T., Korman, A. J., Wake, C. T., Boss, J. M., Kappes, D. J., Fiers, W., Ault, K. A., Gimbrone, M. A., Jr., Strominger, J. L. & Pober, J. S. 1984. Immune interferon activates multiple class II major histocompatibility complex genes and the associated invariant chain gene in human endothelial cells and dermal fibroblasts. *Proc Natl Acad Sci U S A,* 81**,** 4917-21.

Connor, A. A., Denroche, R. E., Jang, G. H., Timms, L., Kalimuthu, S. N., Selander, I., McPherson, T., Wilson, G. W., Chan-Seng-Yue, M. A., Borozan, I., Ferretti, V., Grant, R. C., Lungu, I. M., Costello, E., Greenhalf, W., Palmer, D., Ghaneh, P., Neoptolemos, J. P., Buchler, M., Petersen, G., Thayer, S., Hollingsworth, M. A., Sherker, A., Durocher, D., Dhani, N., Hedley, D., Serra, S., Pollett, A., Roehrl, M. H. A., Bavi, P., Bartlett, J. M. S., Cleary, S., Wilson, J. M., Alexandrov, L. B., Moore, M., Wouters, B. G., McPherson, J. D., Notta, F., Stein, L. D. & Gallinger, S. 2017. Association of Distinct Mutational Signatures With Correlates of Increased Immune Activity in Pancreatic Ductal Adenocarcinoma. *JAMA Oncol,* 3**,** 774-783.

Coulie, P. G., Lehmann, F., Lethe, B., Herman, J., Lurquin, C., Andrawiss, M. & Boon, T. 1995. A mutated intron sequence codes for an antigenic peptide recognized by cytolytic T lymphocytes on a human melanoma. *Proc Natl Acad Sci U S A,* 92**,** 7976-80.

D'Urso, C. M., Wang, Z. G., Cao, Y., Tatake, R., Zeff, R. A. & Ferrone, S. 1991. Lack of HLA class I antigen expression by cultured melanoma cells FO-1 due to a defect in B2m gene expression. *J Clin Invest,* 87**,** 284-92.

Danaher, P., Warren, S., Dennis, L., D'Amico, L., White, A., Disis, M. L., Geller, M. A., Odunsi, K., Beechem, J. & Fling, S. P. 2017. Gene expression markers of Tumor Infiltrating Leukocytes. *J Immunother Cancer,* 5**,** 18.

Davies, H., Glodzik, D., Morganella, S., Yates, L. R., Staaf, J., Zou, X., Ramakrishna, M., Martin, S., Boyault, S., Sieuwerts, A. M., Simpson, P. T., King, T. A., Raine, K., Eyfjord, J. E., Kong, G., Borg, A., Birney, E., Stunnenberg, H. G., van de Vijver, M. J., Borresen-Dale, A. L., Martens, J. W., Span, P. N., Lakhani, S. R., Vincent-Salomon, A., Sotiriou, C., Tutt, A., Thompson, A. M., Van Laere, S., Richardson, A. L., Viari, A., Campbell, P. J., Stratton, M. R. & Nik-Zainal, S. 2017a. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med,* 23**,** 517-525.

Davies, H., Morganella, S., Purdie, C. A., Jang, S. J., Borgen, E., Russnes, H., Glodzik, D., Zou, X., Viari, A., Richardson, A. L., Borresen-Dale, A. L., Thompson, A., Eyfjord, J. E., Kong, G., Stratton, M. R. & Nik-Zainal, S. 2017b. Whole-Genome Sequencing Reveals Breast Cancers with Mismatch Repair Deficiency. *Cancer Res,* 77**,** 4755-4762.

Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. 2017. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science,* 355.

de Bruin, E. C., McGranahan, N., Mitter, R., Salm, M., Wedge, D. C., Yates, L., Jamal-Hanjani, M., Shafi, S., Murugaesu, N., Rowan, A. J., Gronroos, E., Muhammad, M. A., Horswell, S., Gerlinger, M., Varela, I., Jones, D., Marshall, J., Voet, T., Van Loo, P., Rassl, D. M., Rintoul, R. C., Janes, S. M., Lee, S. M., Forster, M., Ahmad, T., Lawrence, D., Falzon, M., Capitanio, A., Harkins, T. T., Lee, C. C., Tom, W., Teefe, E., Chen, S. C., Begum, S., Rabinowitz, A., Phillimore, B., Spencer-Dene, B., Stamp, G., Szallasi, Z., Matthews, N., Stewart, A., Campbell, P. & Swanton, C. 2014. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science,* 346**,** 251-6.

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

Diaz, L. A., Jr., Williams, R. T., Wu, J., Kinde, I., Hecht, J. R., Berlin, J., Allen, B., Bozic, I., Reiter, J. G., Nowak, M. A., Kinzler, K. W., Oliner, K. S. & Vogelstein, B. 2012. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature,* 486**,** 537-40.

Diez-Rivero, C. M., Chenlo, B., Zuluaga, P. & Reche, P. A. 2010. Quantitative modeling of peptide binding to TAP using support vector machine. *Proteins,* 78**,** 63-72.

Dighe, A. S., Richards, E., Old, L. J. & Schreiber, R. D. 1994. Enhanced in vivo growth and resistance to rejection of tumor cells expressing dominant negative IFN gamma receptors. *Immunity,* 1**,** 447-56.

DiLillo, D. J., Matsushita, T. & Tedder, T. F. 2010. B10 cells and regulatory B cells balance immune responses during inflammation, autoimmunity, and cancer. *Ann N Y Acad Sci,* 1183**,** 38-57.

Ding, L., Kim, M., Kanchi, K. L., Dees, N. D., Lu, C., Griffith, M., Fenstermacher, D., Sung, H., Miller, C. A., Goetz, B., Wendl, M. C., Griffith, O., Cornelius, L. A., Linette, G. P., McMichael, J. F., Sondak, V. K., Fields, R. C., Ley, T. J., Mule, J. J., Wilson, R. K. & Weber, J. S. 2014. Clonal architectures and driver mutations in metastatic melanomas. *PLoS One,* 9**,** e111153.

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., Ritchey, J. K., Young, M. A., Lamprecht, T., McLellan, M. D., McMichael, J. F., Wallis, J. W., Lu, C., Shen, D., Harris, C. C., Dooling, D. J., Fulton, R. S., Fulton, L. L., Chen, K., Schmidt, H., Kalicki-Veizer, J., Magrini, V. J., Cook, L., McGrath, S. D., Vickery, T. L., Wendl, M. C., Heath, S., Watson, M. A., Link, D. C., Tomasson, M. H., Shannon, W. D., Payton, J. E., Kulkarni, S., Westervelt, P., Walter, M. J., Graubert, T. A., Mardis, E. R., Wilson, R. K. & DiPersio, J. F. 2012. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature,* 481**,** 506-10.

Drake, C. G., Jaffee, E. & Pardoll, D. M. 2006. Mechanisms of immune evasion by tumors. *Adv Immunol,* 90**,** 51-81.

Dudley, M. E., Wunderlich, J. R., Robbins, P. F., Yang, J. C., Hwu, P., Schwartzentruber, D. J., Topalian, S. L., Sherry, R., Restifo, N. P., Hubicki, A. M., Robinson, M. R., Raffeld, M., Duray, P., Seipp, C. A., Rogers-Freezer, L., Morton, K. E., Mavroukakis, S. A., White, D. E. & Rosenberg, S. A. 2002a. Cancer regression and autoimmunity in patients after clonal repopulation with antitumor lymphocytes. *Science,* 298**,** 850-4.

Dudley, M. E., Wunderlich, J. R., Yang, J. C., Hwu, P., Schwartzentruber, D. J., Topalian, S. L., Sherry, R. M., Marincola, F. M., Leitman, S. F., Seipp, C. A., Rogers-Freezer, L., Morton, K. E., Nahvi, A., Mavroukakis, S. A., White, D. E. & Rosenberg, S. A. 2002b. A phase I study of nonmyeloablative chemotherapy and adoptive transfer of autologous tumor antigen-specific T lymphocytes in patients with metastatic melanoma. *J Immunother,* 25**,** 243-51.

Dunn, G. P., Bruce, A. T., Ikeda, H., Old, L. J. & Schreiber, R. D. 2002. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol,* 3**,** 991-8.

DuPage, M., Mazumdar, C., Schmidt, L. M., Cheung, A. F. & Jacks, T. 2012. Expression of tumour-specific antigens underlies cancer immunoediting. *Nature,* 482**,** 405-9.

Farmer, H., McCabe, N., Lord, C. J., Tutt, A. N., Johnson, D. A., Richardson, T. B., Santarosa, M., Dillon, K. J., Hickson, I., Knights, C., Martin, N. M., Jackson, S. P., Smith, G. C. & Ashworth, A. 2005. Targeting the

DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature,* 434**,** 917-21.

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

Fidler, I. J. 1978. Tumor heterogeneity and the biology of cancer invasion and metastasis. *Cancer Res,* 38**,** 2651-60.

Findlay, J. M., Castro-Giner, F., Makino, S., Rayner, E., Kartsonaki, C., Cross, W., Kovac, M., Ulahannan, D., Palles, C., Gillies, R. S., MacGregor, T. P., Church, D., Maynard, N. D., Buffa, F., Cazier, J. B., Graham, T. A., Wang, L. M., Sharma, R. A., Middleton, M. & Tomlinson, I. 2016. Differential clonal evolution in oesophageal cancers in response to neo-adjuvant chemotherapy. *Nat Commun,* 7**,** 11111.

Fisk, B., Savary, C., Hudson, J. M., O'Brian, C. A., Murray, J. L., Wharton, J. T. & Ioannides, C. G. 1995. Changes in an HER-2 peptide upregulating HLA-A2 expression affect both conformational epitopes and CTL recognition: implications for optimization of antigen presentation and tumor-specific CTL induction. *J Immunother Emphasis Tumor Immunol,* 18**,** 197-209.

Frey, D. M., Droeser, R. A., Viehl, C. T., Zlobec, I., Lugli, A., Zingg, U., Oertli, D., Kettelhack, C., Terracciano, L. & Tornillo, L. 2010. High frequency of tumor-infiltrating FOXP3(+) regulatory T cells predicts improved survival in mismatch repair-proficient colorectal cancer patients. *Int J Cancer,* 126**,** 2635-43.

Fridman, W. H., Pages, F., Sautes-Fridman, C. & Galon, J. 2012. The immune contexture in human tumours: impact on clinical outcome. *Nat Rev Cancer,* 12**,** 298-306.

Fritsch, E. F., Rajasagi, M., Ott, P. A., Brusic, V., Hacohen, N. & Wu, C. J. 2014. HLA-binding properties of tumor neoepitopes in humans. *Cancer Immunol Res,* 2**,** 522-9.

Frosig, T. M., Yap, J., Seremet, T., Lyngaa, R., Svane, I. M., Thor Straten, P., Heemskerk, M. H., Grotenbreg, G. M. & Hadrup, S. R. 2015. Design and validation of conditional ligands for HLA-B*08:01, HLA-B*15:01, HLA-B*35:01, and HLA-B*44:05. *Cytometry A*.

Fu, J., Xu, D., Liu, Z., Shi, M., Zhao, P., Fu, B., Zhang, Z., Yang, H., Zhang, H., Zhou, C., Yao, J., Jin, L., Wang, H., Yang, Y., Fu, Y. X. & Wang, F. S. 2007. Increased regulatory T cells correlate with CD8 T-cell impairment and poor survival in hepatocellular carcinoma patients. *Gastroenterology,* 132**,** 2328-39.

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pages, C., Tosolini, M., Camus, M., Berger, A., Wind, P., Zinzindohoue, F., Bruneval, P., Cugnenc, P. H., Trajanoski, Z., Fridman, W. H. & Pages, F. 2006. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science,* 313**,** 1960-4.

Galon, J., Pages, F., Marincola, F. M., Thurin, M., Trinchieri, G., Fox, B. A., Gajewski, T. F. & Ascierto, P. A. 2012. The immune score as a new possible approach for the classification of cancer. *J Transl Med,* 10**,** 1.

Gao, J., Shi, L. Z., Zhao, H., Chen, J., Xiong, L., He, Q., Chen, T., Roszik, J., Bernatchez, C., Woodman, S. E., Chen, P. L., Hwu, P., Allison, J. P., Futreal, A., Wargo, J. A. & Sharma, P. 2016. Loss of IFN-gamma Pathway Genes in Tumor Cells as a Mechanism of Resistance to Anti-CTLA-4 Therapy. *Cell,* 167**,** 397-404 e9.

Garrido, F., Perea, F., Bernal, M., Sanchez-Palencia, A., Aptsiauri, N. & Ruiz-Cabello, F. 2017a. The Escape of Cancer from T Cell-Mediated Immune Surveillance: HLA Class I Loss and Tumor Tissue Architecture. *Vaccines (Basel),* 5.

Garrido, F., Ruiz-Cabello, F. & Aptsiauri, N. 2017b. Rejection versus escape: the tumor MHC dilemma. *Cancer Immunol Immunother,* 66**,** 259-271.

Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., Nair, V. S., Xu, Y., Khuong, A., Hoang, C. D., Diehn, M., West, R. B., Plevritis, S. K. & Alizadeh, A. A. 2015. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med,* 21**,** 938-945.

Gerlinger, M., Horswell, S., Larkin, J., Rowan, A. J., Salm, M. P., Varela, I., Fisher, R., McGranahan, N., Matthews, N., Santos, C. R., Martinez, P., Phillimore, B., Begum, S., Rabinowitz, A., Spencer-Dene, B., Gulati, S., Bates, P. A., Stamp, G., Pickering, L., Gore, M., Nicol, D. L., Hazell, S., Futreal, P. A., Stewart, A. & Swanton, C. 2014a. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet,* 46**,** 225-33.

Gerlinger, M., McGranahan, N., Dewhurst, S. M., Burrell, R. A., Tomlinson, I. & Swanton, C. 2014b. Cancer: evolution within a lifetime. *Annu Rev Genet,* 48**,** 215-36.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., Nohadani, M., Eklund, A. C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. A. & Swanton, C. 2012a. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med,* 366**,** 883-92.

Gerlinger, M., Santos, C. R., Spencer-Dene, B., Martinez, P., Endesfelder, D., Burrell, R. A., Vetter, M., Jiang, M., Saunders, R. E., Kelly, G., Rioux-Leclercq, N., Stamp, G., Patard, J. J., Larkin, J., Howell, M. & Swanton, C. 2012b. Genome-wide RNA interference analysis of renal carcinoma survival regulators identifies MCT4 as a Warburg effect metabolic target. *J Pathol*.

Gerlinger, M. & Swanton, C. 2010. How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. *Br J Cancer,* 103**,** 1139-43.

Germano, G., Lamba, S., Rospo, G., Barault, L., Magri, A., Maione, F., Russo, M., Crisafulli, G., Bartolini, A., Lerda, G., Siravegna, G., Mussolin, B., Frapolli, R., Montone, M., Morano, F., de Braud, F., Amirouchene-Angelozzi, N., Marsoni, S., D'Incalci, M., Orlandi, A., Giraudo, E., Sartore-Bianchi, A., Siena, S., Pietrantonio, F., Di Nicolantonio, F. & Bardelli, A. 2017. Inactivation of DNA repair

triggers neoantigen generation and impairs tumour growth. *Nature,* 552**,** 116-120.

Gettinger, S. N., Kowanetz, M., Koeppen, H., Wistuba, I. I., Kockx, M., Kadel, E. E., Rizvi, N. A., Spira, A. I., Hirsch, F. R., Boyd, Z., Denker, M., Minn, A., Shames, D. S., Sandler, A., Chen, D. S., Hegde, P. S. & Spigel, D. R. 2015. Molecular, immune and histopathological characterization of NSCLC based on PDL1 expression on tumor and immune cells and association with response to the anti-PDL1 antibody MPDL3280A. *Journal of Clinical Oncology,* 33**,** 3015-3015.

Gfeller, D., Bassani-Sternberg, M., Schmidt, J. & Luescher, I. F. 2016. Current tools for predicting cancer-specific T cell immunity. *Oncoimmunology,* 5**,** e1177691.

Giannakis, M., Mu, X. J., Shukla, S. A., Qian, Z. R., Cohen, O., Nishihara, R., Bahl, S., Cao, Y., Amin-Mansour, A., Yamauchi, M., Sukawa, Y., Stewart, C., Rosenberg, M., Mima, K., Inamura, K., Nosho, K., Nowak, J. A., Lawrence, M. S., Giovannucci, E. L., Chan, A. T., Ng, K., Meyerhardt, J. A., Van Allen, E. M., Getz, G., Gabriel, S. B., Lander, E. S., Wu, C. J., Fuchs, C. S., Ogino, S. & Garraway, L. A. 2016. Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Rep,* 17**,** 1206.

Glodzik, D., Morganella, S., Davies, H., Simpson, P. T., Li, Y., Zou, X., Diez-Perez, J., Staaf, J., Alexandrov, L. B., Smid, M., Brinkman, A. B., Rye, I. H., Russnes, H., Raine, K., Purdie, C. A., Lakhani, S. R., Thompson, A. M., Birney, E., Stunnenberg, H. G., van de Vijver, M. J., Martens, J. W., Borresen-Dale, A. L., Richardson, A. L., Kong, G., Viari, A., Easton, D., Evan, G., Campbell, P. J., Stratton, M. R. & Nik-Zainal, S. 2017. A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. *Nat Genet,* 49**,** 341-348.

Gopalakrishnan, V., Spencer, C. N., Nezi, L., Reuben, A., Andrews, M. C., Karpinets, T. V., Prieto, P. A., Vicente, D., Hoffman, K., Wei, S. C., Cogdill, A. P., Zhao, L., Hudgens, C. W., Hutchinson, D. S., Manzo, T., Petaccia de Macedo, M., Cotechini, T., Kumar, T., Chen, W. S., Reddy, S. M., Sloane, R. S., Galloway-Pena, J., Jiang, H., Chen, P. L., Shpall, E. J., Rezvani, K., Alousi, A. M., Chemaly, R. F., Shelburne, S., Vence, L. M., Okhuysen, P. C., Jensen, V. B., Swennes, A. G., McAllister, F., Sanchez, E. M. R., Zhang, Y., Le Chatelier, E., Zitvogel, L., Pons, N., Austin-Breneman, J. L., Haydu, L. E., Burton, E. M., Gardner, J. M., Sirmans, E., Hu, J., Lazar, A. J., Tsujikawa, T., Diab, A., Tawbi, H., Glitza, I. C., Hwu, W. J., Patel, S. P., Woodman, S. E., Amaria, R. N., Davies, M. A., Gershenwald, J. E., Hwu, P., Lee, J. E., Zhang, J., Coussens, L. M., Cooper, Z. A., Futreal, P. A., Daniel, C. R., Ajami, N. J., Petrosino, J. F., Tetzlaff, M. T., Sharma, P., Allison, J. P., Jenq, R. R. & Wargo, J. A. 2017. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science*.

Greaves, M. & Maley, C. C. 2012. Clonal evolution in cancer. *Nature,* 481**,** 306-13.

Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. & Easton, D. F. 2006. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics,* 173**,** 2187-98.

Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L. B., Tubio, J. M., Papaemmanuil, E., Brewer, D. S., Kallio, H. M., Hognas, G., Annala, M., Kivinummi, K., Goody, V., Latimer, C., O'Meara, S., Dawson, K. J., Isaacs, W., Emmert-Buck, M. R., Nykter, M., Foster, C., Kote-Jarai, Z., Easton, D., Whitaker, H. C., Group, I. P. U., Neal, D. E., Cooper, C. S., Eeles, R. A., Visakorpi, T., Campbell, P. J., McDermott, U., Wedge, D. C. & Bova, G. S. 2015. The evolutionary history of lethal metastatic prostate cancer. *Nature,* 520**,** 353-7.

Guo, B., Cen, H., Tan, X. & Ke, Q. 2016. Meta-analysis of the prognostic and clinical value of tumor-associated macrophages in adult classical Hodgkin lymphoma. *BMC Med,* 14**,** 159.

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Gilks, C. B., Huntsman, D. G., McAlpine, J. N., Aparicio, S. & Shah, S. P. 2014. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res,* 24**,** 1881-93.

Hackl, H., Charoentong, P., Finotello, F. & Trajanoski, Z. 2016. Computational genomics tools for dissecting tumour-immune cell interactions. *Nat Rev Genet,* 17**,** 441-58.

Hacohen, N., Fritsch, E. F., Carter, T. A., Lander, E. S. & Wu, C. J. 2013. Getting personal with neoantigen-based therapeutic cancer vaccines. *Cancer Immunol Res,* 1**,** 11-5.

Hadrup, S. R. & Schumacher, T. N. 2010. MHC-based detection of antigen-specific CD8+ T cell responses. *Cancer Immunol Immunother,* 59**,** 1425-33.

Haffner, M. C., Mosbruger, T., Esopi, D. M., Fedor, H., Heaphy, C. M., Walker, D. A., Adejola, N., Gurel, M., Hicks, J., Meeker, A. K., Halushka, M. K., Simons, J. W., Isaacs, W. B., De Marzo, A. M., Nelson, W. G. & Yegnasubramanian, S. 2013. Tracking the clonal origin of lethal prostate cancer. *The Journal of clinical investigation,* 123**,** 4918-22.

Hanahan, D. & Weinberg, R. A. 2011. Hallmarks of cancer: the next generation. *Cell,* 144**,** 646-74.

Hartmaier, R. J., Charo, J., Fabrizio, D., Goldberg, M. E., Albacker, L. A., Pao, W. & Chmielecki, J. 2017. Genomic analysis of 63,220 tumors reveals insights into tumor uniqueness and targeted cancer immunotherapy strategies. *Genome Med,* 9**,** 16.

Helleday, T., Eshtad, S. & Nik-Zainal, S. 2014. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet,* 15**,** 585-98.

Henderson, S., Chakravarthy, A., Su, X., Boshoff, C. & Fenton, T. R. 2014. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. *Cell Rep,* 7**,** 1833-41.

Herbst, R. S., Soria, J. C., Kowanetz, M., Fine, G. D., Hamid, O., Gordon, M. S., Sosman, J. A., McDermott, D. F., Powderly, J. D., Gettinger, S. N., Kohrt, H. E., Horn, L., Lawrence, D. P., Rost, S., Leabman, M., Xiao,

Y., Mokatrin, A., Koeppen, H., Hegde, P. S., Mellman, I., Chen, D. S. & Hodi, F. S. 2014. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature,* 515**,** 563-7.

Hicklin, D. J., Marincola, F. M. & Ferrone, S. 1999. HLA class I antigen downregulation in human cancers: T-cell immunotherapy revives an old story. *Mol Med Today,* 5**,** 178-86.

Hoang, M. L., Chen, C. H., Sidorenko, V. S., He, J., Dickman, K. G., Yun, B. H., Moriya, M., Niknafs, N., Douville, C., Karchin, R., Turesky, R. J., Pu, Y. S., Vogelstein, B., Papadopoulos, N., Grollman, A. P., Kinzler, K. W. & Rosenquist, T. A. 2013. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci Transl Med,* 5**,** 197ra102.

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., Hassel, J. C., Akerley, W., van den Eertwegh, A. J., Lutzky, J., Lorigan, P., Vaubel, J. M., Linette, G. P., Hogg, D., Ottensmeier, C. H., Lebbe, C., Peschel, C., Quirt, I., Clark, J. I., Wolchok, J. D., Weber, J. S., Tian, J., Yellin, M. J., Nichol, G. M., Hoos, A. & Urba, W. J. 2010. Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med,* 363**,** 711-23.

Hong, M. K., Macintyre, G., Wedge, D. C., Van Loo, P., Patel, K., Lunke, S., Alexandrov, L. B., Sloggett, C., Cmero, M., Marass, F., Tsui, D., Mangiola, S., Lonie, A., Naeem, H., Sapre, N., Phal, P. M., Kurganovs, N., Chin, X., Kerger, M., Warren, A. Y., Neal, D., Gnanapragasam, V., Rosenfeld, N., Pedersen, J. S., Ryan, A., Haviv, I., Costello, A. J., Corcoran, N. M. & Hovens, C. M. 2015. Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nat Commun,* 6**,** 6605.

Hoof, I., Peters, B., Sidney, J., Pedersen, L. E., Sette, A., Lund, O., Buus, S. & Nielsen, M. 2009. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics,* 61**,** 1-13.

Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., Seja, E., Lomeli, S., Kong, X., Kelley, M. C., Sosman, J. A., Johnson, D. B., Ribas, A. & Lo, R. S. 2016. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell,* 165**,** 35-44.

Hunter, C., Smith, R., Cahill, D. P., Stephens, P., Stevens, C., Teague, J., Greenman, C., Edkins, S., Bignell, G., Davies, H., O'Meara, S., Parker, A., Avis, T., Barthorpe, S., Brackenbury, L., Buck, G., Butler, A., Clements, J., Cole, J., Dicks, E., Forbes, S., Gorton, M., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jenkinson, A., Jones, D., Kosmidou, V., Laman, R., Lugg, R., Menzies, A., Perry, J., Petty, R., Raine, K., Richardson, D., Shepherd, R., Small, A., Solomon, H., Tofts, C., Varian, J., West, S., Widaa, S., Yates, A., Easton, D. F., Riggins, G., Roy, J. E., Levine, K. K., Mueller, W., Batchelor, T. T., Louis, D. N., Stratton, M. R., Futreal, P. A. & Wooster, R. 2006. A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy. *Cancer Res,* 66**,** 3987-91.

Jacobs, J. F., Idema, A. J., Bol, K. F., Grotenhuis, J. A., de Vries, I. J., Wesseling, P. & Adema, G. J. 2010. Prognostic significance and mechanism of Treg infiltration in human brain tumors. *J Neuroimmunol,* 225**,** 195-9.

Jamal-Hanjani, M., Wilson, G. A., Horswell, S., Mitter, R., Sakarya, O., Constantin, T., Salari, R., Kirkizlar, E., Sigurjonsson, S., Pelham, R., Kareht, S., Zimmermann, B. & Swanton, C. 2016. Detection of ubiquitous and heterogeneous mutations in cell-free DNA from patients with early-stage non-small-cell lung cancer. *Ann Oncol,* 27**,** 862-7.

Jamal-Hanjani, M., Wilson, G. A., McGranahan, N., Birkbak, N. J., Watkins, T. B. K., Veeriah, S., Shafi, S., Johnson, D. H., Mitter, R., Rosenthal, R., Salm, M., Horswell, S., Escudero, M., Matthews, N., Rowan, A., Chambers, T., Moore, D. A., Turajlic, S., Xu, H., Lee, S. M., Forster, M. D., Ahmad, T., Hiley, C. T., Abbosh, C., Falzon, M., Borg, E., Marafioti, T., Lawrence, D., Hayward, M., Kolvekar, S., Panagiotopoulos, N., Janes, S. M., Thakrar, R., Ahmed, A., Blackhall, F., Summers, Y., Shah, R., Joseph, L., Quinn, A. M., Crosbie, P. A., Naidu, B., Middleton, G., Langman, G., Trotter, S., Nicolson, M., Remmen, H., Kerr, K., Chetty, M., Gomersall, L., Fennell, D. A., Nakas, A., Rathinam, S., Anand, G., Khan, S., Russell, P., Ezhil, V., Ismail, B., Irvin-Sellers, M., Prakash, V., Lester, J. F., Kornaszewska, M., Attanoos, R., Adams, H., Davies, H., Dentro, S., Taniere, P., O'Sullivan, B., Lowe, H. L., Hartley, J. A., Iles, N., Bell, H., Ngai, Y., Shaw, J. A., Herrero, J., Szallasi, Z., Schwarz, R. F., Stewart, A., Quezada, S. A., Le Quesne, J., Van Loo, P., Dive, C., Hackshaw, A., Swanton, C. & Consortium, T. R. 2017. Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl J Med,* 376**,** 2109-2121.

Johnson, B. E., Mazor, T., Hong, C., Barnes, M., Aihara, K., McLean, C. Y., Fouse, S. D., Yamamoto, S., Ueda, H., Tatsuno, K., Asthana, S., Jalbert, L. E., Nelson, S. J., Bollen, A. W., Gustafson, W. C., Charron, E., Weiss, W. A., Smirnov, I. V., Song, J. S., Olshen, A. B., Cha, S., Zhao, Y., Moore, R. A., Mungall, A. J., Jones, S. J., Hirst, M., Marra, M. A., Saito, N., Aburatani, H., Mukasa, A., Berger, M. S., Chang, S. M., Taylor, B. S. & Costello, J. F. 2014. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science,* 343**,** 189-93.

Johnson, L. A., Morgan, R. A., Dudley, M. E., Cassard, L., Yang, J. C., Hughes, M. S., Kammula, U. S., Royal, R. E., Sherry, R. M., Wunderlich, J. R., Lee, C. C., Restifo, N. P., Schwarz, S. L., Cogdill, A. P., Bishop, R. J., Kim, H., Brewer, C. C., Rudy, S. F., VanWaes, C., Davis, J. L., Mathur, A., Ripley, R. T., Nathan, D. A., Laurencot, C. M. & Rosenberg, S. A. 2009. Gene therapy with human and mouse T-cell receptors mediates cancer regression and targets normal tissues expressing cognate antigen. *Blood,* 114**,** 535-46.

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B. & Nielsen, M. 2017. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol,* 199**,** 3360-3368.

Kaplan, D. H., Shankaran, V., Dighe, A. S., Stockert, E., Aguet, M., Old, L. J. & Schreiber, R. D. 1998. Demonstration of an interferon gamma-dependent tumor surveillance system in immunocompetent mice. *Proc Natl Acad Sci U S A,* 95**,** 7556-61.

Karin, M., Cao, Y., Greten, F. R. & Li, Z. W. 2002. NF-kappaB in cancer: from innocent bystander to major culprit. *Nat Rev Cancer,* 2**,** 301-10.

Keats, J. J., Chesi, M., Egan, J. B., Garbitt, V. M., Palmer, S. E., Braggio, E., Van Wier, S., Blackburn, P. R., Baker, A. S., Dispenzieri, A., Kumar, S., Rajkumar, S. V., Carpten, J. D., Barrett, M., Fonseca, R., Stewart, A. K. & Bergsagel, P. L. 2012. Clonal competition with alternating dominance in multiple myeloma. *Blood*.

Kesmir, C., Nussbaum, A. K., Schild, H., Detours, V. & Brunak, S. 2002. Prediction of proteasome cleavage motifs by neural networks. *Protein Eng,* 15**,** 287-96.

Kim, Y., Sidney, J., Pinilla, C., Sette, A. & Peters, B. 2009. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. *BMC Bioinformatics,* 10**,** 394.

Kiyotani, K., Mai, T. H. & Nakamura, Y. 2017. Comparison of exome-based HLA class I genotyping tools: identification of platform-specific genotyping errors. *J Hum Genet,* 62**,** 397-405.

Klco, J. M., Spencer, D. H., Miller, C. A., Griffith, M., Lamprecht, T. L., O'Laughlin, M., Fronick, C., Magrini, V., Demeter, R. T., Fulton, R. S., Eades, W. C., Link, D. C., Graubert, T. A., Walter, M. J., Mardis, E. R., Dipersio, J. F., Wilson, R. K. & Ley, T. J. 2014. Functional heterogeneity of genetically defined subclones in acute myeloid leukemia. *Cancer Cell,* 25**,** 379-92.

Knudson, A. G., Jr. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A,* 68**,** 820-3.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L. & Wilson, R. K. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res,* 22**,** 568-76.

Koopman, L. A., Corver, W. E., van der Slik, A. R., Giphart, M. J. & Fleuren, G. J. 2000. Multiple genetic alterations cause frequent and heterogeneous human histocompatibility leukocyte antigen class I loss in cervical cancer. *J Exp Med,* 191**,** 961-76.

Kosaka, T., Yatabe, Y., Endoh, H., Yoshida, K., Hida, T., Tsuboi, M., Tada, H., Kuwano, H. & Mitsudomi, T. 2006. Analysis of epidermal growth factor receptor gene mutation in patients with non-small cell lung cancer and acquired resistance to gefitinib. *Clin Cancer Res,* 12**,** 5764-9.

Kreso, A., O'Brien, C. A., van Galen, P., Gan, O. I., Notta, F., Brown, A. M., Ng, K., Ma, J., Wienholds, E., Dunant, C., Pollett, A., Gallinger, S., McPherson, J., Mullighan, C. G., Shibata, D. & Dick, J. E. 2013. Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science,* 339**,** 543-8.

Kumar, A., Coleman, I., Morrissey, C., Zhang, X., True, L. D., Gulati, R., Etzioni, R., Bolouri, H., Montgomery, B., White, T., Lucas, J. M., Brown, L. G., Dumpit, R. F., DeSarkar, N., Higano, C., Yu, E. Y., Coleman, R., Schultz, N., Fang, M., Lange, P. H., Shendure, J.,

Vessella, R. L. & Nelson, P. S. 2016a. Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nat Med,* 22**,** 369-78.

Kumar, V., Patel, S., Tcyganov, E. & Gabrilovich, D. I. 2016b. The Nature of Myeloid-Derived Suppressor Cells in the Tumor Microenvironment. *Trends Immunol,* 37**,** 208-220.

Landau, D. A., Carter, S. L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M. S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., Wan, Y., Zhang, W., Shukla, S. A., Vartanov, A., Fernandes, S. M., Saksena, G., Cibulskis, K., Tesar, B., Gabriel, S., Hacohen, N., Meyerson, M., Lander, E. S., Neuberg, D., Brown, J. R., Getz, G. & Wu, C. J. 2013. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell,* 152**,** 714-26.

Landau, D. A., Tausch, E., Taylor-Weiner, A. N., Stewart, C., Reiter, J. G., Bahlo, J., Kluth, S., Bozic, I., Lawrence, M., Bottcher, S., Carter, S. L., Cibulskis, K., Mertens, D., Sougnez, C. L., Rosenberg, M., Hess, J. M., Edelmann, J., Kless, S., Kneba, M., Ritgen, M., Fink, A., Fischer, K., Gabriel, S., Lander, E. S., Nowak, M. A., Dohner, H., Hallek, M., Neuberg, D., Getz, G., Stilgenbauer, S. & Wu, C. J. 2015. Mutations driving CLL and their evolution in progression and relapse. *Nature,* 526**,** 525-30.

Larsen, M. V., Lundegaard, C., Lamberth, K., Buus, S., Lund, O. & Nielsen, M. 2007. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics,* 8**,** 424.

Lavin, Y., Kobayashi, S., Leader, A., Amir, E. D., Elefant, N., Bigenwald, C., Remark, R., Sweeney, R., Becker, C. D., Levine, J. H., Meinhof, K., Chow, A., Kim-Shulze, S., Wolf, A., Medaglia, C., Li, H., Rytlewski, J. A., Emerson, R. O., Solovyov, A., Greenbaum, B. D., Sanders, C., Vignali, M., Beasley, M. B., Flores, R., Gnjatic, S., Pe'er, D., Rahman, A., Amit, I. & Merad, M. 2017. Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. *Cell,* 169**,** 750-765 e17.

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S. & Getz, G. 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature,* 505**,** 495-501.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortes, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D. A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S. & Getz, G. 2013.

Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature,* 499**,** 214-218.

Le, D. T., Uram, J. N., Wang, H., Bartlett, B. R., Kemberling, H., Eyring, A. D., Skora, A. D., Luber, B. S., Azad, N. S., Laheru, D., Biedrzycki, B., Donehower, R. C., Zaheer, A., Fisher, G. A., Crocenzi, T. S., Lee, J. J., Duffy, S. M., Goldberg, R. M., de la Chapelle, A., Koshiji, M., Bhaijee, F., Huebner, T., Hruban, R. H., Wood, L. D., Cuka, N., Pardoll, D. M., Papadopoulos, N., Kinzler, K. W., Zhou, S., Cornish, T. C., Taube, J. M., Anders, R. A., Eshleman, J. R., Vogelstein, B. & Diaz, L. A., Jr. 2015. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med,* 372**,** 2509-20.

Lefranc, M. P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G. & Duroux, P. 2009. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res,* 37**,** D1006-12.

Lennerz, V., Fatho, M., Gentilini, C., Frye, R. A., Lifke, A., Ferel, D., Wolfel, C., Huber, C. & Wolfel, T. 2005. The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. *Proc Natl Acad Sci U S A,* 102**,** 16013-8.

Li, B., Severson, E., Pignon, J. C., Zhao, H., Li, T., Novak, J., Jiang, P., Shen, H., Aster, J. C., Rodig, S., Signoretti, S., Liu, J. S. & Liu, X. S. 2016. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol,* 17**,** 174.

Linnemann, C., van Buuren, M. M., Bies, L., Verdegaal, E. M., Schotte, R., Calis, J. J., Behjati, S., Velds, A., Hilkmann, H., Atmioui, D. E., Visser, M., Stratton, M. R., Haanen, J. B., Spits, H., van der Burg, S. H. & Schumacher, T. N. 2015. High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4+ T cells in human melanoma. *Nat Med,* 21**,** 81-5.

Liu, C., Yang, X., Duffy, B., Mohanakumar, T., Mitra, R. D., Zody, M. C. & Pfeifer, J. D. 2013. ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Res,* 41**,** e142.

Liu, X., Jian, X. & Boerwinkle, E. 2011. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat,* 32**,** 894-9.

Liu, X. S. & Mardis, E. R. 2017. Applications of Immunogenomics to Cancer. *Cell,* 168**,** 600-612.

Lohr, J. G., Stojanov, P., Carter, S. L., Cruz-Gordillo, P., Lawrence, M. S., Auclair, D., Sougnez, C., Knoechel, B., Gould, J., Saksena, G., Cibulskis, K., McKenna, A., Chapman, M. A., Straussman, R., Levy, J., Perkins, L. M., Keats, J. J., Schumacher, S. E., Rosenberg, M., Multiple Myeloma Research, C., Getz, G. & Golub, T. R. 2014. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer Cell,* 25**,** 91-101.

Luksza, M., Riaz, N., Makarov, V., Balachandran, V. P., Hellmann, M. D., Solovyov, A., Rizvi, N. A., Merghoub, T., Levine, A. J., Chan, T. A., Wolchok, J. D. & Greenbaum, B. D. 2017. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature,* 551**,** 517-520.

Lund, O., Karosiene, E., Lundegaard, C., Larsen, M. V. & Nielsen, M. 2013. Bioinformatics identification of antigenic peptide: predicting the specificity of major MHC class I and II pathway players. *Methods Mol Biol,* 960**,** 247-260.

Macintyre, G., Goranova, T., De Silva, D., Ennis, D., Piskorz, A. M., Eldridge, M., Sie, D., Lewsley, L.-A., Hanif, A., Wilson, C., Dowson, S., Glasspool, R. M., Lockley, M., Brockbank, E., Montes, A., Walther, A., Sundar, S., Edmondson, R., Hall, G. D., Clamp, A., Gourley, C., Hall, M., Fotopoulou, C., Gabra, H., Paul, J., Supernat, A., Millan, D., Hoyle, A., Bryson, G., Nourse, C., Mincarelli, L., Navarro Sanchez, L., Ylstra, B., Jimenez-Linan, M., Moore, L., Hofmann, O., Markowetz, F., McNeish, I. A. & Brenton, J. D. 2017. Copy-number signatures and mutational processes in ovarian carcinoma. *bioRxiv*.

Mahmoud, S. M., Paish, E. C., Powe, D. G., Macmillan, R. D., Grainge, M. J., Lee, A. H., Ellis, I. O. & Green, A. R. 2011. Tumor-infiltrating CD8+ lymphocytes predict clinical outcome in breast cancer. *J Clin Oncol,* 29**,** 1949-55.

Maley, C. C., Galipeau, P. C., Finley, J. C., Wongsurawat, V. J., Li, X., Sanchez, C. A., Paulson, T. G., Blount, P. L., Risques, R. A., Rabinovitch, P. S. & Reid, B. J. 2006. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet,* 38**,** 468-73.

Malikic, S., McPherson, A. W., Donmez, N. & Sahinalp, C. S. 2015. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics,* 31**,** 1349-56.

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R. & Campbell, P. J. 2017. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell,* 171**,** 1029-1041 e21.

Marusyk, A. & Polyak, K. 2010. Tumor heterogeneity: causes and consequences. *Biochim Biophys Acta,* 1805**,** 105-17.

Marusyk, A., Tabassum, D. P., Altrock, P. M., Almendro, V., Michor, F. & Polyak, K. 2014. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature,* 514**,** 54-8.

Matsushita, H., Vesely, M. D., Koboldt, D. C., Rickert, C. G., Uppaluri, R., Magrini, V. J., Arthur, C. D., White, J. M., Chen, Y. S., Shea, L. K., Hundal, J., Wendl, M. C., Demeter, R., Wylie, T., Allison, J. P., Smyth, M. J., Old, L. J., Mardis, E. R. & Schreiber, R. D. 2012. Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature,* 482**,** 400-4.

Matsuzaki, J., Gnjatic, S., Mhawech-Fauceglia, P., Beck, A., Miller, A., Tsuji, T., Eppolito, C., Qian, F., Lele, S., Shrikant, P., Old, L. J. & Odunsi, K. 2010. Tumor-infiltrating NY-ESO-1-specific CD8+ T cells are negatively regulated by LAG-3 and PD-1 in human ovarian cancer. *Proc Natl Acad Sci U S A,* 107**,** 7875-80.

McGranahan, N., Favero, F., de Bruin, E. C., Birkbak, N. J., Szallasi, Z. & Swanton, C. 2015. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci Transl Med,* 7**,** 283ra54.

McGranahan, N., Furness, A. J., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., Jamal-Hanjani, M., Wilson, G. A., Birkbak, N. J., Hiley, C.

T., Watkins, T. B., Shafi, S., Murugaesu, N., Mitter, R., Akarca, A. U., Linares, J., Marafioti, T., Henry, J. Y., Van Allen, E. M., Miao, D., Schilling, B., Schadendorf, D., Garraway, L. A., Makarov, V., Rizvi, N. A., Snyder, A., Hellmann, M. D., Merghoub, T., Wolchok, J. D., Shukla, S. A., Wu, C. J., Peggs, K. S., Chan, T. A., Hadrup, S. R., Quezada, S. A. & Swanton, C. 2016. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science,* 351**,** 1463-9.

McGranahan, N., Rosenthal, R., Hiley, C. T., Rowan, A. J., Watkins, T. B. K., Wilson, G. A., Birkbak, N. J., Veeriah, S., Van Loo, P., Herrero, J., Swanton, C. & Consortium, T. R. 2017. Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell,* 171**,** 1259-1271 e11.

McGranahan, N. & Swanton, C. 2015. Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution. *Cancer Cell,* 27**,** 15-26.

Mehta, A. M., Jordanova, E. S., Kenter, G. G., Ferrone, S. & Fleuren, G. J. 2008. Association of antigen processing machinery and HLA class I defects with clinicopathological outcome in cervical carcinoma. *Cancer Immunol Immunother,* 57**,** 197-206.

Meier, B., Cooke, S. L., Weiss, J., Bailly, A. P., Alexandrov, L. B., Marshall, J., Raine, K., Maddison, M., Anderson, E., Stratton, M. R., Gartner, A. & Campbell, P. J. 2014. C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res*.

Merajver, S. D., Frank, T. S., Xu, J., Pham, T. M., Calzone, K. A., Bennett-Baker, P., Chamberlain, J., Boyd, J., Garber, J. E., Collins, F. S. & et al. 1995. Germline BRCA1 mutations and loss of the wild-type allele in tumors from families with early onset breast and ovarian cancer. *Clin Cancer Res,* 1**,** 539-44.

Merlo, L. M., Pepper, J. W., Reid, B. J. & Maley, C. C. 2006. Cancer as an evolutionary and ecological process. *Nat Rev Cancer,* 6**,** 924-35.

Minn, A. J. & Wherry, E. J. 2016. Combination Cancer Therapies with Immune Checkpoint Blockade: Convergence on Interferon Signaling. *Cell,* 165**,** 272-5.

Misale, S., Yaeger, R., Hobor, S., Scala, E., Janakiraman, M., Liska, D., Valtorta, E., Schiavo, R., Buscarino, M., Siravegna, G., Bencardino, K., Cercek, A., Chen, C. T., Veronese, S., Zanon, C., Sartore-Bianchi, A., Gambacorta, M., Gallicchio, M., Vakiani, E., Boscaro, V., Medico, E., Weiser, M., Siena, S., Di Nicolantonio, F., Solit, D. & Bardelli, A. 2012. Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature,* 486**,** 532-6.

Mlecnik, B., Bindea, G., Kirilovsky, A., Angell, H. K., Obenauf, A. C., Tosolini, M., Church, S. E., Maby, P., Vasaturo, A., Angelova, M., Fredriksen, T., Mauger, S., Waldner, M., Berger, A., Speicher, M. R., Pages, F., Valge-Archer, V. & Galon, J. 2016. The tumor microenvironment and Immunoscore are critical determinants of dissemination to distant metastasis. *Sci Transl Med,* 8**,** 327ra26.

Monach, P. A., Meredith, S. C., Siegel, C. T. & Schreiber, H. 1995. A unique tumor antigen produced by a single amino acid substitution. *Immunity,* 2**,** 45-59.

Morgan, R. A., Dudley, M. E., Wunderlich, J. R., Hughes, M. S., Yang, J. C., Sherry, R. M., Royal, R. E., Topalian, S. L., Kammula, U. S., Restifo, N. P., Zheng, Z., Nahvi, A., de Vries, C. R., Rogers-Freezer, L. J., Mavroukakis, S. A. & Rosenberg, S. A. 2006. Cancer regression in patients after transfer of genetically engineered lymphocytes. *Science,* 314**,** 126-9.

Moutaftsi, M., Peters, B., Pasquetto, V., Tscharke, D. C., Sidney, J., Bui, H. H., Grey, H. & Sette, A. 2006. A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus. *Nat Biotechnol,* 24**,** 817-9.

Mroz, E. A. & Rocco, J. W. 2013. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. *Oral Oncol,* 49**,** 211-5.

Mroz, E. A., Tward, A. D., Hammon, R. J., Ren, Y. & Rocco, J. W. 2015. Intra-tumor genetic heterogeneity and mortality in head and neck cancer: analysis of data from the Cancer Genome Atlas. *PLoS Med,* 12**,** e1001786.

Mukherjee, S. 2011. *The Emperor of All Maladies*, 4th Estate.

Muller, M., Gfeller, D., Coukos, G. & Bassani-Sternberg, M. 2017. 'Hotspots' of Antigen Presentation Revealed by Human Leukocyte Antigen Ligandomics for Neoantigen Prioritization. *Front Immunol,* 8**,** 1367.

Mullighan, C. G., Phillips, L. A., Su, X., Ma, J., Miller, C. B., Shurtleff, S. A. & Downing, J. R. 2008. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science,* 322**,** 1377-80.

Murtaza, M., Dawson, S. J., Tsui, D. W., Gale, D., Forshew, T., Piskorz, A. M., Parkinson, C., Chin, S. F., Kingsbury, Z., Wong, A. S., Marass, F., Humphray, S., Hadfield, J., Bentley, D., Chin, T. M., Brenton, J. D., Caldas, C. & Rosenfeld, N. 2013. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature*.

Murugaesu, N., Wilson, G. A., Birkbak, N. J., Watkins, T., McGranahan, N., Kumar, S., Abbassi-Ghadi, N., Salm, M., Mitter, R., Horswell, S., Rowan, A., Phillimore, B., Biggs, J., Begum, S., Matthews, N., Hochhauser, D., Hanna, G. B. & Swanton, C. 2015. Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. *Cancer Discov*.

Naito, Y., Saito, K., Shiiba, K., Ohuchi, A., Saigenji, K., Nagura, H. & Ohtani, H. 1998. CD8+ T cells infiltrated within cancer cell nests as a prognostic factor in human colorectal cancer. *Cancer Res,* 58**,** 3491-4.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J. & Wigler, M. 2011. Tumour evolution inferred by single-cell sequencing. *Nature,* 472**,** 90-4.

Neefjes, J., Jongsma, M. L., Paul, P. & Bakke, O. 2011. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol,* 11**,** 823-36.

Newell, E. W., Sigal, N., Bendall, S. C., Nolan, G. P. & Davis, M. M. 2012. Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8+ T cell phenotypes. *Immunity,* 36**,** 142-52.

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M. & Alizadeh, A. A. 2015. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods,* 12**,** 453-7.

Nguyen, L. T. & Ohashi, P. S. 2015. Clinical blockade of PD1 and LAG3--potential mechanisms of action. *Nat Rev Immunol,* 15**,** 45-56.

Nielsen, M. & Andreatta, M. 2016. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med,* 8**,** 33.

Nielsen, M., Lund, O., Buus, S. & Lundegaard, C. 2010. MHC class II epitope predictive algorithms. *Immunology,* 130**,** 319-28.

Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Roder, G., Peters, B., Sette, A., Lund, O. & Buus, S. 2007. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One,* 2**,** e796.

Nielsen, M., Lundegaard, C., Blicher, T., Peters, B., Sette, A., Justesen, S., Buus, S. & Lund, O. 2008. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. *PLoS Comput Biol,* 4**,** e1000107.

Nielsen, M., Lundegaard, C., Lund, O. & Kesmir, C. 2005. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics,* 57**,** 33-41.

Nielsen, M., Lundegaard, C., Worning, P., Lauemoller, S. L., Lamberth, K., Buus, S., Brunak, S. & Lund, O. 2003. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci,* 12**,** 1007-17.

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., Menzies, A., Martin, S., Leung, K., Chen, L., Leroy, C., Ramakrishna, M., Rance, R., Lau, K. W., Mudie, L. J., Varela, I., McBride, D. J., Bignell, G. R., Cooke, S. L., Shlien, A., Gamble, J., Whitmore, I., Maddison, M., Tarpey, P. S., Davies, H. R., Papaemmanuil, E., Stephens, P. J., McLaren, S., Butler, A. P., Teague, J. W., Jonsson, G., Garber, J. E., Silver, D., Miron, P., Fatima, A., Boyault, S., Langerod, A., Tutt, A., Martens, J. W., Aparicio, S. A., Borg, A., Salomon, A. V., Thomas, G., Borresen-Dale, A. L., Richardson, A. L., Neuberger, M. S., Futreal, P. A., Campbell, P. J. & Stratton, M. R. 2012a. Mutational processes molding the genomes of 21 breast cancers. *Cell,* 149**,** 979-93.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L. B., Martin, S., Wedge, D. C., Van Loo, P., Ju, Y. S., Smid, M., Brinkman, A. B., Morganella, S., Aure, M. R., Lingjaerde, O. C., Langerod, A., Ringner, M., Ahn, S. M., Boyault, S., Brock, J. E., Broeks, A., Butler, A., Desmedt, C., Dirix, L., Dronov, S., Fatima, A., Foekens, J. A., Gerstung, M., Hooijer, G. K., Jang, S. J., Jones, D. R., Kim, H. Y., King, T. A., Krishnamurthy, S., Lee, H. J., Lee, J. Y., Li, Y., McLaren, S., Menzies, A., Mustonen, V., O'Meara, S., Pauporte, I., Pivot, X., Purdie, C. A., Raine, K., Ramakrishnan, K.,

Rodriguez-Gonzalez, F. G., Romieu, G., Sieuwerts, A. M., Simpson, P. T., Shepherd, R., Stebbings, L., Stefansson, O. A., Teague, J., Tommasi, S., Treilleux, I., Van den Eynden, G. G., Vermeulen, P., Vincent-Salomon, A., Yates, L., Caldas, C., van't Veer, L., Tutt, A., Knappskog, S., Tan, B. K., Jonkers, J., Borg, A., Ueno, N. T., Sotiriou, C., Viari, A., Futreal, P. A., Campbell, P. J., Span, P. N., Van Laere, S., Lakhani, S. R., Eyfjord, J. E., Thompson, A. M., Birney, E., Stunnenberg, H. G., van de Vijver, M. J., Martens, J. W., Borresen-Dale, A. L., Richardson, A. L., Kong, G., Thomas, G. & Stratton, M. R. 2016. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature,* 534**,** 47-54.

Nik-Zainal, S. & Morganella, S. 2017. Mutational Signatures in Breast Cancer: The Problem at the DNA Level. *Clin Cancer Res,* 23**,** 2617-2629.

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., Shlien, A., Cooke, S. L., Hinton, J., Menzies, A., Stebbings, L. A., Leroy, C., Jia, M., Rance, R., Mudie, L. J., Gamble, S. J., Stephens, P. J., McLaren, S., Tarpey, P. S., Papaemmanuil, E., Davies, H. R., Varela, I., McBride, D. J., Bignell, G. R., Leung, K., Butler, A. P., Teague, J. W., Martin, S., Jonsson, G., Mariani, O., Boyault, S., Miron, P., Fatima, A., Langerod, A., Aparicio, S. A., Tutt, A., Sieuwerts, A. M., Borg, A., Thomas, G., Salomon, A. V., Richardson, A. L., Borresen-Dale, A. L., Futreal, P. A., Stratton, M. R. & Campbell, P. J. 2012b. The life history of 21 breast cancers. *Cell,* 149**,** 994-1007.

Novellino, L., Renkvist, N., Rini, F., Mazzocchi, A., Rivoltini, L., Greco, A., Deho, P., Squarcina, P., Robbins, P. F., Parmiani, G. & Castelli, C. 2003. Identification of a mutated receptor-like protein tyrosine phosphatase kappa as a novel, class II HLA-restricted melanoma antigen. *J Immunol,* 170**,** 6363-70.

Nowell, P. C. 1976. The clonal evolution of tumor cell populations. *Science,* 194**,** 23-8.

O'Donnell, T., Rubinsteyn, A., Bonsack, M., Riemer, A. & Hammerbacher, J. 2017. MHCflurry: open-source class I MHC binding affinity prediction. *bioRxiv*.

Old, L. J. & Boyse, E. A. 1964. Immunology of Experimental Tumors. *Annu Rev Med,* 15**,** 167-86.

Ortmann, C. A., Kent, D. G., Nangalia, J., Silber, Y., Wedge, D. C., Grinfeld, J., Baxter, E. J., Massie, C. E., Papaemmanuil, E., Menon, S., Godfrey, A. L., Dimitropoulou, D., Guglielmelli, P., Bellosillo, B., Besses, C., Dohner, K., Harrison, C. N., Vassiliou, G. S., Vannucchi, A., Campbell, P. J. & Green, A. R. 2015. Effect of mutation order on myeloproliferative neoplasms. *N Engl J Med,* 372**,** 601-12.

Ott, P. A., Hu, Z., Keskin, D. B., Shukla, S. A., Sun, J., Bozym, D. J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., Chen, C., Olive, O., Carter, T. A., Li, S., Lieb, D. J., Eisenhaure, T., Gjini, E., Stevens, J., Lane, W. J., Javeri, I., Nellaiappan, K., Salazar, A. M., Daley, H., Seaman, M., Buchbinder, E. I., Yoon, C. H., Harden, M., Lennon, N., Gabriel, S., Rodig, S. J., Barouch, D. H., Aster, J. C., Getz, G., Wucherpfennig, K., Neuberg, D., Ritz, J., Lander, E. S., Fritsch, E. F.,

Hacohen, N. & Wu, C. J. 2017. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature,* 547**,** 217-221.

Papaemmanuil, E., Gerstung, M., Malcovati, L., Tauro, S., Gundem, G., Van Loo, P., Yoon, C. J., Ellis, P., Wedge, D. C., Pellagatti, A., Shlien, A., Groves, M. J., Forbes, S. A., Raine, K., Hinton, J., Mudie, L. J., McLaren, S., Hardy, C., Latimer, C., Della Porta, M. G., O'Meara, S., Ambaglio, I., Galli, A., Butler, A. P., Walldin, G., Teague, J. W., Quek, L., Sternberg, A., Gambacorti-Passerini, C., Cross, N. C., Green, A. R., Boultwood, J., Vyas, P., Hellstrom-Lindberg, E., Bowen, D., Cazzola, M., Stratton, M. R., Campbell, P. J. & Chronic Myeloid Disorders Working Group of the International Cancer Genome, C. 2013. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood,* 122**,** 3616-27; quiz 3699.

Pardoll, D. 2003. Does the immune system see tumors as foreign or self? *Annu Rev Immunol,* 21**,** 807-39.

Pardoll, D. M. 2012. The blockade of immune checkpoints in cancer immunotherapy. *Nat Rev Cancer,* 12**,** 252-64.

Park, I. A., Hwang, S. H., Song, I. H., Heo, S. H., Kim, Y. A., Bang, W. S., Park, H. S., Lee, M., Gong, G. & Lee, H. J. 2017. Expression of the MHC class II in triple-negative breast cancer is associated with tumor-infiltrating lymphocytes and interferon signaling. *PLoS One,* 12**,** e0182786.

Pena-Diaz, J., Bregenhorn, S., Ghodgaonkar, M., Follonier, C., Artola-Boran, M., Castor, D., Lopes, M., Sartori, A. A. & Jiricny, J. 2012. Noncanonical mismatch repair as a source of genomic instability in human cells. *Mol Cell,* 47**,** 669-80.

Peng, W., Chen, J. Q., Liu, C., Malu, S., Creasy, C., Tetzlaff, M. T., Xu, C., McKenzie, J. A., Zhang, C., Liang, X., Williams, L. J., Deng, W., Chen, G., Mbofung, R., Lazar, A. J., Torres-Cabala, C. A., Cooper, Z. A., Chen, P. L., Tieu, T. N., Spranger, S., Yu, X., Bernatchez, C., Forget, M. A., Haymaker, C., Amaria, R., McQuade, J. L., Glitza, I. C., Cascone, T., Li, H. S., Kwong, L. N., Heffernan, T. P., Hu, J., Bassett, R. L., Jr., Bosenberg, M. W., Woodman, S. E., Overwijk, W. W., Lizee, G., Roszik, J., Gajewski, T. F., Wargo, J. A., Gershenwald, J. E., Radvanyi, L., Davies, M. A. & Hwu, P. 2016. Loss of PTEN Promotes Resistance to T Cell-Mediated Immunotherapy. *Cancer Discov,* 6**,** 202-16.

Pfeifer, G. P. 2010. Environmental exposures and mutational patterns of cancer genomes. *Genome Med,* 2**,** 54.

Pfeifer, G. P., Denissenko, M. F., Olivier, M., Tretyakova, N., Hecht, S. S. & Hainaut, P. 2002. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene,* 21**,** 7435-51.

Piha-Paul, S. A., Bennouna, J., Albright, A., Nebozhyn, M., McClanahan, T., Ayers, M., Lunceford, J. K. & Ott, P. A. 2016. T-cell inflamed phenotype gene expression signatures to predict clinical benefit from pembrolizumab across multiple tumor types. *Journal of Clinical Oncology,* 34**,** 1536-1536.

Poon, S. L., Pang, S. T., McPherson, J. R., Yu, W., Huang, K. K., Guan, P., Weng, W. H., Siew, E. Y., Liu, Y., Heng, H. L., Chong, S. C., Gan, A.,

Tay, S. T., Lim, W. K., Cutcutache, I., Huang, D., Ler, L. D., Nairismagi, M. L., Lee, M. H., Chang, Y. H., Yu, K. J., Chan-On, W., Li, B. K., Yuan, Y. F., Qian, C. N., Ng, K. F., Wu, C. F., Hsu, C. L., Bunte, R. M., Stratton, M. R., Futreal, P. A., Sung, W. K., Chuang, C. K., Ong, C. K., Rozen, S. G., Tan, P. & Teh, B. T. 2013. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci Transl Med,* 5**,** 197ra101.

Powles, T., Eder, J. P., Fine, G. D., Braiteh, F. S., Loriot, Y., Cruz, C., Bellmunt, J., Burris, H. A., Petrylak, D. P., Teng, S. L., Shen, X., Boyd, Z., Hegde, P. S., Chen, D. S. & Vogelzang, N. J. 2014. MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer. *Nature,* 515**,** 558-62.

Qin, Z., Richter, G., Schuler, T., Ibe, S., Cao, X. & Blankenstein, T. 1998. B cells inhibit induction of T cell-dependent tumor immunity. *Nat Med,* 4**,** 627-30.

Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. 2017. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife,* 6.

Rajasagi, M., Shukla, S. A., Fritsch, E. F., Keskin, D. B., DeLuca, D., Carmona, E., Zhang, W., Sougnez, C., Cibulskis, K., Sidney, J., Stevenson, K., Ritz, J., Neuberg, D., Brusic, V., Gabriel, S., Lander, E. S., Getz, G., Hacohen, N. & Wu, C. J. 2014. Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood,* 124**,** 453-62.

Ribas, A., Robert, C., Hodi, F. S., Wolchok, J. D., Joshua, A. M., Hwu, W.-J., Weber, J. S., Zarour, H. M., Kefford, R., Loboda, A., Albright, A., Kang, S. P., Ebbinghaus, S., Yearley, J., Murphy, E., Nebozhyn, M., Lunceford, J. K., McClanahan, T., Ayers, M. & Daud, A. 2015. Association of response to programmed death receptor 1 (PD-1) blockade with pembrolizumab (MK-3475) with an interferon-inflammatory immune gene signature. *Journal of Clinical Oncology,* 33**,** 3001-3001.

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Consortium, W. G. S., Wilkie, A. O. M., McVean, G. & Lunter, G. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet,* 46**,** 912-918.

Rizvi, N. A., Hellmann, M. D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J. J., Lee, W., Yuan, J., Wong, P., Ho, T. S., Miller, M. L., Rekhtman, N., Moreira, A. L., Ibrahim, F., Bruggeman, C., Gasmi, B., Zappasodi, R., Maeda, Y., Sander, C., Garon, E. B., Merghoub, T., Wolchok, J. D., Schumacher, T. N. & Chan, T. A. 2015. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science,* 348**,** 124-8.

Roberts, E. W., Broz, M. L., Binnewies, M., Headley, M. B., Nelson, A. E., Wolf, D. M., Kaisho, T., Bogunovic, D., Bhardwaj, N. & Krummel, M. F. 2016. Critical Role for CD103(+)/CD141(+) Dendritic Cells Bearing CCR7 for Tumor Antigen Trafficking and Priming of T Cell Immunity in Melanoma. *Cancer Cell,* 30**,** 324-336.

Roberts, S. A., Sterling, J., Thompson, C., Harris, S., Mav, D., Shah, R., Klimczak, L. J., Kryukov, G. V., Malc, E., Mieczkowski, P. A., Resnick,

M. A. & Gordenin, D. A. 2012. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol Cell,* 46**,** 424-35.

Roche, P. A. & Furuta, K. 2015. The ins and outs of MHC class II-mediated antigen processing and presentation. *Nat Rev Immunol,* 15**,** 203-16.

Roh, W., Chen, P. L., Reuben, A., Spencer, C. N., Prieto, P. A., Miller, J. P., Gopalakrishnan, V., Wang, F., Cooper, Z. A., Reddy, S. M., Gumbs, C., Little, L., Chang, Q., Chen, W. S., Wani, K., De Macedo, M. P., Chen, E., Austin-Breneman, J. L., Jiang, H., Roszik, J., Tetzlaff, M. T., Davies, M. A., Gershenwald, J. E., Tawbi, H., Lazar, A. J., Hwu, P., Hwu, W. J., Diab, A., Glitza, I. C., Patel, S. P., Woodman, S. E., Amaria, R. N., Prieto, V. G., Hu, J., Sharma, P., Allison, J. P., Chin, L., Zhang, J., Wargo, J. A. & Futreal, P. A. 2017. Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Sci Transl Med,* 9.

Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. 2015. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell,* 160**,** 48-61.

Rosenberg, S. A. 2012. Raising the bar: the curative potential of human cancer immunotherapy. *Sci Transl Med,* 4**,** 127ps8.

Rosenberg, S. A. & Restifo, N. P. 2015. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science,* 348**,** 62-8.

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. 2016. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol,* 17**,** 31.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Cote, A. & Shah, S. P. 2014. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods,* 11**,** 396-8.

Roth, A., McPherson, A., Laks, E., Biele, J., Yap, D., Wan, A., Smith, M. A., Nielsen, C. B., McAlpine, J. N., Aparicio, S., Bouchard-Cote, A. & Shah, S. P. 2016. Clonal genotype and population structure inference from single-cell tumor sequencing. *Nat Methods,* 13**,** 573-6.

Rottenberg, S., Jaspers, J. E., Kersbergen, A., van der Burg, E., Nygren, A. O., Zander, S. A., Derksen, P. W., de Bruin, M., Zevenhoven, J., Lau, A., Boulter, R., Cranston, A., O'Connor, M. J., Martin, N. M., Borst, P. & Jonkers, J. 2008. High sensitivity of BRCA1-deficient mammary tumors to the PARP inhibitor AZD2281 alone and in combination with platinum drugs. *Proc Natl Acad Sci U S A,* 105**,** 17079-84.

Routy, B., Le Chatelier, E., Derosa, L., Duong, C. P. M., Alou, M. T., Daillere, R., Fluckiger, A., Messaoudene, M., Rauber, C., Roberti, M. P., Fidelle, M., Flament, C., Poirier-Colame, V., Opolon, P., Klein, C., Iribarren, K., Mondragon, L., Jacquelot, N., Qu, B., Ferrere, G., Clemenson, C., Mezquita, L., Masip, J. R., Naltet, C., Brosseau, S., Kaderbhai, C., Richard, C., Rizvi, H., Levenez, F., Galleron, N., Quinquis, B., Pons, N., Ryffel, B., Minard-Colin, V., Gonin, P., Soria, J. C., Deutsch, E., Loriot, Y., Ghiringhelli, F., Zalcman, G., Goldwasser, F., Escudier, B., Hellmann, M. D., Eggermont, A., Raoult, D., Albiges, L., Kroemer, G. & Zitvogel, L. 2017. Gut

microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science*.
Ryland, G. L., Doyle, M. A., Goode, D., Boyle, S. E., Choong, D. Y., Rowley, S. M., Li, J., Australian Ovarian Cancer Study, G., Bowtell, D. D., Tothill, R. W., Campbell, I. G. & Gorringe, K. L. 2015. Loss of heterozygosity: what is it good for? *BMC Med Genomics,* 8**,** 45.
Sahin, U., Derhovanessian, E., Miller, M., Kloke, B. P., Simon, P., Lower, M., Bukur, V., Tadmor, A. D., Luxemburger, U., Schrors, B., Omokoko, T., Vormehr, M., Albrecht, C., Paruzynski, A., Kuhn, A. N., Buck, J., Heesch, S., Schreeb, K. H., Muller, F., Ortseifer, I., Vogler, I., Godehardt, E., Attig, S., Rae, R., Breitkreuz, A., Tolliver, C., Suchan, M., Martic, G., Hohberger, A., Sorn, P., Diekmann, J., Ciesla, J., Waksmann, O., Bruck, A. K., Witt, M., Zillgen, M., Rothermel, A., Kasemann, B., Langer, D., Bolte, S., Diken, M., Kreiter, S., Nemecek, R., Gebhardt, C., Grabbe, S., Holler, C., Utikal, J., Huber, C., Loquai, C. & Tureci, O. 2017. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature,* 547**,** 222-226.
Sahin, U., Tureci, O., Schmitt, H., Cochlovius, B., Johannes, T., Schmits, R., Stenner, F., Luo, G., Schobert, I. & Pfreundschuh, M. 1995. Human neoplasms elicit multiple specific immune responses in the autologous host. *Proc Natl Acad Sci U S A,* 92**,** 11810-3.
Sakuishi, K., Apetoh, L., Sullivan, J. M., Blazar, B. R., Kuchroo, V. K. & Anderson, A. C. 2010. Targeting Tim-3 and PD-1 pathways to reverse T cell exhaustion and restore anti-tumor immunity. *J Exp Med,* 207**,** 2187-94.
Schreiber, R. D., Old, L. J. & Smyth, M. J. 2011. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science,* 331**,** 1565-70.
Schuh, A., Becq, J., Humphray, S., Alexa, A., Burns, A., Clifford, R., Feller, S. M., Grocock, R., Henderson, S., Khrebtukova, I., Kingsbury, Z., Luo, S., McBride, D., Murray, L., Menju, T., Timbs, A., Ross, M., Taylor, J. & Bentley, D. 2012. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood,* 120**,** 4191-6.
Schultz, K. R., Klarnet, J. P., Gieni, R. S., HayGlass, K. T. & Greenberg, P. D. 1990. The role of B cells for in vivo T cell responses to a Friend virus-induced leukemia. *Science,* 249**,** 921-3.
Schulze, K., Imbeaud, S., Letouze, E., Alexandrov, L. B., Calderaro, J., Rebouissou, S., Couchy, G., Meiller, C., Shinde, J., Soysouvanh, F., Calatayud, A. L., Pinyol, R., Pelletier, L., Balabaud, C., Laurent, A., Blanc, J. F., Mazzaferro, V., Calvo, F., Villanueva, A., Nault, J. C., Bioulac-Sage, P., Stratton, M. R., Llovet, J. M. & Zucman-Rossi, J. 2015. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat Genet,* 47**,** 505-511.
Schumacher, T. N. & Schreiber, R. D. 2015. Neoantigens in cancer immunotherapy. *Science,* 348**,** 69-74.
Schwarz, R. F., Ng, C. K., Cooke, S. L., Newman, S., Temple, J., Piskorz, A. M., Gale, D., Sayal, K., Murtaza, M., Baldwin, P. J., Rosenfeld, N.,

Earl, H. M., Sala, E., Jimenez-Linan, M., Parkinson, C. A., Markowetz, F. & Brenton, J. D. 2015. Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. *PLoS Med,* 12**,** e1001789.

Segal, N. H., Parsons, D. W., Peggs, K. S., Velculescu, V., Kinzler, K. W., Vogelstein, B. & Allison, J. P. 2008. Epitope landscape in breast and colorectal cancer. *Cancer Res,* 68**,** 889-92.

Segovia, R., Tam, A. S. & Stirling, P. C. 2015. Dissecting genetic and environmental mutation signatures with model organisms. *Trends Genet,* 31**,** 465-74.

Senbabaoglu, Y., Gejman, R. S., Winer, A. G., Liu, M., Van Allen, E. M., de Velasco, G., Miao, D., Ostrovnaya, I., Drill, E., Luna, A., Weinhold, N., Lee, W., Manley, B. J., Khalil, D. N., Kaffenberger, S. D., Chen, Y., Danilova, L., Voss, M. H., Coleman, J. A., Russo, P., Reuter, V. E., Chan, T. A., Cheng, E. H., Scheinberg, D. A., Li, M. O., Choueiri, T. K., Hsieh, J. J., Sander, C. & Hakimi, A. A. 2016. Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol,* 17**,** 231.

Sensi, M. & Anichini, A. 2006. Unique tumor antigens: evidence for immune control of genome integrity and immunogenic targets for T cell-mediated patient-specific immunotherapy. *Clin Cancer Res,* 12**,** 5023-32.

Sette, A., Vitiello, A., Reherman, B., Fowler, P., Nayersina, R., Kast, W. M., Melief, C. J., Oseroff, C., Yuan, L., Ruppert, J., Sidney, J., del Guercio, M. F., Southwood, S., Kubo, R. T., Chesnut, R. W., Grey, H. M. & Chisari, F. V. 1994. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J Immunol,* 153**,** 5586-92.

Shah, N. P., Nicoll, J. M., Nagar, B., Gorre, M. E., Paquette, R. L., Kuriyan, J. & Sawyers, C. L. 2002. Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. *Cancer Cell,* 2**,** 117-25.

Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., Bashashati, A., Prentice, L. M., Khattra, J., Burleigh, A., Yap, D., Bernard, V., McPherson, A., Shumansky, K., Crisan, A., Giuliany, R., Heravi-Moussavi, A., Rosner, J., Lai, D., Birol, I., Varhol, R., Tam, A., Dhalla, N., Zeng, T., Ma, K., Chan, S. K., Griffith, M., Moradian, A., Cheng, S. W., Morin, G. B., Watson, P., Gelmon, K., Chia, S., Chin, S. F., Curtis, C., Rueda, O. M., Pharoah, P. D., Damaraju, S., Mackey, J., Hoon, K., Harkins, T., Tadigotla, V., Sigaroudinia, M., Gascard, P., Tlsty, T., Costello, J. F., Meyer, I. M., Eaves, C. J., Wasserman, W. W., Jones, S., Huntsman, D., Hirst, M., Caldas, C., Marra, M. A. & Aparicio, S. 2012. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature,* 486**,** 395-9.

Shankaran, V., Ikeda, H., Bruce, A. T., White, J. M., Swanson, P. E., Old, L. J. & Schreiber, R. D. 2001. IFNgamma and lymphocytes prevent

primary tumour development and shape tumour immunogenicity. *Nature,* 410**,** 1107-11.
Sharma, P. & Allison, J. P. 2015. The future of immune checkpoint therapy. *Science,* 348**,** 56-61.
Shen, R. & Seshan, V. E. 2016. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res,* 44**,** e131.
Shi, H., Hugo, W., Kong, X., Hong, A., Koya, R. C., Moriceau, G., Chodon, T., Guo, R., Johnson, D. B., Dahlman, K. B., Kelley, M. C., Kefford, R. F., Chmielowski, B., Glaspy, J. A., Sosman, J. A., van Baren, N., Long, G. V., Ribas, A. & Lo, R. S. 2014. Acquired Resistance and Clonal Evolution in Melanoma during BRAF Inhibitor Therapy. *Cancer Discov,* 4**,** 80-93.
Shukla, S. A., Rooney, M. S., Rajasagi, M., Tiao, G., Dixon, P. M., Lawrence, M. S., Stevens, J., Lane, W. J., Dellagatta, J. L., Steelman, S., Sougnez, C., Cibulskis, K., Kiezun, A., Hacohen, N., Brusic, V., Wu, C. J. & Getz, G. 2015. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat Biotechnol,* 33**,** 1152-8.
Singer, E., Wagner, M. & Woyke, T. 2017. Capturing the genetic makeup of the active microbiome in situ. *ISME J,* 11**,** 1949-1963.
Sivan, A., Corrales, L., Hubert, N., Williams, J. B., Aquino-Michaels, K., Earley, Z. M., Benyamin, F. W., Lei, Y. M., Jabri, B., Alegre, M. L., Chang, E. B. & Gajewski, T. F. 2015. Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science,* 350**,** 1084-9.
Smid, M., Rodriguez-Gonzalez, F. G., Sieuwerts, A. M., Salgado, R., Prager-Van der Smissen, W. J., Vlugt-Daane, M. V., van Galen, A., Nik-Zainal, S., Staaf, J., Brinkman, A. B., van de Vijver, M. J., Richardson, A. L., Fatima, A., Berentsen, K., Butler, A., Martin, S., Davies, H. R., Debets, R., Gelder, M. E., van Deurzen, C. H., MacGrogan, G., Van den Eynden, G. G., Purdie, C., Thompson, A. M., Caldas, C., Span, P. N., Simpson, P. T., Lakhani, S. R., Van Laere, S., Desmedt, C., Ringner, M., Tommasi, S., Eyford, J., Broeks, A., Vincent-Salomon, A., Futreal, P. A., Knappskog, S., King, T., Thomas, G., Viari, A., Langerod, A., Borresen-Dale, A. L., Birney, E., Stunnenberg, H. G., Stratton, M., Foekens, J. A. & Martens, J. W. 2016. Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration. *Nat Commun,* 7**,** 12910.
Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., Walsh, L. A., Postow, M. A., Wong, P., Ho, T. S., Hollmann, T. J., Bruggeman, C., Kannan, K., Li, Y., Elipenahli, C., Liu, C., Harbison, C. T., Wang, L., Ribas, A., Wolchok, J. D. & Chan, T. A. 2014. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med,* 371**,** 2189-99.
Sottoriva, A., Spiteri, I., Piccirillo, S. G., Touloumis, A., Collins, V. P., Marioni, J. C., Curtis, C., Watts, C. & Tavare, S. 2013. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci U S A*.

Spranger, S., Bao, R. & Gajewski, T. F. 2015. Melanoma-intrinsic beta-catenin signalling prevents anti-tumour immunity. *Nature,* 523**,** 231-5.
Starrett, G. J., Luengas, E. M., McCann, J. L., Ebrahimi, D., Temiz, N. A., Love, R. P., Feng, Y., Adolph, M. B., Chelico, L., Law, E. K., Carpenter, M. A. & Harris, R. S. 2016. The DNA cytosine deaminase APOBEC3H haplotype I likely contributes to breast and lung cancer mutagenesis. *Nat Commun,* 7**,** 12918.
Stratton, M. R. 2011. Exploring the genomes of cancer cells: progress and promise. *Science,* 331**,** 1553-8.
Stratton, M. R., Campbell, P. J. & Futreal, P. A. 2009. The cancer genome. *Nature,* 458**,** 719-24.
Street, S. E., Cretney, E. & Smyth, M. J. 2001. Perforin and interferon-gamma activities independently control tumor initiation, growth, and metastasis. *Blood,* 97**,** 192-7.
Sucker, A., Zhao, F., Pieper, N., Heeke, C., Maltaner, R., Stadtler, N., Real, B., Bielefeld, N., Howe, S., Weide, B., Gutzmer, R., Utikal, J., Loquai, C., Gogas, H., Klein-Hitpass, L., Zeschnigk, M., Westendorf, A. M., Trilling, M., Horn, S., Schilling, B., Schadendorf, D., Griewank, K. G. & Paschen, A. 2017. Acquired IFNgamma resistance impairs anti-tumor immunity and gives rise to T-cell-resistant melanoma lesions. *Nat Commun,* 8**,** 15440.
Sveen, A., Loes, I. M., Alagaratnam, S., Nilsen, G., Holand, M., Lingjaerde, O. C., Sorbye, H., Berg, K. C., Horn, A., Angelsen, J. H., Knappskog, S., Lonning, P. E. & Lothe, R. A. 2016. Intra-patient Inter-metastatic Genetic Heterogeneity in Colorectal Cancer as a Key Determinant of Survival after Curative Liver Resection. *PLoS Genet,* 12**,** e1006225.
Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M. & Kohlbacher, O. 2014. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics,* 30**,** 3310-6.
Szollosi, J., Balazs, M., Feuerstein, B. G., Benz, C. C. & Waldman, F. M. 1995. ERBB-2 (HER2/neu) gene copy number, p185HER-2 overexpression, and intratumor heterogeneity in human breast cancer. *Cancer Res,* 55**,** 5400-7.
Tamborero, D., Rubio-Perez, C., Muinos, F., Sabarinathan, R., Piulats, J. M. M., Muntasell, A., Dientsmann, R., Lopez-Bigas, N. & Gonzalez-Perez, A. 2017. A pan-cancer landscape of interactions between solid tumors and infiltrating immune cell populations. *bioRxiv*.
The, M. H. C. s. c. 1999. Complete sequence and gene map of a human major histocompatibility complex. *Nature,* 401**,** 921.
Thomas, L. 1982. On immunosurveillance in human cancer. *Yale J Biol Med,* 55**,** 329-33.
Tirosh, I., Venteicher, A. S., Hebert, C., Escalante, L. E., Patel, A. P., Yizhak, K., Fisher, J. M., Rodman, C., Mount, C., Filbin, M. G., Neftel, C., Desai, N., Nyman, J., Izar, B., Luo, C. C., Francis, J. M., Patel, A. A., Onozato, M. L., Riggi, N., Livak, K. J., Gennert, D., Satija, R., Nahed, B. V., Curry, W. T., Martuza, R. L., Mylvaganam, R., Iafrate, A. J., Frosch, M. P., Golub, T. R., Rivera, M. N., Getz, G., Rozenblatt-Rosen, O., Cahill, D. P., Monje, M., Bernstein, B. E., Louis, D. N., Regev, A. & Suva, M. L. 2016. Single-cell RNA-seq supports a

developmental hierarchy in human oligodendroglioma. *Nature,* 539**,** 309-313.

Toebes, M., Coccoris, M., Bins, A., Rodenko, B., Gomez, R., Nieuwkoop, N. J., van de Kasteele, W., Rimmelzwaan, G. F., Haanen, J. B., Ovaa, H. & Schumacher, T. N. 2006. Design and use of conditional MHC class I ligands. *Nat Med,* 12**,** 246-51.

Topalian, S. L., Hodi, F. S., Brahmer, J. R., Gettinger, S. N., Smith, D. C., McDermott, D. F., Powderly, J. D., Carvajal, R. D., Sosman, J. A., Atkins, M. B., Leming, P. D., Spigel, D. R., Antonia, S. J., Horn, L., Drake, C. G., Pardoll, D. M., Chen, L., Sharfman, W. H., Anders, R. A., Taube, J. M., McMiller, T. L., Xu, H., Korman, A. J., Jure-Kunkel, M., Agrawal, S., McDonald, D., Kollia, G. D., Gupta, A., Wigginton, J. M. & Sznol, M. 2012a. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *The New England journal of medicine,* 366**,** 2443-54.

Topalian, S. L., Hodi, F. S., Brahmer, J. R., Gettinger, S. N., Smith, D. C., McDermott, D. F., Powderly, J. D., Carvajal, R. D., Sosman, J. A., Atkins, M. B., Leming, P. D., Spigel, D. R., Antonia, S. J., Horn, L., Drake, C. G., Pardoll, D. M., Chen, L., Sharfman, W. H., Anders, R. A., Taube, J. M., McMiller, T. L., Xu, H., Korman, A. J., Jure-Kunkel, M., Agrawal, S., McDonald, D., Kollia, G. D., Gupta, A., Wigginton, J. M. & Sznol, M. 2012b. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N Engl J Med,* 366**,** 2443-54.

Tran, E., Robbins, P. F., Lu, Y. C., Prickett, T. D., Gartner, J. J., Jia, L., Pasetto, A., Zheng, Z., Ray, S., Groh, E. M., Kriley, I. R. & Rosenberg, S. A. 2016. T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. *N Engl J Med,* 375**,** 2255-2262.

Tran, E., Turcotte, S., Gros, A., Robbins, P. F., Lu, Y. C., Dudley, M. E., Wunderlich, J. R., Somerville, R. P., Hogan, K., Hinrichs, C. S., Parkhurst, M. R., Yang, J. C. & Rosenberg, S. A. 2014. Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. *Science,* 344**,** 641-5.

Tumeh, P. C., Harview, C. L., Yearley, J. H., Shintaku, I. P., Taylor, E. J., Robert, L., Chmielowski, B., Spasic, M., Henry, G., Ciobanu, V., West, A. N., Carmona, M., Kivork, C., Seja, E., Cherry, G., Gutierrez, A. J., Grogan, T. R., Mateus, C., Tomasic, G., Glaspy, J. A., Emerson, R. O., Robins, H., Pierce, R. H., Elashoff, D. A., Robert, C. & Ribas, A. 2014. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature,* 515**,** 568-71.

Turajlic, S., Litchfield, K., Xu, H., Rosenthal, R., McGranahan, N., Reading, J. L., Wong, Y. N. S., Rowan, A., Kanu, N., Al Bakir, M., Chambers, T., Salgado, R., Savas, P., Loi, S., Birkbak, N. J., Sansregret, L., Gore, M., Larkin, J., Quezada, S. A. & Swanton, C. 2017. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol,* 18**,** 1009-1021.

Turke, A. B., Zejnullahu, K., Wu, Y. L., Song, Y., Dias-Santagata, D., Lifshits, E., Toschi, L., Rogers, A., Mok, T., Sequist, L., Lindeman, N. I., Murphy, C., Akhavanfard, S., Yeap, B. Y., Xiao, Y., Capelletti, M., Iafrate, A. J., Lee, C., Christensen, J. G., Engelman, J. A. & Janne, P.

A. 2010. Preexistence and clonal selection of MET amplification in EGFR mutant NSCLC. *Cancer Cell,* 17**,** 77-88.

Uchi, R., Takahashi, Y., Niida, A., Shimamura, T., Hirata, H., Sugimachi, K., Sawada, G., Iwaya, T., Kurashige, J., Shinden, Y., Iguchi, T., Eguchi, H., Chiba, K., Shiraishi, Y., Nagae, G., Yoshida, K., Nagata, Y., Haeno, H., Yamamoto, H., Ishii, H., Doki, Y., Iinuma, H., Sasaki, S., Nagayama, S., Yamada, K., Yachida, S., Kato, M., Shibata, T., Oki, E., Saeki, H., Shirabe, K., Oda, Y., Maehara, Y., Komune, S., Mori, M., Suzuki, Y., Yamamoto, K., Aburatani, H., Ogawa, S., Miyano, S. & Mimori, K. 2016. Integrated Multiregional Analysis Proposing a New Model of Colorectal Cancer Evolution. *PLoS Genet,* 12**,** e1005778.

Van Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., Sucker, A., Hillen, U., Foppen, M. H. G., Goldinger, S. M., Utikal, J., Hassel, J. C., Weide, B., Kaehler, K. C., Loquai, C., Mohr, P., Gutzmer, R., Dummer, R., Gabriel, S., Wu, C. J., Schadendorf, D. & Garraway, L. A. 2015. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science,* 350**,** 207-211.

van den Broek, M. E., Kagi, D., Ossendorp, F., Toes, R., Vamvakas, S., Lutz, W. K., Melief, C. J., Zinkernagel, R. M. & Hengartner, H. 1996. Decreased tumor surveillance in perforin-deficient mice. *J Exp Med,* 184**,** 1781-90.

van der Bruggen, P., Traversari, C., Chomez, P., Lurquin, C., De Plaen, E., Van den Eynde, B., Knuth, A. & Boon, T. 1991. A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science,* 254**,** 1643-7.

Van Loo, P., Nordgard, S. H., Lingjaerde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., Borresen-Dale, A. L. & Kristensen, V. N. 2010. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A,* 107**,** 16910-5.

van Rooij, N., van Buuren, M. M., Philips, D., Velds, A., Toebes, M., Heemskerk, B., van Dijk, L. J., Behjati, S., Hilkmann, H., El Atmioui, D., Nieuwland, M., Stratton, M. R., Kerkhoven, R. M., Kesmir, C., Haanen, J. B., Kvistborg, P. & Schumacher, T. N. 2013. Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *J Clin Oncol,* 31**,** e439-42.

Vetizou, M., Pitt, J. M., Daillere, R., Lepage, P., Waldschmitt, N., Flament, C., Rusakiewicz, S., Routy, B., Roberti, M. P., Duong, C. P., Poirier-Colame, V., Roux, A., Becharef, S., Formenti, S., Golden, E., Cording, S., Eberl, G., Schlitzer, A., Ginhoux, F., Mani, S., Yamazaki, T., Jacquelot, N., Enot, D. P., Berard, M., Nigou, J., Opolon, P., Eggermont, A., Woerther, P. L., Chachaty, E., Chaput, N., Robert, C., Mateus, C., Kroemer, G., Raoult, D., Boneca, I. G., Carbonnel, F., Chamaillard, M. & Zitvogel, L. 2015. Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science,* 350**,** 1079-84.

Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., Wheeler, D. K., Gabbard, J. L., Hix, D., Sette, A. & Peters, B. 2015. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res,* 43**,** D405-12.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., Jr. & Kinzler, K. W. 2013. Cancer genome landscapes. *Science,* 339**,** 1546-58.

Wagle, N., Emery, C., Berger, M. F., Davis, M. J., Sawyer, A., Pochanard, P., Kehoe, S. M., Johannessen, C. M., Macconaill, L. E., Hahn, W. C., Meyerson, M. & Garraway, L. A. 2011. Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. *J Clin Oncol,* 29**,** 3085-96.

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

Warren, R. L., Choe, G., Freeman, D. J., Castellarin, M., Munro, S., Moore, R. & Holt, R. A. 2012. Derivation of HLA types from shotgun sequence datasets. *Genome Med,* 4**,** 95.

Weston-Bell, N., Gibson, J., John, M., Ennis, S., Pfeifer, S., Cezard, T., Ludwig, H., Collins, A., Zojer, N. & Sahota, S. S. 2013. Exome sequencing in tracking clonal evolution in multiple myeloma following therapy. *Leukemia,* 27**,** 1188-91.

Wolchok, J. D., Kluger, H., Callahan, M. K., Postow, M. A., Rizvi, N. A., Lesokhin, A. M., Segal, N. H., Ariyan, C. E., Gordon, R. A., Reed, K., Burke, M. M., Caldwell, A., Kronenberg, S. A., Agunwamba, B. U., Zhang, X., Lowy, I., Inzunza, H. D., Feely, W., Horak, C. E., Hong, Q., Korman, A. J., Wigginton, J. M., Gupta, A. & Sznol, M. 2013. Nivolumab plus ipilimumab in advanced melanoma. *N Engl J Med,* 369**,** 122-33.

Woo, E. Y., Yeh, H., Chu, C. S., Schlienger, K., Carroll, R. G., Riley, J. L., Kaiser, L. R. & June, C. H. 2002. Cutting edge: Regulatory T cells from lung cancer patients directly inhibit autologous T cell proliferation. *J Immunol,* 168**,** 4272-6.

Xie, C., Yeo, Z. X., Wong, M., Piper, J., Long, T., Kirkness, E. F., Biggs, W. H., Bloom, K., Spellman, S., Vierra-Green, C., Brady, C., Scheuermann, R. H., Telenti, A., Howard, S., Brewerton, S., Turpaz, Y. & Venter, J. C. 2017. Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. *Proc Natl Acad Sci U S A,* 114**,** 8059-8064.

Yap, T., Gerlinger, M., Futreal, A., Pustzai, L. & Swanton, C. 2012. Intratumour Heterogeneity: Seeing the wood for the trees. *Science Translational Medicine,* In press 2012.

Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L. B., Larsimont, D., Davies, H., Li, Y., Ju, Y. S., Ramakrishna, M., Haugland, H. K., Lilleng, P. K., Nik-Zainal, S., McLaren, S., Butler, A., Martin, S., Glodzik, D., Menzies, A., Raine, K., Hinton, J., Jones, D., Mudie, L. J., Jiang, B., Vincent, D., Greene-Colozzi, A., Adnet, P. Y., Fatima, A., Maetens, M., Ignatiadis, M., Stratton, M. R., Sotiriou, C., Richardson, A. L., Lonning, P. E., Wedge, D. C. & Campbell, P. J. 2015. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med,* 21**,** 751-9.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and

medium sized insertions from paired-end short reads. *Bioinformatics,* 25**,** 2865-71.

Yip, S., Miao, J., Cahill, D. P., Iafrate, A. J., Aldape, K., Nutt, C. L. & Louis, D. N. 2009. MSH6 mutations arise in glioblastomas during temozolomide therapy and mediate temozolomide resistance. *Clin Cancer Res,* 15**,** 4622-9.

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., Lawrence, M. S., Zhang, C. Z., Wala, J., Mermel, C. H., Sougnez, C., Gabriel, S. B., Hernandez, B., Shen, H., Laird, P. W., Getz, G., Meyerson, M. & Beroukhim, R. 2013. Pan-cancer patterns of somatic copy number alteration. *Nat Genet,* 45**,** 1134-1140.

Zaretsky, J. M., Garcia-Diaz, A., Shin, D. S., Escuin-Ordinas, H., Hugo, W., Hu-Lieskovan, S., Torrejon, D. Y., Abril-Rodriguez, G., Sandoval, S., Barthly, L., Saco, J., Homet Moreno, B., Mezzadra, R., Chmielowski, B., Ruchalski, K., Shintaku, I. P., Sanchez, P. J., Puig-Saus, C., Cherry, G., Seja, E., Kong, X., Pang, J., Berent-Maoz, B., Comin-Anduix, B., Graeber, T. G., Tumeh, P. C., Schumacher, T. N., Lo, R. S. & Ribas, A. 2016. Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *N Engl J Med,* 375**,** 819-29.

Zhang, J., Fujimoto, J., Zhang, J., Wedge, D. C., Song, X., Zhang, J., Seth, S., Chow, C. W., Cao, Y., Gumbs, C., Gold, K. A., Kalhor, N., Little, L., Mahadeshwar, H., Moran, C., Protopopov, A., Sun, H., Tang, J., Wu, X., Ye, Y., William, W. N., Lee, J. J., Heymach, J. V., Hong, W. K., Swisher, S., Wistuba, II & Futreal, P. A. 2014. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science,* 346**,** 256-9.

Zhao, F., Sucker, A., Horn, S., Heeke, C., Bielefeld, N., Schrors, B., Bicker, A., Lindemann, M., Roesch, A., Gaudernack, G., Stiller, M., Becker, J. C., Lennerz, V., Wolfel, T., Schadendorf, D., Griewank, K. & Paschen, A. 2016. Melanoma Lesions Independently Acquire T-cell Resistance during Metastatic Latency. *Cancer Res,* 76**,** 4347-58.